Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Journal of Computational Physics 225 (2007) 74–99
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

An accelerated algorithm for 2D simulations of the quantum ballistic transport in nanoscale MOSFETs

N. Ben Abdallah [a], M. Mouis [b], C. Negulescu [a,*]

[a] *Mathématiques pour l'Industrie et la Physique, UMR CNRS 5640, Université Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse Cedex, France*
[b] *IMEP (UMR CNRS/INPG/UJF), 23 rue des Martyrs, BP 257, F-38016 Grenoble Cedex, France*

Received 21 December 2005; received in revised form 6 November 2006; accepted 16 November 2006
Available online 26 January 2007

Abstract

An accelerated algorithm for the resolution of the coupled Schrödinger/Poisson system, with open boundary conditions, is presented. This method improves the sub-band decomposition method (SDM) introduced in [N. Ben Abdallah, E. Polizzi, Subband decomposition approach for the simulation of quantum electron transport in nanostructures, J. Comput. Phys. 202 (1) (2005) 150–180]. The principal feature of the here presented model consists in an inexpensive and fast resolution of the Schrödinger equation in the transport direction, due to the application of WKB techniques. Oscillating WKB basis elements are constructed and replace the piecewise polynomial interpolation functions used in SDM. This procedure is shown to reduce considerably the computational time, while keeping a good accuracy.

Keywords: Schrödinger–Poisson equation; Open boundary conditions; Subband model; WKB approximation; Confinement effects; Quantum tunneling

1. Introduction

MOSFET size dimensions continue to decrease rapidly towards the sub-10 nm range. The interest for such a scaling is multiple: more functionality, higher operating speeds, reduced power consumption. To support this effort, it is of primary importance to develop modeling and simulation tools that are adequate for the description of ultra small devices. The objective of this paper is the presentation of a very efficient and inexpensive method for the simulation of nanoscale MOSFET devices. This method was previously introduced by the authors in a condensed form in [20].

The scaling of devices beyond a certain limit enhances the importance of some physical phenomena, with correlated consequences on device modeling techniques. The electron transport becomes near-ballistic and quantum effects, as interferences, tunneling and confinements, can no more be neglected. Quantum ballistic transport models are thus adequate to describe what can be ultimately expected in these devices. The interested

---

[*] Corresponding author. Tel.: +33 5 61 55 86 25.
E-mail addresses: naoufel@mip.ups-tlse.fr (N. Ben Abdallah), mouis@enserg.fr (M. Mouis), negulesc@mip.ups-tlse.fr (C. Negulescu).

0021-9991/$ - see front matter
doi:10.1016/j.jcp.2006.11.028

reader can find more physical details in [1,3,10,12]. There is a great amount of work dedicated to semiconductor device simulations, either by the non-equilibrium Green's function formalism [11,16,17,26,27] or by a finite element/difference resolution of the Schrödinger–Poisson system [4,5,8,13,18,22]. The approach presented in this paper is based on the resolution of the self-consistent Schrödinger–Poisson equation with open boundary conditions, which enable the current flow.

In previous works of Ben Abdallah and Polizzi [8], respectively Laux et al. [18], the Schrödinger–Poisson equation is solved self-consistently via standard variational formulations in the whole definition/simulation domain. The first approach uses the original QTBM boundary conditions [19], whereas the second one extends these boundary conditions to simulate devices far from equilibrium. The disadvantage of these approaches is that they are very time consuming due to the expensive resolution of the 3D or 2D Schrödinger equation. In a recent work of Ben Abdallah and Polizzi [7], a sub-band decomposition method (SDM) was proposed to reduce the numerical cost for the resolution of the 2D Schrödinger equation. The SDM method is based on the fact that in modern devices the electron gas is confined in one or more directions and that consequently the dimension of the propagation space is reduced. Hence the resolution of the Schrödinger equation in the whole 2D domain is replaced by 1D eigenvalue problems in the confined (or transversal) direction and a system of coupled 1D Schrödinger equations projected on the transport (or longitudinal) direction. It is important to remark that the SDM method retains the coupling effects in both directions of space.

The goal of the present paper is to propose a new powerful model, further named SDM/WKB, which develops the SDM method to provide a relatively inexpensive way to solve the 2D Schrödinger equation. This new method reduces once again considerably the simulation time by accelerating the resolution of the longitudinal coupled 1D Schrödinger system through the use of WKB techniques. The WKB approximation, often called also semi-classical approximation, is a powerful technique to treat problems involving two different scales. It is used in [9] for the Helmholtz equation and in [15,25] for the Schrödinger equation with the objective to investigate the high frequency asymptotics ($\hbar \to 0$) of the just mentioned equations. In the present work, however, we are interested in an approximation method for the Schrödinger equation, suitable for fixed, arbitrary wavelengths and equally accurate independently on the Planck constant $\hbar$ and the electron energy E (see [23] for the Helmholtz equation). Indeed, for slowly varying potentials this approximation is good not only for high frequencies. In [6] the WKB approximation is used in this manner by Ben Abdallah and Pinaud for the 1D simulation of a resonant tunneling diode (RTD). The present paper combines the two methods, SDM and WKB, with the objective to lower the numerical burden for two dimensional applications, as the MOSFET devices.

The principal idea is the following. The SDM method uses the conventional finite element approach to solve the 1D Schrödinger equation in the transport direction. The "10 degrees of freedom per wavelength"-rule requires thus a refined mesh size in this direction to accurately approximate the highly oscillating wave functions, which correspond to high injection energies. The original contribution of the SDM/WKB method consists in replacing the linear or polynomial interpolation functions by more elaborated, oscillating interpolation functions. These oscillating basis functions are determined by means of the WKB plane-wave Ansatz and posses a frequency close to that of the unknown wave-function. Therefore, accurate results can be obtained with much coarser grids, leading for this reason to a highly reduced simulation time.

The outline of this paper is the following. Section 2 starts by briefly presenting the SDM method, in order to introduce afterwards the new SDM/WKB method, which is the core of this work. Section 3 compares numerically (simulation time, accuracy) the two methods with a standard variational resolution method for the Schrödinger equation. Moreover, it enumerates some interesting points concerning the implementation of the SDM/WKB method. Finally, Section 4 is devoted to the discussion of the simulation results obtained with the SDM/WKB method for several devices in order to validate the results from a physical point of view. To keep this paper simple, we preferred to present the extensive mathematical and numerical analysis of this model in a separate future work [2,21].

2. Description of the SDM/WKB model

The purpose of this section is to introduce the original 3D SDM/WKB transport model, developed for the simulation of the ballistic quantum electron transport in nanoscale MOSFETs. This paper deals with the stationary problem.

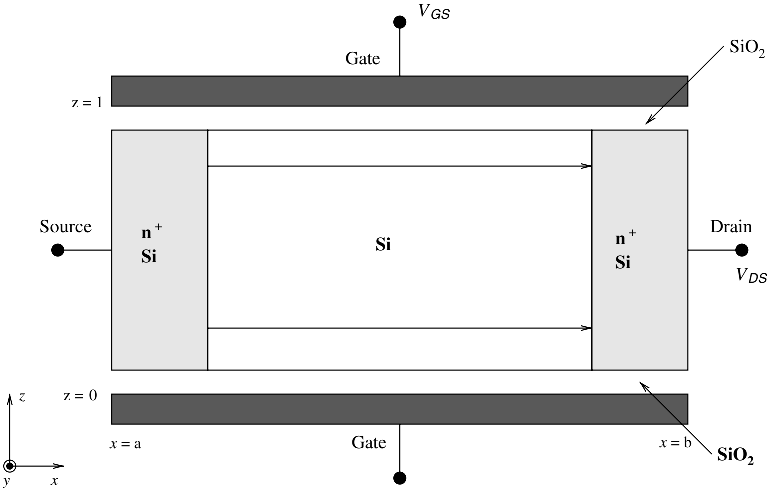

Fig. 1. Schematic representation of the modeled device.

The double-gate MOSFET, which has been used as a test structure for this numerical study, is schematically represented in Fig. 1. It is a symmetrical device, consisting in a Si-channel separated by thin SiO_2 insulator layers from the two gate contacts. A drain–source voltage V_{DS} is applied to create an electron flow in the x-direction, whereas a gate–source voltage V_{GS} controls the conductivity of the Si-channel. Electrons are injected from the source and the drain reservoirs. These two reservoirs are assumed to be in a thermal equilibrium state, characterized by different Fermi energy levels.

The electron gas is supposed to be confined in the direction z. It is exactly upon this important feature, that the SDM method is based. The confinement induces the discretization of the energy continuum and thus the creation of energy-subbands in the transverse direction. The electron motion is allowed in the remaining two directions. The y-direction is assumed to be translational invariant, such that the x-direction shall be considered as the transport direction. Because of the translational invariance in the y-direction, the simulation domain is limited to a 2D domain, $(x,z) \in [a,b] \times [0,1]$, including the two oxide layers and parts of the source and the drain regions. The source and the drain are modeled by highly doped n^+-regions, whereas the Si-channel is assumed to be undoped ($10^{16}\,\mathrm{cm}^{-3}$ residual doping level).

We account in this model for the anisotropic crystal structure of Si, illustrated by six equivalent conduction band ellipsoids (see Fig. 2). This gives rise to two different effective masses, the longitudinal heavy effective mass $m_l = 0.98 * m_e$ and the transversal light effective mass $m_t = 0.19 * m_e$, where m_e stands for the electron rest mass. Note that as the oxide layer is included into the simulated domain, the effective masses depend on the variable z.

Let us give now a brief and necessary overview of the SDM method, introduced in [7]. This overview shall help to understand better the here proposed extension SDM/WKB, which is the essence of this work and is based on the WKB approximation [6].

2.1. The SDM method

The coupled Schrödinger–Poisson equation is the most appropriate model to describe the quantum, ballistic electron transport. Due to the non-linearity, an iterative method (Gummel iteration) will be used for the resolution of this system, consisting of two blocs: the resolution of the Schrödinger equation and the resolution of the Poisson equation.

Quantum mechanically the electron evolution is described by means of the stationary Schrödinger equation

$$(H - qV(x,z))\Psi_E(x,y,z) = E\Psi_E(x,y,z), \tag{2.1}$$

with $(x,y,z) \in [a,b] \times \mathbb{R} \times [0,1]$ and where H is the Hamiltonian, defined by

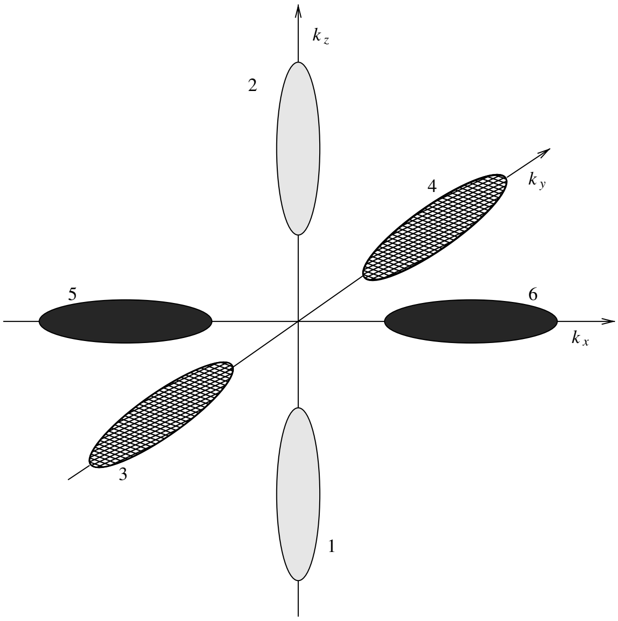

Fig. 2. The six equivalent conduction band ellipsoids of Si.

$$H = -\frac{\hbar^2}{2}\left(\frac{1}{m_x(z)}\Delta_x + \frac{1}{m_y(z)}\Delta_y\right) - \frac{\hbar^2}{2}\frac{\partial}{\partial z}\left(\frac{1}{m_z(z)}\frac{\partial}{\partial z}\right). \tag{2.2}$$

We denoted by Ψ_E the complex-valued wave function, depending on the electron injection energy E. Furthermore, $\hbar$ is the Planck constant, q the elementary electron charge and m_x, m_y, m_z are the z-dependent effective masses in the x-, y- respect. z-direction. The electrostatic potential V depends only on the confinement and the transport direction and is split into a given exterior potential V_e, and a self-consistent one V_s. The self-consistent potential is solution of the Poisson equation

$$\begin{cases} -\Delta V_s(x,z) = -qn(x,z), \\ \partial_{\mathbf{n}} V_s(a,\cdot) = \partial_{\mathbf{n}} V_s(b,\cdot) = 0, \\ V_s(\cdot,0) = V_g^0, \quad V_s(\cdot,1) = V_g^1, \end{cases} \tag{2.3}$$

where $\partial_{\mathbf{n}}$ denotes the normal derivative to the boundary and $V_g^0,\ V_g^1$ are the applied gate voltages. The Neumann boundary conditions at the device/lead interface are more appropriate in ballistic devices than the Dirichlet boundary conditions, in order to ensure the charge neutrality in the leads. The electron density n is derived on its turn from the wave function Ψ_E, as follows:

$$n = \int |\Psi_E(x,y,z)|^2 f_{\mathrm{FD}}(E)\,\mathrm{d}E, \tag{2.4}$$

with f_{FD} the Fermi–Dirac distribution function. The exact expression is given in (2.15) and (2.16).

Due to the translational invariance of the problem in the y-direction, the wave function Ψ_E can be decomposed as

$$\Psi_E(x,y,z) = \mathrm{e}^{\mathrm{i}k_y y}\psi_\epsilon(x,z) \quad \text{with } E = \epsilon + \frac{(\hbar k_y)^2}{2\overline{m}_y}, \tag{2.5}$$

where ψ_ϵ is solution of the 2D Schrödinger equation in $[a,b]\times[0,1]$

$$-\frac{\hbar^2}{2}\frac{1}{m_x(z)}\Delta_x\psi_\epsilon(x,z) - \frac{\hbar^2}{2}\frac{\partial}{\partial z}\left(\frac{1}{m_z(z)}\frac{\partial}{\partial z}\psi_\epsilon(x,z)\right) - qV(x,z)\psi_\epsilon(x,z) = \epsilon\psi_\epsilon(x,z). \tag{2.6}$$

We denoted by k_y the wave vector in the y-direction and by $\overline{m}_y$ the mean value of m_y over the whole (x,z)-domain, with ψ_ϵ as weight function

$$\overline{m}_y[\psi_\epsilon] := \frac{\int_a^b \int_0^1 m_y(z)|\psi_\epsilon(x,z)|^2 \, \mathrm{d}z \, \mathrm{d}x}{\int_a^b \int_0^1 |\psi_\epsilon(x,z)|^2 \, \mathrm{d}z \, \mathrm{d}x}. \tag{2.7}$$

The resolution of the coupled Schrödinger–Poisson system (2.3)–(2.6) by means of standard variational methods (FEM) was the subject of several papers [18,8], but is numerically very expensive. The SDM method, introduced in [7], was proposed in order to reduce the numerical cost for the resolution of the 2D Schrödinger equation (2.6). This method replaces the resolution of the full 2D Schrödinger equation by the resolution of 1D problems in the confinement direction and 1D Schrödinger equations in the transport direction, both directions remaining still coupled. In other words, it consists in expanding ψ_ϵ in the $L^2_z(0,1)$ orthonormal basis $\{\chi_i(z;x)\}$ as

$$\psi_\epsilon(x,z) = \sum_i \varphi^i_\epsilon(x)\chi_i(z;x), \tag{2.8}$$

where the transversal wave functions χ_i are solutions of the 1D eigenvalue problem in the confinement direction z

$$\begin{cases} -\frac{\hbar^2}{2}\frac{\partial}{\partial z}\left(\frac{1}{m_z(z)}\frac{\partial}{\partial z}\chi_i(z;x)\right) - qV(x,z)\chi_i(z;x) = E_i(x)\chi_i(z;x), \\ \chi_i(\cdot;x) \in H^1_0(0,1), \quad \int_0^1 \chi_i(z;x)\chi_j(z;x)\,\mathrm{d}z = \delta_{ij}, \end{cases} \tag{2.9}$$

and the longitudinal wave functions φ^i_ϵ are the solutions of the coupled 1D Schrödinger equations in the transport direction x

$$\begin{cases} -\frac{\mathrm{d}^2}{\mathrm{d}x^2}\varphi^i_\epsilon(x) - 2\sum\limits_{j=1}^{\infty} a_{ij}(x)\frac{\mathrm{d}}{\mathrm{d}x}\varphi^j_\epsilon(x) - \sum\limits_{j=1}^{\infty}\left(b_{ij}(x) + \frac{2}{\hbar^2}c_{ij}(x)(\epsilon - E_j(x))\right)\varphi^j_\epsilon(x) = 0 \\ +\text{open boundary conditions.} \end{cases} \tag{2.10}$$

The index i, called transversal mode, indicates the different subbands. E_i and χ_i represent the potential energy and the transversal wave function of the electrons belonging to the ith sub-band. The coupling terms between the different subbands, a_{ij}, b_{ij} and c_{ij}, are given by

$$a_{ij}(x) := \int_0^1 \chi_i(z;x)\frac{\partial}{\partial x}\chi_j(z;x)\,\mathrm{d}z, \tag{2.11}$$

$$b_{ij}(x) := \int_0^1 \chi_i(z;x)\frac{\partial^2}{\partial x^2}\chi_j(z;x)\,\mathrm{d}z, \tag{2.12}$$

$$c_{ij}(x) := \int_0^1 m_x(z)\chi_i(z;x)\chi_j(z;x)\,\mathrm{d}z. \tag{2.13}$$

Both Eqs. (2.9) and (2.10) are solved, in the SDM method, by standard finite element methods. In the next section, we will show that considerable gain in simulation time can be obtained by using instead of the standard FEM, a "WKB finite element method" (based on a WKB approach) for the resolution of the 1D equations (2.10).

The open boundary conditions completing the Schrödinger equation (2.10) are calculated by assuming that the potential is only dependent on the variable z outside the device, in the two electron leads (thermal equilibrium condition). The thus obtained plane wave solutions in the leads are linked via continuity conditions with the solutions inside the channel, giving rise to the following so-called open boundary conditions (current-carrying conditions) [14,19]

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}x}\varphi^i_\epsilon(a) = \mathbf{i}k^a_{x,i}(2\alpha^a_{i,\epsilon} - \varphi^i_\epsilon(a))\delta_{i\leqslant I_a(\epsilon)} + k^a_{x,i}\varphi^i_\epsilon(a)\delta_{i>I_a(\epsilon)}, \\ \frac{\mathrm{d}}{\mathrm{d}x}\varphi^i_\epsilon(b) = -\mathbf{i}k^b_{x,i}(2\alpha^b_{i,\epsilon} - \varphi^i_\epsilon(b))\delta_{i\leqslant I_b(\epsilon)} - k^b_{x,i}\varphi^i_\epsilon(b)\delta_{i>I_b(\epsilon)}, \end{cases} \tag{2.14}$$

where $\alpha^r_{i,\epsilon}$ denotes the amplitude of the incoming wave, penetrating the device from the source $r = a$ or from the drain $r = b$, in the ith transversal mode and with the injection energy ϵ. The corresponding longitudinal wave vector $k^r_{x,i}$ is given by

$$k^r_{x,i}(\epsilon) := \frac{\sqrt{2\overline{m}_x(\epsilon - E_i(r))}}{\hbar},$$

where $\overline{m}_x$ stands for the mean value of the effective mass m_x, defined by

$$\overline{m}_x[\chi_i(\cdot, r)] := \int_0^1 m_x(z)|\chi_i(z;r)|^2 \, \mathrm{d}z.$$

We have chosen the transversal wave function as weight, to account for the nonuniform electron distribution in the z-direction. Moreover, $I_r(\epsilon)$ is the number of propagating modes

$$I_r(\epsilon) := \sup\{i \in \mathbb{N}/\epsilon > E_i(r)\}.$$

In other words, the boundary conditions (2.14) express the fact that a wave coming from a reservoir and penetrating the device at $x = a$ or $x = b$ is partially reflected and transmitted by the potential barrier. The transmitted part can be an evanescent wave. We shall denote in the following the solution of (2.10), (2.14) corresponding to one incoming wave from the reservoir r_0, in the transversal mode i_0, with the energy ϵ and amplitude one by $\varphi_{r_0,i_0,\epsilon}$ and the corresponding 2D wave by $\psi_{r_0,i_0,\epsilon}$. In this case we have for the amplitude $\alpha^r_{i,\epsilon} = \delta_{i,i_0}\delta_{r,r_0}$. The electron density for a given effective mass configuration is now given as the superposition of densities of these scattering states

$$n_{m_x,m_y,m_z}(x,z) = 2\sum_{r_0=a,b}\sum_{i_0=1}^{\infty}\int_0^{\infty}|\psi_{r_0,i_0,\epsilon}(x,z)|^2\left(\int_{-\infty}^{\infty} f_{\mathrm{FD}}(E;\mu_{r_0})\frac{\mathrm{d}k_y}{2\pi}\right)\frac{\mathrm{d}k_x}{2\pi}, \tag{2.15}$$

where the Fermi–Dirac distribution function f_{FD} characterizes the electron injection in the device and is given by the formula

$$f_{\mathrm{FD}}(E,\mu_{r_0}) := \frac{1}{1 + \exp\frac{E - q\mu_{r_0}}{k_{\mathrm{B}}T}}.$$

The electron temperature is denoted by T, k_{B} is the Boltzmann constant and the chemical potential μ_{r_0} is deduced from

$$\mu_{r_0} = \mu - V_{r_0},$$

with μ the chemical potential at equilibrium, when no drain–source voltage V_{DS} is applied, and V_{r_0} is the applied potential at $x = r_0$. The factor 2 in (2.15) comes from the Pauli principle. The energy-wavevector relations are

$$\epsilon(r_0, i_0, k_x) = E_{i_0}(r_0) + \frac{(\hbar k_x)^2}{2\overline{m}_x}, \quad E(r_0, i_0, k_x, k_y) = E(\epsilon, k_y) = \epsilon + \frac{(\hbar k_y)^2}{2\overline{m}_y}.$$

Knowledge of the total electron density, defined as the sum of the contributions corresponding to the different effective mass configurations

$$n(x,z) = 2(n_{m_{\mathrm{t}},m_{\mathrm{t}},m_{\mathrm{l}}} + n_{m_{\mathrm{t}},m_{\mathrm{l}},m_{\mathrm{t}}} + n_{m_{\mathrm{l}},m_{\mathrm{t}},m_{\mathrm{t}}}), \tag{2.16}$$

will now allow the potential to be calculated by solving (2.3) and to restart the Schrödinger–Poisson iteration. Once the Schrödinger–Poisson iteration is accomplished, the electron current is computed as

$$j_{m_x,m_y,m_z}(x,z) = 2\sum_{r_0=a,b}\sum_{i_0=1}^{\infty}\int_0^{\infty}\frac{q\hbar}{\overline{m}_x}\mathscr{I}m\{\overline{\psi}_{r_0,i_0,\epsilon}(x,z)\nabla\psi_{r_0,i_0,\epsilon}(x,z)\}\left(\int_{-\infty}^{\infty} f_{\mathrm{FD}}(E;\mu_{r_0})\frac{\mathrm{d}k_y}{2\pi}\right)\frac{\mathrm{d}k_x}{2\pi}, \tag{2.17}$$

and, in the same manner, the total current is given by

$$j(x,z) = 2(j_{m_{\mathrm{t}},m_{\mathrm{t}},m_{\mathrm{l}}} + j_{m_{\mathrm{t}},m_{\mathrm{l}},m_{\mathrm{t}}} + j_{m_{\mathrm{l}},m_{\mathrm{t}},m_{\mathrm{t}}}). \tag{2.18}$$

Due to the fact that at a fixed temperature the probability to find an electron at high energies is insignificant, we shall take into account in the numerical approximation only a finite number of transversal modes, denoted by M, such that the index i_0 runs over $i_0 = 1, \ldots, M$ in (2.15) and (2.17).

The advantage of the SDM method is that it reduces the initial 2D Schrödinger equation (2.6) to the resolution of 1D eigenvalue problems in the confined direction (2.9) and many 1D Schrödinger equations projected on the transport direction (2.10). Thus the size of the linear system to be solved at the end is reduced from $N_x \times N_z$ for the initial 2D model to essentially $N_x \times M$ for the SDM method, where N_x and N_z are the number of grid points in the transport respectively confined direction, and M is the number of transversal modes taken into consideration. This leads to a considerable gain in the simulation time.

To summarize this method, Fig. 3 presents a block diagram with the various levels of computation.

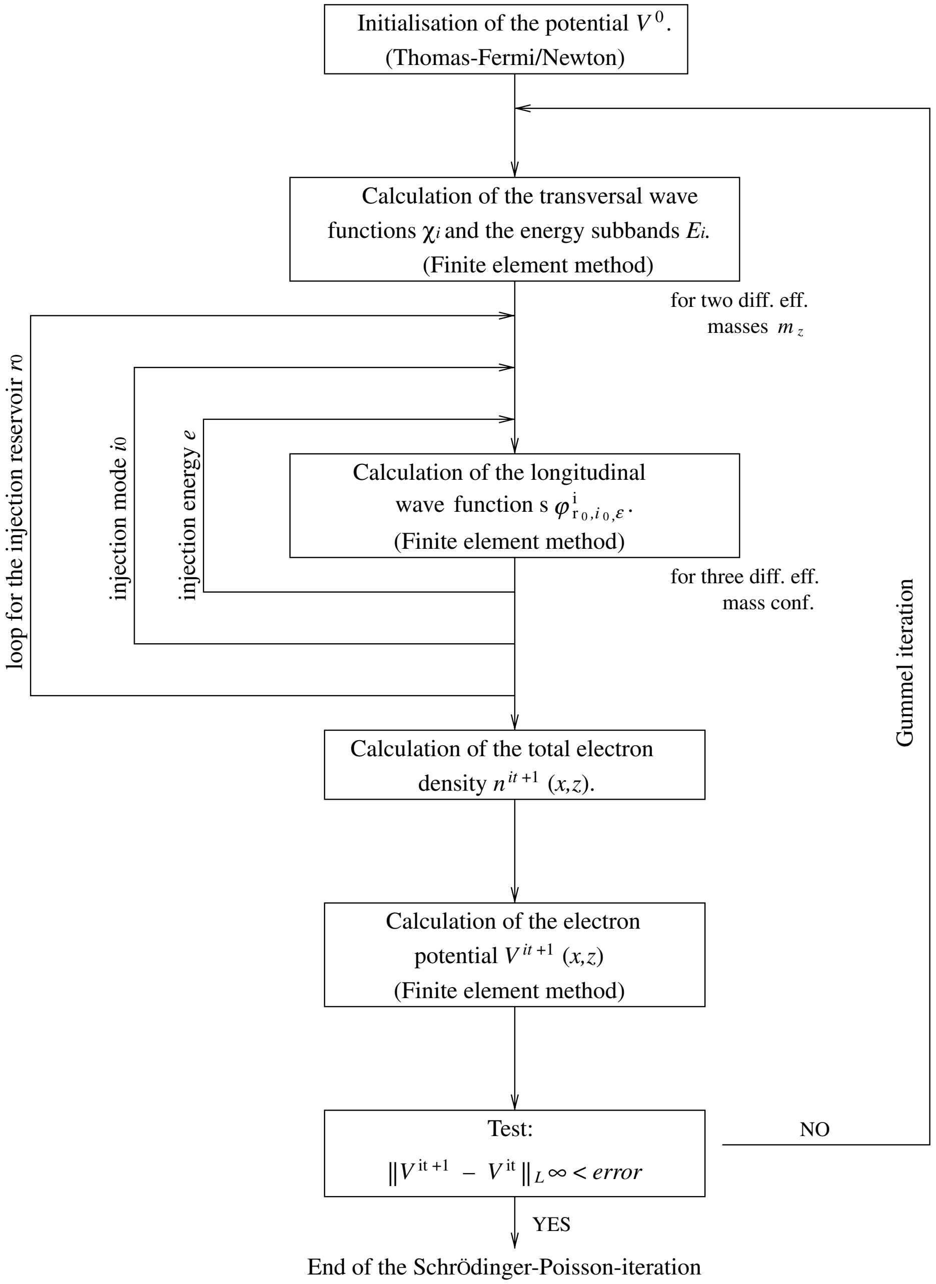

Fig. 3. The SDM algorithm.

2.2. The WKB hat functions – the SDM/WKB method

The new and essential part of our work is presented in this section, introducing the SDM/WKB method. To reduce once more the numerical cost of the simulation, we improved the previous SDM model through the use of WKB techniques. In the following we call this model SDM/WKB. The idea behind is to accelerate the resolution of the 1D Schrödinger equation (2.10), (2.14) by reducing the number of grid points in the transport direction. In this manner we will achieve an important gain in the computational time, as the 1D Schrödinger equation has to be solved several times, one for each value ϵ of the discretized energy spectrum. For this remark that the high energy wave functions are strongly oscillating, and as such a refined mesh size is needed for the SDM model to accurately account for these functions ("10 degrees of freedom"-rule). Indeed, the SDM approach uses finite elements, which consist in connecting nodal points by piecewise linear or polynomial approximation functions. The basic idea of the SDM/WKB method is that rather than using polynomial interpolation functions, oscillating interpolation functions are used, which incorporate a priori knowledge about the solution. Due to the fact that the frequency of these oscillating functions is close to that of the wave-function itself, this method permits to obtain accurate results with much coarser grids. The oscillating interpolation functions are deduced by a WKB Ansatz and will be called WKB hat functions.

Starting from the coupled 1D Schrödinger equations (2.10) and considering only a finite number M of subbands, we can write this system under vectorial form as

$$-\hbar^2\Phi''(x) - 2\hbar^2 A(x)\Phi'(x) - \hbar^2 B(x)\Phi(x) - C(x)\Phi(x) = 0, \quad x \in [a,b], \tag{2.19}$$

with $\Phi := (\varphi^i_\epsilon)^M_{i=1}$; $A := (a_{ij})^M_{i,j=1}$; $B := (b_{ij})^M_{i,j=1}$ and $C := (2c_{ij}(\epsilon - E_j))^M_{i,j=1}$. The idea is to seek for an approximate solution of (2.19) in the form of a plane wave with a phase function S and modulated by an amplitude function $\vec{e}$ as follows:

$$\Phi(x) = \mathrm{e}^{\frac{\mathbf{i}}{\hbar}S(x)}\vec{e}(x) \quad \text{with} \quad \vec{e}(x) = \alpha(x)\vec{u}(x) \quad \text{where } |\vec{u}(x)| = 1 \quad \forall x,\ \forall\hbar. \tag{2.20}$$

Inserting this Ansatz in (2.19) and neglecting terms in $\hbar^2$ yields

$$-2\mathbf{i}\hbar S'\alpha'\vec{u} - 2\mathbf{i}\hbar S'\alpha\vec{u}' - \mathbf{i}\hbar S''\alpha\vec{u} + (S')^2\alpha\vec{u} - 2\mathbf{i}\hbar S'\alpha A\vec{u} - \alpha C\vec{u} = 0.$$

The resolution of this equation is possible, if we consider the decomposition

$$-2\mathbf{i}\hbar S'\vec{u}' + (S')^2\vec{u} - 2\mathbf{i}\hbar S'A\vec{u} - C\vec{u} = 0, \quad 2S'\alpha' + S''\alpha = 0. \tag{2.21}$$

Notice that in the present 2D case, S and $\vec{e}$ will depend on $\hbar$, which differs from the pure 1D case, investigated in [6]. The second equation of system (2.21) can be rewritten as

$$(S'_\hbar(\alpha_\hbar)^2)' = 0,$$

implying thus immediately for $S'_\hbar \neq 0$

$$\alpha_\hbar(x) = \frac{c}{\sqrt{|S'_\hbar(x)|}}, \quad c \in \mathbb{C}.$$

In order to solve the first equation, we expand $\vec{u}_\hbar$ and $S_\hbar$ in powers of $\hbar$

$$\vec{u}_\hbar(x) = \vec{u}^0(x) + \hbar\vec{u}^1(x) + \cdots; \quad S_\hbar(x) = S^0(x) + \hbar S^1(x) + \cdots. \tag{2.22}$$

Substituting these expansions in (2.21) and comparing the terms of the same order in $\hbar$, results in a sequence of equations to be solved to determine $\vec{u}^0_\hbar, \vec{u}^1_\hbar, S^0_\hbar, S^1_\hbar, \ldots$. Limiting us to the zeroth and first order terms, we obtain the equations

$$\left(\frac{\mathrm{d}}{\mathrm{d}x}S^0\right)^2(x)u^0(x) = C(x)u^0(x), \tag{2.23}$$

$$-2\mathbf{i}S^{0'}u^{0'} + (S^{0'})^2u^1 + 2S^{0'}S^{1'}u^0 - 2\mathbf{i}S^{0'}Au^0 - Cu^1 = 0. \tag{2.24}$$

For notational simplicity we omitted the index $\hbar$ and the vector arrows. The matrix C is close to a real, diagonal matrix. This follows from (2.13) and the fact that $\chi_i \in H_0^1(0,1)$ are orthogonal functions. From the perturbation theory, we deduce that C is diagonalizable and has M simple, real eigenvalues $\lambda_j(x)$. Denote by $u_j^0(x) \in \mathbb{R}^M$ the corresponding eigenvectors with $|u_j^0(x)| = 1$. Then we get from (2.23)

$$(S_j^0)'(x) = \begin{cases} \pm\sqrt{\lambda_j(x)} & \text{for } \lambda_j(x) \geqslant 0, \\ \pm\mathbf{i}\sqrt{|\lambda_j(x)|} & \text{for } \lambda_j(x) < 0, \end{cases} \qquad j = 1, \ldots, M.$$

Let us denote by P the transformation matrix $P(x) := (u_1^0(x)| \cdots |u_M^0(x))$, such that $P^{-1}CP$ is the diagonal matrix containing the eigenvalues of C. Consider next a fixed transversal mode $l \in \{1, \ldots, M\}$ and let us compute the first order terms S_l^1, u_l^1 corresponding to S_l^0 and u_l^0. For this multiply (2.24) by P^{-1}, take the lth coordinate and use (2.23). This leads to

$$S_l^{1'}(x) = \mathbf{i}\tilde{a}_{l,l}(x) + \mathbf{i}(P^{-1}u_l^{0'}(x))_l,$$

where $\tilde{A} = P^{-1}AP$. Taking the jth coordinate, with $j \neq l$, we get

$$(P^{-1}u_l^1)_j(x) = 2\mathbf{i}\frac{S_l^{0'}(x)}{\lambda_l(x) - \lambda_j(x)}[\tilde{a}_{jl}(x) + (P^{-1}u_l^{0'})_j].$$

To deduce u_l^1, remark that

$$u_l^1 = \sum_{j=1}^{M} \tau_j u_j^0 \quad \text{with } \tau_j := (P^{-1}u_l^1)_j.$$

It remains thus to determine τ_l. For this we shall take advantage of $u_l^0 \cdot \mathscr{R}e u_l^1 = 0$, fact which follows from the property $|u_\hbar(x)| = 1\ \forall\hbar$. This implies

$$\mathscr{R}e\tau_l = 2\mathscr{I}m(S_l^{0'})\sum_{j\neq l}\frac{1}{\lambda_l(x) - \lambda_j(x)}[\tilde{a}_{jl}(x) + (P^{-1}u_l^{0'})_j](u_j^0 \cdot u_l^0).$$

The imaginary part of τ_l can be chosen arbitrary, leading thus to the expression of u_l^1

$$u_l^1(x) = 2\mathbf{i}S_l^{0'}\sum_{j\neq l}\frac{1}{\lambda_l(x) - \lambda_j(x)}[\tilde{a}_{jl}(x) + (P^{-1}u_l^{0'})_j](u_j^0 - (u_j^0 \cdot u_l^0)u_l^0).$$

The numerical results however have shown that this part of the amplitude function increases only the simulation cost, without any gain in the accuracy of the approximate solution. For this reason and to simplify the following calculus, we shall omit u^1 and stop the expansion (2.22) at the zeroth order term u^0.

So far we can write an approximate solution of Eq. (2.19) in the form

$$\Phi(x) = T(x)\xi(x), \tag{2.25}$$

with

$$T(x) = \left(u_1^0(x)\mathrm{e}^{-\int_{x_0}^{x}[\tilde{a}_{1,1}(t)+(P^{-1}u_1^{0'})_1]\,\mathrm{d}t}\Big| \cdots \Big|u_M^0(x)\mathrm{e}^{-\int_{x_0}^{x}[\tilde{a}_{M,M}(t)+(P^{-1}u_M^{0\,\prime})_M]\,\mathrm{d}t}\right),$$

and

$$\xi_j(x) = \begin{cases} \frac{c_j}{\sqrt[4]{\lambda_j(x)}}\mathrm{e}^{\frac{\mathbf{i}}{\hbar}\int_{x_0}^{x}\sqrt{\lambda_j(t)}\,\mathrm{d}t} + \frac{\mathrm{d}_j}{\sqrt[4]{\lambda_j(x)}}\mathrm{e}^{-\frac{\mathbf{i}}{\hbar}\int_{x_0}^{x}\sqrt{\lambda_j(t)}\,\mathrm{d}t} & \text{for } \lambda_j(\cdot) > \delta, \\ \frac{c_j}{\sqrt[4]{|\lambda_j(x)|}}\mathrm{e}^{\frac{1}{\hbar}\int_{x_0}^{x}\sqrt{|\lambda_j(t)|}\,\mathrm{d}t} + \frac{\mathrm{d}_j}{\sqrt[4]{|\lambda_j(x)|}}\mathrm{e}^{-\frac{1}{\hbar}\int_{x_0}^{x}\sqrt{|\lambda_j(t)|}\,\mathrm{d}t} & \text{for } \lambda_j(\cdot) < -\delta, \end{cases} \tag{2.26}$$

where $c_j, d_j \in \mathbb{C}$ are constants and x_0 is a reference point. These expressions are not valid close to a turning point $\lambda_j(x) = 0$. A threshold value $\delta > 0$ was therefore introduced and the above solutions are expressed in intervals far from the turning points. This procedure will require a matching of WKB solutions, which are, each of them, valid in their respective regions.

As you can see from (2.26), we have essentially oscillating solutions in classically allowed regions and exponential solutions in classically forbidden regions. With the help of this formula, we shall construct now basis functions, which incorporate a priori knowledge about the solution. The aim is to create interpolation spaces with better approximation properties than the linear interpolation spaces, usually used in standard finite element methods. For this, let us discretize the interval $[a,b]$ of the transport direction into several intervals $I_n := (x_n, x_{n+1})$. By straightforward algebraic computations, formula (2.26) can be rewritten in I_n in the form

$$\xi_j(x) = w_n^j(x)\xi_j^n + v_n^j(x)\xi_j^{n+1}, \quad \xi_j^n := \xi_j(x_n) \tag{2.27}$$

with

$$w_n^j(x) := \begin{cases} \alpha_n^j(x) f_n^j(x) & \text{for } |\lambda_j| \geqslant \delta, \\ \frac{x_{n+1}-x}{x_{n+1}-x_n} & \text{for } |\lambda_j| < \delta, \end{cases} \qquad v_n^j(x) := \begin{cases} \beta_n^j(x) f_{n+1}^j(x) & \text{for } |\lambda_j| \geqslant \delta, \\ \frac{x-x_n}{x_{n+1}-x_n} & \text{for } |\lambda_j| < \delta, \end{cases}$$

$$\alpha_n^j(x) := \begin{cases} -\frac{\sin \sigma_{n+1}^j(x)}{\sin \gamma_n^j} & \text{for } \lambda_j \geqslant \delta, \\ -\frac{\sinh \sigma_{n+1}^j(x)}{\sinh \gamma_n^j} & \text{for } \lambda_j \leqslant -\delta; \end{cases} \qquad \beta_n^j(x) := \begin{cases} \frac{\sin \sigma_n^j(x)}{\sin \gamma_n^j} & \text{for } \lambda_j \geqslant \delta, \\ \frac{\sinh \sigma_n^j(x)}{\sinh \gamma_n^j} & \text{for } \lambda_j \leqslant -\delta, \end{cases}$$

$$\sigma_n^j(x) := \frac{1}{\hbar}\int_{x_n}^{x} \sqrt{|\lambda_j(t)|}\,\mathrm{d}t; \quad \gamma_n^j := \frac{1}{\hbar}\int_{x_n}^{x_{n+1}} \sqrt{|\lambda_j(t)|}\,\mathrm{d}t; \quad f_n^j(x) := \sqrt[4]{\frac{|\lambda_j(x_n)|}{|\lambda_j(x)|}}.$$

In the neighborhood of a turning point, we have matched the different approximation functions by linear interpolation functions. This can be explained by looking at the pure 1D case, where the second derivative of the wave function φ'' vanishes at a turning point, implying thus a linear evolution of φ close to this point. The functions α_n^j and β_n^j are what we call WKB basis or hat functions. These are highly oscillating functions, with a frequency close to that of the wave function, which is given by

$$\boldsymbol{\lambda}_{\epsilon,i}(x) = \frac{2\pi\hbar}{\sqrt{2m_x(\epsilon - E_i(x))}}.$$

In the limit $\Delta x \ll \boldsymbol{\lambda}$ they reduce to the usual linear interpolation functions. Fig. 4 compares a "WKB hat-function" and a standard linear hat-function.

Thus we posses at this stage an appropriate finite dimensional approximation space. Projecting the wavefunction on this space, by means of (2.25) and (2.27), and choosing a discretization method for the Schrödinger equation, permits to get a scheme for the unknown values ξ^n. For example, a finite element method can be chosen as discretization method, as done in [21]. In the present paper, we discretized the Schrödinger equation by means of the following "finite volume method", deduced by integrating (2.19) over $(x_{i-1/2}, x_{i+1/2})$

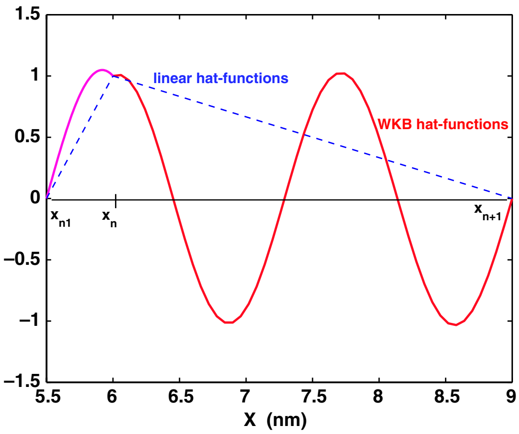

Fig. 4. A WKB hat-function and a standard linear hat-function on an irregular grid.

$$\Phi'(x_{n+1/2}) - \Phi'(x_{n-1/2}) = -2\int_{x_{n-1/2}}^{x_{n+1/2}} A(x)\Phi'(x)\,\mathrm{d}x - \int_{x_{n-1/2}}^{x_{n+1/2}} \left(B(x) + \frac{1}{\hbar^2}C(x)\right)\Phi(x)\,\mathrm{d}x. \tag{2.28}$$

Replacing in this equation Φ and its derivatives by the projections in the WKB approximation space yields a numerical scheme in the unknowns ξ_j^n. Close to the boundary we consider the intervals $[x_1, x_{1+1/2}]$, respectively $[x_{N-1/2}, x_N]$ and include the following boundary conditions into the corresponding expressions

$$\begin{cases} \Phi'(x_1) = D\Phi(x_1) + v_a, \\ \Phi'(x_N) = E\Phi(x_N) + v_b, \end{cases}$$

with

$$D := \begin{pmatrix} -\mathbf{i}k^a_{x,1}\delta_{1\leqslant I_a} + k^a_{x,1}\delta_{1>I_a} & & 0 \\ & \ddots & \\ 0 & & -\mathbf{i}k^a_{x,M}\delta_{M\leqslant I_a} + k^a_{x,M}\delta_{M>I_a} \end{pmatrix}; \quad v_a := 2\mathbf{i}(k^a_{x,i}\alpha^a_{i,\epsilon})_{i=1}^M,$$

and

$$E := \begin{pmatrix} \mathbf{i}k^b_{x,1}\delta_{1\leqslant I_b} - k^b_{x,1}\delta_{1>I_b} & & 0 \\ & \ddots & \\ 0 & & \mathbf{i}k^b_{x,M}\delta_{M\leqslant I_b} - k^b_{x,M}\delta_{M>I_b} \end{pmatrix}; \quad v_b := -2\mathbf{i}(k^b_{x,i}\alpha^b_{i,\epsilon})_{i=1}^M.$$

The resulting block tridiagonal linear system is solved by standard methods.

3. Numerical investigation of the SDM/WKB method

3.1. Comparison of the two methods: SDM and SDM/WKB

In order to show the efficiency of the new method SDM/WKB, an extensive comparison with the SDM method has been performed for the case of a MOSFET with a 5 nm large and 10 nm long channel. An analytic solution of the non-linear Schrödinger–Poisson system (2.1)–(2.4) does not exist. Nevertheless a reference solution is computed by means of a standard finite element method on a very fine mesh ($N_x = 540, N_z = 210$). This solution can be seen as trustful (somehow exact), validated by several papers [18,8]. The efficiency of the SDM method compared to standard methods was shown in [7] and is not the objective of this work. The aim of the present section is to show the power of the new method as compared to the SDM method. The results obtained with both methods, SDM and SDM/WKB, on various fine/coarse meshes are thus investigated according to accuracy (with respect to the reference solution) and simulation time.

In Fig. 5a a plot of the sheet charge density versus x obtained with the SDM/WKB method on a grid of $N_x \times N_z = 13 \times 60$ points (full line) is compared to the analogous plot of the reference solution (broken line). Fig. 5b represents the corresponding cross-sections of the potential energy along the channel, near the Si/SiO$_2$ interface. As one can observe that the agreement between full lines and broken lines is satisfactory, the SDM/WKB method reproduces quite well the macroscopic quantities, as the charge density and the potential energy.

To illustrate more precisely the efficiency of the SDM/WKB method, we plot in Fig. 6 the pointwise relative errors of both methods (with respect to the reference solution) and this for the sheet charge density and the potential energy represented above. The grids chosen for the presented results consist of 13 × 60 points for the SDM/WKB method and 72 × 60 points for SDM. These grids were selected in such a manner to get similar relative errors for the two different methods, more or less 5% in the L^∞ norm, so that we can compare the corresponding simulation times. We observe that for approximately the same precision, the simulation time is significantly reduced with the SDM/WKB method by a factor of about 2 compared to SDM. More details about the gain in the simulation time can be extracted from Table 1. The total simulation time corresponds to the computation of a current–voltage characteristic for a fixed gate potential V_{GS}, that means the Schrödinger–Poisson system is solved several times, for 11 different drain–source potentials $V_{\mathrm{DS}} = 0\ \mathrm{V}, \ldots, 0.5\ \mathrm{V}$.

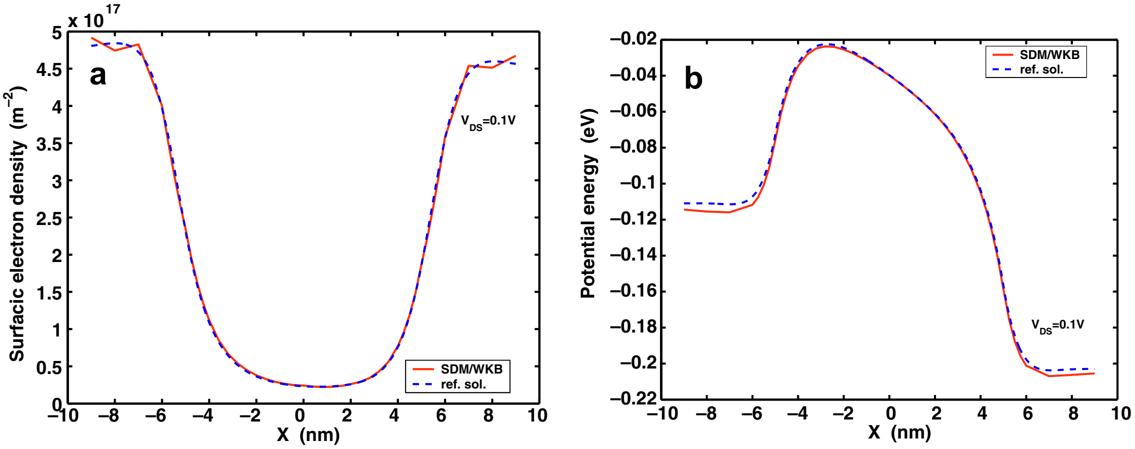

Fig. 5. (a) Plot of the sheet charge density for $V_{\mathrm{GS}} = 0.1$ V and $V_{\mathrm{DS}} = 0.1$ V. Full lines: SDM/WKB (13 × 60 points); broken lines: reference solution. (b) Cross-sections of the potential energy at 1 nm from the interface Si/SiO$_2$ for $V_{\mathrm{GS}} = 0.1$ V and $V_{\mathrm{DS}} = 0.1$ V.

It is of interest to notice that the most time consuming part of the simulation is the resolution of the total 1D Schrödinger equation. This is a consequence of the fact that the 1D longitudinal Schrödinger equation (2.10) has to be solved several times, one for each injection configuration (j_0, r_0, ϵ). Thus it is not surprising that reducing via the WKB approximation the computational time for the resolution of a single Schrödinger equation, leads to a considerable gain in the total simulation time.

Let us now compare the accuracy and the simulation time of the two considered methods for the same mesh. We represent in the following two figures the pointwise relative errors for the surfacic density obtained with the SDM/WKB method (Fig. 7a), respectively, with the SDM method (Fig. 7b) for various meshes, the relative error being computed with respect to the reference solution. Table 2 gives the corresponding total simulation times and mean relative errors for the two methods, dependent on the number of grid points in the x-direction. For example, on a mesh of 13 × 60 points the SDM/WKB method is 8 times more accurate in the L^∞ norm than the SDM method (resp. 5 times more accurate in the L^1 norm), but 3 times less rapid. Generally we can say that for the same mesh, the SDM method is less accurate but faster than the SDM/WKB method.

To conclude, the WKB approximation has been successfully used in a SDM method to further reduce computation time and allow extensive simulation of 2D quantum transport in nanoscale MOSFETs. We performed in this section only a numerical validation of the SDM/WKB method, a physical discussion being

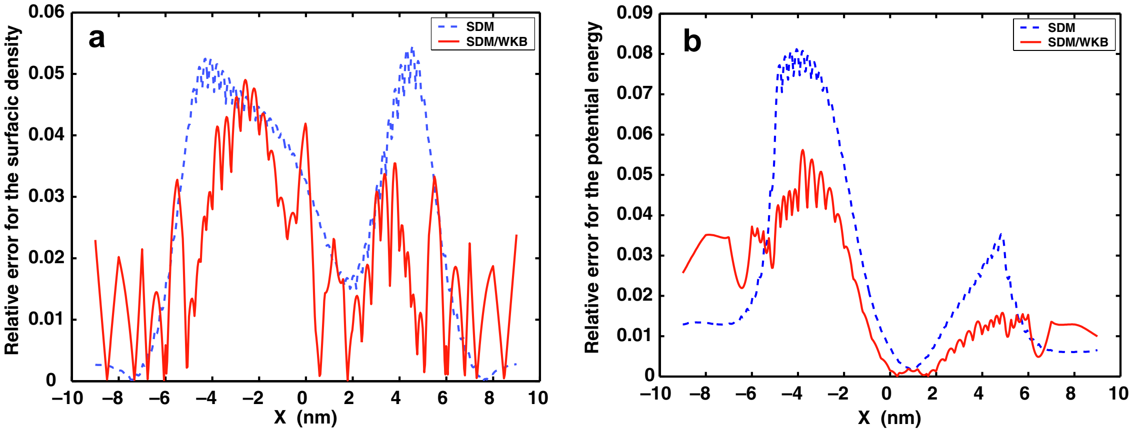

Fig. 6. Relative errors of the SDM (72 × 60 points) resp. SDM/WKB (13 × 60 points) methods with respect to the reference solution, for the (a) sheet charge density, $V_{\mathrm{GS}} = 0.1$ V, $V_{\mathrm{DS}} = 0.1$ V; (b) potential energy at 1 nm from the interface Si/SiO$_2$, $V_{\mathrm{GS}} = 0.1$ V, $V_{\mathrm{DS}} = 0.1$ V.

Table 1
Comparison of the simulation times of SDM and SDM/WKB

	Nbr. grid pts. in x	Total sim. time	Poisson equation	Eigenval. problem	Schrödinger 1D $1 \times (j_0, r_0, k_x)$	Schrödinger 1D total
SDM	72	89 min 32 s	0.04 s	0.01 s	0.05 s	86 s
WKB	13	46 min 28 s	0.02 s	0.01 s	0.03 s	42 s

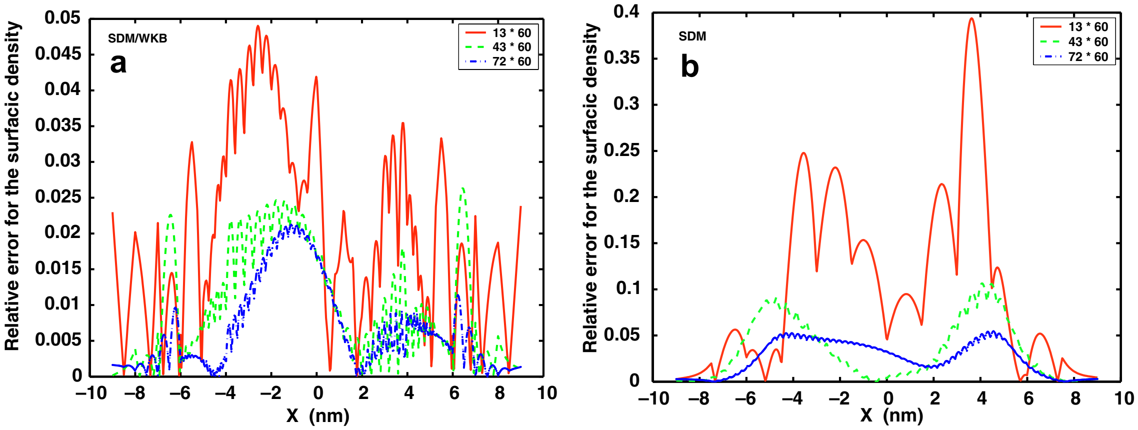

Fig. 7. Relative errors for the sheet charge density obtained with the two methods SDM/WKB (a) and SDM (b) as compared to the reference solution, for different meshes and the potential configuration $V_{GS} = 0.1$ V, $V_{DS} = 0.1$ V.

pursued in the last section. A rigorous numerical analysis of the method is deferred to future work, due to its complexity and importance. The 1D oscillating case is treated in [21], giving an error estimate in terms of the mesh size h and independent on the wavelength λ. The 1D general case is in progress [2] and finally the full 2D case will be the objective of an ulterior paper.

3.2. Some numerical considerations and problems

We shall discuss in this section some points that have to be considered with a certain attention in the implementation of the SDM/WKB method. Other interesting remarks concerning the resolution of the Poisson equation, the linearization of the coupled Schrödinger–Poisson system by means of the Gummel iteration, the initial guess of the electrostatic potential, etc. can be found in [7], since this steps of the algorithm are common to both SDM and SDM/WKB models.

(1) Table 1 shows that a mesh of 13 grid points in the transport direction is used for the resolution of the 1D coupled Schrödinger equations (2.10). However, in order to compute accurately the derivatives of the transversal wave-functions χ_i with respect to x, needed for the computation of the coupling terms (2.11), (2.12) and the current density (2.17), we have to introduce a second mesh in the transport direction, which is much finer. Thus we are let to use two different discretizations of this direction. The eigenvalue problem (2.9) shall be solved for each “parameter” x belonging to the fine grid (43 grid points in our algorithm), whereas the Schrö-

Table 2
Comparison of the total simulation times and mean relative errors of SDM and SDM/WKB for different meshes

$N_x \times N_z$	Total sim. time SDM/WKB	Mean rel. error SDM/WKB	Total sim. time SDM	Mean rel. error SDM
13 × 60	46 min 28 s	0.0203	17 min 21 s	0.1009
43 × 60	155 min 58 s	0.0101	35 min 57 s	0.0339
72 × 60	305 min 3 s	0.0067	89 min 32 s	0.0255

dinger equations (2.10) are solved on the coarser grid (13 grid points). It remains to interpolate the results from the coarse grid to the fine one. For this, we remark that once we have solved the tridiagonal linear system corresponding to the discretization scheme (2.28), in order to deduce the values ξ^n at the coarse grid nodal points, we can determine by means of (2.25) and (2.27) the values of the wave-function and of its derivatives anywhere in the interior of a coarse grid interval. In this manner we obtain the charge density and consequently the electrostatic potential on the fine mesh in order to restart the Schrödinger–Poisson iteration. This "multi-grid" procedure enables to exploit the advantages of the WKB approximation for the resolution of (2.10), which were mentioned in the previous section, without loosing at the same time on precision for other computations, as for example for the derivation of the transversal wave-functions with respect to x. Note, moreover, that the eigenvalue problem (2.9) is solved only few times in a Schrödinger–Poisson loop (two times for a fixed parameter x), in contrast to the 1D Schrödinger equations, which have to be solved several times, one for each electron injection energy. For this reason, the just described procedure does not increase much the computational cost.

(2) The choice of the "multi-grid" strategy in the transport direction, mentioned in the first point, has also another reason. The fine grid is not only necessary for the derivation of the transversal wave-functions χ_i, but also for the regularity with respect to the x-direction of the first two transversal waves $\chi_1^{m_l}$ and $\chi_2^{m_l}$, corresponding to the heavy effective mass m_l in the confinement direction.

Indeed the particularity of the first two subbands associated to m_l is that they are very close (see Fig. 11). Consequently, the computer treats them as a double eigenvalue, such that the corresponding eigenvectors $\chi_1^{m_l}$ and $\chi_2^{m_l}$ are chosen arbitrarily as orthonormal eigenvectors of a 2D eigenspace. These so obtained eigenvectors can be rather different from the correct ones, corresponding to the two separate eigenvalues. In [7] a numerical procedure was introduced to cope with this failure. This procedure consists in the following steps. Starting from a point x_j with given $\chi_1^{m_l}(z;x_j)$ and $\chi_2^{m_l}(z;x_j)$, we compute independently $\widehat{\chi_1^{m_l}}(z;x_{j+1})$ and $\widehat{\chi_2^{m_l}}(z;x_{j+1})$ by solving (2.9) at x_{j+1}. Then we rotate these so obtained eigenvectors in such a manner to get the regular wave-functions $\chi_1^{m_l}$ and $\chi_2^{m_l}$ in the x-direction. In other words, we obtain the new $\chi_1^{m_l}(z;x_{j+1})$ and $\chi_2^{m_l}(z;x_{j+1})$ by the following rotation:

$$\begin{pmatrix} \chi_1(z;x_{j+1}) \\ \chi_2(z;x_{j+1}) \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} \widehat{\chi}_1(z;x_{j+1}) \\ \widehat{\chi}_2(z;x_{j+1}) \end{pmatrix},$$

with the rotation angle θ chosen so that $\chi_1(z;x_{j+1})$ (resp. $\chi_2(z;x_{j+1})$) is close to $\chi_1(z;x_j)$ (resp. $\chi_2(z;x_j)$). This leads to the condition

$$\int_0^1 \chi_1(z;x_{j+1})\chi_2(z;x_j)\,\mathrm{d}z = 0,$$

implying

$$\tan(\theta) = -\frac{\int_0^1 \widehat{\chi}_1(z;x_{j+1})\chi_2(z;x_j)\,\mathrm{d}z}{\int_0^1 \widehat{\chi}_2(z;x_{j+1})\chi_2(z;x_j)\,\mathrm{d}z}.$$

It is now obvious that in order to get a precise rotation angle θ, we need a fine mesh in the x-direction.

Another point to be treated with care in the just recalled numerical procedure is the choice of the starting point x_*. It is convenient to begin at a point x_*, at which the first two eigenvalues $E_1^{m_l}(x_*)$ and $E_2^{m_l}(x_*)$ are the most distant possible. This occurs rather in the middle of the interval, as can be observed in Fig. 11, so that starting from this point, we shall carry out the rotation technique stepwise to the left and to the right.

(3) This point concerns the truncation of the injection energy integral in the calculus of the charge density (2.15) and current density (2.17).The upper limit of the electron injection energy ϵ is fixed at a temperature of $T = 300$ K to $E_{\mathrm{sup}} = E_{\mathrm{F}}^r + 4k_{\mathrm{B}}T$ for gate potentials $V_{\mathrm{GS}} \geqslant 0$ V. Here we denoted by E_{F}^r the Fermi level of the source reservoir if $r = a$ or drain reservoir if $r = b$. In the case of $V_{\mathrm{GS}} \geqslant 0$ V, the number of particles injected with an energy superior to E_{sup} is very small compared to the number of particles contributing to the electron current and can thus be neglected. This comes from the fact, that the Fermi–Dirac statistic is very small if $E - E_{\mathrm{F}} > 4k_{\mathrm{B}}T$. Indeed, we have for $E - E_{\mathrm{F}} > 4k_{\mathrm{B}}T$

Table 3
Upper electron injection energy limits for different gate potentials

V_{GS}	0 V	−0.1 V	−0.2 V	−0.3 V
E_{sup}	$E_{\mathrm{F}}^{r}+4k_{\mathrm{B}}T$	$E_{\mathrm{F}}^{r}+6k_{\mathrm{B}}T$	$E_{\mathrm{F}}^{r}+8k_{\mathrm{B}}T$	$E_{\mathrm{F}}^{r}+10k_{\mathrm{B}}T$

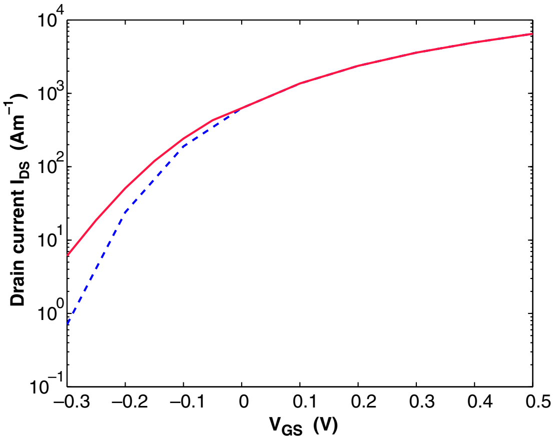

Fig. 8. Current versus gate potential characteristics in log. scale, for $V_{\mathrm{DS}}=0.5$ V. Full line: The maximal injection energy E_{sup} is rectified with respect to V_{GS} as given in Table 3. Broken line: E_{sup} is fixed to $E_{\mathrm{F}}^{r}+4k_{\mathrm{B}}T$.

$$f_{\mathrm{FD}}(E)=\frac{1}{1+\exp\frac{E-E_{\mathrm{F}}}{k_{\mathrm{B}}T}}\leqslant\frac{1}{1+e^{4}}\sim e^{-4}=1.8\times10^{-2}\ll1.$$

However, for gate potentials inferior to 0 V the number of particles contributing to the electron transport by thermionic emission or quantum tunneling is rather small so that the contribution of the particles injected with an energy superior to $E_{\mathrm{F}}^{r}+4k_{\mathrm{B}}T$ becomes significant. The size of the injection energy interval has thus to be adjusted to the gate potential and it is necessary to consider larger injection energy intervals for $V_{\mathrm{GS}}<0$ V. Numerically, we determined the upper energy limit for a fixed $V_{\mathrm{GS}}<0$ V, by comparing the obtained currents (at $V_{\mathrm{DS}}=0.5$ V) for different E_{sup}. The upper limit E_{sup} is increased until the changes in the current become negligible (relative error of 5%). The values determined by this manner are shown in Table 3.

As an example, Fig. 8 illustrates the current versus gate potential characteristic obtained with a fixed maximal injection energy of $E_{\mathrm{sup}}=E_{\mathrm{F}}^{r}+4k_{\mathrm{B}}T$ (broken line) and that one obtained with the values of Table 3 (full line). Remark that the difference between the two characteristics is significant.

It is important to remark that increasing the maximal electron injection energy, one has to verify if the number of subbands taken into consideration is large enough. Tests have shown that the choice of $M=12$ subbands is sufficient even for $E_{\mathrm{sup}}=E_{\mathrm{F}}^{r}+10k_{\mathrm{B}}T$.

4. Discussion of the numerical results obtained with SDM/WKB

The aim of this last section is to analyze from a physical point of view the numerical results obtained with SDM/WKB. Experimental results are at present not available for such ultra-small devices of 10 nm channel length. Advances in lithography permitted to shrunken down the channel dimensions to 20–30 nm, but there is still some work to do. Comparable simulation results were generated by means of other methods, like the standard full 2D variational methods [18,8], the SDM method [7] or the Green's function approach [24,27].

4.1. Parameters for the considered MOSFET

The various parameters describing the structure of the simulated double gate MOSFET, illustrated in Fig. 9, are presented in Table 4.

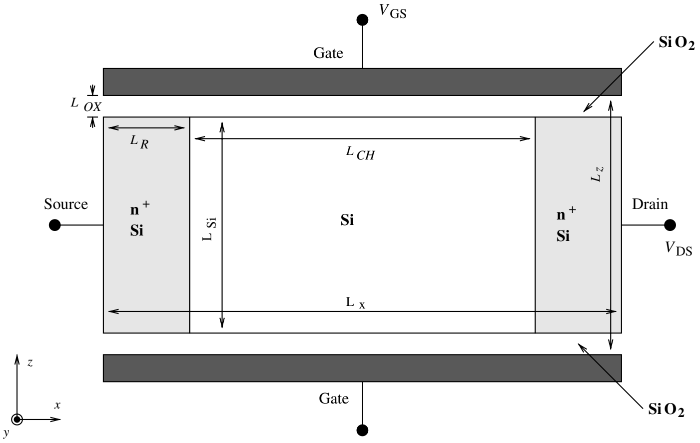

Fig. 9. The DG NMOSFET.

Table 4
Specifications of the modeled device

Parameter	Value	Parameter	Value
L_x	18 nm	m_l	$0.98 \times m_e$
L_z	12 nm, 7 nm, 5 nm	m_t	$0.19 \times m_e$
L_{OX}	1 nm	T	300 K
L_{Si}	10 nm, 5 nm, 3 nm	n^+	$10^{20}\ \text{cm}^{-3}$
L_R	4 nm	V_{GS}	$-0.3, \ldots, 0.5$ eV
L_{CH}	10 nm	V_{DS}	$0, \ldots, 0.5$ eV
m_{SiO_2}	$0.5 \times m_e$	E_C	3.15 eV

4.2. Analysis of the 10 nm × 10 nm MOSFET

We present in this section the essential simulation results of the 10 nm × 10 nm MOSFET transistor, obtained with the SDM/WKB model.

In the ballistic regime, the MOSFET schematically operates as follows (Fig. 10): Electrons are injected from the two thermal equilibrium reservoirs (the source and the drain) across a potential energy barrier. Electrons having an energy higher than this energy barrier are transmitted from the source to the drain or inversely by thermionic emission, whereas electrons with a smaller energy can pass only by quantum mechanical tunneling. The gate voltage modulates the height of this potential barrier.

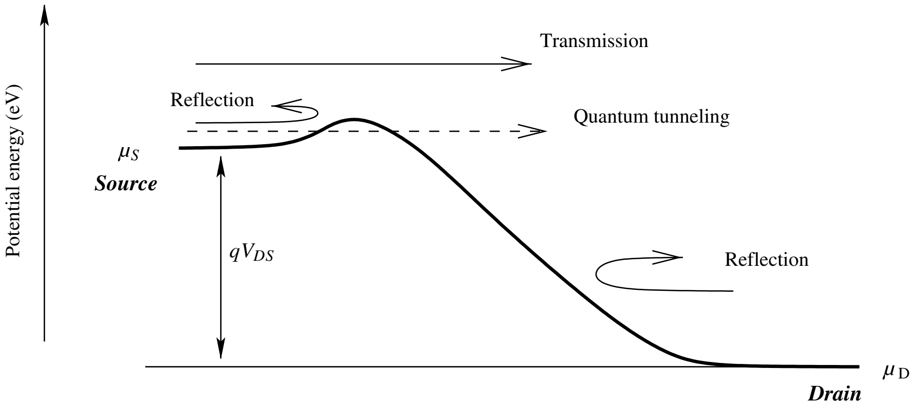

Fig. 10. Schematic mechanism of a MOSFET.

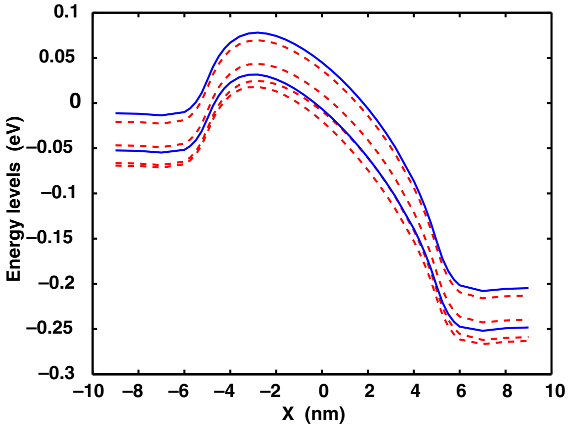

Fig. 11. Conduction energy subbands for $V_{\mathrm{DS}} = 0.2$ V and $V_{\mathrm{GS}} = 0$ V. The full lines correspond to the first two energy subbands associated to m_{t} (primed subbands), whereas the broken lines are the first four energy subbands associated to m_{l} (unprimed subbands).

Due to the anisotropic nature of the conduction band of Si, the energy discretization in the confined direction results in two sets of subbands, associated to two different effective mass values. The lower subbands (and thus firstly populated) are associated with the heavy longitudinal effective mass m_{l} in the confined direction z, and the higher ones are associated with the lighter transversal mass m_{t} in direction z (Fig. 11).

Note that the subbands are closely spaced if the Silicon film is thick, and as such several subbands have to be taken into account in the model, in order to simulate correctly the electron transfer within the device. In particular, the first two subbands associated to m_{l} are very close, fact which requires a slightly different numerical procedure for the determination of the corresponding transversal waves $\chi_1^{m_{\mathrm{l}}}, \chi_2^{m_{\mathrm{l}}}$ (see Section 3 and [7]). Fig. 12 illustrates the profile of the first m_{t}-subband for different drain–source, respectively, different gate voltages.

Fig. 13 represents the electron density and potential energy of the examined MOSFET ($V_{\mathrm{DS}} = 0.5$ V, $V_{\mathrm{GS}} = 0.1$ V). The energy potential barrier between source and channel is visible. With a 10 nm Si thickness, the lower potential energy near the Si/SiO$_2$ interface induces the formation of two inversion layers, where the electron density is more raised. The gate potential V_{GS} modulates the height of this barrier and thus the number of free electrons in the channel. For transistors with ultrathin Si-bodies (<3 nm), the two inversion layers merge to a single one, as it will be discussed in Section 4.3.

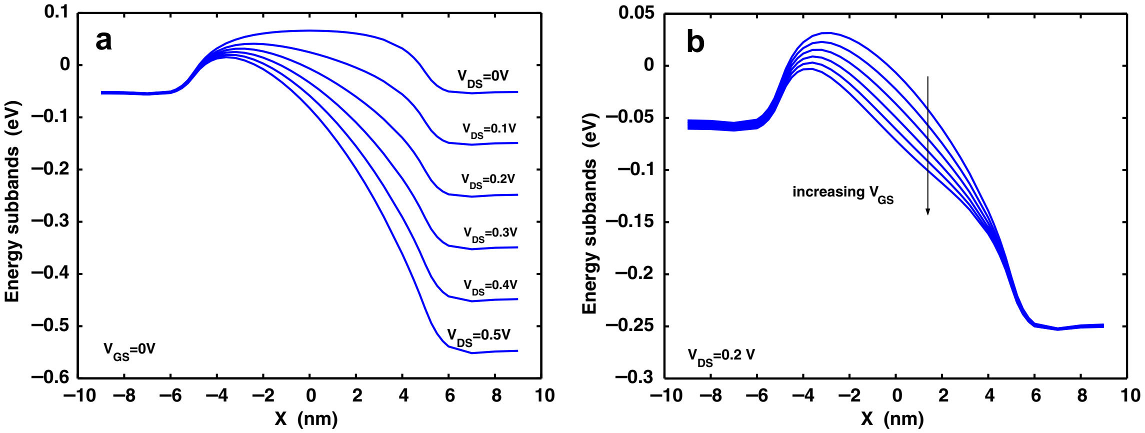

Fig. 12. Subband deformation with the variation of the drain–source voltage (a) respectively gate voltage (b). Represented is the first m_{t}-subband.

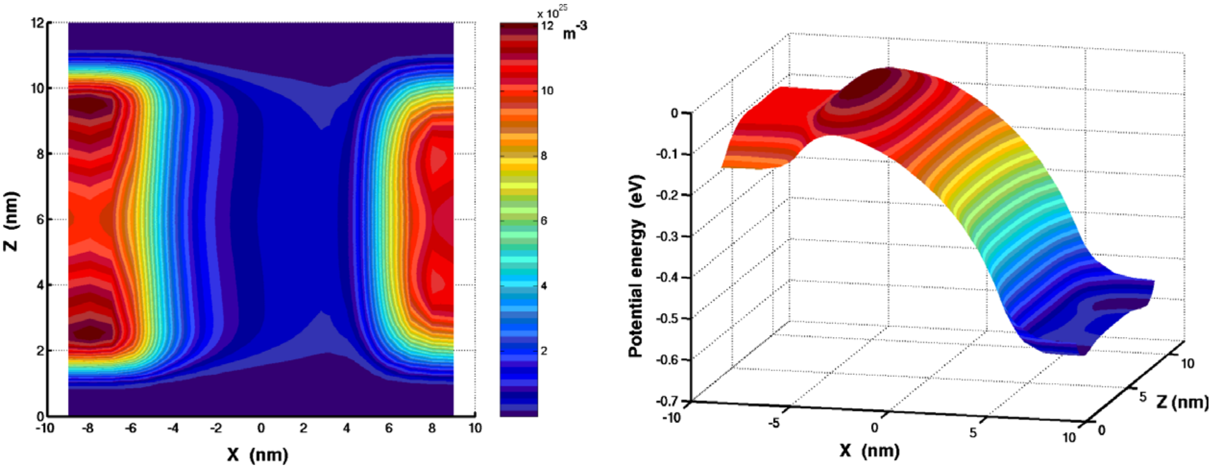

Fig. 13. Electron density (left) and potential energy (right) for $V_{\mathrm{DS}} = 0.5$ V and $V_{\mathrm{GS}} = 0.1$ V.

Fig. 14 shows cross-sections of the charge density and potential energy for various values of the drain–source potential and for a fixed gate potential. Remark the influence of the drain–source potential on the length of the potential barrier. As the drain bias voltage V_{DS} increases, the source barrier shrinks not only in height, measured by the DIBL (drain induced barrier lowering), but also in length. This implies a higher contribution of the tunneling charge to the drain current I_{DS}.

The current density is presented in Fig. 15. Note that the existence of the two channels close to the device/oxide interface (double channel conduction) (Fig. 15 (right)). In Fig. 15 (left), which differs from the right figure by the value of the gate potential, one can observe the pinch-off point near the end of the channel. At this point the electrostatic potential equals the saturation potential $V_{\mathrm{sat}} = V_{\mathrm{GS}} - V_{\mathrm{T}}$, where V_{T} is the threshold voltage. In the region beyond this point conduction is moved from the interfaces to the volume of the film. With the increase of the drain bias V_{DS}, the pinch-off point moves towards the source.

It has been checked that the current given by

$$I_{\mathrm{DS}}(x) = \int_0^1 j(x,z) \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} \mathrm{d}z$$

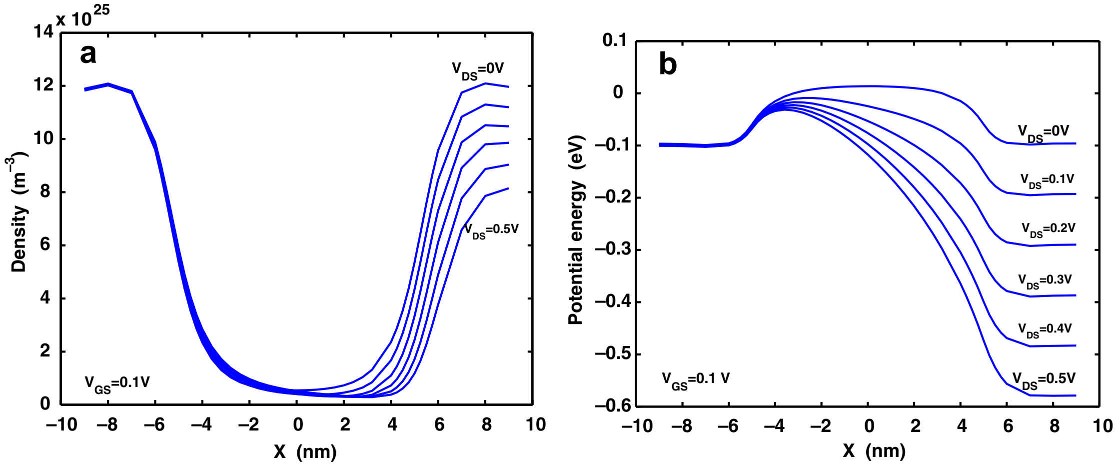

Fig. 14. Cross-sections of the electron density (a) respectively potential energy (b) at 1 nm from the interface Si/SiO$_2$ for $V_{\mathrm{GS}} = 0.1$ V and various drain–source potentials V_{DS}.

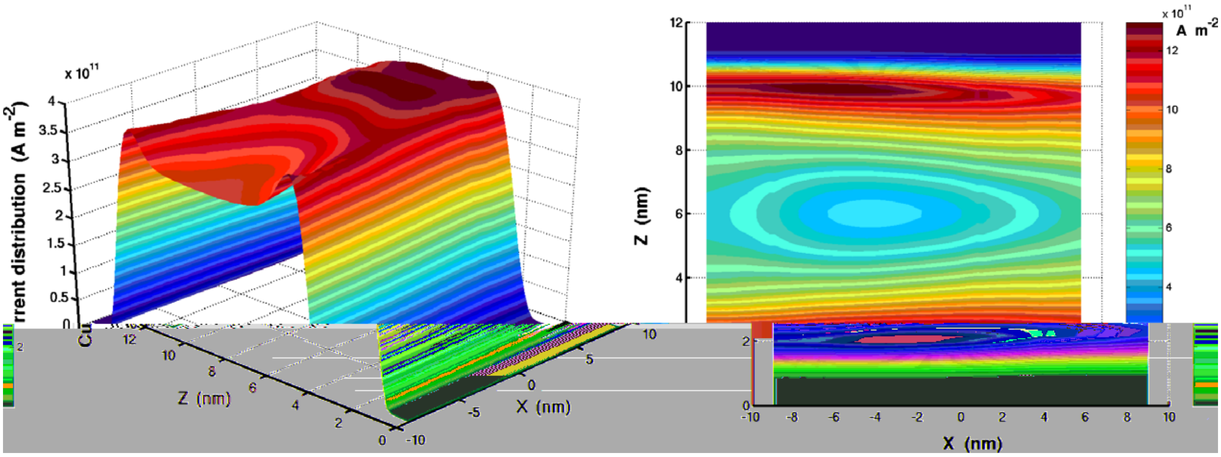

Fig. 15. Distribution of the current density for $V_{\mathrm{GS}} = 0.1$ V and $V_{\mathrm{DS}} = 0.5$ V (left) respectively $V_{\mathrm{GS}} = 0.5$ V and $V_{\mathrm{DS}} = 0.5$ V (right).

is constant along the MOSFET. *I–V* characteristics are shown in Fig. 16. As expected with an aspect ratio of 1 between channel length and film thickness, rather large short channel effects are observed (large DIBL and low $g_{\mathrm{m}}/g_{\mathrm{d}}$). As the channel length of the modeled device is rather short (<1 μm), the drain current I_{DS} increases slowly with I_{DS} beyond the saturation value I_{sat}, such that no saturation can be observed.

The repartition of the charge density, average velocity (Fig. 17) and current density (Fig. 18) between the three valley orientations shows interesting features as both the spatial variations and the amplitudes are affected by the valley orientations.

Electrons occupying the lower subbands (configuration $m_{\mathrm{l}z}$: electrons respond with the longitudinal effective mass in the direction z) are highly localized against the interface and are much more numerous, whereas the electrons corresponding to the configurations $m_{\mathrm{l}x}$ and $m_{\mathrm{l}y}$ are less confined and in smaller number than the first ones. Due to the different effective masses in the transport direction, the electrons have also different velocities. Electrons characterized by the heavy effective mass in the transport direction (configuration $m_{\mathrm{l}x}$) are less rapid compared to those characterized by the light effective mass in this direction (configurations $m_{\mathrm{l}y}$ and $m_{\mathrm{l}z}$). This results in a different contribution to current, both in absolute value and spatial distribution, for the three valley orientations.

Fig. 19 illustrates how the average velocity increases with the applied drain–source potential V_{DS}.

A 2D plot of the kinetic energy and average velocity (in the transport direction) is presented in Fig. 20.

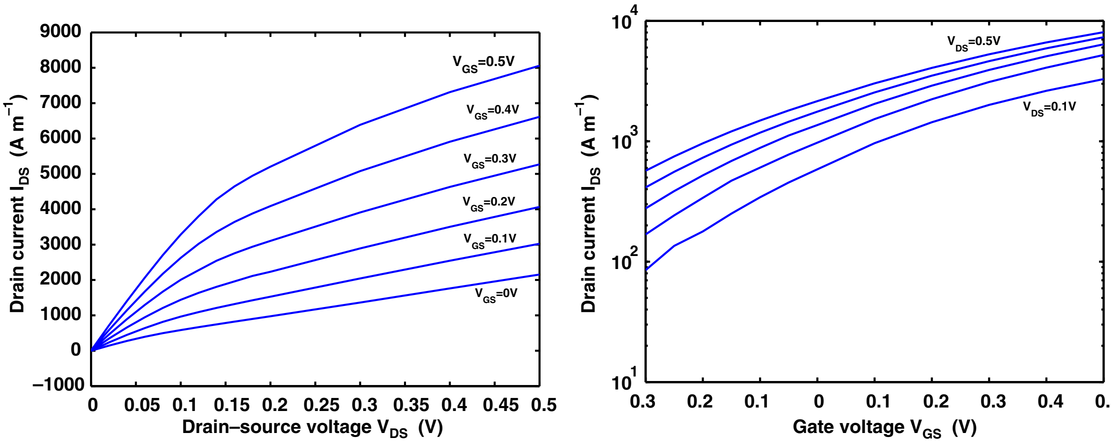

Fig. 16. Current–voltage characteristics. Left: Current versus drain–source potential V_{DS}. Right: Current versus gate potential V_{GS} in logarithmic scale.

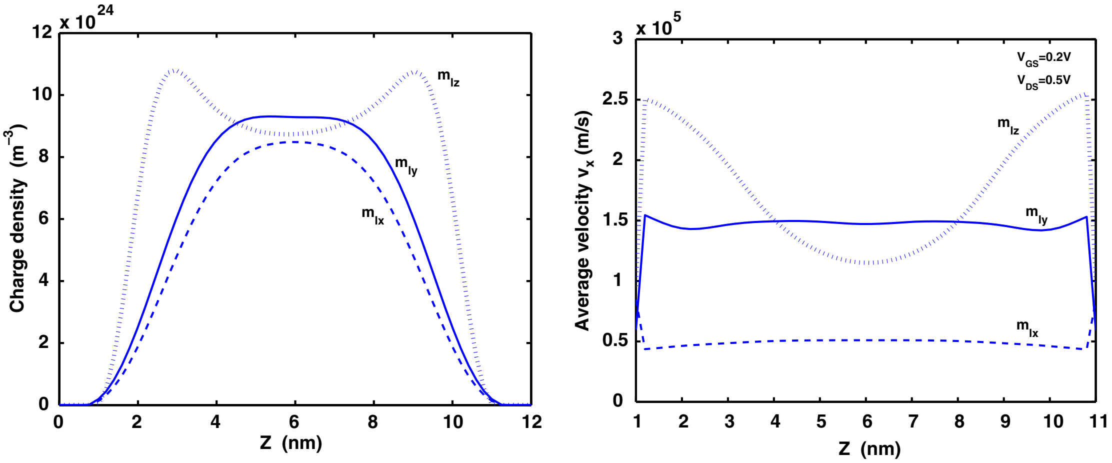

Fig. 17. Cross-sections of the charge density (left) and the x-component of the average velocity $\vec{v}_x$ (right) close to the drain, for $V_{\mathrm{GS}} = 0.2$ V and $V_{\mathrm{DS}} = 0.5$ V. Represented are the plots for the three different mass configurations.

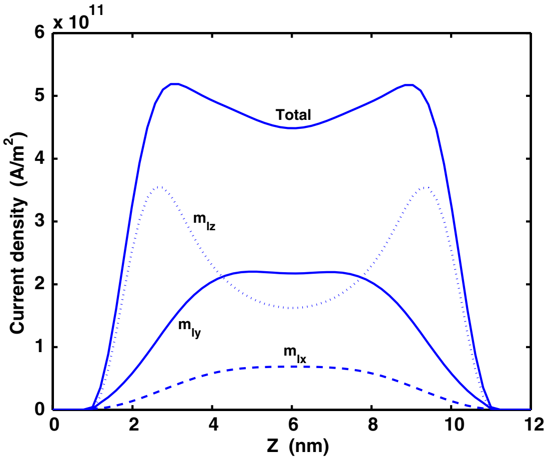

Fig. 18. Cross-sections of the current density $\vec{j}_x$, corresponding to the three different effective mass configurations and taken close to the drain. $V_{\mathrm{GS}} = 0.2$ V; $V_{\mathrm{DS}} = 0.5$ V.

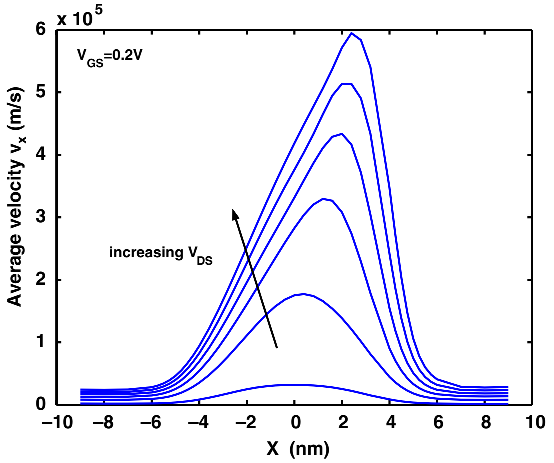

Fig. 19. Profiles of the average velocity $\vec{v}_x$ versus position, for different drain–source voltages and fixed $V_{\mathrm{GS}} = 0.2$ V.

4.3. Influence of the silicon thickness

We will devote this section on the comparison of MOSFET devices with different body thicknesses.

Fig. 21 shows cross-sections of the charge density for three different body thicknesses. As the transistor body gets thinner, the two inversion layers progressively merge into only one inversion region in the volume of the silicon film (volume inversion). Although the peak density gets higher as body thickness decreases, the integrated charge is reduced. This is consistent with the well known threshold shift induced by quantum confinement.

For thinner films, the spacing between energy subbands increases so that fewer subbands need to be simulated. Fig. 22 shows the electron occupancy of the subbands for the three compared devices, where we summed over the three valley orientations. It can be observed that in very thin devices a one subband approximation is sufficiently accurate.

Finally, the influence of body thickness on *I–V* characteristics is shown in Fig. 23.

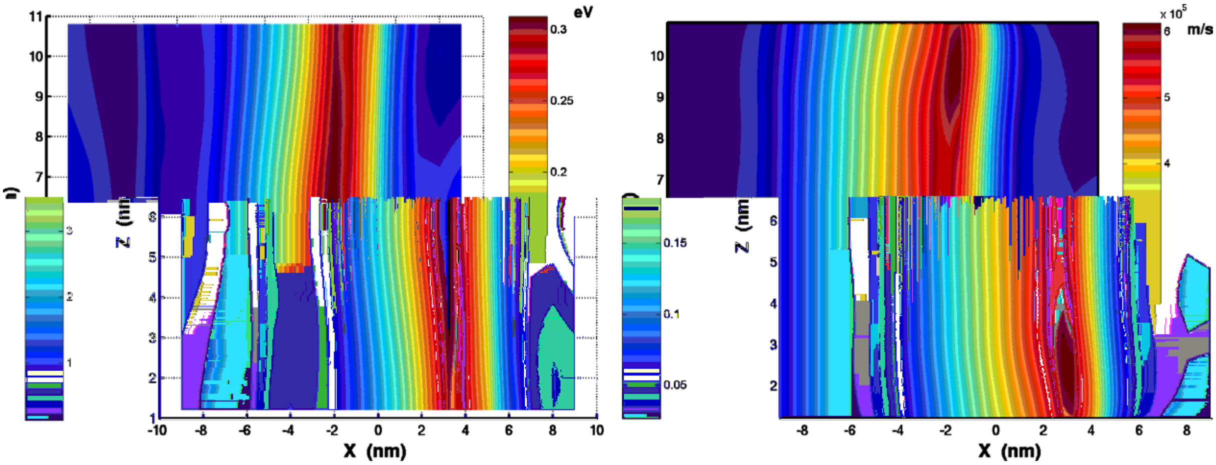

Fig. 20. 2D plots representing the x-component of the electron kinetic energy (left) respectively average velocity (right), for $V_{\mathrm{DS}} = 0.5$ V and $V_{\mathrm{GS}} = 0.1$ V.

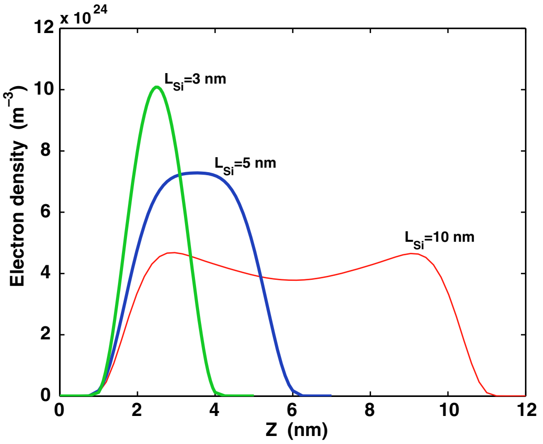

Fig. 21. Cross sections of the electron density in the middle of the device ($X = 0$ nm) for different body thicknesses and for $V_{\mathrm{GS}} = 0.1$ V and $V_{\mathrm{DS}} = 0.1$ V.

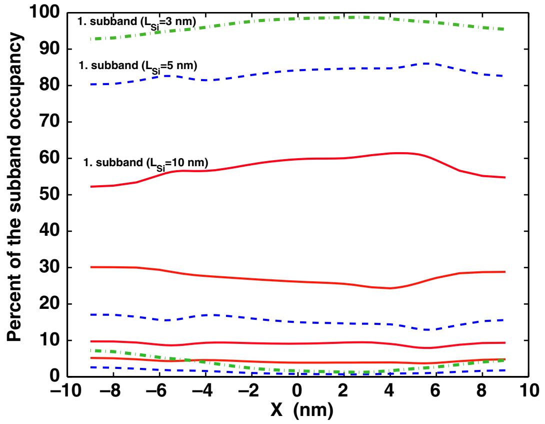

Fig. 22. Percentage of the subband occupancy for the three different body thicknesses. The full lines represent the 10 nm large MOSFET, the broken lines the 5 nm and the broken-dotted lines the 3 nm device. $V_{GS} = 0$ V, $V_{DS} = 0.2$ V.

4.4. The tunneling effect

Analogous to standard full 2D variational methods, the two methods SDM/WKB and SDM have the attribute of accounting for quantum effects in both directions, in the confinement as well as in the transport direction. They are therefore able to analyze tunneling effects between the source and the drain. For instance, Fig. 24 shows how the energy spectrum of the electrons injected from the source in the first unprimed subbands (m_l in the confinement direction) evolves along the channel for a 5 nm × 10 nm device. This spectroscopy has been done for a low V_{GS} bias voltage and medium drain voltage. It shows that the leakage current which flows through the device is strongly influenced by electrons which are tunneling through the source/channel barrier. This tunneling current leads to a larger drain induced barrier lowering (DIBL), as extracted from current voltage characteristics, compared to the internal barrier modulation by drain voltage. In particular, the barrier lowering is of the order of 15 meV when the drain voltage increases from 0.2 V to 0.5 V. This leads, at a temperature of 300 K, to a current increase by a factor of 1.79. The transfer characteristics show however a current variation of a factor of about 2.03. This discrepancy is due to the tunneling of electrons through the source/channel barrier. These tunneling electrons bring a non negligible contribution to

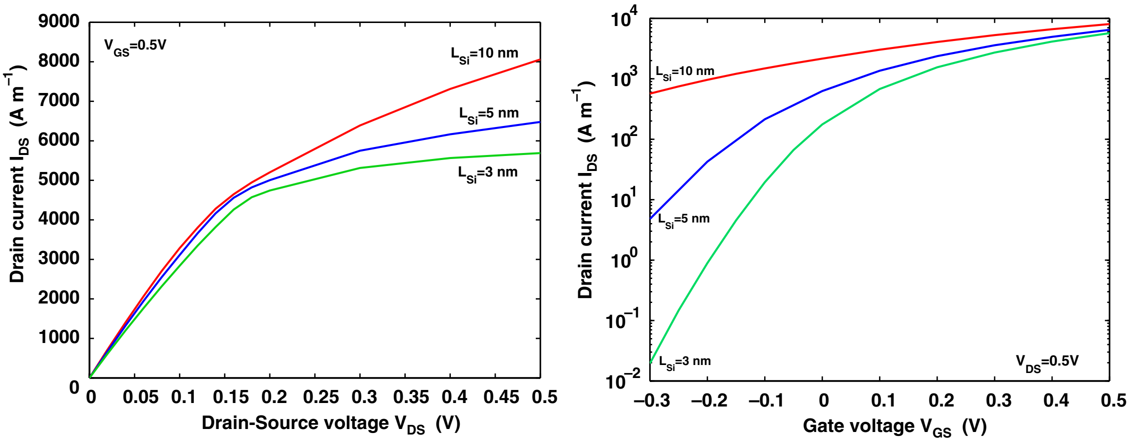

Fig. 23. Comparison of the characteristics for the three different body thicknesses. Left: Current versus drain–source potential V_{DS}, for $V_{GS} = 0.5$ V; Right: Current versus gate potential V_{GS}, for $V_{DS} = 0.5$ V (log. scale).

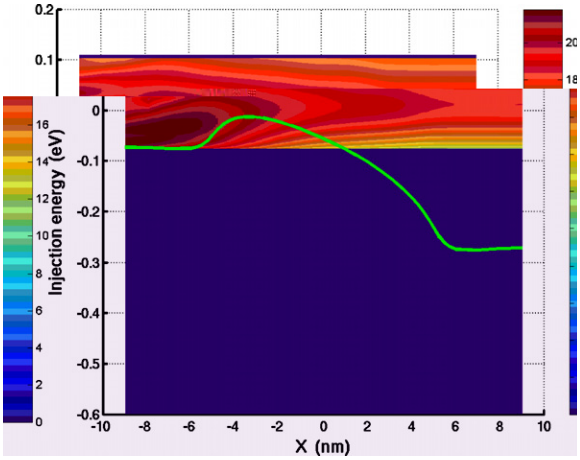

Fig. 24. Source-injected electron population (configuration $m_x = m_l$) in log. scale and the tunneling effect beneath the potential barrier (full line) for $V_{DS} = 0.2$ V and $V_{GS} = 0.1$ V.

drain current compared to electrons injected thermionically above the barrier. Similar tunneling figures can be found in [24].

4.5. A different device design

We present in this section the numerical results obtained with the SDM/WKB method applied on a trapezoidal device design. The aim is to show the flexibility of SDM/WKB and to compare and analyze the performances of different concept devices.

The geometry of the considered MOSFET transistor is represented in Fig. 25. The thickness of the oxide layer is kept constant all along the structure and equal 1 nm. The channel length remains unchanged, equal 10 nm, as for the previous device, whereas the channel width varies from 5 nm at the source side to 3 nm at the drain side. The other parameters are conserved as for the rectangular MOSFET. The difference in the numerical algorithm consists in the fact that the width of the device is no more constant and equal L_z, but depends on the x-coordinate, hence it is denoted by $L_z(x)$. For each point x,

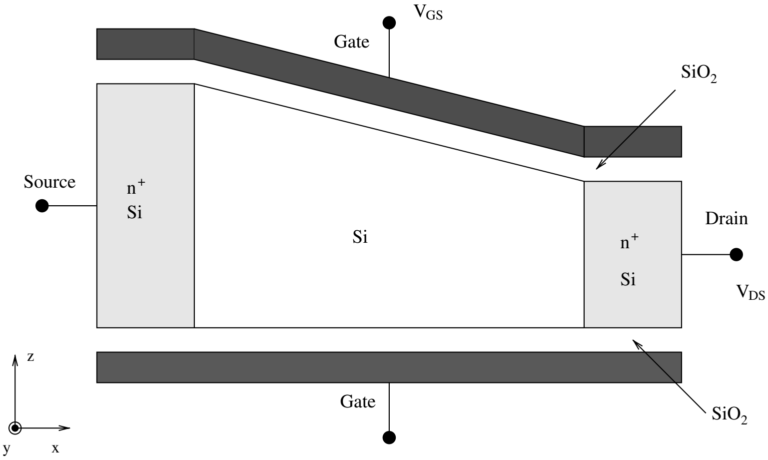

Fig. 25. Schematic representation of the trapezoidal MOSFET device.

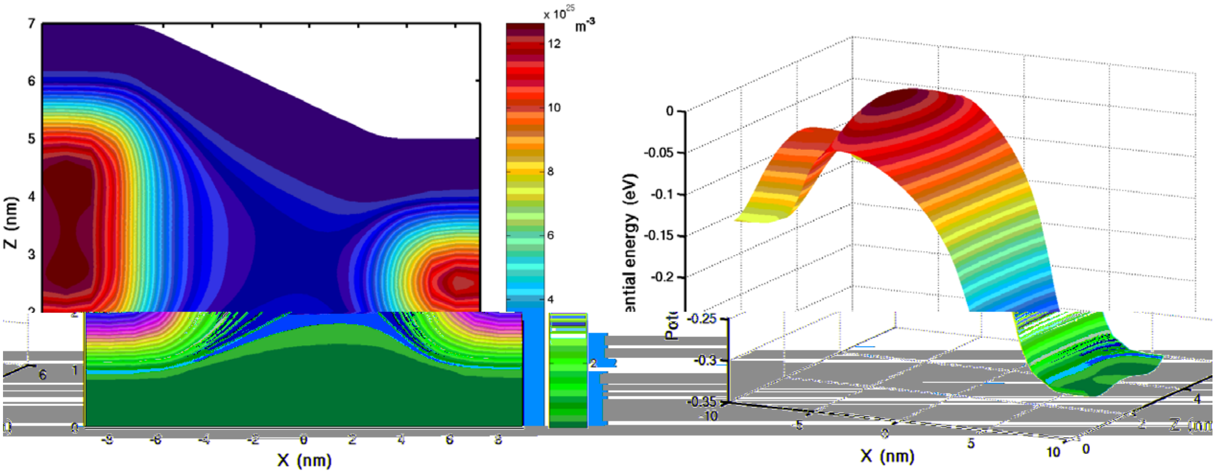

Fig. 26. Electron density (left) and potential energy (right) for $V_{DS} = 0.2$ V and $V_{GS} = 0.1$ V.

the same number of grid points is taken in the confinement direction in the channel, giving thus rise to a non-uniform grid.

Fig. 26 represents the electron density and the potential energy for the just described device. As expected, the two inversion layers, which are formed at the source side, merge towards a single channel at the drain side. Moreover, Fig. 27 (left) illustrates that the energy subbands get more and more distanced as they come closer to the drain. This results in a preference of the electrons for the first subband in the proximity of the drain. In fact, Fig. 27 (right) represents the electron occupancy of the first three subbands. It can be observed that the electrons jump towards the first subband as they come nearer to the drain, that means to the narrower region of the transistor.

Finally, the comparison of the characteristics of the considered trapezoidal device with that ones corresponding to the rectangular MOSFETs of a Si thickness of 5 nm, respectively, 3 nm illustrates that the performances of the trapezoidal device are placed between the two respective rectangular MOSFETs (see Fig. 28).

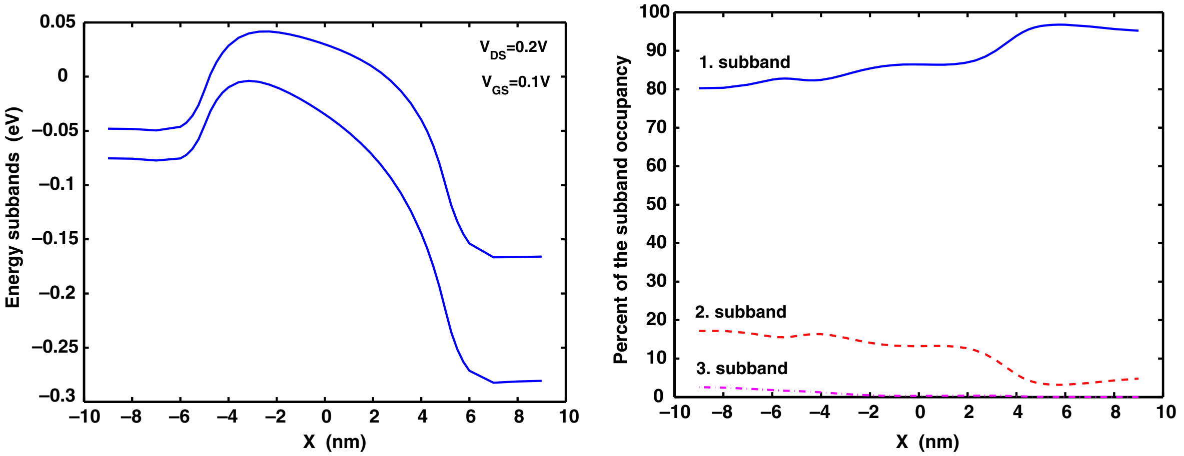

Fig. 27. Left: First two conduction energy subbands associated to m_l, for $V_{DS} = 0.2$ V and $V_{GS} = 0.1$ V. Right: Percentage of the subband occupancy for the trapezoidal MOSFET, for $V_{DS} = 0.2$ V and $V_{GS} = 0.1$ V.

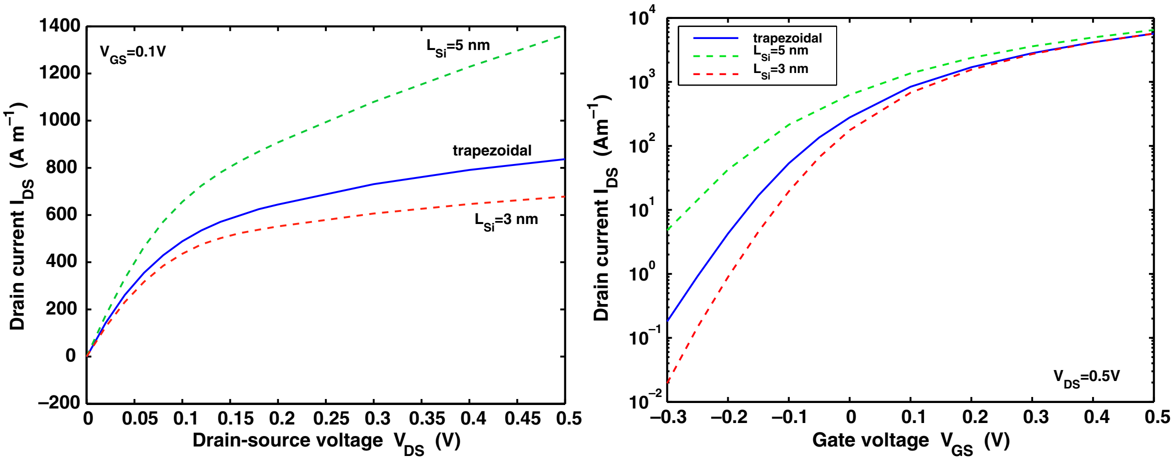

Fig. 28. Comparison of the current versus drain–source voltage characteristics for $V_{\mathrm{GS}} = 0.1$ V (left) and current versus gate voltage characteristics for $V_{\mathrm{DS}} = 0.5$ V (right) for different device designs.

5. Conclusion

A quantum ballistic transport model, which combines two methods, the SDM method and the WKB approximation, was introduced and investigated in this paper. The approach is based on the self-consistent resolution of the Schrödinger–Poisson system with quantum transmitting boundary conditions and was applied for the simulation of nanoscale double-gate MOSFETs. The present study has shown numerically the efficiency of the SDM/WKB method as compared to SDM and to standard methods. Accurate results have been obtained with significantly reduced computational time. A rigorous numerical analysis is postponed to an ulterior work (see [2,21] for the first part).

Acknowledgments

This work has been supported by the HYKE network (Hyperbolic and Kinetic Equations: Asymtotics, Numerics, Analysis), Ref. HPRN-CT-2002-00282, and by the ACI Nouvelles Interfaces des Mathématiques MOQUA (ACINIM 176-2004), funded by the French Ministry of Research.

References

[1] T. Ando, B. Fowler, F. Stern, Electronic properties of two-dimensional systems, Rev. Mod. Phys. 54 (1982) 437–672.
[2] A. Arnold, N. Ben Abdallah, C. Negulescu (in preparation).
[3] G. Bastard, Wave Mechanics Applied to Semiconductor Heterostructures, Les éditions de physique, 1996.
[4] N. Ben Abdallah, On a multidimensional Schrödinger–Poisson scattering model for semiconductors, J. Math. Phys. 41 (3–4) (2000) 4241–4261.
[5] N. Ben Abdallah, P. Degond, P.A. Markowich, On a one-dimensional Schrödinger–Poisson scattering model, ZAMP 48 (1997) 35–55.
[6] N. Ben Abdallah, O. Pinaud, Multiscale simulation of transport in an open quantum system: resonances and WKB interpolation, J. Comput. Phys. 213 (1) (2006) 288–310.
[7] N. Ben Abdallah, E. Polizzi, Subband decomposition approach for the simulation of quantum electron transport in nanostructures, J. Comput. Phys. 202 (1) (2005) 150–180.
[8] N. Ben Abdallah, E. Polizzi, Self-consistent three dimensional models for quantum ballistic transport in open systems, Phys. Rev. B 66 (2002) 245301.
[9] J.-D. Benamou, O. Lafitte, I. Solliec, R. Sentis, A geometric optics method for high-frequency electromagnetic fields computations near fold caustics. II. The energy, J. Comp. Appl. Math. 167 (1) (2004) 91–134.
[10] S. Datta, Electronic Transport in Mesoscopic Systems, Cambridge University Press, Cambridge, UK, 1997.
[11] S. Datta, Nanoscale device modelling: the Green's function method, Superlattices Microstruct. 28 (2000) 253–278.
[12] J.H. Davies, The Physics of Low Dimensional Semiconductors, Cambridge University Press, Cambridge, 1998.

[13] M.V. Fischetti, Theory of electron transport in small semiconductor devices using the Pauli master equation, J. Appl. Phys. 83 (1998) 270–291.

[14] W.R. Frensley, Boundary conditions for open quantum systems driven far from equilibrium, Rev. Mod. Phys. 62 (3) (1990) 745–791.

[15] L. Gosse, N.J. Mauser, Multiphase semiclassical approximation of an electron in a one-dimensional crystalline lattice. III. From ab initio models to WKB for Schrödinger–Poisson, J. Comput. Phys. 211 (1) (2006) 326–346.

[16] G. Klimeck, R. Lake, R.C. Bowen, W.R. Frensley, T. Moise, Quantitative resonant tunneling diode simulation, J. Appl. Phys. 81 (1997) 3207.

[17] R. Lake, G. Klimeck, R.C. Bowen, D. Jovanovic, Single and multi-band modeling of quantum electron transport through layered semiconductor devices, J. Appl. Phys. 81 (1997) 7845.

[18] S.E. Laux, A. Kumar, M.V. Fischetti, Analysis of quantum ballistic electron transport in ultrasmall silicon devices including space-charge and geometric effects, J. Appl. Phys. 95 (10) (2004) 5545–5582.

[19] C.S. Lent, D.J. Kirkner, The quantum transmitting boundary method, J. Appl. Phys. 67 (1990) 6353–6359.

[20] C. Negulescu, N. Ben Abdallah, E. Polizzi, M. Mouis, Simulation schemes in 2D nanoscale MOSFETs: a WKB based method, J. Comp. Elect. 3 (2004) 397–400.

[21] C. Negulescu, Numerical analysis of a multiscale finite element scheme for the resolution of the stationary Schrödinger equation, Numer. Math. (submitted).

[22] F. Nier, A stationary Schrödinger–Poisson system arising from the modelling of electronic devices, Forum Math. 2 (5) (1990) 489–510.

[23] E. Perrey-Debain, O. Laghrouche, P. Bettess, J. Trevelyan, Plane-wave basis finite elements and boundary elements for three-dimensional wave scattering, Philos. Trans. R. Soc. Lond. A 362 (1816) (2004) 561–577.

[24] Z. Ren, R. Venugopal, S. Goasguen, S. Datta, M.S. Lundstrom, NanoMOS 2.0: a two-dimensional simulator for quantum transport in nanoscale MOSFETs, IEEE Trans. Electron. Devices 50 (9) (2003) 1914–1925.

[25] C. Sparber, P.A. Markowich, N.J. Mauser, Wigner functions versus WKB-methods in multivalued geometrical optics, Asymptot. Anal. 33 (2) (2003) 153–187.

[26] A. Svizhenko, M. Anatram, T.R. Govinda, B. Biegel, R. Venugopal, Two dimensional quantum mechanical modeling of nanotransistors, J. Appl. Phys. 91 (2002) 2343.

[27] R. Venugopal, Z. Ren, S. Datta, M.S. Lundstrom, D. Jovanovic, Simulating quantum transport in nanoscale MOSFETs: real vs. mode space approaches, J. Appl. Phys. 92 (2002) 3730–3739.