Study of the Boundary Conditions of the Wigner Function Computed by Solving the Schrödinger Equation

Andrea Savio* and Alain Poncet†
Lyon Institute of Nanotechnology – INSA Lyon
7 avenue Jean Capelle, 69621 Villeurbanne, France
*andrea.savio@insa-lyon.fr †alain.poncet@insa-lyon.fr

***Abstract*—In this work, we compute the Wigner distribution from wavefunctions that are generated by solving the Schrödinger equation. Our goal is to propose an avenue of research that may help better understand certain limitations of deterministic Wigner transport equation solvers, such as negative electron densities or limited charge drops in presence of potential barriers. We evaluate the numerical accuracy required by the Schrödinger solver to compute the Wigner function and compare the performance of an analytic and a numerical solver applied to a constant potential profile, as well as to single- and double-barrier one-dimensional structures. Then, we investigate how the Wigner function boundary conditions vary in these same structures as the contact length increases. We also investigate the range of the wave vector grid required to accurately compute the charge from the Wigner function. Finally, we carry out the same study on biased structures.**

***Keywords*—Quantum transport, Schrödinger equation, Wigner function**

I. Introduction

THe constant drive toward increasing integration densities is pushing the size of electronic devices down, ever closer to the nanometer scale. As an example of this trend, the oxide barrier in ULSI MOS transistors was expected to shrink to a thickness below one nanometer by the 45 nm technology node, prior to the introduction of high-permittivity dielectrics. Another example is the channel length of this same type of transistors, which is expected to drop below 10 nm within the next few technology process generations, according to the current ITRS roadmap [2].

As the size of electronic devices approaches the nanometer scale, quantum phenomena begin to affect the charge carriers' distributions and currents, and TCAD simulators must be capable of accounting for these phenomena in order to model device operation accurately. Instead, current commercial simulation software mostly implements classical models such as the drift-diffusion, thermo- and hydrodynamic ones [3]. These models are derived from the phase-state Boltzmann Transport Equation (BTE) and do not take into account the wave nature of charge carriers. Quantum effects are in general treated only tangentially, to simulate parasitic phenomena such as tunneling currents and confinement levels. In order to be capable of accurately simulating next-generation electronic devices, TCAD software needs to implement full-quantum models. This would enable engineers not only to better predict and characterize parasitic effects in current devices, but also to explore innovative quantum-based designs, such as Resonant Tunneling Diodes (RTD) and quantum dots.

The Schrödinger Equation (SE) is the starting point for a number of approaches that model quantum phenomena. In its one-dimensional (1D) transient form, this equation represents carriers as wavefunctions $\psi(x,k,t)$ of energy $E(k)$, which propagate through a lattice potential energy $U(x)$. x indicates the real space and k the wave vector space. The carriers are given an effective mass $m^* = m_r m_0$, where m_r is the relative mass and m_0 the electron mass in vacuum. The SE is thus given by:

$$i\hbar\frac{\partial}{\partial t}\psi + \frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial x^2}\psi = U\psi \qquad (1)$$

$\hbar$ is the reduced Planck constant. Although the SE can be solved analytically or numerically through a number of different schemes, it remains ill-suited to simulate carrier transport. A major shortcoming is that it is difficult to match the electron wavefunction to measurable physical quantities at the boundaries of a system. It is also problematic to account for parasitic phenomena such as carrier-carrier interactions.

One way to address these shortcomings is to use the Wigner Function (WF) instead of the SE to compute charge densities and currents. The WF is a quasi phase-space distribution function that is obtained by solving the Wigner Transport Equation (WTE), which is itself derived from the SE. The WTE was first studied by Wigner [4], and was implemented numerically much later by Kluksdahl, to simulate quantum tunneling [5], [6], and by Frensley, to study a 1D RTD device [7], [8]. Frensley's implementation uses a first-order differentiation scheme and assumes a constant effective mass across the structure. Higher-order schemes were later studied by Jensen and Buot [9]–[14], while Tsuchiya [15] and Gullapalli [16], [17] applied a varying effective mass. Implementations on RTD devices are also studied by Miller [18] and Wu [19]. Biegel compares various differentiation and self-consistency schemes and applies them to the simulations of RTD devices [20]. Grubin looks at the resolution of the transient WTE [21], while Nedjalkov analyzes the issue of interactions [22]. Yamada studies a 3D mixed self-consistent scheme applied to a silicon nanowire transistor, by solving the SE across the device's cross-section and the WTE along the transport direction, using a differentiation scheme up to the third order [23]. Finally, Kefi-Ferhane simulates a thin 2D MOS transistor by applying a WTE solver along the channel and a Schrödinger solver perpendicularly [24].

In this paper, we discuss a number of issues that we encountered when implementing the 1D deterministic numerical WTE solver described by Frensely. In order to better understand these issues, we study a method to compute the WF directly from the SE, rather than by solving the WTE. We hope that this approach may give us better insight into the nature of the WF and help us in the future in addressing the problems encountered. In this paper, we present some significant initial results, as we look at the WF boundary conditions in unbiased and biased structures and estimate the minimum contact length that has to be applied to a device in a simulation. In addition, we investigate the minimum range that has to be used for the wave vector mesh in order to accurately compute carrier densities from the WF.

II. Derivation of the Wigner equation

The WTE is derived from the SE by calculating the Density Matrix Function (DMF) and then carrying out a variable change and a Fourier Transform (FT). In the formulae that follow, the transient nature of the wavefunction is implied. The DMF $\rho(r,s)$ is derived by correlating the wavefunction on (r,s) couples of points in real space. In the case of a 1D structure with entry and exit contacts (the “Emitter” and “Collector” respectively), this gives [8]:

$$\begin{aligned}\rho(r,s) =& \\ &\frac{2m^*_{\text{Emitter}}k_BT}{h^2}\int_0^\infty \psi(r)\overline{\psi(s)}f_{\text{FD}}(E(k))\,dk \\ +&\frac{2m^*_{\text{Collector}}k_BT}{h^2}\int_{-\infty}^0 \psi(r)\overline{\psi(s)}f_{\text{FD}}(E(k))\,dk\end{aligned} \tag{2}$$

h is the Planck constant, k_B the Boltzmann constant, and T the absolute temperature, which is set to 300 K in all the simulations presented in this work. The first term of the formula represents wavefunctions incident at the emitter, i.e., with a positive wave vector k, while the second represents wavefunctions incident at the collector, i.e., with a negative wave vector. The wavefunctions are weighed by the carrier energy spectrum density, which is given by a Fermi-Dirac Distribution (FDD) $f_{\text{FD}}(k)$ integrated over transverse momenta:

$$f_{\text{FD}}(k) = \ln\left[1+\exp\left(-\frac{E(k)-E_F}{k_BT}\right)\right] \tag{3}$$

E_F is the Fermi energy level at the contact. Assuming a parabolic band, the carrier energy E is given by:

$$E(k) = \frac{\hbar^2k^2}{2m^*_{\text{Contact}}} \tag{4}$$

By applying the SE to the DMF and then carrying out the following variable change:

$$\begin{aligned}r = x+y/2 \quad &, \quad s = x-y/2 \\ u(x,y) = \rho(x&+y/2, x-y/2)\end{aligned} \tag{5}$$

the Liouville - von Neumann Transport Equation (LNTE) is derived [8], [25]:

$$\frac{\partial}{\partial t}u - i\frac{\hbar}{m^*}\frac{\partial}{\partial y}\left(\frac{\partial u}{\partial x}\right) + \frac{i}{\hbar}\left[U\left(x+\frac{y}{2}\right) - U\left(x-\frac{y}{2}\right)\right]u = 0 \tag{6}$$

The WTE is derived by applying a FT to the LNTE [8]:

$$\begin{aligned}\left(\frac{\partial f_W}{\partial t}\right)_{\text{Scattering}} =& \underbrace{\frac{1}{2\pi\hbar}\int_{-\infty}^{\infty}[\delta U(x,k-k')f_W]\,dk'}_{\text{Drift term}} \\ +& \underbrace{\frac{\hbar k}{m^*}\frac{\partial f_W}{\partial x}}_{\text{Diffusion term}} + \underbrace{\frac{\partial f_W}{\partial t}}_{\text{Transient term}}\end{aligned} \tag{7}$$

In this formula, $f_W(x,k,t)$ designates the WF. The formula includes a scattering term that accounts for carrier interactions. Note that the FT transforms the space variable y into the wave vector k. $\delta U(x,k-k')$ is the non-local potential, given by [8]:

$$\delta U(x,k) = 2\int_0^\infty \sin(ky)\left[U\left(x+\frac{y}{2}\right) - U\left(x-\frac{y}{2}\right)\right]dy \tag{8}$$

The WF can be either computed by solving the WTE, or calculated directly from the DMF [4], [8]:

$$f_W(x,k) = \int_{-\infty}^{\infty} e^{-iky}\rho\left(x+\frac{y}{2}, x-\frac{y}{2}\right)dy \tag{9}$$

The charge $n(x)$ can be computed from either the DMF or the WF [8]:

$$n(x) = \rho(x,x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f_W(x,k)\,dk \tag{10}$$

III. WTE implementation issues

At present, commercial simulation software typically deals with quantum parasitic phenomena by applying ad-hoc models to the areas of a device that are most affected. As the device size decreases and these areas become relatively larger, full-quantum simulators might eventually come to replace current classical models. Even after repeated shrinks, however, some regions in a device might still behave classically (e.g., the contacts), and quantum models should therefore be capable of smoothly handling the transition between quantum and classical transport.

In the specific case of WTE solvers, when simulating sufficiently large devices with negligible quantum effects, the values of the charge and the current should be consistent with those obtained by solving the BTE. On smaller devices, as quantum effects begin to appear, the simulated characteristics should be consistent with those yielded by other, comparable quantum models.

Fig. 1 tests these consistency constraints by showing an edge-case classical structure that is sufficiently small to let quantum effects begin to appear. The simulated structure is an abrupt silicon $N^+P^+N^+$ double junction. Each region is 15 nm long; the device is not biased. The figure displays the electron

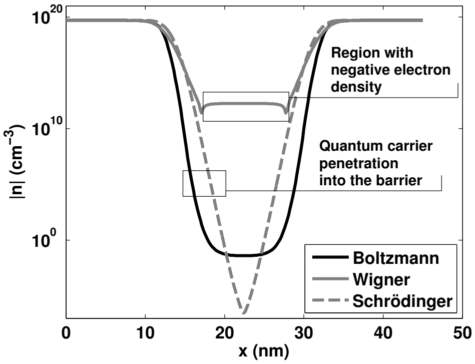


Fig. 1. Absolute-value electron densities obtained on a $N^+P^+N^+$ silicon structure with self-consistent Boltzmann, Wigner and Schrödinger solvers. The dopant profiles are abrupt and each region is 15 nm long. $N^+ = P^+ = 5{\times}10^{19}$ cm^{-3}.

density plots obtained by solving the WTE, the BTE and the Schrödinger equations self-consistently with the same mesh numerical parameters. The details concerning the Schrödinger solver implementation are discussed in Section IV. For the time being, note that this solver should be considered as the one providing the most accurate results when quantum effects are taken into account.

The preponderant classical nature of the simulated device can be seen from the large drop in the electron density (over 20 decades) given by both the BTE and Schrödinger solvers. Quantum effects result in a less abrupt slope in the middle-region valley in the Schrödinger plot compared to the BTE one. This can be explained by taking into account the penetration of the electron wave packet into the P^+ barrier. The slope in the WTE plot is consistent with the Schrödinger one, which indicates that quantum effects are correctly accounted for. However, the WTE plot shows a glaring artifact, as the electron density in the middle of the valley takes negative values, which has no physical sense. Moreover, the minimum electron density in the WTE plot is 10 decades higher than that in the Schrödinger one.

The shape of the WTE plot seems to suggest that the WTE solver cannot handle a drop in the charge by more than a few decades. This is consistent with the literature on the WTE, which mostly presents simulations displaying limited drops in charge and current. For example, the peak/valley current ratio simulated in RTD devices is generally lower than one decade [20], while the nanowire transistor simulated by Yamada has an $I_{\text{On}}/I_{\text{Off}}$ ratio of about 100 [23]. If the WTE solver is indeed accurate only for small variations of the simulated electric macroscopic quantities, it could be problematic to use it to simulate devices with a mixed quantum and classical character.

Investigating these issues with WTE solvers is somewhat problematic, due to memory constraints. Indeed, as the WTE contains an integral term, it is implemented numerically as a block matrix [20], where the number of non-zero coefficients increases with the square of the mesh density in the wave vector space. The rapidly-growing memory footprint thus limits the resolution at which the WF can be calculated, as well as the ability to investigate its properties. By computing the WF from the SE, rather than from the WTE, we are able to reduce this footprint; this enables us to define much denser meshes than those used in a WTE solver, and thus to thoroughly investigate the issue of boundary conditions in single- and double-barrier, classical and quantum 1D structures.

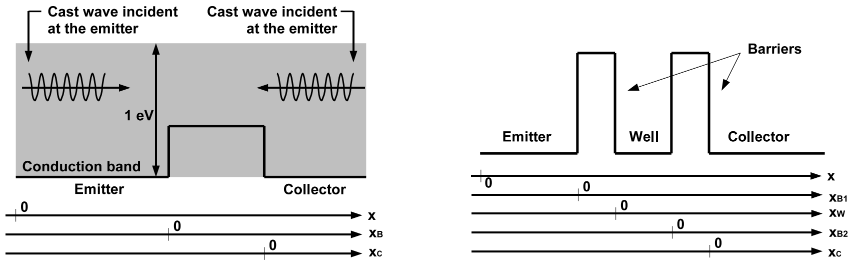
IV. Computation of the WF with a Schrödinger solver in unbiased structures

In order to test whether the WTE solver can accurately compute high charge and current drops, one could try to increase the x- and k-grid resolutions (N_x and N_k respectively). However, memory constraints rapidly limit this technique, as the size of the drift term matrix increases proportionally to $N_x N_k^2$ [20]. In fact, it would be more efficient to apply the solver to only a small region of interest where quantum phenomena take place, e.g., between the two ends of a potential barrier. The memory overhead of meshing the contact regions could thus be avoided. However, this technique poses the problem of what boundary conditions to apply to the solver as, according to Frensley, a FDD can be used only if the boundaries are distant from the quantum region [8]. Moreover, to the best of our knowledge, no study has yet been conducted to evaluate the minimum contact length at which equilibrium conditions can be applied.

In order to investigate the matter of the minimum contact length, we implement a solver that computes the WF directly from the SE. Its ultimate purpose in future studies will be to apply its solution to the WTE as a boundary condition. This section focuses on developing an accurate implementation of this solver, and on assessing its limitations. The Schrödinger solver algorithm is implemented as follows: first, a number of wavefunctions are cast into a given potential profile; then (2) is applied to compute the DMF, and finally (9) is used to calculate the WF. The solver thus works on three different 1D grids to account for the x, y and k variables. N_x, N_y and N_k denote the respective grid resolutions. The details of the numerical implementation, including grid spacing, are discussed further at the end of this section.

The solver has been applied to three potential profiles at zero bias: (a) a constant potential over a length of 50 nm; (b) one with two contacts separated by a 7 nm-thick rectangular barrier and (c) one with two contact regions and two 1 nm-long barriers separated by a 3 nm well. The optimal length of the contact regions is discussed in Section VI. In all three cases, the Fermi level is set equal to the conduction band energy at both contacts. For the constant potential profile, a relative mass of 0.5 is taken. For the other two, two different combinations are tested: (a) a relative mass of 0.5 and a barrier height of 1.5 eV, which are representative of a generic silicon/nitride structure, and (b) a relative mass of 0.067 and a barrier height of 0.3 eV, which are representative of a generic III-V structure. The solver used in this work is not self-consistent, as the purpose of this study is simply to compute the WF from a given potential profile. However, self-consistency can be implemented, as for the plot in Fig. 1.

Fig. 2 illustrates how wavefunctions are computed. Starting at the emitter contact, plane wavefunctions $\psi_{\text{Emitter}}(x,k)$ of the form:



(a) (b)

Fig. 2. Schematic view of the SE solver implementation on a single- (a) and a double-barrier (b) potential profile. Wavefunctions are cast into the structure from the emitter and collector contacts, within an 1 eV-wide range from the conduction band upwards. The x, x_B, x_W and x_C axes, as defined in (11) and (13)-(15), are displayed and their origins are marked.

$$\psi_{\text{Emitter}}(x,k) = e^{ikx} + b_{\text{Emitter}}(k)e^{-ikx} \tag{11}$$

are cast into the device, where k is the positive wave vector and the carrier energy E is given by (4). These wavefunctions are used to compute the first integral term in (2). Each wavefunction contains a normalized incident component with a positive wave vector and a reflected component with a negative wave vector and a complex reflection coefficient $b_{\text{Emitter}}(k)$. For each incident wave vector, the wavefunction is computed at each node on the x-grid through either an analytic or a numerical scheme.

The analytic scheme applies the transfer-matrix method. In short, this method is composed of four steps: first, the structure is divided into separate regions, namely the contacts, the barriers and the well. Then, the wavefunction is calculated symbolically in each region by solving the SE. As the SE is a second-degree differential equation relative to space, its solution in region i is the linear combination of two functions $f_i(x,k)$ and $g_i(x,k)$, and has the form:

$$\psi(x,k) = a_i(k)f_i(x,k) + b_i(k)g_i(x,k) \tag{12}$$

$a_i(k)$ and $b_i(k)$ are the wavefunction coefficients in the region. In the third step, these coefficients are evaluated by setting up a system of two equations at each interface between adjacent regions; these equations express the continuity of the wavefunction and of its first derivative at the interface. When all interfaces are accounted for, solving the overall system yields the wavefunction coefficients in each region. Finally, the wavefunction can be evaluated at all points.

The potential profile in each region needs to be regular enough so that there exists a symbolic solution for the SE. This is possible for instance if the potential is constant, as discussed further in this section, or if it varies linearly, as seen in Section VII; in the latter case, the solutions are given by Airy functions, which can be evaluated to a high precision with appropriate numerical libraries. With a flat or linear potential profile, the wavefunction can be calculated symbolically at all points in the structure, and to evaluate it numerically as a last step.

In the barriers, for a constant potential $U(x) = U_B$ the solutions of the SE take the form:

$$\psi_{\text{Emitter}}(x_B,k) = \begin{cases} a_{\text{Barrier}}(k)e^{ik_Bx_B} + b_{\text{Barrier}}(k)e^{-ik_Bx_B} & E(k) > U_B \\ a_{\text{Barrier}}(k)x_B + b_{\text{Barrier}}(k) & E(k) = U_B \\ a_{\text{Barrier}}(k)e^{k_Bx_B} + b_{\text{Barrier}}(k)e^{-k_Bx_B} & E(k) < U_B \end{cases} \tag{13}$$

In this formula, $k_B = \sqrt{(2m^*|E(k) - U_B|)}/\hbar$ and x_B is the x coordinate with the origin set at the left foot of the barrier. In the case of a double barrier, the solution in the well is:

$$\psi_{\text{Well}}(x_W,k) = a_{\text{Well}}(k)e^{ikx_W} + b_{\text{Well}}(k)e^{-ikx_W} \tag{14}$$

The origin of x_W is set at the left end of the well. Finally, the solution at the collector contact is of the type:

$$\psi_{\text{Collector}}(x_C,k) = a_{\text{Collector}}(k)e^{ikx_C} \tag{15}$$

This wavefunction has no negative wave vector, as no wave is incident at the collector. The origin of x_C is set at the start of the collector region.

The numerical scheme solves the SE by applying the Numerov method [26]. Once again, the process can be divided into four steps. The computation starts at the collector, where the wavefunction has only one component. This component is temporarily normalized, i.e., $a_{\text{Collector}}$ is set equal to 1. This makes it possible to use (15) to evaluate the wavefunction at the first two nodes in the collector region:

$$\begin{gathered} \psi_{\text{Collector Normalized}}(x_C = 0) = 1 \\ \psi_{\text{Collector Normalized}}(x_C = \Delta x) = e^{ik\Delta x} \end{gathered} \tag{16}$$

Then, from these two data points, the Numerov method is applied to compute the wavefunction backwards into the structure. When the emitter contact is reached, the following step consists in evaluating the wavefunction coefficients a_{Emitter} and b_{Emitter} in this region. At the emitter boundary, the function and its first derivative are given by (11) and have the form:

$$\begin{gathered} \psi_{\text{Emitter}}(x = 0) = a_{\text{Emitter}} + b_{\text{Emitter}} \\ \frac{\partial \psi_{\text{Emitter}}}{\partial x}(x = 0) = ika_{\text{Emitter}} - ikb_{\text{Emitter}} \end{gathered} \tag{17}$$

Note that a_{Emitter} is not equal to 1 because the wavefunctions are normalized at the collector instead of the emitter, for the time being. The numerical value of the first derivative is computed by applying a differentiation scheme [26] centered on the emitter boundary node. Having computed both the wavefunction and its derivative, the system (17) can be solved for the two coefficients. Finally, all wavefunction values calculated across the structure are divided by a_{Emitter}, so that they are correctly normalized.

According to Pang [26], the Numerov method can be considered to be of order O(N_x^4). However, its present application limits convergence to the order O(N_x^2), because of the evaluation of the wavefunction derivative at the emitter contact. Once the wavefunction packet incident at the emitter has been evaluated, normalized wavefunctions are similarly cast into the structure from the collector, in order to compute the second integral term in (2).

The numerical implementation of (1), (2) and (9) requires a certain care in order to avoid aliasing of the discrete FT. To be consistent with Frensley's scheme, the condition $\Delta y = 2\Delta x$ is imposed, where Δx and Δy are the x- and y-grid spacings respectively. In this way, the x-grid vertices can be reused for the y-grid. In addition, in (9), the k-grid is implemented symmetrically to $k = 0$, such that [8]:

$$k_i = \{-k_{\text{Max}} + (i + 1/2)\Delta k\}_{i=0..(N-1)} \tag{18}$$

where Δk is the k-grid spacing. The y-grid is defined as:

$$y_i = \{-y_{\text{Max}} + i\Delta y\}_{i=0..(N-1)} \tag{19}$$

k_{Max} and y_{Max} are the half-ranges of the two grids, which have the same number of nodes $N = N_y = N_k$. k_{Max} and y_{Max} are related by:

$$y_{\text{Max}} = \pi/\Delta k = \pi N/(2k_{\text{Max}}) \tag{20}$$

Note that this last condition cannot be fully satisfied, because y_{Max} is rounded to the nearest node on the x-grid. The error on y_{Max} is, however, equal to $\Delta x/2$ at most, i.e., less than 1% for $N > 50$. Also note that, in order to fully mesh the y-grid, the x-grid must be extended by $y_{\text{Max}}/2$ beyond the emitter and collector contacts. The k-grid defined in (18) should be reused in (2) in order to calculate the wave vectors that are cast into the SE. However, we observed that this is not necessary: in [1], we determine that 500 vectors spaced linearly over a 8 eV wide range from the conduction band upwards are sufficient to compute the DMF and WF without significant aliasing. In a subsequent study, we determine that even lower values (250 vectors over a 1 eV range) can be used [27], [28].

V. Comparison of the analytic and numerical solvers

Fig. 3 shows the WF computed on the emitter node for a constant potential profile. For such a profile, (2) and (9) resolve to:

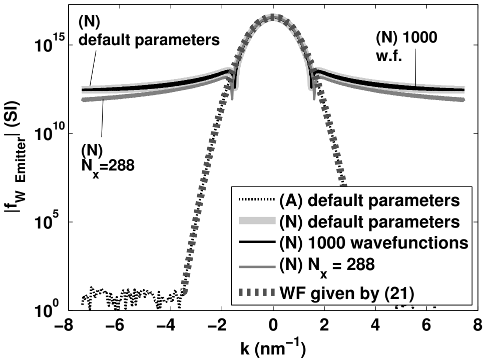


Fig. 3. Absolute WF vs. wave vector plots at $x = 0$, obtained by applying the analytic and numerical SE solver. Default simulation parameters: $m_r = 0.5$, 250 wavefunctions, $k_{\text{Max}} = 10$ nm^{-1}, $N_x = 144$, $N_k = 1000$.

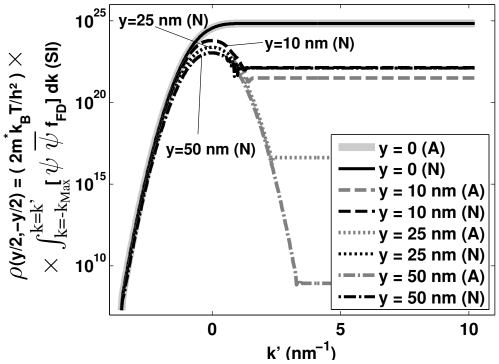
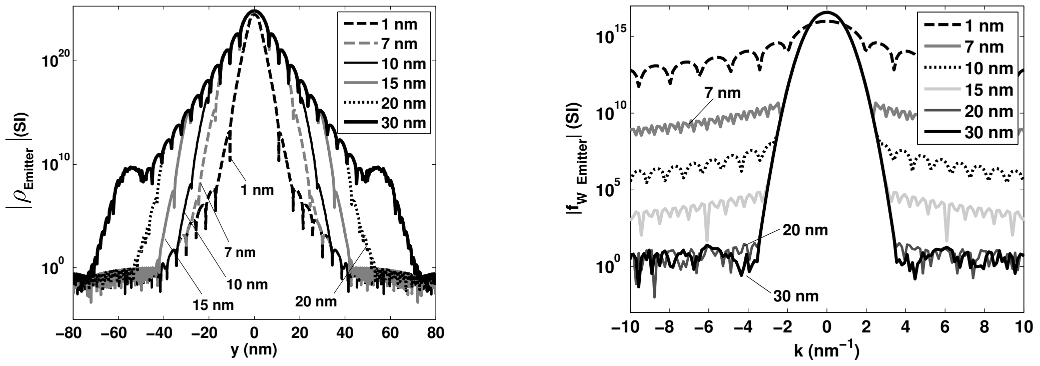


Fig. 4. $\rho(y/2, -y/2)$ integral computed between $-k_{\text{Max}}$ and k' at $x = 0$ for different values of y. The letter A denotes the analytic SE solver, while N denotes the numerical one. Note that beyond $y = 10$ nm the value of the integral for the numerical solver remains constant, as the plots for $y = 10$, 25 and 50 nm coincide. Default simulation parameters: $m_r = 0.5$, 250 wavefunctions, $k_{\text{Max}} = 10$ nm^{-1}, $N_x = 144$, $N_k = 1000$.

$$\begin{aligned} \psi(x,k) &= e^{ikx} \\ \rho\left(x+\tfrac{y}{2}, x-\tfrac{y}{2}\right) &= \frac{2m^* k_B T}{h^2}\int_{-\infty}^{\infty} e^{iky} f_{\text{FD}}(E(k))\, dk \\ f_W(x,k) &= \frac{m^* k_B T}{\pi\hbar^2} f_{\text{FD}}(E(k)) \end{aligned} \tag{21}$$

The first two plots in the figure are generated by casting 250 wavefunctions. A 1000-point k-grid is then used to calculate the WF. The first plot is obtained from the analytic SE solver. Consistently with (21), the WF is proportional to a FDD over a range of 15 decades, i.e., within machine precision of the IEEE 754 double data types used in the computations. This result confirms that FT aliasing is negligible, even though the number of wavefunctions is only half the k-grid resolution.

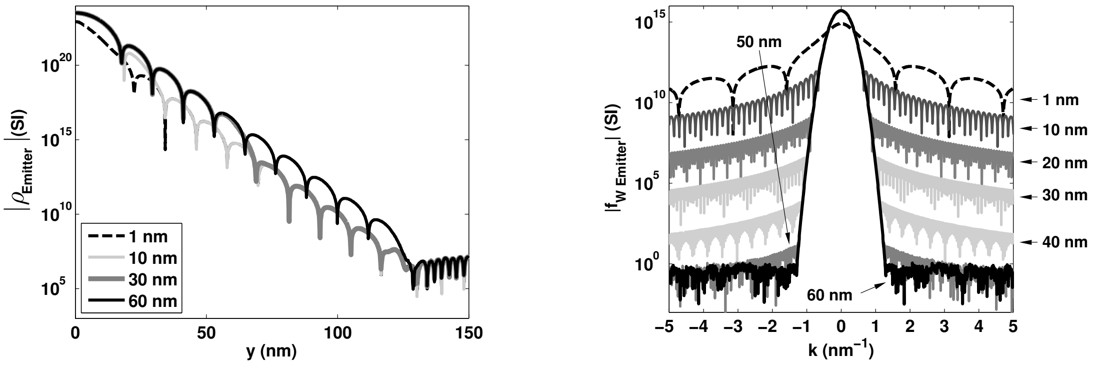
Three plots in the figure show a much smaller drop: these are obtained from the numerical SE solver. Because aliasing is negligible, the lobe-like artifacts observed in these plots can only be caused by numerical error. The figure shows that doubling the number of wavefunctions or the x-grid resolution does not significantly reduce this error. Note however that the lobes have only a minor impact on the charge, as they



(a)

(b)

Fig. 5. Absolute-value DMF (a) and WF (b) computed at the emitter contact of a silicon-based single-barrier structure with a varying contact length. For a contact length between 20 and 30 nm, the WF is equal within machine precision to that obtained with a constant potential profile.



(a)

(b)

Fig. 6. Absolute-value DMF (a) and WF (b) computed at the emitter contact of a III-V-based single-barrier structure with a varying contact length. For a contact length between 50 and 60 nm, the WF is equal within machine precision to that obtained with a constant potential profile.

begin to separate from the central FDD peak about 4 decades below the WF maximum. The charges given by the analytic and numerical solvers are thus consistent within 1%.

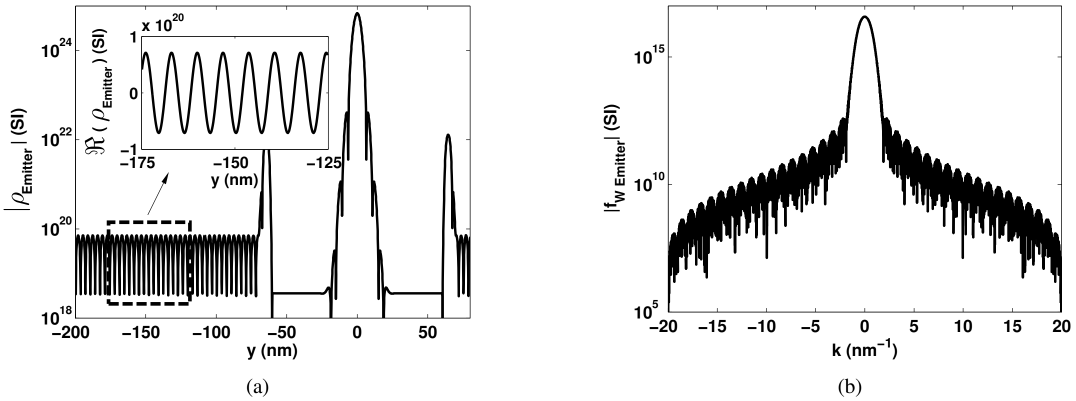
Fig. 4 provides some insight into the nature of the numerical error, by highlighting the computation of the DMF $\rho(y/2, -y/2)$ at $x = 0$. The different plots show how the value of the partially-computed ρ integral varies as the upper bound k' increases up to its maximum value k_{Max}. The plots obtained from the numerical and analytic solvers are compared. For $y > 10$ nm, the value of the integral peaks at $k' = 0$ and then drops as k' increases. For $y = 50$ nm, the drop for the analytic solver spans about 15 decades. By looking at this specific plot, one realizes that the integral is subject to a variation of 15 orders of magnitude as it is computed. This means that, in order for the DMF to be calculated accurately, the integrand, i.e., the wavefunctions, must be evaluated to a relative accuracy of the same order.

The analytic solver is shown to be capable of this level of accuracy, as the three plots obtained for $y > 10$ nm drop to clearly distinct values. On the contrary, the numerical solver is not, because the same three plots are indistinguishable once they drop by only two decades below their peak. In the case of $y = 50$ nm, the accuracy of the numerical solver should be improved by more than 10 orders of magnitude to match the analytic one. Because of the insufficient accuracy of the numerical solver, all simulations presented in the following sections are based on the analytic one.

VI. Study of the minimum contact length and of the wave vector grid range in unbiased structures

The first part of this section presents a study of the length of the emitter and collector contact regions. Its purpose is to determine the minimum contact length where the WF at the boundaries is equal to the equilibrium FDD within machine precision. The second part discusses the range of the wave vector grid that has to be applied to the WF in order to accurately evaluate the charge densities at all points in a given structure.

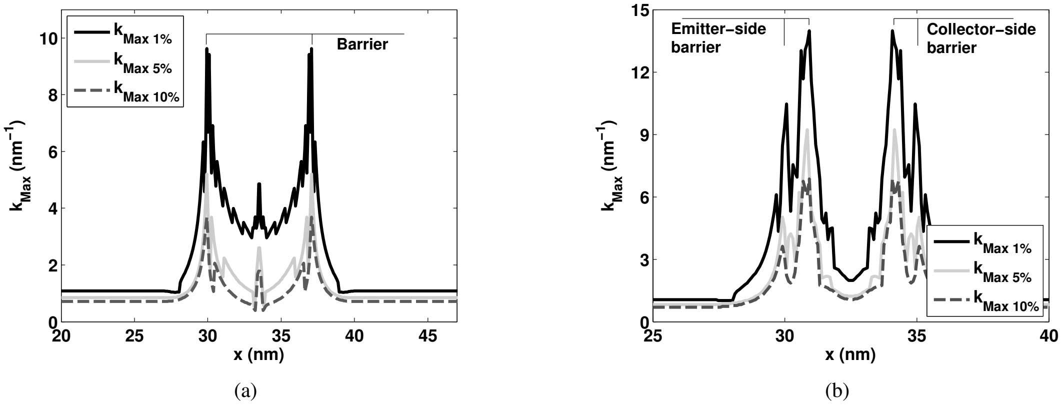
Fig. 5 shows the DMF and the WF computed at the emitter contact of a silicon-based single-barrier structure with varying contact lengths. The plots show that a minimum length between 20 and 30 nm should be used. Fig. 6 shows similar plots on a III-V based structure. In this case, the WF at the boundary converges to a FDD profile for a contact length of



(a)

(b)

Fig. 7. Absolute-value DMF (a) and WF (b) computed at the emitter contact of a silicon-based double-barrier structure with a contact length of 30 nm.



(a)

(b)

Fig. 8. Minimum wave vector grid half-range, as defined in (18), required to compute the charge within a threshold error of 1, 5 and 10% in silicon-based single- (a) and double-barrier (b) structures. The contact length is 30 nm. As k_{Max} is almost constant in the contact regions, these are left out of the figures.

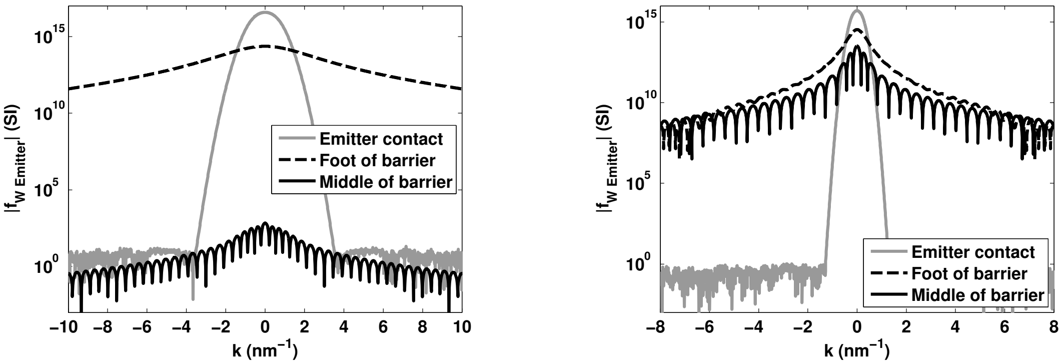
about 60 nm. As we explain in [27], this is due to the smaller relative mass in the III-V materials.

As for double barriers, Fig. 7 shows the DMF and WF computed at the emitter boundary of a silicon device. While in a single-barrier structure the plot drops to the level of numerical noise within less than 100 nm from the origin of the y-axis, in the case of a double-barrier it keeps oscillating with an amplitude that remains about constant over several hundred nanometers. The WF plot is also quite different from its single-barrier equivalent, as oscillating lobe-like artifacts separate from the central FDD peak about 4 decades below its maximum. This behavior is unaffected by the contact length. In [1], we hypothesized that this behavior was caused by an inaccurate computation of the wavefunctions. However, further analysis in [27] shows that the oscillations in the DMF are due to the very sharp peaks in the transmission spectrum that occur when the well resonates. While the amplitude of the oscillations is expected to drop eventually as y increases, it does so very slowly: indeed, the y-grid range would have to be extended by orders of magnitude in order for the oscillations to fall below the level of numerical noise, which is not feasible due to computational resource constraints.

Aside from the contact length, there is another parameter that has a significant effect on the accuracy of a simulation, namely, the range of the wave vector grid that is applied to the WTE solver. Indeed, because the charge is computed by integrating the WF over the wave vector space, a too-narrow range can cause it to be underestimated. Here, we evaluate the minimum ranges that have to be used in order to compute the charge within error margins of 1, 5 and 10%. The reference charge used in the error computation is evaluated by applying the very large range $k_{\text{Max}} = 20\ \text{nm}^{-1}$. Fig. 8 shows how the wave vector range varies across a silicon-based structure for each error threshold. Note that, even if the double-barrier WF is affected by lobes as shown in Fig. 7, these cause an error in the integrated charge by less than 1% and are thus not expected to significantly affect the minimum range plots, at least for the 5 and 10% error thresholds.

For a single barrier, the minimum range plots show distinct spikes at the barrier end points. In fact, Fig. 9 shows that the WF at the foot of the barrier has a much gentler slope than in the contact region, thus requiring a wider wave vector range to accurately evaluate the charge. Also note that the WF in the middle of the barrier is much more oscillatory, which means that a finer mesh has to be applied. A similar behavior is observed in the double-barrier structure, as the plots reach their maximum values at the two ends of each barrier.

Finally, Fig. 10 shows a similar behavior in a III-V based



(a)

(b)

Fig. 9. Absolute-value WF computed at the emitter contact, as well as at the foot and in the middle point of the barrier in silicon- (a) and III-V-based (b) single-barrier structures. The lengths of the contact regions are 30 and 60 nm respectively.

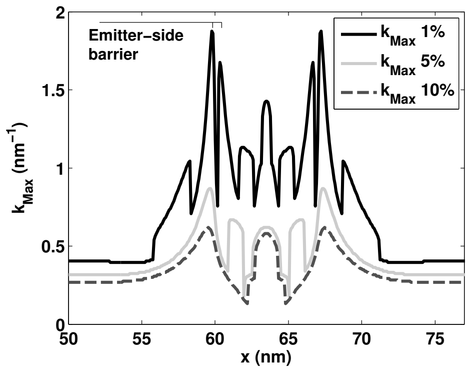


Fig. 10. Minimum wave vector grid half-range required to compute the charge within a threshold error of 1, 5 and 10% in a III-V-based double-barrier structure. The contact length is 60 nm.

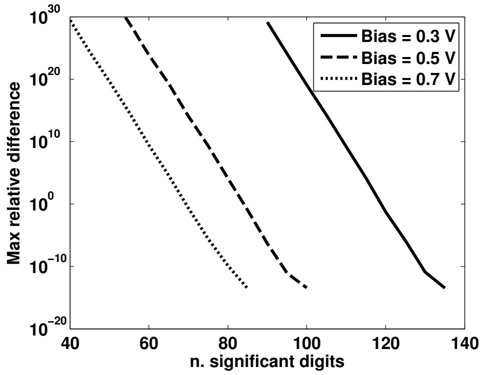
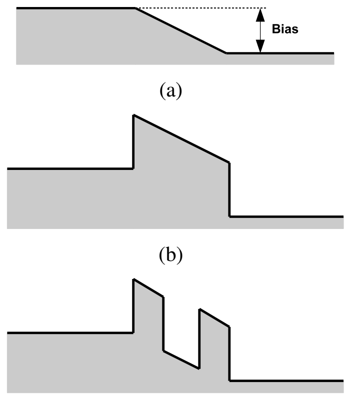


Fig. 12. Relative error in the wavefunction module for a single-barrier biased silicon structure, as a function of the number of significant digits of numerical precision used in the computation and at different bias values. The values plotted indicate the maximum error measured on 250 wavefunctions distributed over a range of 1 eV. The error is relative to reference wavefunctions calculated with a numerical precision of 200 significant digits: the reference wavefunction values and those at a lower precision are both converted to standard double precision, then the maximum relative error between their absolute values is evaluated.



(a)

(b)

(c)

Fig. 11. Schematic view of the potential profiles with a bias applied for a structure with no barrier (a) and one with a single (b) and a double (c) barrier.

double-barrier device, but with lower peaks. This is consistent with (3) and (4), where the lower effective mass in III-V devices results in narrower FDD and WF values at the contacts, as seen in Fig. 8.

VII. Computation of the WF with a Schrödinger solver in biased structures

In this section, we look at the WF at the emitter and collector boundaries of structures where a bias is applied. Our goal is to observe whether lobes appear, and to make a rough estimate of their height. Three configurations are studied: the devices either have no barrier, or one, or two, as shown in Fig. 11. Each structure is composed of 30 nm long contacts that enclose a middle region which has a thickness of 7 nm in the no-barrier and single-barrier configurations and of 5 nm in the double-barrier one. In the single-barrier device, the middle region contains the barrier; in the double-barrier one, it contains both the barriers, which are 1 nm thick, and the well, which is 3 nm thick. The applied material parameters are representative of a silicon-based structure.

The potential profile is constant at the contacts and falls linearly in the middle region. This simplification makes it possible to solve the SE symbolically using Airy functions.

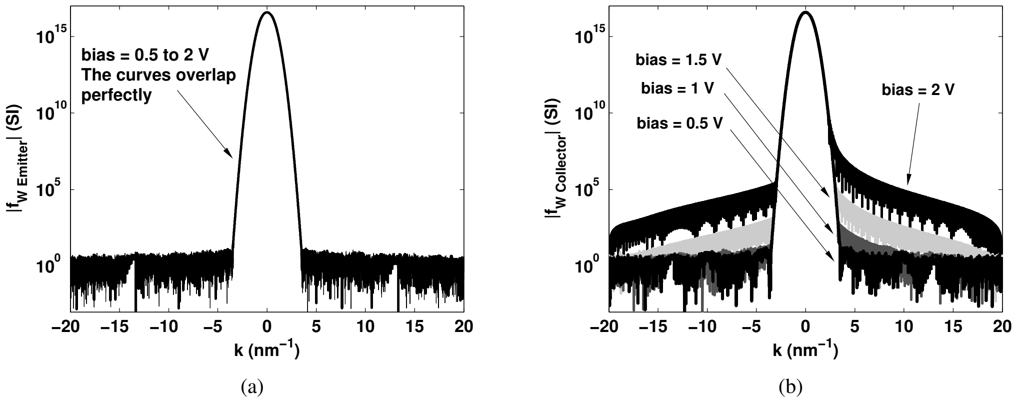


Fig. 13. WF at the emitter (a) and collector (b) of a single-barrier silicon structure, at different bias points.

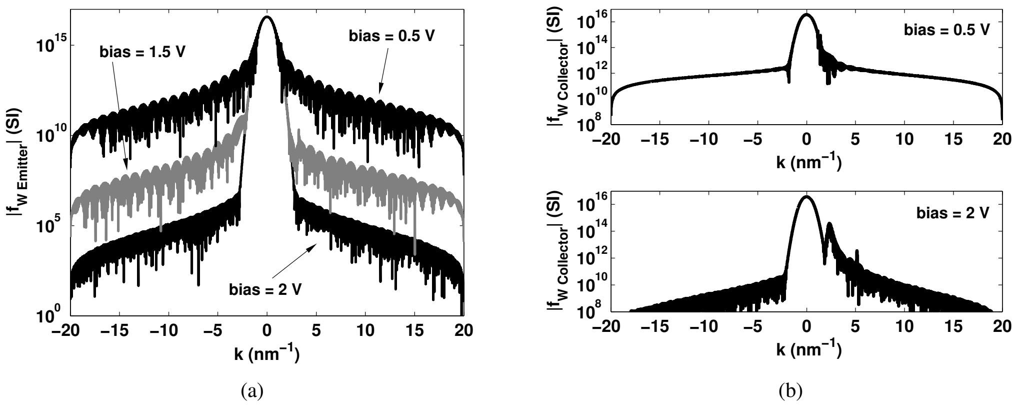
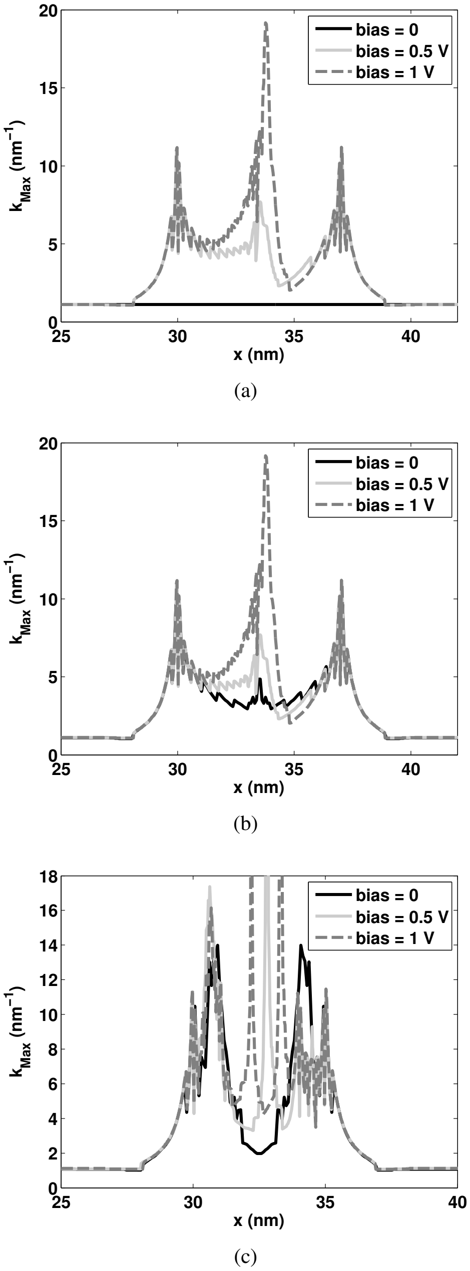


Fig. 14. WF at the emitter (a) and collector (b) of a double-barrier silicon structure, at different bias points.

The symbolic wavefunctions formulae are very complex and are not optimized to reduce the numerical error. When evaluating them, it is therefore usually necessary to use a higher computing precision than the IEEE 754 standard machine double one. The MPMATH arbitrary-precision Python library is used [29]. Fig. 12 displays the level of numerical precision needed to calculate wavefunctions in single-barrier structures. It plots the numerical error in the wavefunction module at different working precisions, relative to a reference of 200 significant digits. The wavefunctions are computed across the structures at different bias points, by casting 250 incident wave vectors within a 1 eV energy range. The modules of the wavefunctions are then computed, and they are finally converted to double machine precision. The plots in the figure trace the maximum relative difference compared to the 200-digit reference. The plots all end as the relative difference reaches a value of about 10^{-15}: this is due to the conversion to double precision, which does not allow to measure relative errors smaller than about 15 decades. The end point of each plot indicates the minimum working precision that is needed to compute the wavefunctions to double data type accuracy.

It can be seen that the required minimum precision is much greater than for unbiased structures, where standard double precision suffices. For double-barrier structures, the numerical precision required in the computation is similar to that for a single barriers; on the contrary, for no-barrier structures, double precision is once again sufficient. It should be stressed that optimizing the symbolic wavefunction formulae to reduce numerical error propagation might be beneficial in lowering the minimum required working precision; nevertheless, standard machine precision may still not be enough.

Fig. 13 shows the WF computed at the emitter and collector boundaries on a biased silicon single-barrier structure. The plots are virtually indistinguishable from those obtained on a device with no barrier. The emitter WF looks very similar to the symmetric FDD profile obtained on an unbiased device. In fact, the left side ($k < 0$) of the curve bulges a little less than the right one; this effect is however difficult to spot visually and can only be seen by looking at the numerical values of the WF. An asymmetric WF curve is expected, as it indicates a current flow between the emitter and the collector. In this structure, however, the high and thick barrier insulates well the two contacts. This occurs even at high bias points, as the different curves overlap almost perfectly. On the other hand, on a no-barrier structure, wavefunctions that are incident at the emitter can propagate freely towards the collector, once again independently of the bias point. On the collector side, the asymmetry is more evident, yet it still occurs many orders



(a)

(b)

(c)

Fig. 15. Minimum wave vector grid half-range required to compute the charge within a threshold error of 1% in silicon-based no- (a), single- (b) and double-barrier (c) structure.

of magnitude below the peak.

Fig. 14 plots the WF at the emitter and collector boundaries of a double-barrier structure. On the emitter side, one notices that the asymmetry between the left and right lobes is more evident, especially at the 1.5 V bias point. On the collector side, the asymmetry of the WF plot is again very evident, and lobes again appear about 5 decades below the peak. The plots in III-V-based devices are similar.

Fig. 15 plots the minimum wave vector grid range required to compute the charge within an error margin of 1% at bias voltages of 0, 0.5 and 1 V in the three structure types. As a bias is applied, the plots loose their symmetry. Similarly to the unbiased structures, they peak at the two ends of the middle region; the left peak (high-bias contact) is however higher in general than the right one (low-bias contact). The difference in height is especially marked in the case of the double-barrier structure, although it is no greater than 30%. In both single- and double-barrier structures, the difference in the heights of the peaks with and without an applied bias is also quite small.

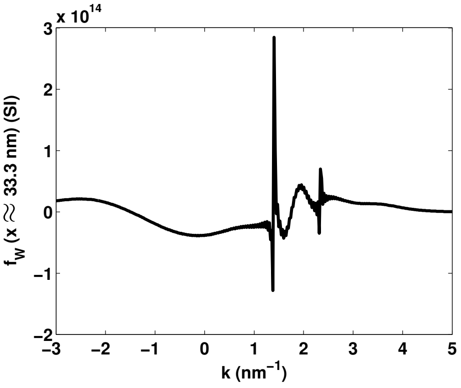


Fig. 16. WF in a double-barrier silicon structure with a 1 V bias at the x-node where the minimum wave vector range required to compute the charge within a threshold error of 1% is highest and close to 60 nm^{-1}.

The biggest difference compared to the no-bias plots occurs next to the middle point of the structure, where a very high peak appears. This peak varies considerably with the applied bias and is especially high on the double-barrier structure, where it goes up to 60 nm^{-1} at a bias of 1 V. In fact, in the case of the double-barrier device, two distinct peaks appear near the middle of the structure, with the right one being higher. Fig. 16 shows the WF at the point where the peak is highest: while the plot is markedly asymmetric, it oscillates about the origin, with the negative and positive areas being very similar in size and canceling each other out when the charge integral is evaluated. Similar trends are observed when applying III-V material parameters.

As the WF is computed from the wavefunctions, scattering effects are not accounted for. Normally, interactions have the effect dissipating the charge carriers' energy, thus screening the electric field [30]. It is thus possible that the high-energy peaks observed in Fig. 15 may be considerably lower if interactions are simulated, as the distribution profiles would be pushed back toward the origin of the wave vector space.

VIII. Conclusion

In this work, we have studied the Wigner Function (WF) by computing it directly from the Schrödinger Equation (SE), rather than by solving the Wigner Transport Equation (WTE). We have shown that, in order to accurately compute the WF over a large wave vector range, an extremely high machine precision is required, often higher than the IEEE 754 double data type. In fact, the numerical solver which we implemented, despite being able to computing charge densities well within 1% accuracy, generated lobes on the WF curves that could not be eliminated by applying denser grids.

This difficulty in computing the WF accurately comes from the density matrix integration step, followed by the application of the Fourier transform. Because these two operations are an inherent part of the WTE, the problems encountered in their implementation might explain those met in solving the WTE itself, namely the negative charge densities and the low charge drop-offs in presence of large barriers. These points have to be investigated further, and the SE solver may be of help, as it allows to compute accurate boundary conditions in the quantum region, at least for a single rectangular barrier structure. These boundary conditions can then be applied to the WTE solver, thus eliminating the memory overhead of meshing large contact regions.

This approach has also helped investigate some basic geometry parameters and numerical implementation conditions that must be applied to the WTE solver in order to accurately simulate 1D single- and double-barrier structures. As for the device geometry, the minimum contact length has been investigated. In single-barrier structures, it has been found that the WF at the boundaries follows a Fermi-Dirac Distribution (FDD) if the contact length is greater than 30 nm in silicon-based devices and about 60 nm in III-V-based ones; in double-barrier ones, the WF separates from the FDD reference profile a few decades below the peak and forms oscillating lobes that are not significantly affected by either the contact length or the numerical computing precision. As for the simulation numerical parameters, this work investigates the range of the WF in the wave vector space required to accurately compute the charge. In silicon-based structures, this range is estimated between 10 and 15 nm^{-1} for an error smaller than 1%, and five to ten times lower in III-V equivalent structures.

This work also presents WF plots in silicon biased structures. The potential profiles used are simplified in order to allow for a symbolic solution of the SE, and carrier interactions are not taken into account. Nevertheless, it is still possible to observe some trends at different bias points, namely the increasing asymmetry in the lobes for positive and negative wave vectors in the WF plots at the collector of the different structures.

References

[1] A. Savio and A. Poncet, "Study of the Wigner function computed by solving the Schrödinger equation," in *Proc. 4th International Conference on Quantum, Nano and Micro Technologies (ICQNM 10)*, 2010, pp. 59–64.

[2] *Process Integration, Devices, & Structures (PIDS)*, International Technology Roadmap for Semiconductors, 2009. [Online]. Available: http://www.itrs.net/Links/2009ITRS/2009Chapters_2009Tables/2009Tables_FOCUS_C_ITRS.xls

[3] *Sentaurus Device User Guide*, Synopsys, Inc., Mountain View, CA, Sep. 2008, Version A-2008.09.

[4] E. Wigner, "On the quantum correction for thermodynamic equilibrium," *Phys. Rev.*, vol. 40, no. 5, pp. 749–759, 1932.

[5] N. Kluksdahl, W. Pötz, U. Ravaioli, and D. K. Ferry, "Wigner function study of a double quantum barrier resonant tunnelling diode," *Superlattices and Microstructures*, vol. 3, no. 1, pp. 41–45, 1987.

[6] N. Kluksdahl, A. M. Kriman, D. K. Ferry, and C. Ringhofer, "Self-consistent study of the resonant-tunneling diode," *Phys. Rev. B*, vol. 39, no. 11, pp. 7720–7735, 1989.

[7] W. R. Frensley, "Transient response of a tunneling device obtained from the Wigner function," *Phys. Rev. Lett.*, vol. 57, no. 22, pp. 2853–2856, 1986.

[8] W. R. Frensley, "Wigner-function model of a resonant-tunneling semiconductor device," *Phys. Rev. B*, vol. 36, no. 3, pp. 1570–1580, 1987.

[9] K. L. Jensen and F. A. Buot, "Numerical simulation of transient response and resonant-tunneling characteristics of double-barrier semiconductor structures as a function of experimental parameters," *J. Appl. Phys.*, vol. 65, no. 12, pp. 5248–5250, 1989.

[10] K. L. Jensen and F. A. Buot, "The effects of scattering on current-voltage characteristics, transient response, and particle trajectories in the numerical simulation of resonant tunneling diodes," *J. Appl. Phys.*, vol. 67, no. 12, pp. 7602–7607, 1990.

[11] K. L. Jensen and F. A. Buot, "Numerical simulation of intrinsic bistability and high-frequency current oscillations in resonant tunneling structures," *Phys. Rev. Lett.*, vol. 66, no. 8, pp. 1078–1081, 1991.

[12] F. Jensen, K.L. Buot, "The methodology of simulating particle trajectories through tunneling structures using a Wigner distribution approach," *IEEE Trans. Electron Devices*, vol. 38, no. 10, pp. 2337–2347, 1991.

[13] K. L. Jensen and A. K. Ganguly, "Numerical simulation of field emission and tunneling: A comparison of the Wigner function and transmission coefficient approaches," *J. Appl. Phys.*, vol. 73, no. 9, pp. 4409–4427, 1993.

[14] F. A. Buot and K. L. Jensen, "Lattice Weyl-Wigner formulation of exact many-body quantum-transport theory and applications to novel solid-state quantum-based devices," *Phys. Rev. B*, vol. 42, no. 15, pp. 9429–9457, 1990.

[15] H. Tsuchiya, M. Ogawa, and T. Miyoshi, "Simulation of quantum transport in quantum devices with spatially varying effective mass," *IEEE Trans. Electron Devices*, vol. 38, no. 6, pp. 1246–1252, 1991.

[16] K. K. Gullapalli and D. P. Neikirk, "Incorporating spatially varying effective-mass in the Wigner-Poisson model for AlAs/GaAs resonant-tunneling diodes," in *Proc. 3rd International Workshop on Computational Electronics*, Portland, OR, 1994, pp. 171–174.

[17] K. K. Gullapalli, D. R. Miller, and D. P. Neikirk, "Simulation of quantum transport in memory-switching double-barrier quantum-well diodes," *Phys. Rev. B*, vol. 49, no. 4, pp. 2622–2628, 1994.

[18] D. R. Miller and D. P. Neikirk, "Simulation of intervalley mixing in double-barrier diodes using the lattice Wigner function," *Appl. Phys. Lett.*, vol. 58, no. 24, pp. 2803–2805, 1991.

[19] G. Y. Wu and K.-P. Wu, "Electron transport in a resonant-tunneling diode under the effect of a transverse magnetic field: A quantum theory in the Wigner formalism," *J. Appl. Phys.*, vol. 71, no. 3, pp. 1259–1264, 1992.

[20] B. A. Biegel and J. D. Plummer, "Comparison of self-consistency iteration options for the Wigner function method of quantum device simulation," *Phys. Rev. B*, vol. 54, no. 11, pp. 8070–8082, 1996.

[21] H. Grubin and R. Buggeln, "RTD relaxation oscillations, the time dependent Wigner equation and phase noise," *J. Comput. Electron.*, vol. 1, no. 1, pp. 33–37, 2002.

[22] M. Nedjalkov, H. Kosina, R. Kosik, and S. Selberherr, "A Wigner equation with quantum electron-phonon interaction," *Microelectron. Eng.*, vol. 63, no. 1, pp. 199–203, 2002.

[23] Y. Yamada, H. Tsuchiya, and M. Ogawa, "Quantum transport simulation of silicon-nanowire transistors based on direct solution approach of the Wigner transport equation," *IEEE Trans. Electron Devices*, vol. 56, no. 7, pp. 1396–1401, 2009.

[24] J. Kefi-Ferhane and A. Poncet, "Deterministic simulation of transport in MOSFETs by computing Wigner, Poisson and Schrödinger equations," *Phys. Status Solidi A*, vol. 201, no. 11, pp. 2518–2521.

[25] A. Jüngel, *Transport equations for semiconductors*. Berlin: Springer, 2009.

[26] T. Pang, *An introduction to computational physics*, 2nd ed. New York: Cambridge University Press, 2006.

[27] A. Savio and A. Poncet, "Study of the Wigner function at the device boundaries in 1D single- and double-barrier structures," *J. Appl. Phys.*, accepted for publication.

[28] A. Savio and A. Poncet, "Study of the Wigner function computed by solving the Schrödinger equation," in *15th International Conference on Simulation of Semiconductor Processes and Devices (SISPAD 10)*, 2010.

[29] F. Johansson. (2010) MPMATH Python library for arbitrary-precision floating-point arithmetic. [Online]. Available: http://code.google.com/p/mpmath/

[30] W. R. Frensley, "Boundary conditions for open quantum systems driven far from equilibrium," *Rev. Mod. Phys.*, vol. 62, no. 3, pp. 745–791, 1990.