

Ballistic quantum transport using the contact block reduction (CBR) method

An introduction

Stefan Birner · Christoph Schindler · Peter Greck ·
Matthias Sabathil · Peter Vogl

Published online: 3 October 2009
© Springer Science+Business Media LLC 2009

Abstract The contact block reduction (CBR) method is a variant of the nonequilibrium Green's function formalism and can be used to describe quantum transport in the ballistic limit very efficiently. We present a numerical implementation of a charge self-consistent version of the CBR algorithm. We show in detail how to calculate the electronic properties of open quantum systems such as the transmission function, the local density of states and the carrier density. Several 1D and 2D examples are provided to illustrate the key points. The CBR method is a very powerful tool to tackle the challenge of calculating transport in the ballistic limit for 3D devices of arbitrary shape and with an arbitrary number of contacts.

Keywords Ballistic quantum transport · Nonequilibrium Green's function formalism · NEGF · Transmission function · Landauer-Büttiker formalism · Device simulation

1 Introduction

Since electronic devices have been shrinking steadily to nanometer dimensions, quantum transport is increasingly becoming a topic of interest not only to physicists but

also to the electrical engineering community [1]. The nonequilibrium Green's function (NEGF) formalism (e.g. [2]) provides a rigorous framework for the development of quantum device models. Here, we describe one of its implementations—the contact block reduction (CBR) method [3]. It can be used to describe quantum transport in the ballistic limit very efficiently. Our aim in this article is to make the Green's function formalism in the limit of ballistic quantum transport accessible to a more general audience. Thus, a detailed description of the underlying algorithm is given and numerical examples are provided as concrete illustrations. As it is very important to perform charge self-consistent calculations, we also give details on how to solve the nonlinear system of coupled Schrödinger and Poisson equations. Interested readers should be able to reproduce these results by setting up their own computer program. All results presented in the figures of this article can be reproduced with the software that is provided as an Online Resource [4].

2 Ballistic quantum transport

A conductor shows nonohmic behavior if its dimensions are smaller than certain characteristic lengths: The mean free path and the phase-relaxation length of the electron [5]. If the length of a conductor becomes shorter than the mean free path, the conductance approaches a limiting value. This classical ballistic limit has still nothing to do with quantum mechanics. Quantum mechanics does not become important until the dimensions of the conductor are smaller than the phase-relaxation length and interference-related effects come into play. In present day high-mobility semiconductor heterostructures such as modulation doped GaAs/AlGaAs heterojunctions or quantum wells, mean free paths and

S. Birner (✉) · C. Schindler · P. Greck · M. Sabathil · P. Vogl
Walter Schottky Institut and Physics Department, Technische
Universität München, Am Coulombwall 3, 85478 Garching,
Germany
e-mail: stefan.birner@nextnano.de

Present address:
M. Sabathil
OSRAM Opto Semiconductors GmbH, Leibnizstr. 4,
93055 Regensburg, Germany

phase-relaxation lengths of several μm are relatively easy to obtain at low temperatures. Thus ballistic quantum transport plays an important role in many mesoscopic transport experiments.

The theoretical approach that has proven to be most useful in describing mesoscopic transport was introduced by Landauer [6, 7] in 1988. A generalization to multiterminal devices in magnetic fields was proposed by Büttiker [8, 9] and is generally referred to as the Landauer-Büttiker (LB) formalism. It is equivalent to the nonequilibrium Green's function formalism in the limit of no inelastic or elastic scattering. The essential idea behind the LB formalism is that the current through a ballistic conductor is determined by the probabilities of the electrons to be transmitted or reflected. The contacts of the conductor are assumed to be large electron reservoirs in equilibrium, so that each contact can be described by its own Fermi distribution with a chemical potential μ . The difference between the chemical potentials in the contacts is equal to the externally applied bias voltage. By the Landauer-Büttiker formula, these relations are expressed as follows

$$I_{\lambda\lambda'} = \frac{g_s e}{h} \int T_{\lambda\lambda'}(E) [f(E, \mu_\lambda) - f(E, \mu_{\lambda'})] dE, \quad (1)$$

where $I_{\lambda\lambda'}$ is the current between contact λ and contact λ' , $T_{\lambda\lambda'}(E)$ is the corresponding energy dependent transmission function between these contacts, μ_λ and $\mu_{\lambda'}$ are the chemical potentials in these contacts, E is the energy, h is Planck's constant, e is the positive elementary charge, and $g_s = 2$ is the spin degeneracy of the electrons.

$$f(E, \mu_\lambda) = \frac{1}{1 + \exp[(E - \mu_\lambda)/(k_B T)]} \quad (2)$$

is the equilibrium Fermi-Dirac distribution function inside contact λ , k_B is Boltzmann's constant and T is the temperature. Thus the Landauer-Büttiker formalism reduces the problem of calculating the ballistic current in a mesoscopic device to the determination of the transmission probabilities of an open device connected to reservoirs. We emphasize that (1) has been simplified here. It generally involves an integration over all quantum numbers that characterize the lead states [10]. We suppressed their momentum dependence (which is, however, included in the calculations as described further below) to keep things as simple as possible and assume conservation of spin, energy E and parallel momentum. We also assume a parabolic dispersion of the bands so that the integration over the parallel momentum can be simplified. The current in any device can be calculated via the transmission function as long as the propagation of the electrons through the device is coherent. Coherent means that there are no phase-breaking scattering processes involved. Elastic scattering processes can be taken into account within the CBR method if they can directly be included into the Hamiltonian \mathbf{H}^0 of the closed system. For

extremely small devices or very low temperatures, where the elastic and inelastic scattering lengths exceed the geometrical device size, the assumption of coherent transport is justified in most cases. If it is necessary to include the momentum and energy relaxation of the carriers (e.g. elastic and inelastic scattering of electrons with phonons), a full NEGF calculation is required which is however very challenging for 2D and 3D devices due to its enormous computational effort. In principle, it is possible to include inelastic scattering via Büttiker probes [11, 12] where additional contacts in the interior of the device are used to model the coupling of the carriers to a phonon bath. However, a satisfactory integration of dissipative processes into the CBR method—which maintains its computational efficiency—has not been found yet.

Several numerical methods have been developed to determine the transmission coefficient for quantum devices via the scattering matrix, e.g. the transfer matrix method [13, 14], the quantum transmitting boundary method [15], the R-matrix method [16] and the recursive Green's function method [17, 18]. In this article, we will describe in detail how to obtain the transmission function $T_{\lambda\lambda'}(E)$ by means of the contact block reduction (CBR) method, where the transmission is calculated from the retarded single particle Green's function. In passing, we note that the transmission function not only determines the electrical current because the electronic component of heat currents can be calculated with a Landauer formula similar to (1) [19]. Optimizing the thermoelectric coefficients in devices by quantum-engineering the transmission function is an interesting topic in thermoelectrics research.

3 The contact block reduction (CBR) method— An overview

The CBR method is a very efficient Green's function technique which has been developed by Mamaluy *et al.* [3]. It can be used to calculate the electronic properties of open quantum systems such as the transmission function, the local density of states, and the carrier density in the ballistic limit for 1D, 2D and 3D devices of arbitrary shape and with an arbitrary number of contacts. We start with a device that is discretized in real space on N_T total grid points. It can be characterized by a corresponding Hamiltonian matrix \mathbf{H}^0 of size N_T . The device has no contacts and is thus termed a *closed* system. It has sharp resonant energies (eigenvalues of \mathbf{H}^0) and the electrons are described by wave functions (eigenfunctions of \mathbf{H}^0). We now add contacts to the device and divide the total number of grid points into N_C contact grid points and N_D interior device grid points

$$N_T = N_C + N_D. \quad (3)$$

The contact grid points N_C are the boundary grid points of the device that overlap with the leads. Connecting the device to contacts leads to a broadening of the resonant energies: The discrete energy spectrum transforms into a continuous density of states. This is described by the broadening matrix $\Gamma(E)$. It depends on energy E and has the same size as \mathbf{H}^0 . It can directly be calculated from the self-energy $\Sigma(E)$. This self-energy matrix is added to the Hamiltonian to account for the new boundary conditions due to the contacts (see Sect. 4.3.1 for details). It is non-Hermitian, thus leading to complex eigenvalues. In fact, the imaginary part of the eigenvalues is the origin of the broadening of the density of states and introduces a finite lifetime to the eigenstates. Consequently, the device wave functions leak out into the contacts (*open* device). As Σ also depends on energy, it is more convenient to look at the device from another point of view. Rather than asking for the eigenenergies of the system, it is more appropriate to ask: How does the open device *respond* to incident electrons that have a certain energy E ? In the ballistic case, all observables of interest can be obtained from the retarded Green’s function \mathbf{G}^R of the open device. It is defined as

$$\mathbf{G}^R(E) = \left(E\mathbf{1} - \mathbf{H}^0 - \Sigma \right)^{-1}, \tag{4}$$

where $\mathbf{1}$ is the identity matrix. It can be expressed in terms of the retarded Green’s function \mathbf{G}^0 of the closed device via the Dyson equation

$$\mathbf{G}^R = \mathbf{A}^{-1}\mathbf{G}^0 = \left(\mathbf{1} - \mathbf{G}^0\Sigma \right)^{-1} \mathbf{G}^0, \tag{5}$$

$$\mathbf{G}^0(E) = \left(E\mathbf{1} - \mathbf{H}^0 + i0^+ \right)^{-1} = \sum_n \frac{|\psi_n\rangle \langle \psi_n|}{E - E_n + i0^+}. \tag{6}$$

The last expression (spectral representation) shows how to write the retarded Green’s function \mathbf{G}^0 in terms of the eigenenergies and wave functions of the closed device Hamiltonian (see Sect. 4.1). $|\psi_n\rangle \langle \psi_n|$ is the dyadic product where $\langle \psi_n|$ is a row vector and $|\psi_n\rangle$ is a column vector (bra-ket or Dirac notation), each of size N_T . In a numerical implementation of this equation, the infinitesimally small positive imaginary number $i0^+$ can be ignored if one ensures that $E \neq E_n$. Additionally, if the wave functions ψ_n are real, the retarded Green’s function of the closed device is real. Thus it is identical to the advanced Green’s function of the closed device. (The conjugate transpose (\dagger) of the retarded Green’s function is called the advanced Green’s function $\mathbf{G}_C^A = \mathbf{G}_C^{R\dagger}$.) We call \mathbf{G}^0 just the *Green’s function of the closed device* and omit $i0^+$ and the term *retarded* in the following for simplicity.

Once the self-energy matrix has been calculated (see Sect. 4.3.1), the evaluation of the retarded Green’s function \mathbf{G}^R of the open device requires—in general—the inversion of a large matrix \mathbf{A} whose size is proportional to the num-

ber N_T of total grid points of the device. Even in two spatial dimensions, this can be a quite demanding task.

The essence of the contact block reduction method consists in the decomposition of the retarded Green’s function into blocks such that the transmission function of the open device can be calculated by inverting only small matrices: The retarded Green’s function can be ‘reduced to the contact block’ \mathbf{G}_C^R . The contact block (index C) consists of all device grid points that are in contact with the leads. This number is orders of magnitude smaller than the number of device grid points. This explains the astonishing efficiency of this approach and makes it possible to address quantum transport in 3D devices. The CBR method has been applied to calculate the transport in 3D structures, like quantum dots [20], quantum interference devices such as a quantum logic gate [21] or nano-FinFETs [22]. The latter requires to include the Poisson equation in order to guarantee charge self-consistency (self-consistent CBR [23], see Sect. 8). The CBR method has been extended to describe systems of two interacting particles for the study of two-qubit devices [24]. It has also been extended to more sophisticated band structure models, like the $\mathbf{k} \cdot \mathbf{p}$ method in order to describe hole transport in quantum wires [25] and to tight-binding methods [26]. It has been integrated into the nextnano software package [27] which is available online [4].

In this article we describe in detail how to calculate the transmission function $T(E)$ and the local density of states $\rho(\mathbf{x}, E)$ from the Green’s function matrix \mathbf{G}^R of the open system

$$\mathbf{G}^R = \begin{pmatrix} \mathbf{G}_C^R & \mathbf{G}_{CD}^R \\ \mathbf{G}_{DC}^R & \mathbf{G}_D^R \end{pmatrix}. \tag{7}$$

This matrix has been subdivided into four blocks, a submatrix within the contact block (C) and another one within the interior of the device (D). The other two correlate the contact grid points to the device grid points (CD and DC). To obtain the transmission function, it is only necessary to evaluate the upper left part—the contact block

$$\mathbf{G}_C^R = \mathbf{A}_C^{-1}\mathbf{G}_C^0. \tag{8}$$

For calculating the local density of states, additionally the lower left part

$$\mathbf{G}_{DC}^R = \mathbf{G}_{DC}^0\mathbf{B}_C^{-1} \tag{9}$$

has to be evaluated. Thus for each energy E of interest, the two matrices

$$\mathbf{A}_C = \mathbf{1}_C - \mathbf{G}_C^0\Sigma_C \tag{10}$$

and

$$\mathbf{B}_C = \mathbf{1}_C - \Sigma_C\mathbf{G}_C^0 \tag{11}$$

have to be inverted where $\mathbf{1}_C$ is the identity matrix of dimension N_C . The dimension of these matrices is very small

and is determined by the number N_C of grid points that connect the device to the contacts. For one-dimensional devices $N_C = 2$, so both matrices are of size 2×2 . In the ballistic case, the self-energy matrix Σ is nonzero only at the contact grid points and can thus be reduced to Σ_C . This is the reason why only small parts of the Green's functions have to be evaluated. The transmission function determines the current through the device, and from the local density of states, the charge density can be derived. This is all one needs to describe quantum transport in arbitrary devices within the ballistic limit, i.e. for situations where incoherent scattering can be ignored.

4 The CBR method for one-dimensional devices

In this section we describe the contact block reduction (CBR) method for simple one-dimensional devices where the device geometry is assumed to be translationally invariant in the (x, y) plane. Current transport is assumed to be along the z direction. We choose the 1D case in order to highlight the main points of the CBR method, avoiding therefore to include the additional, more complicated features coming into play when one deals with two-dimensional and three-dimensional devices described in Sect. 5. We try to avoid reproducing the equations and the arguments of the original CBR papers [3, 25] and adopt the more straightforward approach to focus specifically on the aspects with respect to a numerical implementation.

In a one-dimensional device one can only have two leads (i.e. contacts) in total ($L = 2$). These leads are located at the leftmost and rightmost boundary points of the device and each lead λ contains exactly one grid point ($N_\lambda = 1$) that connects the lead to the device, i.e. the total number of (relevant) lead grid points is thus equal to $N_C = \sum_{\lambda=1}^L N_\lambda = 2$. This simplifies the CBR method substantially because the dimension of the CBR contact matrices is exactly equal to $N_C = 2$. This means that for the calculation of the transmission coefficient $T(E)$ (see (1)), for each energy E only a small square matrix of size $N_C = 2$ has to be inverted. A further simplification is that each lead has only one mode. In a 2D or 3D simulation, each lead consists of several lead grid points connected to the device ($N_\lambda > 1$). The number of lead grid points corresponds to the number of lead modes (see Sect. 5), i.e. each lead has N_λ modes. In a 1D simulation, the CBR algorithm is then implemented as follows:

4.1 Energy levels and wave functions of the device Hamiltonian (closed system)

First, we calculate the energy levels and the wave functions of the device Hamiltonian without taking the leads into account. This Hamiltonian \mathbf{H}^0 is then identical to the Hamiltonian of the closed system. We use a standard approach to

solve the Schrödinger equation, namely the envelope function approximation assuming a parabolic dispersion (single-band effective mass equation).

The Schrödinger equation for a semiconductor structure grown along the z direction and homogeneous along the x and y directions reads

$$\mathbf{H}_{\mathbf{k}_\parallel}^0 \Psi_n(z, \mathbf{k}_\parallel) = E_n(\mathbf{k}_\parallel) \Psi_n(z, \mathbf{k}_\parallel). \quad (12)$$

The wave function $\Psi_n(z, \mathbf{k}_\parallel)$ can be factorized into a solution $\psi_n(z, \mathbf{k}_\parallel)$ along the z direction, and a plane wave $e^{i\mathbf{k}_\perp \cdot \mathbf{r}}$ in the (x, y) plane

$$\Psi_n(z, \mathbf{k}_\parallel) = \psi_n(z, \mathbf{k}_\parallel) e^{i\mathbf{k}_\perp \cdot \mathbf{r}}. \quad (13)$$

In the following we ignore the dependence of $\psi_n(z, \mathbf{k}_\parallel)$ on the parallel momentum \mathbf{k}_\parallel . Then the envelope functions $\psi_n(z)$ of the n_T quantized states are obtained as the solutions of the one-dimensional Schrödinger equation ($n = 1, \dots, n_T$ where $n_T = N_T$):

$$\mathbf{H}^0 \psi_n(z) = E_n \psi_n(z), \quad (14)$$

$$\left[-\frac{\hbar^2}{2} \frac{\partial}{\partial z} \left(\frac{1}{m_\perp(z)} \frac{\partial}{\partial z} \right) + V(z) \right] \psi_n(z) = E_n \psi_n(z). \quad (15)$$

$m_\perp(z)$ is the effective mass tensor component along the z direction, \hbar is Planck's constant divided by 2π , $V(z) = E_c(z) = E_{c,0}(z) - e\phi(z)$ is the spatially varying potential energy (conduction band edge profile), $E_{c,0}(z)$ represents the conduction band edge profile of the particular band of interest including band offsets at material interfaces and $\phi(z)$ is the electrostatic potential which is obtained from solving Poisson's equation (see Sect. 8.1). It includes the external bias potential and the internal potential resulting from mobile charge carriers and ionized impurities.

We discretize this equation with a finite differences method on a uniform grid using Neumann boundary conditions at the left and right device boundaries. At these points the device is in contact to the leads, once they are added to form the open system. It has been found that Neumann boundary conditions at the contact grid points are much better suited for the CBR method than Dirichlet boundary conditions [3]. This will become clear from (21) where the projection of the wave functions at the contact grid points is used. It is important to have the "closed eigenfunctions" mimicking as much as possible the real, resonant open system wave functions. Dirichlet conditions do not even allow the carriers to enter or leave the device area because of their vanishing wave function amplitudes at the contact grid points. In contrast, Neumann conditions are better representing the open system because of their finite wave function amplitudes at these points.

The discretized sparse, square and Hermitian (in most cases even real and symmetric) Hamiltonian matrix of size

equal to the number of total device grid points N_T has to be diagonalized numerically to yield the eigenvalues and eigenvectors. The eigenenergies E_n correspond to the energies of the electron along the z direction. The total energy of the electron includes the parallel momentum of the electron due to $\mathbf{k}_{\parallel} = (k_x, k_y)$

$$E_n(\mathbf{k}_{\parallel}) = E_n + \frac{\hbar^2}{2m_{\parallel}} \mathbf{k}_{\parallel}^2, \tag{16}$$

where m_{\parallel} derives from the mass tensor components in the (x, y) plane. For more detailed information on how to solve (15) numerically, we refer to e.g. [28].

The one-dimensional envelope functions ψ_n are usually normalized to 1

$$\int \psi_n^*(z) \psi_n(z) dz = \sum_{i=1}^{N_T} \psi_{n,i}^* \psi_{n,i} \Delta_i = 1, \tag{17}$$

where $\psi_{n,i}$ is the amplitude of the wave function at grid point i , and Δ_i the corresponding grid spacing along the z direction. If the latter has units of [nm], then the wave functions ψ_n have units of [nm^{-1/2}]. The sum includes all grid points N_T (see (3)) of the total device because here we consider the closed system where we do not have to distinguish between interior device grid points N_D and contact grid points N_C . The latter are the boundary grid points of the device that overlap with the leads but are still part of the device, and are thus part of the closed system device Hamiltonian \mathbf{H}^0 . In principle, the wave functions could have been calculated using a nonuniform grid spacing. However, in the following sections we assume that the grid spacing Δ_i is homogeneous for all grid points. This allows us to renormalize the wave functions so that they become dimensionless. This is achieved by dividing ψ_n by the norm $1/\sqrt{\Delta}$. Then the normalization reads

$$\sum_{i=1}^{N_T} \psi_{n,i}^* \psi_{n,i} = 1. \tag{18}$$

Incomplete set of eigenstates We want to emphasize that the actual number n_{α} of eigenvalues and wave functions needed to get meaningful results within the CBR method can be much smaller than the total number n_T of eigenfunctions of the Hamiltonian matrix. The energy $E_{n_{\alpha}}$ of the highest eigenvector taken into account is the cutoff energy. It should be significantly above the energy interval of interest in order to get reliable results (see Figs. 4 and 6). For a 2D and 3D simulation, using such an incomplete set of eigenstates will drastically improve the computational performance as only n_{α} eigenstates have to be calculated ($n_{\alpha} \approx 10\%$ of n_T). This fact makes it attractive to use fast, iterative solvers [29] for calculating only a small number of eigenstates of these

sparse matrices. In 1D, where it is not a computational challenge to calculate all eigenvalues of the spectrum, exact solvers [30] might be preferable.

4.2 Projection of device eigenfunctions onto lead modes

This part is very easy for a one-dimensional simulation where for each of the two leads only one lead mode exists. One simply has to store—for each eigenvalue n —the values of the wave functions $\psi_{n,i}$ at the leftmost grid point ($i = 1$) and at the rightmost grid point ($i = N_T$)

$$\text{Lead 1: } \chi_n^{\lambda=1} = \psi_{n,1} \quad (\text{left boundary}), \tag{19}$$

$$\text{Lead 2: } \chi_n^{\lambda=2} = \psi_{n,N_T} \quad (\text{right boundary}). \tag{20}$$

For each eigenvalue n these projected eigenvector amplitudes χ_n^{λ} are stored in a vector of size $N_C = 2$

$$\chi_n = \begin{pmatrix} \chi_n^{\lambda=1} \\ \chi_n^{\lambda=2} \end{pmatrix}. \tag{21}$$

4.3 Setup energy interval and calculate properties for each energy E_i

We are interested in the transmission coefficient $T_{12}(E)$ from lead $\lambda = 1$ (left contact) to lead $\lambda = 2$ (right contact) for all energy values E in the energy interval of interest ($E_{\min} < E < E_{\max}$). To do this, we divide this energy interval into N_E energy grid points and calculate for each the transmission coefficient $T_{12}(E_i)$ from lead 1 to lead 2 for the energy value $E_i = E_{\min} + (i - 1) \Delta_E$ where $\Delta_E = \frac{(E_{\max} - E_{\min})}{N_E - 1}$ is the energy grid spacing and $i = [1, \dots, N_E]$. It is worth mentioning that the properties for different energies E_i can be computed in parallel. This intrinsic parallelism is very useful for implementing parallel computing schemes on multicore CPUs, especially for large two- and three-dimensional simulations. For each energy E_i the following matrices have to be calculated:

- self-energy matrix $\Sigma_C(E_i)$
- broadening matrix $\Gamma_C(E_i)$
- Green’s function $\mathbf{G}_C^0(E_i)$ of the closed device
- retarded Green’s function $\mathbf{G}_C^R(E_i)$ of the open device

For the latter, a square matrix of dimension N_C has to be inverted (for each energy E_i). The subscript C (contact) indicates that all quantities are reduced contact block matrices of size N_C , i.e. relatively small matrices that have to be evaluated only at the boundary points where the device overlaps with the contact grid points. In a 1D simulation, $N_C = 2$, so that only 2×2 matrices occur. The energy E_i corresponds to the energy E_z of the electron along the z direction because in 1D the transmission coefficient is a function of the energy E_z only: $T(E_i) = T(E_z)$. The energy due to the parallel momentum of the electron does not have to be considered for

calculating T . However, one should keep in mind that the total energy of the electron is given by

$$E_{\text{total}} = E_z + \frac{\hbar^2}{2m_{\parallel}} \mathbf{k}_{\parallel}^2, \tag{22}$$

which becomes relevant when calculating the density and the current.

4.3.1 Self-energy matrix Σ_C

Σ_C is the contact self-energy matrix which represents the coupling of the device to the leads. The self-energy matrix in a real space representation is nonzero only at the boundary points of the device which are in contact with the leads. In a mode space representation (see Sect. 5.2) the self-energy matrix Σ is a diagonal matrix. In 1D the contact self-energy matrix Σ_C has only two nonzero entries on the diagonal ($\Sigma_{\lambda=1}$ and $\Sigma_{\lambda=2}$)

$$\Sigma_C = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}. \tag{23}$$

We assume that each lead is represented by a semi-infinite one-dimensional wire described by a one-band effective mass Hamiltonian. The potential energy E_C^λ of this contact Hamiltonian is equal to the conduction band edge energy of the corresponding grid point at the left or right device boundary. Then the contact self-energy Σ_λ for lead λ is given by [1]

$$\Sigma_\lambda = -t \exp(ik^\lambda \Delta), \tag{24}$$

where t is the kinetic coupling matrix element (also called constant intersite coupling element). It is defined as

$$t = \frac{\hbar^2}{2m} \frac{1}{\Delta^2}, \tag{25}$$

where m is the effective electron mass of the contact, and Δ is the grid spacing of the contact grid point along the propagation direction z . The wave vector $k^\lambda(E_i)$ of lead λ has to be calculated for each energy E_i from the lead dispersion $E(k^\lambda)$. The dispersion of a discrete lattice is given by

$$E(k^\lambda) = E^\lambda + 2t (1 - \cos(k^\lambda \Delta)), \tag{26}$$

where we assume the lead to be discretized with the same grid spacing Δ . Thus the corresponding wave vector k^λ is obtained as follows

$$k^\lambda(E_i) = \frac{1}{\Delta} \arccos\left(\frac{E_i - E^\lambda}{2t} - 1\right). \tag{27}$$

$\arccos(x)$ is the inverse function of the trigonometric $\cos(x)$ function which must be expressed using the complex logarithm

$$\arccos(x) = -i \ln\left(x + i\sqrt{1-x^2}\right) \tag{28}$$

in order to allow for complex k vectors. For real wave vectors, the self-energy Σ_λ corresponds to a traveling plane wave (equation (24)) with a particular energy. The response of the open system to an incident electron wave tells us if this electron wave will be reflected or transmitted. Complex wave vectors, on the other hand, give rise to exponentially rising (unphysical) or decaying waves. Here, we only consider the decaying evanescent waves. In 1D, the conduction band edge energy E_C^λ at the corresponding lead λ has to be taken for the energy E^λ . In general, the relation for the wave numbers k will differ at each contact. In 1D, there is only one mode for each lead, so only one k vector for each lead has to be calculated (for each energy). Consequently, the contact self-energy Σ_λ is a scalar for each lead but in general it is a matrix whose size is determined by the number of contact grid points of this lead (or lead modes taken into account, respectively). In this work, the concept of self-energy only describes the coupling of the device to the leads. However, this concept is far more general and can be used to describe all kinds of interactions, e.g. scattering processes that can be included in more advanced NEGF algorithms [2].

4.3.2 Broadening matrix Γ_C

The broadening matrix Γ_C is the anti-Hermitian part of Σ_C and corresponds to the broadened density of states in the device. It has units of energy and is given by

$$\Gamma_C = i\left(\Sigma_C - \Sigma_C^\dagger\right). \tag{29}$$

The eigenstates of the closed system Hamiltonian correspond to sharp energy levels, and thus they have an infinite lifetime: An electron in one of these states will stay there forever. In contrast, the broadening matrix Γ_C describes the leakage of the eigenstates into the contacts. Consequently, this will lead to a finite lifetime of the electronic states in the device.

4.3.3 Green's function \mathbf{G}_C^0 of the closed device

The reduced contact block matrix $\mathbf{G}_C^0(E_i)$ can be written in terms of the projected wave functions χ_n of the decoupled device Hamiltonian \mathbf{H}^0 at the contact grid points

$$\mathbf{G}_C^0(E_i) = \sum_{n=1}^{n_\alpha} \frac{|\psi_{n,C}\rangle \langle \psi_{n,C}|}{E_i - E_n} = \sum_{n=1}^{n_\alpha} \frac{\chi_n \chi_n^T}{E_i - E_n}. \tag{30}$$

Here we use the spectral representation in terms of the wave function amplitudes $\psi_{n,C}$ of the closed device Hamiltonian at the contacts to calculate the Green’s function \mathbf{G}_C^0 . $\chi_n \chi_n^T$ is the dyadic product where χ_n is a column vector and χ_n^T its transpose, i.e. a row vector, each of size N_C containing the projection of the wave function amplitude onto the two lead grid points (see Sect. 4.2). In 1D this dyadic product leads to a square matrix of dimension $N_C = 2$. Only for the exact solution, one has to take into account all n_T eigenstates ($n_\alpha = n_T$). For 2D and 3D simulations, n_α is typically chosen to be much smaller ($\approx 10\%$ of all eigenstates), making use of an incomplete set of eigenstates. To guarantee an optimal use of the CBR method, the value of n_α should be chosen as small as possible to minimize computational effort. However, one has to ensure that it is still large enough in order to get meaningful results for the energy interval of interest (see Sect. 4.1).

The matrix \mathbf{G}_C^0 can further be understood by decomposing it into submatrices

$$\mathbf{G}_C^0 = \begin{pmatrix} \mathbf{G}_{\lambda\lambda}^0 & \mathbf{G}_{\lambda\lambda'}^0 \\ \mathbf{G}_{\lambda'\lambda}^0 & \mathbf{G}_{\lambda'\lambda'}^0 \end{pmatrix} = \begin{pmatrix} \mathbf{G}_{11}^0 & \mathbf{G}_{12}^0 \\ \mathbf{G}_{21}^0 & \mathbf{G}_{22}^0 \end{pmatrix}. \tag{31}$$

The submatrix $\mathbf{G}_{\lambda\lambda'}^0$ couples lead λ to lead λ' . In 1D this submatrix is a scalar because χ_n^λ is a scalar

$$\mathbf{G}_{\lambda\lambda'}^0(E_i) = \sum_{n=1}^{n_\alpha} \frac{|\psi_{n,C}^\lambda\rangle\langle\psi_{n,C}^{\lambda'}|}{E_i - E_n} = \sum_{n=1}^{n_\alpha} \frac{\chi_n^\lambda \chi_n^{\lambda'}}{E_i - E_n}. \tag{32}$$

4.3.4 Retarded Green’s function \mathbf{G}_C^R of the open device

In order to calculate the transmission coefficient, we first have to evaluate the retarded Green’s function \mathbf{G}_C^R within the contact region from the Dyson equation

$$\mathbf{G}_C^R = \mathbf{A}_C^{-1} \mathbf{G}_C^0, \tag{33}$$

$$\mathbf{A}_C = \mathbf{1}_C - \mathbf{G}_C^0 \mathbf{\Sigma}_C, \tag{34}$$

where $\mathbf{1}_C$ is the identity matrix and $\mathbf{G}_C^0 \mathbf{\Sigma}_C$ is a simple matrix multiplication. \mathbf{G}_C^R is a small submatrix of size 2×2 of the open device’s retarded Green’s function \mathbf{G}^R within the contact regions (see (7)). \mathbf{G}^R has the size of the total number of grid points N_T and is thus a very large matrix for 2D and 3D devices. The direct evaluation of the retarded Green’s function requires the inversion of a large matrix of dimension N_T which is practically impossible for a 3D device, and can be quite demanding even in two spatial dimensions. The essence of the CBR method consists in realizing that for the calculation of the transmission function, only the small part \mathbf{G}_C^R is needed. The determination of this small submatrix from \mathbf{G}^0 and $\mathbf{\Sigma}$ actually requires only the inversion of a matrix that is proportional to the number of grid points N_C that connect the device with the leads.

The inversion of the matrix \mathbf{A}_C to obtain \mathbf{A}_C^{-1} is the central part of the CBR algorithm because in a 2D or 3D simulation, most of the CPU time is consumed here. The inversion can be performed by a standard inversion routine from a numerical library (e.g. LAPACK routine ZGESV [30] which is also available from precompiled libraries that make efficient use of multicore processor architectures). For a matrix of dimension N_C , this usually requires of the order of $(N_C)^{2.8}$ to $(N_C)^3$ operations. Luckily, N_C is generally very small because the number of contact grid points is much smaller than the number of device grid points.

4.3.5 Transmission coefficient

Finally, we calculate for each energy the transmission coefficient $T_{\lambda\lambda'}(E_i)$ from the broadening matrix $\mathbf{\Gamma}_C$ and the retarded Green’s function \mathbf{G}_C^R within the contact region

$$T_{\lambda\lambda'}(E_i) = \text{Tr} \left(\mathbf{\Gamma}_C^\lambda \mathbf{G}_C^R \mathbf{\Gamma}_C^{\lambda'} \mathbf{G}_C^{R\dagger} \right) \quad (\lambda \neq \lambda'), \tag{35}$$

where \dagger indicates the conjugate transpose. The three matrix multiplications only have to be performed for the relevant elements that contribute to the trace of the square matrix of dimension N_C . The elements of the small matrix \mathbf{G}_C^R completely determine the transmission function from lead $\lambda = 1$ to lead $\lambda' = 2$. $\mathbf{\Gamma}_C^\lambda$ is the broadening matrix for lead λ defined analogously to (29). It is nonzero only at the contact points of the relevant lead. In the basis we employ here, it is a diagonal matrix. In fact, for a 1D simulation only one element of this matrix is nonzero and the calculation of the transmission coefficient involves the multiplication of four scalars, two of them are due to the nonzero entries of the broadening matrices of the leads, and the other two originate from the offdiagonal elements of the reduced matrix \mathbf{G}_C^R .

4.4 Transmission function of a double barrier structure (1D Example)

As a simple intuitive example we present in Fig. 1 the calculated transmission coefficient $T(E)$ as a function of energy for a double barrier structure with varying barrier widths of 2 nm, 4 nm and 10 nm (barrier height 100 meV, barrier separation 10 nm, effective mass $m = 0.067m_0$, grid spacing 0.5 nm, device length 50 nm). At 25 meV there is a peak where the double barrier becomes transparent, i.e. $T(E) = 1$. This is exactly the energy that matches the resonant state in the well. The inset shows the conduction band edge profile and the probability density of this quasi-bound resonant state for the case of 10 nm barrier widths where the resonant state hardly couples to the two leads. In the opposite case of strong coupling of this resonant state to the leads (2 nm barrier widths), the local density of states (LDOS) $\rho(z, E)$ around this resonant state broadens, leading to a

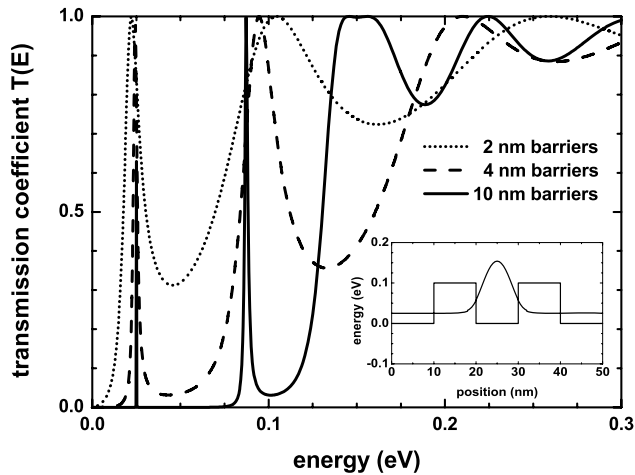


Fig. 1 Calculated transmission coefficient $T(E)$ as a function of energy for a double barrier structure with varying barrier widths of 2 nm, 4 nm and 10 nm (barrier height 100 meV, barrier separation 10 nm). At 25 meV there is a peak where the double barrier becomes transparent, i.e. $T(E) = 1$. This is exactly the energy that matches the resonant state in the well. The inset shows the conduction band edge profile and the probability density of this quasi-bound resonant state for the case of 10 nm barrier widths

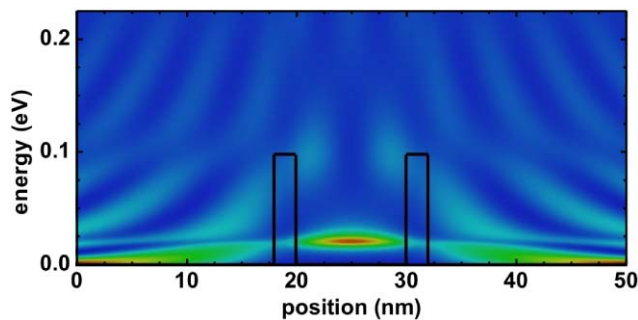


Fig. 2 (Color online) Calculated local density of states $\rho(z, E)$ for a double barrier structure (barrier widths 2 nm, barrier height 100 meV, barrier separation 10 nm). The conduction band edge profile is indicated by the thick solid line. The resonant state inside the double barrier is very broad with respect to energy because it couples strongly to the leads at the left and right boundaries. This is in contrast to the situation for the 10 nm barriers (not shown) where due to the large barrier widths the resonant state is quasi-bound, i.e. with a very sharp and high density of states at the resonance energy because of the very weak coupling to the contacts. Red (blue) color indicates high (low) density of states

broadening of the peaks in the transmission coefficient. This is shown in Fig. 2 where the LDOS is plotted as a function of position and energy for the 2 nm case. The red (blue) color indicates high (low) density of states. This is in contrast to the situation for the 10 nm barriers (not shown) where due to the large barrier widths the resonant state is quasi-bound, i.e. with a very sharp and high density of states at the resonance energy because of the very weak coupling to the contacts. If the energy grid is not fine enough, very sharp resonances can

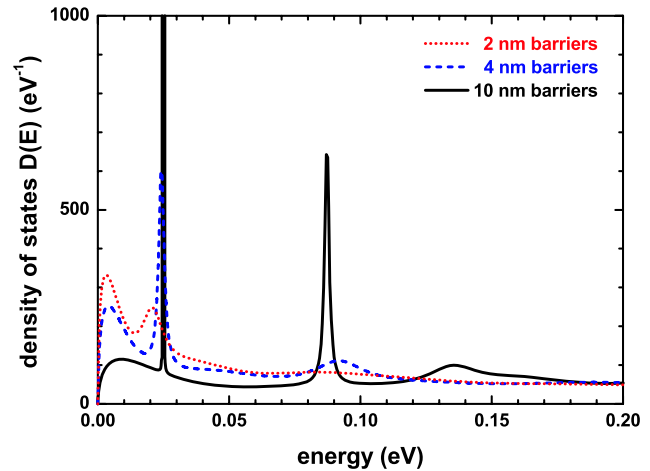


Fig. 3 Calculated density of states $D(E)$ for the double barrier structures. The first peak in the DOS for the 10 nm barrier structure differs substantially from the other two structures because it is extremely sharp and high. The second peak in the DOS at 87 meV due to the second confined well state is only visible for the 10 nm structure. This is consistent to the transmission coefficient (see Fig. 1) which shows a sharp maximum only for the 10 nm structure at this energy

be missed in a numerical calculation. This is the reason why we used an energy grid spacing of 0.5 meV. However, this grid spacing is still not fine enough to get perfect transmission ($T = 1$) for the first peak of the 10 nm barrier structure. Only if the energy grid point exactly matches the resonance energy, the peak would be well resolved. As it is very instructive to investigate the local density of states in different parts of the device, we will show in Sect. 6 how to calculate it with the CBR method.

The calculated density of states (DOS) for the 2 nm, 4 nm and 10 nm double barrier structures is shown in Fig. 3. The DOS corresponds to the LDOS integrated over position. The first peak in the DOS for the 10 nm barrier structure differs substantially from the other two structures because it is extremely sharp and high. It is actually much higher than the figure suggests because its maximum is not included on this scale. The second peak in the DOS at 87 meV due to the second confined well state is only visible for the 10 nm structure. This is consistent to the transmission coefficient (see Fig. 1) which shows a sharp maximum only for the 10 nm structure at this energy.

Figure 4 shows the calculated transmission coefficient $T(E)$ of the 2 nm double barrier structure highlighting the CBR feature of using an incomplete set of eigenstates (10%, 40% and 100% of the eigenstates of the closed device Hamiltonian). Even if only 10% of the eigenstates are used, the first resonance can nicely be reproduced. The cutoff energy in this case is at 180 meV which explains the sudden drop in $T(E)$ for energies exceeding this value. Using 40% of the eigenstates, the main features in the energy interval of

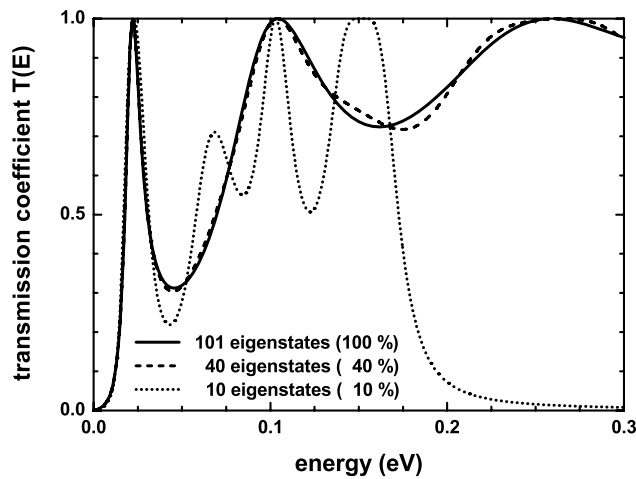


Fig. 4 Calculated transmission coefficient $T(E)$ of a double barrier structure (barrier widths of 2 nm) showing the CBR feature of using an incomplete set of eigenstates (10%, 40% and 100% of the eigenstates of the closed device Hamiltonian). Even if only 10% of the eigenstates are used, the first resonance can nicely be reproduced. The cutoff energy in this case is at 180 meV which explains the sudden drop in $T(E)$ for energies exceeding this value. Using 40% of the eigenstates, the main features in the energy interval of interest can be reproduced very well

interest can be reproduced very well because only the low-energy part of the retarded Green’s function is relevant for the transmission function near the band edge.

In Figs. 1–4 the energy E actually refers to the energy E_z of the electron along the z direction.

5 The CBR method for two- and three-dimensional devices

This section has the same structure as the one for one-dimensional devices. We only mention the differences with respect to the 1D devices. The most important aspect is, that now we have to deal with lead modes. For 2D devices the contacts are one-dimensional lines with one-dimensional eigenfunctions. An example is shown in Fig. 7 that is further discussed in Sect. 5.4. For 3D devices the contacts are two-dimensional surfaces leading to two-dimensional lead eigenfunctions. To obtain these lead eigenfunctions χ_m^λ a corresponding one- or two-dimensional Schrödinger equation has to be solved for each lead. The 1D Schrödinger equation is identical to (14), and the normalization of the wave functions has to be done consistently to the device wave functions. The dimension of the contact Hamiltonian matrix is given by the number N_λ of contact grid points connecting this lead to the device. The total number of modes m^λ of this lead is then also equal to N_λ .

5.1 Energy levels and wave functions of the device Hamiltonian (closed system)

For 2D and 3D devices, the corresponding two-dimensional and three-dimensional Schrödinger equations are solved for the closed system. At the device boundary grid points where the device is in contact to the leads, Neumann boundary conditions are employed along the propagation direction, i.e. perpendicular to the lead line (2D device) or lead surface (3D device). For all other device boundary grid points that are not connected to leads, Dirichlet boundary conditions are taken. Obviously, at these Dirichlet points any possible leakage currents are zero which is reasonable if the barriers are sufficiently high. If leakage currents are important, additional contacts can be placed behind the barriers. Usually, only a small number n_α of the total number n_T of eigenvectors has to be calculated.

5.2 Projection of device eigenfunctions onto lead modes

In general, the vector χ_n of (21) has now the following structure

$$\chi_n = \begin{pmatrix} \chi_{n,m=1}^{\lambda=1} \\ \dots \\ \chi_{n,m=m^1}^{\lambda=1} \\ \dots \\ \chi_{n,m=1}^{\lambda=L} \\ \dots \\ \chi_{n,m=m^L}^{\lambda=L} \end{pmatrix}, \tag{36}$$

and takes into account that for each lead λ ($\lambda = 1, \dots, L$) several lead modes m ($m = 1, \dots, m^\lambda$) exist. The components of the vector χ_n are calculated by projecting the parts of the device eigenvectors $\psi_{n,C}$ (real space representation) that are in contact to the leads into the basis of the orthogonal lead eigenfunctions (mode space representation). For each eigenvalue n , the amplitude of the wave function $\psi_{n,i}^\lambda$ at the contact grid point i is projected onto the amplitude of the lead eigenfunction $\chi_{m,i}^\lambda$ of mode m at this contact grid point

$$\chi_{n,m}^\lambda = \sum_i \langle \psi_{n,i}^\lambda | \chi_{m,i}^\lambda \rangle. \tag{37}$$

The sum runs over all contact grid points i of the relevant lead λ .

Incomplete set of lead modes Within the mode space basis the self-energy matrix Σ_C is diagonal and can be truncated at the cutoff energy. We want to emphasize that the actual number m_α^λ of lead eigenvalues and lead wave functions needed to get meaningful results within the CBR method

can be much smaller than the total number m^λ of lead modes of this lead Hamiltonian matrix. Neglecting nonpropagating high energy modes reduces the size of the contact block matrices that have to be inverted for each energy. The new size is then given by $N_{C,m}$ which is also the size of the vector χ_n that now only takes into account the modes up to m_α^λ for each lead. For an exact solution all modes m^λ have to be included. The energy $E_{m_\alpha^\lambda}$ of the highest lead eigenvector taken into account is the cutoff energy for this lead. It should be significantly above the energy interval of interest in order to get reliable results. This lead mode cutoff energy should have about the same value as for the expansion of \mathbf{G}_C^0 (see (30)). For a 2D and 3D simulation, using such an incomplete set of lead modes will significantly improve the computational performance. For further details on the transformation into the subspace of the propagating lead modes for a 2D or 3D device, we refer to Sect. 5. “Mode space reduction in single-band case” of [25].

5.3 Setup energy interval and calculate properties for each energy E_i

For a 3D simulation, the energy E_i corresponds to the total electron energy whereas for a 2D simulation $E_i = E_{x,y}$ with $E_{x,y}$ being the energy of the electron in the (x, y) plane

$$E_{\text{total}} = E_{x,y} + \frac{\hbar^2}{2m_{\parallel}} k_{\parallel}^2. \tag{38}$$

Here, we assume the device to be translationally invariant along the z direction ($k_{\parallel} = k_z, m_{\parallel} = m_z$).

5.3.1 Self-energy matrix Σ_C

Within the basis of the orthogonal lead eigenfunctions (mode space representation) the self-energy matrix Σ_C is diagonal

$$\Sigma_C = \begin{pmatrix} \Sigma_{\lambda=1} & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \Sigma_{\lambda=L} \end{pmatrix}. \tag{39}$$

For each lead a small diagonal submatrix Σ_λ of dimension m_α^λ has to be calculated. Its components are the contact self-energies Σ_λ^m for each mode m ($m = 1, \dots, m_\alpha^\lambda$) of the relevant lead

$$\Sigma_\lambda^m = -t \exp(ik_m^\lambda \Delta). \tag{40}$$

Therefore a wave vector k_m^λ for each lead and for each transverse mode m has to be calculated (for each energy E_i)

$$k_m^\lambda(E_i) = \frac{1}{\Delta} \arccos\left(\frac{E_i - E_m^\lambda}{2t} - 1\right). \tag{41}$$

Obviously, the propagation direction can now be along the x, y or z direction, depending on the orientation of the lead with respect to the device. Then for the contact boundary grid point the corresponding mass tensor component and the grid spacing Δ along the appropriate propagation direction has to be taken for t . Equation (41) can be derived from the dispersion of a discrete lattice

$$E(k_m^\lambda) = E_m^\lambda + 2t(1 - \cos(k_m^\lambda \Delta)), \tag{42}$$

where E_m^λ is the eigenenergy of the m th mode of lead λ .

5.4 Transmission function of a 2D structure with several barriers (2D Example)

As a simple 2D illustration we take the same example as presented in [3]. The structure consists of three leads with a Gaussian shaped barrier of height 1.0 eV in the middle and a double barrier in the upper part of the device with a height of 0.4 eV. The device has a width of 20 nm and is discretized with 41 grid points in each direction leading to a Hamiltonian matrix of dimension $N_T = 1681$ (grid spacing 0.5 nm). For further details we refer to the original publication [3]. Figure 5 shows the conduction band edge profile and the square of the wave function of the 26th eigenstate which is a resonance state of the device where the transmission coefficient $T_{13}(E)$ between lead 1 and lead 3 shows a local maximum at around 0.18 eV (see Fig. 6). This corresponds to resonant tunneling in the upper path where the electron tunnels through the double barrier. The first peak at

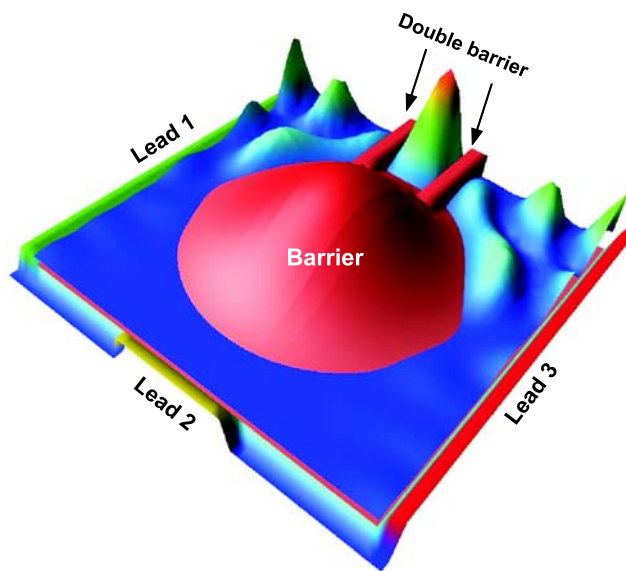


Fig. 5 Conduction band profile (barrier of 2D Gaussian shape and a double barrier of height 0.4 eV) of a 2D device that is connected to three leads. Indicated is the square of the wave function of the 26th eigenstate which is a resonance state of the device where the transmission coefficient $T_{13}(E)$ between lead 1 and lead 3 shows a local maximum at around 0.18 eV (see Fig. 6)

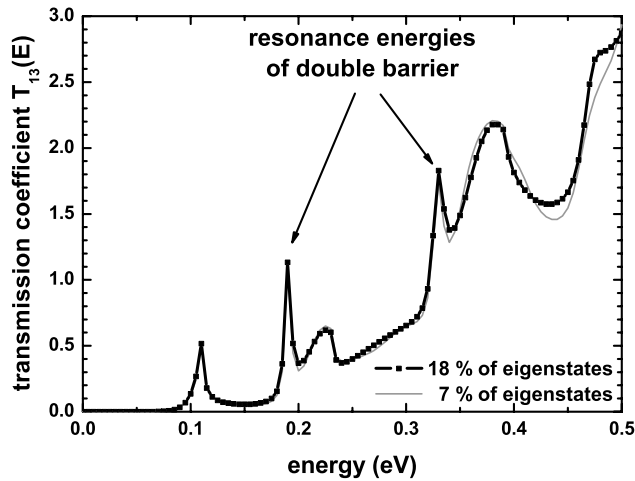


Fig. 6 Transmission coefficient $T_{13}(E)$ between lead 1 and lead 3 using an incomplete set of eigenstates of 7% (thin line) and 18% (thick solid line with squares) of the 2D device Hamiltonian of the closed system for the structure presented in Fig. 5. The first peak at 0.11 eV is a resonance due to electrons traveling the path below the Gaussian shaped barrier, the next two peaks are resonances where the electrons travel the other path where they tunnel through the double barrier

0.11 eV in the calculated transmission coefficient is not due to a resonance of the double barrier—as one might first be tempted to guess. It is related to the electron travelling the lower path around the Gaussian shaped barrier. Such information can be obtained by the visualization of the relevant wave functions or local density of states at this energy (not shown). This example demonstrates that even for very simple structures, it is vital to have access to calculated quantum mechanical properties in order to characterize the peaks correctly.

Figure 6 shows that using an incomplete set of eigenstates of only 7% (118 of 1681) of the 2D device Hamiltonian of the closed system is sufficient to calculate the transmission coefficient up to energies of 0.4 eV. In 1D devices the transmission function cannot exceed the value of 1. For 2D and 3D devices the maximum value of the transmission function is obtained if each of the m^λ lead modes in one lead transmits perfectly to the other lead. So in our example where the leads 1 and 3 each have 41 modes, the maximum of the transmission can certainly exceed $T = 1$ but the upper limit is $T = 41$. Figure 7 shows the calculated lead modes (eigenfunctions of the one-dimensional Schrödinger equation) of lead no. 1 of the same structure. The conduction band edge profile at the contact grid points (squares) is not constant due to the Gaussian shaped barrier in the center of the device that extends to the contacts. Shown are the lowest four eigenenergies (thin, constant lines) and their corresponding probability amplitudes $|\chi_m^{\lambda=1}|^2$ that are shifted with their eigenenergies. The lead modes have been calculated by discretizing the 1D Schrödinger equation with a grid spacing of 0.5 nm and 41 grid points, using Dirichlet boundary conditions. The

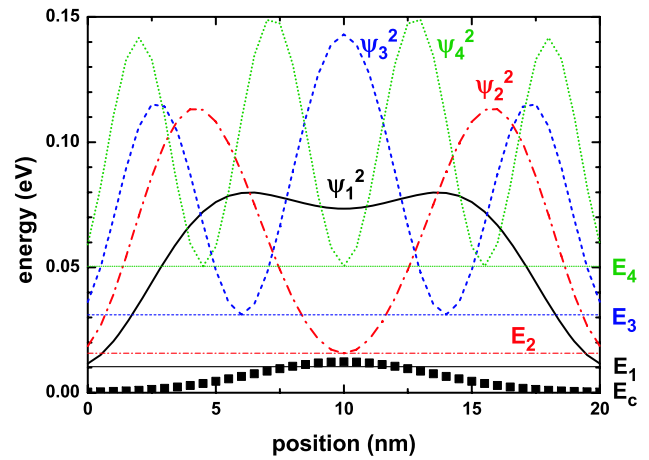


Fig. 7 Calculated lead modes (eigenfunctions of the one-dimensional Schrödinger equation) of lead no. 1 of the same structure as in Fig. 5. The conduction band edge profile E_c at the contact grid points (squares) is not constant due to the Gaussian shaped barrier in the center of the device. Shown are the lowest four eigenenergies (thin, constant lines) and their corresponding probability amplitudes that are shifted with their eigenenergies

lead modes of lead no. 3 are identical because the structure is symmetric.

6 Local density of states

In order to obtain the local density of states (LDOS) for each energy E_i , some additional steps are required. In the following the index C indicates that the matrices have size N_C . If an incomplete set of lead modes has been used (see Sect. 5.2) all these matrices are in fact smaller and have only a size of $N_{C,m}$. However, for better readability we now omit the index m and write only C.

We need the lower left part G_{DC}^R of the retarded Green’s function that correlates the device and the contacts (see (7)). It is obtained from the corresponding Dyson equation

$$G_{DC}^R = G_{DC}^0 B_C^{-1}, \tag{43}$$

$$B_C = I_C - \Sigma_C G_C^0, \tag{44}$$

where I_C is the identity matrix and $\Sigma_C G_C^0$ is a simple matrix multiplication of two small matrices that have been calculated already (see Sects. 4.3.1 and 4.3.3). The matrices G_{DC}^R and G_{DC}^0 are not square matrices. As they correlate the interior device grid points with the leads, they have the dimension $N_D \times N_C$. They are represented within a mixed real space and lead mode space representation. The Green’s function G_{DC}^0 of the closed device can be expressed in terms of the following spectral representation

$$G_{DC}^0 = \sum_{n=1}^{n_\alpha} \frac{|\psi_{n,D}\rangle \langle \psi_{n,C}|}{E_i - E_n}, \tag{45}$$

which reads in mixed real space (index z) and lead mode space (index m)

$$\mathbf{G}_{\text{DC}}^0(z, m) = \sum_{n=1}^{n_\alpha} \frac{\langle z | \psi_{n,\text{D}} \rangle \langle \psi_{n,\text{C}} | m \rangle}{E_i - E_n} \quad (46)$$

$$= \sum_{n=1}^{n_\alpha} \frac{\psi_{n,\text{D}} \chi_n^T}{E_i - E_n} = \Psi_{\text{D}} \mathbf{M}. \quad (47)$$

In fact, this only involves a matrix multiplication $\Psi_{\text{D}} \mathbf{M}$ where the dimensions of the matrices are

$$\begin{aligned} \mathbf{G}_{\text{DC}}^0 &: N_{\text{D}} \times N_{\text{C}}, \\ \Psi_{\text{D}} &: N_{\text{D}} \times n_\alpha, \\ \mathbf{M} &: n_\alpha \times N_{\text{C}}. \end{aligned} \quad (48)$$

The matrix Ψ_{D} contains in columns $1, \dots, n_\alpha$ the wave vectors $\psi_{n,\text{D}}$ of the eigenstate n of the closed device Hamiltonian

$$\Psi_{\text{D}} = (\psi_1 \quad \dots \quad \psi_{n_\alpha}). \quad (49)$$

The matrix \mathbf{M} is defined as

$$\mathbf{M} = \begin{pmatrix} M_1^T \\ \dots \\ M_{n_\alpha}^T \end{pmatrix}, \quad (50)$$

where we store for each eigenvalue E_n the following vector

$$M_n = \frac{1}{E_i - E_n} \chi_n. \quad (51)$$

This is actually the same as (21) (or (36), respectively) apart from the coefficient $1/(E_i - E_n)$. \mathbf{M} is of dimension $n_\alpha \times N_{\text{C}}$ where $n_\alpha \leq n_{\text{T}}$ is the number of eigenvalues taken into account, and N_{C} is the dimension of the χ_n vectors in the mode space representation.

The inverted matrix $\mathbf{B}_{\text{C}}^{-1}$ can be obtained using the same implementation of the inversion algorithm analogously to Sect. 4.3.4. The matrix multiplications involving large matrices ((43) and (47)) can be performed efficiently using standard routines from numerical libraries (e.g. BLAS [31] routines ZGEMM and DGEMM, respectively).

The local density of states $\rho(z, E)$ is the diagonal (divided by 2π) of a more general concept called the spectral function $A(E) = A(z, z', E)$. The local density of states ρ^λ for each lead λ (lead connected local density of states) can easily be calculated at each grid point z from the retarded Green's function $\mathbf{G}_{\text{DC}}^{\text{R}}(z, m)$ (in mixed real space and mode space representation) and the broadening matrix $\mathbf{\Gamma}$

$$\rho^\lambda(z, E_i) = \frac{1}{2\pi} A_\lambda(z, z, E_i) \quad (52)$$

$$= \frac{1}{2\pi} \langle z | \mathbf{G}^{\text{R}} \mathbf{\Gamma}^\lambda \mathbf{G}^{\text{R}\dagger} | z \rangle \quad (53)$$

$$= \frac{1}{2\pi} \sum_{m=1}^{m_\alpha^\lambda} |\mathbf{G}_{\text{DC}}^{\text{R}}(z, m)|^2 \mathbf{\Gamma}_{mm}^\lambda. \quad (54)$$

$\mathbf{\Gamma}^\lambda$ is diagonal in the subspace of the propagating lead modes. The sum runs over all modes m of lead λ . For a 1D simulation, there is only one mode, and thus $\mathbf{G}_{\text{DC}}^{\text{R}}$ is a $N_{\text{D}} \times 2$ matrix and $\mathbf{\Gamma}_{mm}^\lambda$ is the diagonal of $\mathbf{\Gamma}^\lambda$, i.e. only a vector of length 2 has to be stored. The LDOS must have units of $1/\text{energy} \cdot 1/\text{length}$ ($\text{eV}^{-1} \text{nm}^{-1}$) in a 1D simulation (2D: $1/\text{energy} \cdot 1/\text{area}$, 3D: $1/\text{energy} \cdot 1/\text{volume}$). As we normalized the wave functions to be dimensionless, the calculated LDOS has to be divided by the grid spacing Δ for consistency. The total local density of states is simply the sum over the LDOS of each lead

$$\rho(z, E_i) = \sum_{\lambda=1}^L \rho^\lambda(z, E_i). \quad (55)$$

So far we calculated the local density of states only at the interior device grid points N_{D} . All of the equations in this section apply equally well to the contact grid points N_{C} if one replaces the subscript D with the subscript C. In a numerical implementation, one simply has to use in (47) the wave vectors $\psi_{n,\text{T}}$ of the total device

$$\mathbf{G}_{\text{TC}}^0(z, m) = \Psi_{\text{T}} \mathbf{M}, \quad (56)$$

and to replace in (43) and (54) $\mathbf{G}_{\text{DC}}^{\text{R}}$ by $\mathbf{G}_{\text{TC}}^{\text{R}}$ to obtain the local density of states $\rho^\lambda(z, E_i)$ for both interior and contact grid points simultaneously. The matrix Ψ_{T} is stored in memory and has been obtained from numerically solving the Schrödinger equation (15).

Density of states The density of states (DOS) $D^\lambda(E_i)$ for each lead can be obtained by integrating the local density of states $\rho^\lambda(z, E_i)$ for each energy E_i over the spatial coordinate z

$$D^\lambda(E_i) = \int \rho^\lambda(z, E_i) dz = \Delta \sum_{z=1}^{N_{\text{T}}} \rho_z^\lambda(E_i). \quad (57)$$

Thus the DOS $D^\lambda(E_i)$ can easily be obtained by adding the components of the vector that stores $\rho_z^\lambda(E_i)$ and multiplying this sum by the grid spacing Δ . The total density of states is then simply the sum over the DOS for each lead

$$D(E_i) = \sum_{\lambda=1}^L D^\lambda(E_i). \quad (58)$$

The DOS is in units $1/\text{energy}$.

7 Density

The charge density can be calculated via the density matrix or via the local density of states [23]. We recommend to use the local density of states. This is favorable within a self-consistent scheme, since it allows for the use of a predictor-corrector scheme (see Sect. 8.2) to improve the convergence. From the lead connected local density of states $\rho^\lambda(\mathbf{x}, E)$, the local energy resolved carrier density $n_E^\lambda(\mathbf{x}, E)$ for each lead λ is obtained by occupying each level with the distribution function $f(E, \mu_\lambda)$ of the corresponding lead

$$n_E^\lambda(\mathbf{x}, E) = g_s g_v \rho^\lambda(\mathbf{x}, E) f_{dD}(E, \mu_\lambda), \tag{59}$$

where $g_s = 2$ is the spin degeneracy and g_v is the valley degeneracy. The latter is relevant when treating electrons that are in the X or L valleys, like in AlAs, silicon or germanium. In higher dimensions or if these bands are split due to strain, usually for each valley a separate Schrödinger equation has to be solved. Depending on the simulation dimension ($d = 1, 2, 3$) the appropriate Fermi function f_{dD} has to be used which takes into account the \mathbf{k}_\parallel vectors that occur in 1D and 2D simulations. For a device that is homogeneous along the x and y directions (1D simulation) it is given at a particular energy E_z by

$$f_{1D}(E_z, \mu) = \frac{m_\parallel k_B T}{2\pi \hbar^2} \ln \left(1 + e^{-(E_z - \mu)/k_B T} \right), \tag{60}$$

where $m_\parallel(z)$ is the effective mass tensor component in the (x, y) plane of the respective valley (which generally varies with position z and thus has to be averaged over the spatial coordinates weighted with the local density of states for each energy). f_{1D} is in units of 1/area.

The analogous equations for 2D and 3D devices are

$$f_{2D}(E_{x,y}, \mu) = \frac{1}{2} \sqrt{\frac{m_\parallel k_B T}{2\pi \hbar^2}} F_{-1/2}((\mu - E_{x,y})/k_B T) \tag{61}$$

where $m_\parallel(x, y)$ is the effective mass tensor component along the homogenous z direction, and $F_{-1/2}$ is the Fermi-Dirac integral of order $-1/2$ which can be evaluated efficiently using approximation formulas [32]. f_{2D} is in units of 1/length. In 3D the usual Fermi function is used which is of course dimensionless

$$f_{3D}(E, \mu) = \frac{1}{1 + \exp((E - \mu)/k_B T)}. \tag{62}$$

From the lead connected local energy resolved density $n_E^\lambda(\mathbf{x}, E)$, the local carrier density $n^\lambda(\mathbf{x})$ for each lead λ is obtained by integrating over the energy E

$$n^\lambda(\mathbf{x}) = \int n_E^\lambda(\mathbf{x}, E) dE. \tag{63}$$

The total density is the sum over the contributions from all leads

$$n(\mathbf{x}) = \sum_{\lambda=1}^L n^\lambda(\mathbf{x}). \tag{64}$$

The units are 1/volume in all dimensions.

In the explanations above we introduced the term *energy resolved density*. For 1D (2D) simulations this energy $E = E_z$ ($E = E_{x,y}$) did not take into account the energy due to $\mathbf{k}_\parallel \neq 0$. The total energy of the electron is given by (22) for the 1D case and by (38) for the 2D case. It is necessary to include these \mathbf{k}_\parallel contributions into the energy resolved density to get meaningful plots. This is done by first evaluating the local density of states $\rho(z, E_{\text{total}})$ for the total energy, and then occupying the LDOS by the usual Fermi function (62). In 1D simulations, information about the system under study can be obtained by plotting the energy resolved electron density $n(\mathbf{x}, E_{\text{total}})$ and the energy resolved electron density $n^\lambda(\mathbf{x}, E_{\text{total}})$ for each lead. These are two-dimensional plots like the local density of states. The density can be split into two parts, one originating from the left lead, and one from the right lead (see Fig. 10). In 2D simulations the plot of the energy resolved electron density or local density of states is a three-dimensional plot. This makes it difficult to analyze these quantities in 3D simulations where they are four-dimensional. Thus one can only plot slices through these 4D data.

8 Self-consistent CBR algorithm

The self-consistent solution of the ballistic transport properties of an open device requires the repeated solution of the Schrödinger and Poisson equations due to the coupling via the potential and the quantum mechanical density. Also, the lead modes are calculated self-consistently using the potential at the contacts, obtained from the solution of the Poisson equation. In principle, it is possible to simply iterate the solution of these equations, and with enough damping this will lead to yield a converged result. To improve the convergence of a highly nonlinear set of coupled equations, such as the Schrödinger-Poisson problem, the Newton algorithm is usually the first choice. But since the exact Jacobian cannot be derived analytically and a numerical evaluation would be too costly, the simple adaption of this method is not feasible. For the case of a closed system this problem has been solved using a predictor-corrector approach [33]. The aim of this highly efficient method is to find a good approximation for the quantum density as a function of the electrostatic potential where an expression for the Jacobian is known. Within this approximation the nonlinear Poisson equation can efficiently be solved using the Newton scheme resulting in a

predictor update for the electrostatic potential and the carrier density. If this approximation is close enough to the real quantum density, only very few updates will be necessary to yield a converged solution. This means that for each bias step, the Schrödinger equation has to be solved less than approximately 10 times until the potential and the density are sufficiently converged. In the *nin*-resistor example presented below, the Schrödinger equation had to be solved only 2–3 times for each bias step.

8.1 Poisson equation

The Poisson equation describes the electrostatics within the device and reads

$$\nabla \cdot [\varepsilon_0 \varepsilon_r(\mathbf{x}) \nabla \phi(\mathbf{x})] = -\rho(\mathbf{x}), \quad (65)$$

where ε_0 is the permittivity of vacuum and ε_r is the material dependent relative permittivity at position \mathbf{x} . The charge density distribution $\rho(\mathbf{x})$ within a semiconductor device is given by

$$\rho(\mathbf{x}) = e [-n(\mathbf{x}) + p(\mathbf{x}) + N_D^+(\mathbf{x}) - N_A^-(\mathbf{x})], \quad (66)$$

where n and p are the electron and hole densities, and N_D^+ and N_A^- are the ionized donor and acceptor concentrations, respectively. The electron and hole densities can be calculated classically within the Thomas-Fermi approximation or quantum mechanically if quantum confinement effects are important. In this work, we only take into account one conduction band and calculate the electron density n quantum mechanically as described in Sect. 7. We only consider fully ionized donors N_D^+ and neglect all other contributions to the density. The Poisson equation is discretized on a uniform grid with a finite differences method. It is solved numerically by an iterative Newton-Raphson scheme. More details about the numerical solution of the Poisson equation can be found in [18, 28]. For both equilibrium and nonequilibrium calculations, we use Neumann boundary conditions for the Poisson equation which implies a vanishing electric field at the boundaries

$$\frac{\partial \phi}{\partial z} = 0. \quad (67)$$

This is the recommendation for ballistic devices [1]. An alternative would be to use Dirichlet boundary conditions for nonequilibrium simulations [18]. Here, one first has to determine the electrostatic potential in equilibrium (built-in potential) using zero-field (Neumann) boundary conditions. The electrostatic potential at the boundaries is then fixed with respect to the chemical potentials taking into account the previously calculated built-in potential at the boundaries. For both boundary conditions, the chemical potentials at the contacts are fixed and correspond to the applied bias. Further

boundary conditions are summarized in [34]. These include the concept of a drifted Fermi distribution function in the leads that accounts for a net current flow in those leads.

8.2 Predictor-corrector approach

A fast and robust iterative method for obtaining self-consistent solutions to the coupled system of Schrödinger and Poisson equations is very important. Basically, a simple expression describing the dependence of the quantum electron density on the electrostatic potential is required ($\partial \rho / \partial \phi$). This expression is then used to implement an iteration scheme, based on a predictor-corrector type approach, for the solution of the coupled system of differential equations. Within the CBR method, a predictor-corrector approach can easily be applied making use of the previously calculated local density of states by modifying (59) slightly to get the energy resolved density for the predictor potential. This predictor density $n_{E,p}^\lambda(\mathbf{x}, E, \Delta_\phi)$ is then given by

$$n_{E,p}^\lambda = g_s g_v \rho^\lambda(\mathbf{x}, E) f_{dD}(E - e\Delta_\phi(\mathbf{x}), \mu_\lambda). \quad (68)$$

The idea behind this approximation is that to first order the wave functions, and therefore the local density of states $\rho^\lambda(\mathbf{x}, E)$, remain unchanged for small deviations in the potential. Only the eigenenergies are adjusted locally to small changes in the electrostatic potential $\Delta_\phi(\mathbf{x})$. This is achieved by using $E - e\Delta_\phi(\mathbf{x})$ instead of E as the new argument for the Fermi function f_{dD} .

The charge density used in the Poisson equation is a function of the electrostatic potential ('nonlinear' equation). The nonlinear Poisson equation can be solved very fast using a predictor density. This density avoids the time-consuming procedure of solving the Schrödinger equation many times. Once the new electrostatic potential for the predictor density has been obtained, the new quantum mechanical density, i.e. the new local density of states for this potential can be evaluated. This procedure is iterated until convergence of both the electrostatic potential and the quantum mechanical charge density is achieved.

The nonlinear Poisson equation itself is solved by a Newton-Raphson method where the functional

$$\mathbf{F} = \mathbf{A} \cdot \phi + \rho = 0 \quad (69)$$

is minimized. Here, \mathbf{A} represents the discretized Poisson matrix and ρ is a vector representing the charge density for each grid point. The Newton algorithm finds an electrostatic potential vector $\phi^{j+1} = \phi^j + \Delta_\phi$ such that the magnitude of the residuum vector \mathbf{F} becomes smaller than a certain small threshold of ε . The electrostatic potential ϕ^j of the j th iteration step is kept fixed within the Newton method. The index j refers to the outer Schrödinger-Poisson iteration and counts how often the Schrödinger equation has to be solved

until convergence is obtained. Once the Newton algorithm has converged to a correction $\Delta\phi$, the Schrödinger equation (i.e. the CBR algorithm) is solved for the updated electrostatic potential ϕ^{j+1} . The new local density of states is then input to the next iteration of the Newton algorithm.

For the Newton correction, the Jacobi matrix \mathbf{J} is needed. It is simply the Poisson matrix plus the derivative of the density with respect to the potential

$$\mathbf{J} = \frac{\partial \mathbf{F}}{\partial \phi} = \mathbf{A} + \frac{\partial \rho}{\partial \phi} = \mathbf{A} + \frac{\partial n_p}{\partial \phi}. \tag{70}$$

Thus within the CBR method, the derivative of the predictor density n_p with respect to the potential is needed. This derivative is available using (68) and the derivative of the Fermi functions f_{dD} with respect to $\Delta\phi$.

The iteration approach presented in this section simplifies the numerical implementation of the nonlinear Schrödinger-Poisson problem significantly. In addition, it provides excellent convergence speed and stability. Further details about it can be found in [23].

8.3 Self-adapting energy grid

For the numerical implementation of a self-consistent scheme using a continuous density of states, the energy grid is of high importance. To integrate the carrier density, we discretize the local density of states in energy space and then employ a simple numerical integration by summing up the values for each energy step weighted by the Fermi distribution and the energy grid spacing ΔE . Since the DOS is a very spiky function with peaks corresponding to highly localized states due to the onset of the conduction band edges at the contacts (1D) or due to the propagating lead mode energies (2D/3D), it is very important for the convergence of the self-consistent CBR algorithm to have these features properly resolved. Additional peaks arise from quasi-bound states, like for instance in the double barrier structure (10 nm barrier widths) as discussed in Sect. 4.4. Usually the main structural features in the DOS are due to the lead modes. If quasi-bound states are the dominant features in the DOS, one could use the information about their energy levels (which is available within the CBR method) to optimize the energy grid. Thus we need an energy grid that is self-adapting to the density of states which varies for each iteration. Otherwise, a well converged self-consistent solution is not possible unless a lot of energy grid points are used.

In 2D simulations of e.g. a double gate MOSFET where the channel acts as a one-dimensional wire, the peaks show a $1/\sqrt{(E_i - E_m^\lambda)}$ dependence, where E_m^λ is the peak energy arising due to the onset of the lead modes (2D/3D). The peaks in our 1D *nin*-resistor example (Fig. 11) show also a $1/\sqrt{(E_i - E_m^\lambda)}$ dependence, where $E_m^\lambda = E_c^\lambda$ is the peak energy arising due to the onset of the conduction band edges

at the contacts (1D). The integral over the peak is thus very poor when using an energy grid with constant grid spacing (uniform energy grid, see Sect. 4.3), since the relative distance between the nearest energy grid point E_i and the peak energy E_m^λ is arbitrary. Additionally, the lead mode energy is slightly shifted with each iteration step, leading to a varying integration error during the self-consistent cycle, which is a sure kill for any self-consistent algorithm. Thus a solution to this problem is to use the physical information we have about the system and employ a self-adapting energy grid that resolves each known (i.e. relevant) peak m with a local energy grid of a few tens of energy grid points that is fixed to the lead mode energy E_m^λ . Additionally, extra points are distributed in the space between the peaks to obtain a smooth enough energy grid. An exponential grid type is recommended since it provides a good resolution of the $1/\sqrt{E}$ behavior of the peaks. In order to avoid singularities the energy grid points are not allowed to match exactly the eigenenergies of the closed system. For each peak, the first grid point is set slightly below the onset of the peak and then each grid point i is set with increasing energy grid spacing

$$\Delta_{E^i} = g \Delta_{E^{i-1}} = g^i \Delta_{E^0}, \tag{71}$$

starting with the initial grid spacing $\Delta_{E^0} = 0.1$ meV, and a grid factor $g = 1.2$, for instance. A grid factor of $g = 1.0$ leads to a locally linear grid which has been found to be not as efficient as the exponential grid. The parameters that specify the energy grid are the total number of energy grid points, the maximum number of peaks taken into account, the number of energy grid points in the local grid around a peak, and the grid factor. The minimum value of the energy grid should be slightly below the minimum of the conduction band edges of the contacts, the maximum value should not be higher than $E_{\max} = 0.25t$, where t is defined analogously to (25).

Figure 11 demonstrates that the peaks in the LDOS and DOS of a simple *nin*-structure (see Sect. 8.5) are well resolved, and that for other regions in the energy interval less grid points are fully sufficient. The energy grid consists of 300 grid points including the extra points used to resolve the onsets of the two peaks at the conduction band edges of the contacts. Importantly, the integration error is reduced compared to the uniform grid and remains constant within the iteration, since the grid is locally fixed to the shifting lead mode energies (or conduction band edges in a 1D simulation). The convergence behavior of a uniform grid with an order of magnitude more energy grid points is very similar for the first iteration steps. The achieved convergence is measured by a residuum which is a very small number. Compared to the self-adapting grid, the uniform grid reaches a bottom at the residuum, which cannot be reduced further. This is due to the fluctuating integration error. In contrast,

the self-adapting energy grid guarantees satisfying convergence.

As the contact block matrices have to be inverted for each energy, the computational time depends linearly on the total number of energy grid points. Therefore, a numerical implementation of an optimized energy grid is very important for an efficient use of the CBR method.

8.4 Extracting the quasi-Fermi level

For all calculations presented in this paper, the extraction of the quasi-Fermi level was not necessary because only one conduction band has been involved. For equilibrium solutions, we so far assumed that the Fermi level (chemical potential) is constant and fixed to $E_F = 0$ eV, allowing the semiconductor band edges to adjust according to the electrostatic potential as calculated from the Poisson equation (see Sect. 8.1). For nonequilibrium calculations where the device is under bias, one could extract a spatially varying quasi-Fermi level $E_F(z)$ in order to get meaningful (or to avoid artificially wrong) charge densities for all other bands that are not treated quantum mechanically with the CBR method, e.g. hole bands or higher lying electron bands. This might be necessary for the self-consistent CBR algorithm under high bias conditions, where for each iteration the quasi-Fermi levels have to be obtained self-consistently. The reason is that the equation for the classical densities needs a reasonable value for the local quasi-Fermi level. In nonequilibrium calculations electrons and holes can be described by different quasi-Fermi levels ($E_{F,n}(z)$, $E_{F,p}(z)$, respectively). The quasi-Fermi level for electrons can be obtained by finding (e.g. using a bisection algorithm) for each grid point z the appropriate local quasi-Fermi level $E_{F,n}(z)$ that corresponds to the actual electron density at this grid point (and similar for the holes). These Fermi levels would lie in between the chemical potentials of the left and right contact which are kept fixed in a nonequilibrium calculation. Rather than occupying the lead connected local density of states with the chemical potential of the relevant lead (equation (59)), one would occupy the total local density of states (equation (55)) at position z by taking an average ($E_{F,n}(z)$) of the chemical potentials of all leads.

Bound states treatment Electronic states that are below the conduction band edges of the contacts do not get occupied within a ballistic algorithm. All higher lying states contribute via the local density of states to the quantum mechanical density. It is not a realistic treatment to ignore the lower lying bound states as they usually get filled through scattering events. Therefore the density originating from the bound states obviously contributes to the electrostatics of the device and should be included into the Poisson equation. An example calculation of a quantum well that is completely

empty within a ballistic calculation but gets filled once scattering is included has been discussed in detail in [35]. As the probability densities of the electronic states are available (see Sect. 4.1), one could use this information and occupy the states that are below the conduction band edges of the contacts locally with a self-consistently determined quasi-Fermi level. This is the standard approach usually employed in Schrödinger-Poisson solvers. Here, however, for energies where the LDOS from the ballistic calculation is available, the CBR density is used instead. So the total density has two contributions, one from the bound states and one from the CBR density. Another approach how to include bound states is described in [23].

8.5 *nin*-resistor (1D example)

As a simple example to illustrate the self-consistent CBR method, we choose a *nin*-structure where quantum confinement effects are not relevant. Hence, the equilibrium solution can be easily checked against the standard approach for calculating the carrier concentration in semiconductor devices. This classical density is obtained using Fermi-Dirac integrals and the effective density of states of the conduction (and valence) bands. We emphasize that in Figs. 8 and 9 we only used quantum mechanical densities calculated with the CBR method (see Sect. 7). In equilibrium, the CBR approach leads to the same conduction band profile and the same carrier densities as the classical approach (not shown). The *nin*-structure consists of GaAs and has a length of 80 nm. The doping profile is symmetric with a donor concentration of $N_D^+ = 1 \cdot 10^{18} \text{ cm}^{-3}$ (fully ionized). The 35 nm

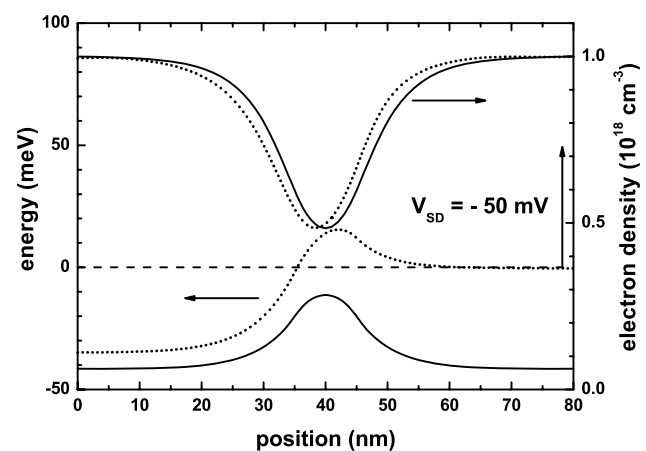


Fig. 8 Conduction band edge profiles and electron densities of a symmetric *nin*-structure calculated with the self-consistent CBR method. Solid lines are equilibrium results, dotted lines correspond to an applied bias of $V_{SD} = -50$ mV at the right contact. The chemical potential in equilibrium is equal to $\mu = 0$ meV (dashed line). Under bias, the chemical potential of the right contact is increased by 50 meV, indicated by the vertical arrow. The doping profile is symmetric ($N_D^+ = 1 \cdot 10^{18} \text{ cm}^{-3}$)

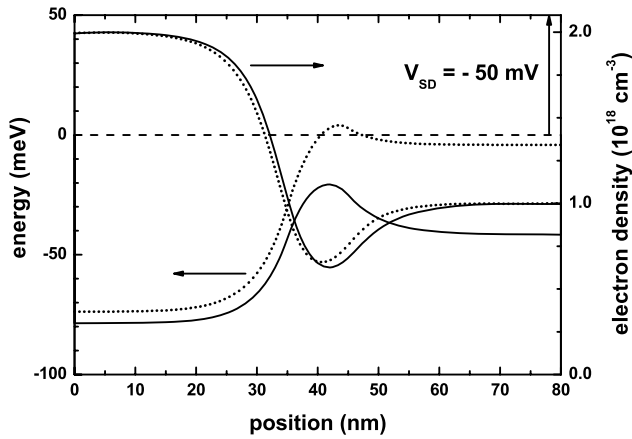


Fig. 9 Same as Fig. 8 but for the n^+in -structure that has an asymmetric doping profile (0–35 nm: $N_D^+ = 2 \cdot 10^{18} \text{ cm}^{-3}$, 45–80 nm: $N_D^+ = 1 \cdot 10^{18} \text{ cm}^{-3}$)

wide n -type doped regions at the source and drain sides are separated by a 10 nm wide intrinsic region in the middle. For comparison, we also study an asymmetrically doped n^+in -structure where the left n -type region has a doping concentration of $N_D^+ = 2 \cdot 10^{18} \text{ cm}^{-3}$ (0–35 nm) and the right doping region has a concentration of $N_D^+ = 1 \cdot 10^{18} \text{ cm}^{-3}$ (45–80 nm). The temperature is set to 300 K. The device is discretized with a grid spacing of 1 nm. A self-adapting energy grid of 300 energy grid points is used. For each bias point, it is sufficient to solve the Schrödinger equation only 2–3 times to get well converged results. This shows that this nin -resistor is well suited as a test case to benchmark an efficient implementation of both the self-adapting energy grid, and the predictor-corrector algorithm.

Figures 8 and 9 show the conduction band edge profiles and electron densities of the symmetric nin - and asymmetric n^+in -structures, respectively, calculated with the self-consistent CBR method. The solid lines are equilibrium results and the dotted lines correspond to an applied bias of $V_{SD} = -50 \text{ mV}$ at the right contact. The chemical potential in equilibrium is equal to $\mu = 0 \text{ meV}$ (dashed line). Under bias, the chemical potential of the right contact is increased by 50 meV, indicated by the vertical arrows. As a consequence of the zero-field boundary conditions for the Poisson equation, the band edges are flat at the contacts. However, for the symmetric nin -structure the difference in the conduction band edges at the left and right contact is smaller than the actual difference in the chemical potentials. The same holds for the asymmetric n^+in -structure if one takes the built-in potential (of the equilibrium calculation) into account. The reason for this behavior is as follows (see Chap. 11.4 ‘Where is the voltage drop’ of [1]): In ballistic simulations a fraction of the density of states at one contact is always controlled by the contact at the other end. Making the end regions of the device longer will not change this

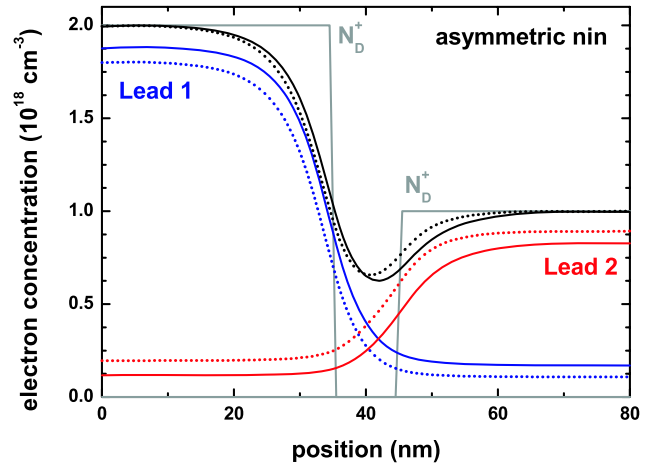


Fig. 10 (Color online) Individual electron densities due to the left (blue lines) and the right (red lines) contact are shown for the asymmetric n^+in -structure. The total density is identical to Fig. 9 (black lines). Solid lines are equilibrium results, dotted lines correspond to the applied bias $V_{SD} = -50 \text{ mV}$. The donor concentration profile N_D^+ is also shown. Raising the chemical potential at the right contact increases (decreases) the density due to the occupation of the corresponding local density of states of the right (left) contact

situation. This can easily be understood by separately visualizing the electron densities that originate from the left and right contacts. This is shown in Fig. 10 where the individual electron densities due to the left (blue lines) and the right (red lines) lead are shown for the asymmetric n^+in -structure. The total density is identical to Fig. 9 (black lines). Solid lines are equilibrium results, dotted lines correspond to the applied bias $V_{SD} = -50 \text{ mV}$. The donor concentration profile N_D^+ is also shown. Raising the chemical potential at the right contact increases the density due to the occupation of the corresponding local density of states of this contact (lead connected local density of states, see Sect. 6). Consequently, the density due to the other lead must decrease to guarantee global charge neutrality. There are two ways for the density to decrease, one is changing the chemical potential of the relevant lead (which is not possible as it is fixed due to the boundary condition), the other possibility is to adjust the electrostatic potential, and thus the conduction band edge. The latter situation corresponds to zero-field boundary conditions (Neumann). This explains why Dirichlet boundary conditions are inappropriate for ballistic devices. For quantum cascade laser (QCL) simulations where the doping concentration is low, Neumann boundary conditions seem to be a natural choice where one allows the derivative of the potential at the left and right boundaries

$$\frac{\partial \phi}{\partial z} = \text{const} \tag{72}$$

to adjust self-consistently under the condition of global charge neutrality, i.e. requiring equal slope at the bound-

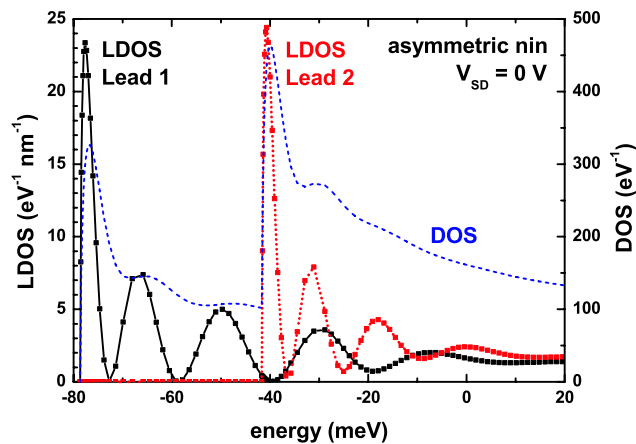


Fig. 11 (Color online) Local density of states (LDOS) at the left (Lead 1, black solid line) and right (Lead 2, red dotted line) contact of the asymmetric n^+in -structure in equilibrium. The self-adapting energy grid is able to resolve the peaks in the LDOS sufficiently accurate (300 energy grid points in total). Also shown is the density of states (DOS, blue dashed line) which is the integrated LDOS over the position (sum over all lead contributions). The DOS has peaks at the onset of the conduction bands edges at the left and right contacts

aries. The slope is adjusted in such a way that the potential drop across the device equals the bias voltage that is defined by the difference between the chemical potentials in the contacts [2]. This will lead to finite electric fields at the boundaries that correspond to the applied electric field in the QCL.

Figure 11 shows the local density of states at the left (Lead 1, black solid line) and right (Lead 2, red dotted line) contact of the asymmetric n^+in -structure in equilibrium, i.e. one-dimensional slices at the first ($z = 1$) and last ($z = N_T$) grid point of the two-dimensional LDOS $\rho^\lambda(z, E_i)$ plot of lead λ (equation (54)). The self-adapting energy grid is able to resolve the peaks in the LDOS sufficiently accurate (300 energy grid points in total). This is very important in a self-consistent algorithm to ensure converged results for the electron density which has to be integrated over energy (equation (63)). Also shown is the density of states (DOS, blue dashed line) which is the integrated LDOS over the position for each lead, and then the contributions of each lead are added (equation (58)). The DOS has peaks at the onset of the conduction bands edges at the left and right contacts. Note that the energy axis corresponds to the energy E_z along the z direction and not to the total energy E_{total} which includes the integration over \mathbf{k}_\parallel . The spin degeneracy factor is included in this figure.

The linear regime of the current-voltage characteristics of the symmetric (solid line) and asymmetric (dotted line) nin -structures has been calculated with the self-consistent CBR method and is shown in Fig. 12. For the asymmetric n^+in -resistor the applied voltage corresponds to reverse bias operation. In comparison to the symmetric nin -structure, the

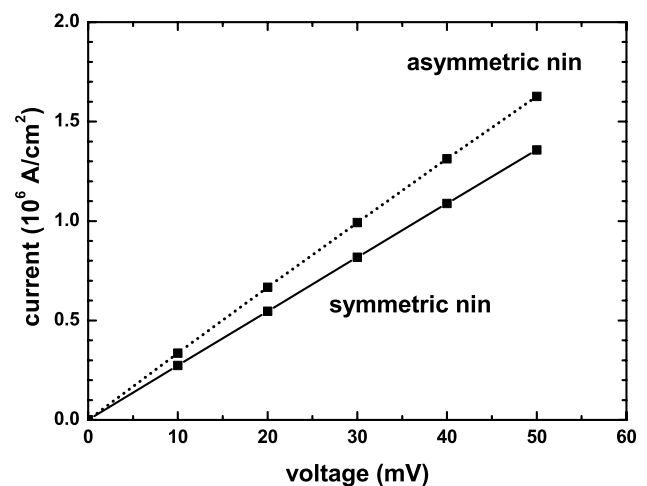


Fig. 12 Linear regime of the current-voltage characteristics of the symmetric (solid line) and asymmetric (dotted line) nin -structures calculated with the self-consistent CBR method at a temperature of 300 K

asymmetric resistor shows a higher current density because the effective barrier width and the effective barrier height due to the intrinsic region is reduced. In this small device, the limiting case of ballistic quantum transport is a suitable approximation. For such low biases, the calculated current density of the ballistic calculations do not deviate strongly from calculations that include both scattering and more advanced lead models (not shown). The main reason is the absence of confined states below the conduction band edges of the leads. These states get only filled if scattering is present, and can then influence the charge carrier distribution significantly. In such a case, a ballistic quantum transport model is not adequate. When modeling resonant tunneling devices and especially quantum cascade lasers, it is very important to include scattering. The latter might be designed based on e.g. resonant conditions with longitudinal optical phonons (LO-phonon scattering).

9 Current

The transmission function $T(E)$ can be computed, once the band edge profile of the device has been obtained by means of a charge self-consistent calculation. The ballistic current from lead λ to lead λ' can be calculated based on the Landauer-Büttiker formula (see (1)). This equation has to be adjusted for 1D and 2D simulations if the transmission coefficient is a function of E_z or $E_{x,y}$, respectively, rather than of the total energy. In 1D, the integration has to be performed over E_z , and the Fermi function $f(E, \mu)$ has to be replaced with the corresponding Fermi function $f_{1D}(E_z, \mu)$ given in (60). In 2D, the integration has to be performed over $E_{x,y}$ and (61) shows the appropriate Fermi function

$f_{2D}(E_{x,y}, \mu)$. The Fermi functions include the corresponding units, so the current in 1D is given in units of [A/m²], in 2D in [A/m] and in 3D in [A].

If more than two leads are present in the device, then for the total current through a particular lead λ the contributions from all other leads λ' have to be summed up

$$I_{\lambda} = \sum_{\lambda'=1}^L I_{\lambda\lambda'} \quad (\lambda' \neq \lambda). \quad (73)$$

10 Conclusions

The contact block reduction method is a variant of the non-equilibrium Green's function formalism. In this article we presented a numerical implementation of the CBR method in detail. Charge self-consistent calculations can be performed very efficiently even for 3D structures by means of the CBR approach. Once the potential profile of a device with an arbitrary number of contacts has been obtained, the ballistic current can be calculated based on the Landauer-Büttiker formula. The presented model is suitable for extremely small devices or very low temperatures, where the incoherent scattering lengths exceed the geometrical device size.

Acknowledgements Financial support of the German Excellence Initiative via the 'Nanosystems Initiative Munich' (NIM), the Austrian Science Fund FWF (SFB IR-ON), and the Deutsche Forschungsgemeinschaft (SFB 631) is gratefully acknowledged.

References

- Datta, S.: Quantum Transport: Atom to Transistor. Cambridge University Press, Cambridge (2005)
- Kubis, T., Yeh, C., Vogl, P., Benz, A., Fasching, G., Deutsch, C.: Theory of non-equilibrium quantum transport and energy dissipation in terahertz quantum cascade lasers. *Phys. Rev. B* **79**, 195323 (2009)
- Mamaluy, D., Sabathil, M., Vogl, P.: Efficient method for the calculation of ballistic quantum transport. *J. Appl. Phys.* **93**, 4628 (2003)
- The nextnano software can be obtained from <http://www.nextnano.de> and <http://www.wsi.tum.de/nextnano>. A demo that includes a Windows executable and the input files of the CBR examples presented in the figures of this article can be downloaded from this link: <http://www.nextnano.de/customer/downloadCBR.php> (Online Resource)
- Datta, S.: Electronic Transport in Mesoscopic Systems. Cambridge University Press, Cambridge (1995)
- Landauer, R.: Spatial variation of currents and fields due to localized scatterers in metallic conduction. *IBM J. Res. Develop.* **32**, 306 (1988)
- Landauer, R.: Conductance from transmission: common sense points. *Phys. Scr.* **T42**, 110 (1992)
- Büttiker, M.: Four-terminal phase-coherent conductance. *Phys. Rev. Lett.* **57**, 1761 (1986)
- Büttiker, M.: Symmetry of electrical conduction. *IBM J. Res. Develop.* **32**, 317 (1988)
- Di Carlo, A., Vogl, P., Pötz, W.: Theory of Zener tunneling and Wannier-Stark states in semiconductors. *Phys. Rev. B* **50**, 8358 (1994)
- Büttiker, M.: Small normal-metal loop coupled to an electron reservoir. *Phys. Rev. B* **32**, 1846 (1985)
- Venugopal, R., Paulsson, M., Goasguen, S., Datta, S., Lundstrom, M.S.: A simple quantum mechanical treatment of scattering in nanoscale transistors. *J. Appl. Phys.* **93**, 5613 (2003)
- Kane, E.O.: Tunneling Phenomena in Solids. Plenum, New York (1969), ed. by E. Burstein and S. Lundqvist
- Schulman, J.N., Chang, Y.C.: Reduced Hamiltonian method for solving the tight-binding model of interfaces. *Phys. Rev. B* **27**, 2346 (1983)
- Lent, C., Kirkner, D.: The quantum transmitting boundary method. *J. Appl. Phys.* **67**, 6353 (1990)
- Smrčka, L.: R-matrix and the coherent transport in mesoscopic systems. *Superlattices Microstruct.* **8**, 221 (1990)
- Ferry, D.K., Goodnick, S.M.: Transport in Nanostructures. Cambridge University Press, Cambridge (1997)
- Lake, R., Klimeck, G., Bowen, R.C., Jovanovic, D.: Single and multiband modeling of quantum electron transport through layered semiconductor devices. *J. Appl. Phys.* **81**, 7845 (1997)
- Kim, R., Datta, S., Lundstrom, M.S.: Influence of dimensionality on thermoelectric device performance. *J. Appl. Phys.* **105**, 034506 (2009)
- Sabathil, M., Birner, S., Mamaluy, D., Vogl, P.: Efficient computational method for ballistic currents and application to single quantum dots. *J. Comput. Electron.* **2**, 269 (2003)
- Sabathil, M., Mamaluy, D., Vogl, P.: Prediction of a realistic quantum logic gate using the contact block reduction method. *Semicond. Sci. Technol.* **19**, S137 (2004)
- Khan, H.R., Mamaluy, D., Vasileska, D.: Quantum transport simulation of experimentally fabricated nano-FinFET. *IEEE Trans. Electron Devices* **54**, 784 (2007)
- Vasileska, D., Mamaluy, D., Khan, H.R., Raleva, K., Goodnick, S.M.: Semiconductor device modeling. *J. Comput. Theor. Nanosci.* **5**, 1 (2008)
- Zibold, T., Vogl, P., Bertoni, A.: Theory of semiconductor quantum-wire-based single- and two-qubit gates. *Phys. Rev. B* **76**, 195301 (2007)
- Mamaluy, D., Vasileska, D., Sabathil, M., Zibold, T., Vogl, P.: Contact block reduction method for ballistic transport and carrier densities of open nanostructures. *Phys. Rev. B* **71**, 245321 (2005)
- Ryu, H., Klimeck, G.: Contact block reduction method for ballistic quantum transport with semi-empirical sp³d⁵s* tight binding band models. In: 9th International Conference on Solid-State and Integrated-Circuit Technology (ICSICT 2008), p. 349 (2008)
- Birner, S., Zibold, T., Andlauer, T., Kubis, T., Sabathil, M., Trellick, A., Vogl, P.: nextnano: General Purpose 3-D Simulations. *IEEE Trans. Electron Devices* **54**, 2137 (2007)
- Tan, I.-H., Snider, G.L., Chang, L.D., Hu, E.L.: A self-consistent solution of Schrödinger-Poisson equations using a nonuniform mesh. *J. Appl. Phys.* **68**, 4071 (1990)
- ARPACK—ARnoldi PACKage, The ARPACK source code is available from <http://www.netlib.org/arpack>. Lehoucq, R.B., Sorensen, D.C., Yang, C., ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods. SIAM, Philadelphia (1998)
- LAPACK—Linear Algebra PACKage, The LAPACK source code is available from <http://www.netlib.org/lapack>. Optimized libraries containing the LAPACK routines are included in several compiler suites
- BLAS—Basic Linear Algebra Subprograms, The BLAS source code is available from <http://www.netlib.org/blas>. Optimized libraries containing the BLAS routines are included in several compiler suites

32. Antia, H.M.: Rational function approximations for Fermi-Dirac integrals. *Astrophys. J. Suppl.* **84**, 101 (1993)
33. Trellakis, A., Galick, A.T., Pacelli, A., Ravaioli, U.: Iteration scheme for the solution of the two-dimensional Schrödinger-Poisson equations in quantum structures. *J. Appl. Phys.* **81**, 7880 (1997)
34. Laux, S.E., Kumar, A., Fischetti, M.V.: Analysis of quantum ballistic electron transport in ultrasmall silicon devices including space-charge and geometric effects. *J. Appl. Phys.* **95**, 5545 (2004)
35. Kubis, T., Trellakis, A., Vogl, P.: Self-consistent quantum transport theory of carrier capture in heterostructures. In: Saraniti, M., Ravaioli, U. (eds.) *Proceedings of the 14th International Conference on Nonequilibrium Carrier Dynamics in Semiconductors (HCIS 14, Chicago, USA)*. Springer Proceedings in Physics, vol. 110, p. 369. Springer, Berlin (2005)