Materials Science and Engineering R 38 (2002) 181–236

R Reports: *A Review Journal*

www.elsevier.com/locate/mser

Computational electronics

Dragica Vasileska, Stephen M. Goodnick*

Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287-5706, USA

Abstract

In this review article we give an overview of the basic techniques used in the field of computational electronics related to the simulation of state-of-the art devices fabricated in a variety of device technologies. We begin with a review of the electronic band structure and the associated dynamics of the carriers under external fields, followed by a discussion of the basic equations governing transport in semiconductors, and leading to the description of the Monte Carlo method for the solution of the Boltzmann transport equation and the simplified hydrodynamic and drift-diffusion models. We also give an overview of field solvers for both high-frequency and low-frequency application, followed by a description of particle-based simulation tools for both low and high-frequency applications. The need of more sophisticated simulation tools that go beyond the Boltzmann transport picture is also addressed, and is followed with the description of the two approaches that allow successful and rather inexpensive (in terms of needed CPU time) incorporation of quantum-mechanical space-quantization effects into existing semi-classical device simulators: the effective-potential approach and the quantum hydrodynamic model. Examples derived from our own research are given throughout the text to illustrate the usefulness and the limitations of the computational techniques discussed in this review.

Keywords: Device simulation; Band structure; Boltzmann transport equation; Monte Carlo method; Hydrodynamic model; Poisson equation solvers; Maxwell's equations; Particle-based simulations; Effective potential; Quantum hydrodynamic model

1. Introduction

As the density of integrated circuits continues to increase, there is a resulting need to shrink the dimensions of the individual devices of which they are comprised. Smaller circuit dimensions reduce the overall die area, thus allowing for more transistors on a single die without negatively impacting the cost of manufacturing. As semiconductor feature sizes shrink into the nanometer scale regime, device behavior becomes increasingly complicated as new physical phenomena at short dimensions occur and limitations in material properties are reached. In addition to the problems related to the actual operation of ultra-small devices, the reduced feature sizes require more complicated and time-consuming manufacturing processes. This fact signifies that a pure trial-and-error approach to device optimization will become impossible since it is both too time-consuming and expensive. Since computers are considerably cheaper resources, simulation is becoming an indispensable tool for the device engineer. Besides offering the possibility to test hypothetical devices which have not (or could not) yet been manufactured, simulation offers unique insight into device behavior by allowing the observation of phenomena that can not be measured on real devices. Computational electronics in this context refers to the physical simulation of semiconductor devices in terms of charge transport and the corresponding electrical behavior. It is related to, but usually separate from process simulation, which deals with various physical processes, such as material growth, oxidation,

* Corresponding author. Tel.: +1-480-965-2030; fax: +1-480-965-3837.
E-mail address: stephen.goodnick@asu.edu (S.M. Goodnick).

0927-796X/02/$ – see front matter
PII: S0927-796X(02)00039-6

impurity diffusion, etching and metal deposition inherent in device fabrication leading to integrated circuits. Device simulation is distinct from another important aspect of computer-aided design (CAD), device modeling, which deals with compact behavioral models for devices and sub-circuits relevant for circuit simulation in commercial packages, such as SPICE [1].

1.1. Issues in semiconductor device scaling

For silicon metal oxide semiconductor field effect transistors (MOSFETs), in conventional device scaling, the device size is scaled in all dimensions, resulting in smaller oxide thickness, junction depth, channel length, channel width and isolation spacing. Advances in lithography have driven device dimensions to the deep submicrometer range, where gate lengths are drawn at 0.1 μm and below. The Semiconductor Industry Association (SIA) projects that by the end of 2009, leading edge production devices will employ 25 nm gate lengths and have oxide thickness of 1.5 nm or less [2]. In fact, laboratory MOSFET devices with gate lengths down to 15 nm have been reported at the time of the present review, which exhibit excellent *I–V* characteristics [3]. Beyond that, there has been extensive work over the past decade related to nanoelectronic or quantum scale devices which operate on very different principles from conventional MOSFET devices, but may allow the continued scaling beyond the end of the current scaling road map [4]. This trend has been motivated by the fact that the performance of the scaled device in the 25 nm regime is itself problematic [5], as discussed later.

For example, to enhance device performance, the gate oxide thickness has to be aggressively scaled. However, as the gate oxide thickness approaches 1 nm through scaling, tunneling through the gate oxide results in unacceptably large off-state currents, dramatically increasing quiescent power consumption [6] and rendering the device impractical for analog applications due to unacceptable noise levels. Another consequence of scaling is that the stack of layered materials that comprise electronic devices is becoming more like a continuum of interfaces rather than a stack of bulk thin films. Therefore, topology effects arising from surface (interface)-to-surface (interface) interactions now dominate the formation of potential barriers at interfaces. The interface inhomogeneity effects include morphological and compositional inhomogeneities. Morphological inhomogeneities, typically manifested as atomic-scale roughness, are often responsible for increased leakage currents in MOSFET gates. Fluctuations in the elemental distribution are expressions of compositional inhomogeneities. For finite dimensions and number of atoms, interface domains cannot be represented as superpositions of a few homogeneous thin film regions. Instead, the challenge of characterizing this complex system requires accurate atomic level information about the three-dimensional structure, geometry and composition of atomic-scale interfaces.

Yet another issue that will pose serious problems on the operation of future ultra-small devices is related to the substrate doping used to gain control of the electrophysical properties of the semiconductor and the operational parameters of electronic devices by control of the type, concentration and distribution of impurities. The distribution of dopants is traditionally treated as continuum in semiconductor physics, which implies: (a) the number of impurity atoms is small as compared to the total number of atoms in the semiconductor matrix; and (b) the impurity atoms distribution is statistically uniform, while the position of an individual atom in the lattice is not defined, e.g. is random. The assumption of statistical uniformity requires large number of atoms, which is not the case in, for example, a 25 nm MOSFET device in which one has less than 100 dopant atoms in the junction region. In these future ultra-small devices, the number and location of each dopant atom will play an important role in determining the overall device behavior. The challenge of precisely placing small number of dopants may represent an insurmountable barrier, which could end conventional MOSFET scaling.

To date, the manifestation of the discrete nature of dopants, as device sizes shrink, has only been addressed in regard to statistical device-to-device variation of transistor parameters. There are, however, more fundamental aspects of the problem directly related to the nature of semiconductors that are based on two fundamental 'concentration' parameters—the intrinsic concentration and the effective density of states. There are two fundamental issues that one needs to consider in this regard: single dopant effects and deterministic doping effects. As previously described, in classical physics, the doping of semiconductor materials has been treated only as a macroscopic phenomena level, e.g. for large number of atoms. From this statistical perspective, doping determines the Fermi level in the semiconductor. In this macroscopic picture, the contribution of one extra dopant/electron to the system will not induce a significant change in the potential distribution. However, the situation changes for very small volume semiconducting materials, where the effective doping concentration increase and resistivities vary with decreasing device dimensions. Regarding deterministic doping effects, the exact position of a single dopant atom will influence the device properties. This scenario represents the transition between a stochastic approach to junction engineering and the precise control of dopant atom location, distribution, numbers and type that should be applied to all issues of source/drain/channel engineering, such as dopant solubility and activation in the nano-scale, optimization of channel doping profile, design of shallow junctions and abruptness of the source and drain regions.

Quantum-mechanical effects, due to spatial quantization in the device channel region, are also expected to play significant role in the operation of nano-scale devices. To understand this issue, one has to consider the operation of a MOSFET device based on two fundamental aspects: (1) the channel charge induced by the gate at the surface of the substrate; and (2) the carrier transport from source to drain along the channel. Quantum effects in the surface potential will have a profound impact on both, the amount of charge which can be induced by the gate electrode through the gate oxide and the profile of the channel charge in the direction perpendicular to the surface (the transverse direction). The critical parameter in this direction is the gate oxide thickness, which for a 25 nm MOSFET device is, as noted earlier, on the order of 1 nm. Another aspect, which determines device characteristics is the carrier transport along the channel (lateral direction). Because of the two-dimensional confinement of carriers in the channel, the mobility (or microscopically speaking, the carrier scattering) will be different from the three-dimensional case. Theoretically speaking, the two-dimensional mobility should be larger than its three-dimensional counterpart due to reduced density of states function, i.e. reduced number of final states the carriers can scatter into. It is important to note, however, that in the smallest size devices, carriers experience very little or no scattering at all (ballistic limit), which makes this second issue less critical when modeling nano-scale devices.

To summarize, the most important consequence of space-quantization effects in the MOSFET channel is the larger average displacement of the carriers from the interface proper. This gives rise to an effective oxide thickness increase and a total gate capacitance degradation. Additional degradation of the total gate capacitance arises from polydepletion effects, thus leading to additional capacitance component in series with the oxide and the inversion layer capacitances. Both effects, for different technology generations, are demonstrated in Fig. 1. Both Fermi–Dirac (F–D) and Maxwell–Boltzmann (M–B) statistics have been used for the classical charge description and F–D statistics for the case of the quantum-mechanical charge description. The quantum-mechanical charge description gives approximately a 20% degradation of the total gate capacitance when compared to the classical charge description model for the end of the road map devices with oxide thickness between 1 and 1.5 nm. Further capacitance degradation, of up to 40% for the smallest oxide thickness, is due to the depletion effect in the poly(Si) gate. The total gate capacitance degradation, on the other hand, when combined with the quantum-mechanical band-gap widening effect and reduced density of states for a quasi-two-dimensional system, gives rise to a reduction of the sheet electron density. This, in turn, increases the threshold voltage and, at the same time, degrades the device

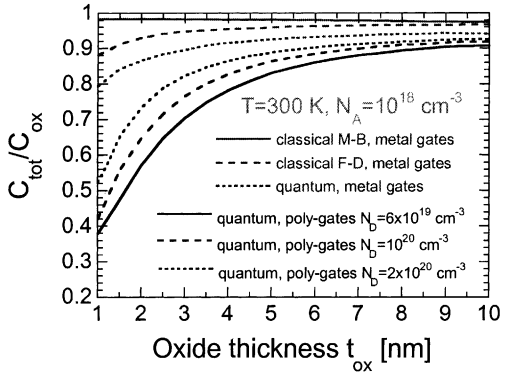


Fig. 1. Total gate capacitance as a function of the oxide thickness.

transconductance. Hence, to properly describe the operation of future ultra-small devices, it becomes mandatory to incorporate quantum-mechanical and polydepletion effects into the device simulators mentioned in Section 1.2 and described in more details in Sections 5 and 6 of this review article.

1.2. Computational electronics

In Section 1.1, we have detailed how the physics of device behavior becomes increasingly complicated as nano-scale dimensions are approached. The goal of computational electronics is to provide simulation tools with the level of sophistication necessary to capture the essential physics while at the same time minimizing the computational burden so that results may be obtained within a reasonable time frame. Fig. 2 illustrates the main components of semiconductor device simulation at any level. There are two main kernels, which must be solved self-consistently with one another, the transport equations governing charge flow and the fields driving charge flow. Both are coupled strongly to one another and hence must be solved simultaneously. The fields arise from external sources, as well as the charge and current densities which act as sources for the time varying electric and magnetic fields obtained from the solution of Maxwell's equations. Under appropriate conditions, only the quasi-static electric fields arising from the solution of Poisson's equation are necessary.

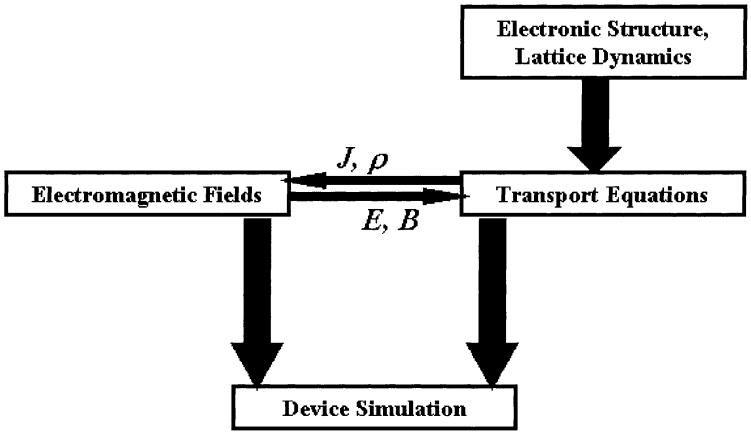


Fig. 2. Schematic description of the device simulation sequence.

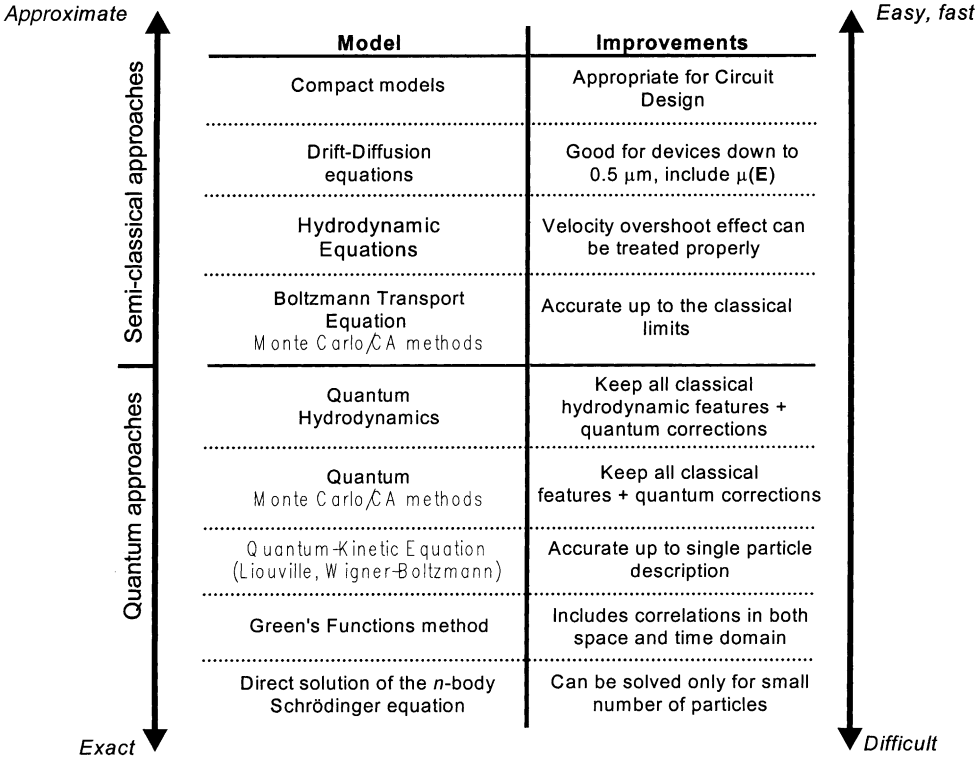


Fig. 3. Illustration of the hierarchy of transport models.

The fields in turn are driving forces for charge transport as illustrated in Fig. 3 for the various levels of approximation within a hierarchical structure ranging from compact modeling at the top to an exact quantum-mechanical description at the bottom. It is now common practice in industry to use either drift-diffusion or standard hydrodynamic models in trying to understand the operation of as-fabricated devices, by adjusting any number of phenomenological parameters (e.g. mobility, impact ionization coefficient, etc.). However, such tools do not have predictive capability for ultra-small structures, for which it is necessary to relax some of the approximations in the Boltzmann transport equation (BTE). Therefore, one needs to move downward to the quantum transport area in the hierarchical map of transport models shown in Fig. 3 in which, at the very bottom we have the Green's function approach. The latter is the most exact, but at the same time the most difficult of all. In contrast to, for example, the Wigner function approach (which is Markovian in time), the Green's functions method allows one to consider simultaneously correlations in space and time, both of which are expected to be important in nano-scale devices. However, the difficulties in understanding the various terms in the resultant equations and the enormous computational burden needed for its actual implementation, make the usefulness in understanding quantum effects in actual devices of limited values. For example, the only successful utilization of the Green's function approach commercially is the nano-electronics modeling (NEMO) simulator [7], which is effectively one-dimensional and is primarily applicable to resonant tunneling diodes.

From the discussion it follows that, contrary to the recent technological advances discussed in Section 1.1, the present state of the art in device simulation is currently lacking in the ability to treat these new challenges in scaling of device dimensions from conventional down to quantum scale devices. For silicon devices with active regions below 0.2 μm in diameter, macroscopic transport descriptions based

on drift-diffusion models (Fig. 3) are clearly inadequate. As already noted, even standard hydrodynamic models do not usually provide a sufficiently accurate description since they neglect significant contributions from the tail of the phase space distribution function in the channel regions [8,9]. Within the requirement of self-consistently solving the coupled transport-field problem in this emerging domain of device physics, there are several computational challenges, which limit this ability. One is the necessity to solve both the transport and the Poisson's equations over the full three-dimensional domain of the device (and beyond if one includes radiation effects). As a result, highly efficient algorithms targeted to high-end computational platforms (most likely in a multi-processor environment) are required to fully solve even the appropriate field problems. The appropriate level of approximation necessary to capture the proper non-equilibrium transport physics relevant to a future device model is an even more challenging problem both computationally and from a fundamental physics framework.

In this review article, we give an overview of the basic techniques used in the field of computational electronics related to device simulation. As shown schematically in Fig. 2, we begin with a review of the electronic band structure of semiconductors and the associated dynamics of carriers under external fields (Section 2). This allows one to calculate relevant material parameters, such as effective masses, effective density-of-states function, etc. Afterwards, we present a discussion of the basic equations governing transport in semiconductors, leading to the description of the Monte Carlo (MC) method for the solution of the semi-classical BTE (Section 3.1) and the hydrodynamic and drift-diffusion models for device simulation, that follow from moments of the BTE (Section 3.2). In Sections 4.1 and 4.2, we give an overview of field solvers for both high frequency (solution of the Maxwell equations) and low frequency (solution of quasi-static Poisson equation) applications, respectively. Some key elements of particle-based simulation, such as grid size and time step criteria, charge assignment scheme, inclusion of the short-range electron-electron and electron-ion interactions, etc. are described in Section 5.1. In Section 5.2 we give an overview of commercially available drift-diffusion/hydrodynamics device simulators. The simulation of the optoelectronic and high frequency devices via the solution of the full set of Maxwell's equations coupled with a MC transport kernel is discussed in Section 5.3. The inclusion of quantum corrections into particle-based simulators, using the effective potential approach, is discussed in Section 6.1. A brief description of the quantum hydrodynamic (QHD) model for device simulation and its application to modulation-doped high-electron mobility transistors (HEMTs) is given in Section 6.2. A thorough review of the quantum transport approaches listed in the hierarchy of Fig. 3 will be discussed elsewhere.

2. Electronic band structure calculation

The basis for discussing transport in semiconductors is the underlying electronic band structure of the material arising from the solution of the many-body Schrödinger equation in the presence of the periodic potential of the lattice, which is discussed in a host of solid state physics textbooks. The electronic solutions in the presence of the periodic potential of the lattice are in the form of Bloch functions

$$\psi_{n,k} = u_n(\boldsymbol{k})\mathrm{e}^{\mathrm{i}kr}, \tag{1}$$

where $\boldsymbol{k}$ is the wave vector and n labels the band index corresponding to different solutions for a given wave vector. The cell-periodic function, $u_n(\boldsymbol{k})$, has the periodicity of the lattice and modulates the traveling wave solution associated with free electrons.

Electronic band structure calculation methods can be grouped into two general categories [10]. The first category consists of ab initio methods, such as Hartree–Fock or density functional theory (DFT), which calculate the electronic structure from first principles, i.e. without the need for

empirical fitting parameters. In general, these methods utilize a variational approach to calculate the ground state energy of a many-body system, where the system is defined at the atomic level. The original calculations were performed on systems containing a few atoms. Today, calculations are performed using approximately 1000 atoms but are computationally expensive, sometimes requiring massively parallel computers.

In contrast to ab initio approaches, the second category consists of empirical methods, such as the orthogonalized plane wave (OPW) [11], tight-binding (TB) [12] (also known as the linear combination of atomic orbitals (LCAO) method), the $\boldsymbol{k}\cdot\boldsymbol{p}$ method [13] and the local [14] or the non-local [15] empirical pseudo-potential method (EPM). These methods involve empirical parameters to fit experimental data such as the band-to-band transitions at specific high-symmetry points derived from optical absorption experiments. The appeal of these methods is that the electronic structure can be calculated by solving a one-electron Schrödinger wave equation (SWE). Thus, empirical methods are computationally less expensive than ab initio calculations and provide a relatively easy means of generating the electronic band structure. Due to their wide spread usage, in the rest of this section we will review some of the most commonly used ones, namely the EPM, the TB and the $\boldsymbol{k}\cdot\boldsymbol{p}$ method. In the descriptions presented below, more emphasis will be placed on the description of the EPM (Section 2.1). The TB (Section 2.2) and the $\boldsymbol{k}\cdot\boldsymbol{p}$ (Section 2.3) methods are only briefly reviewed, followed by a brief description of the carrier dynamics is given in Section 2.4.

Before proceeding with the description of the various empirical band structure methods, it is useful to introduce the spin–orbit interaction Hamiltonian. The effects of spin–orbit coupling are most easily considered by regarding the spin–orbit interaction energy H_{SO} as a perturbation. In its most general form, H_{SO} operating on the wave functions ψ_k is then given by

$$H_{\mathrm{SO}} = \frac{\hbar}{4m^2c^2}[\nabla V \times \boldsymbol{p}]\cdot\sigma, \tag{2}$$

where V is the potential energy term of the Hamiltonian and σ the Pauli spin tensor. It can also be written in the following form as an operator on the cell-periodic function

$$H_{\mathrm{SO}} = \frac{\hbar}{4m^2c^2}[\nabla V \times \boldsymbol{p}]\cdot\sigma + \frac{\hbar^2}{4m^2c^2}[\nabla V \times \boldsymbol{k}]\cdot\sigma. \tag{3}$$

The first term is $\boldsymbol{k}$-independent and is analogous to the atomic spin–orbit splitting term. The second term is proportional to $\boldsymbol{k}$ and is the additional spin–orbit energy coming from the crystal momentum. Rough estimates indicate that the effect of the second term on the energy bands is less than 1% of the effect of the first term. The relatively greater importance of the first term comes from the fact that the velocity of the electron in its atomic orbit is very much greater than the velocity of a wave packet made up of wave vectors in the neighborhood of $\boldsymbol{k}$.

2.1. The empirical pseudo-potential method

2.1.1. Introductory comments

The concept of pseudo-potentials was introduced by Fermi [16] to study high-lying atomic states. Afterwards, Hellman proposed that pseudo-potentials be used for calculating the energy levels of the alkali metals [17]. The wide spread usage of pseudo-potentials did not occur until the late 1950s, when activity in the area of condensed matter physics began to accelerate. The main advantage of using pseudo-potentials is that only valence electrons have to be considered. The core

electrons are treated as if they are frozen in an atomic-like configuration. As a result, the valence electrons are thought to move in a weak one-electron potential.

The pseudo-potential method is based on the OPW method due to Herring [11]. In this method, the crystal wave function ψ_k is constructed to be orthogonal to the core states. This is accomplished by expanding ψ_k as a smooth part of symmetrized combinations of Bloch functions φ_k, augmented with a linear combination of core states. This is expressed as

$$\psi_k = \varphi_k + \sum_t b_{k,t} \Phi_{k,t}, \tag{4}$$

where $b_{k,t}$ are orthogonalization coefficients and $\Phi_{k,t}$ are core wave functions. For Si-14, the summation over t in Eq. (4) is a sum over the core states $1s^2\ 2s^2\ 2p^6$. Since the crystal wave function is constructed to be orthogonal to the core wave functions, the orthogonalization coefficients can be calculated, thus yielding the final expression

$$\psi_k = \varphi_k - \sum_t \langle \Phi_{k,t} | \varphi_k \rangle \Phi_{k,t}. \tag{5}$$

To obtain a wave equation for φ_k, the Hamiltonian operator

$$H = \frac{p^2}{2m} + V_C, \tag{6}$$

is applied to Eq. (5), where V_C is the attractive core potential and the following wave equation results

$$\left(\frac{p^2}{2m} + V_C + V_R \right) \varphi_k = E \varphi_k, \tag{7}$$

where V_R represents a short-range, non-Hermitian repulsion potential, of the form

$$V_R = \sum_t \frac{(E - E_t) \langle \Phi_{k,t} | \varphi_k \rangle \Phi_{k,t}}{\varphi_k}. \tag{8}$$

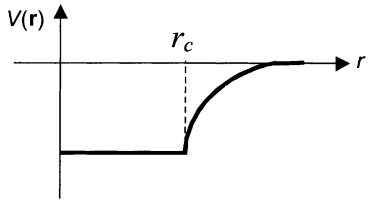
E_t in Eq. (8) represents the atomic energy eigenvalue and the summation over t represents a summation over the core states. The result given in Eq. (7) can be thought of as wave equation for the pseudo-wave function, φ_k, but the energy eigenvalue E corresponds to the true energy of the crystal wave function ψ_k. Furthermore, as a result of the orthogonalization procedure, the repulsive potential V_R, which serves to cancel the attractive potential V_C, is introduced into the pseudo-wave function Hamiltonian. The result is a smoothly varying pseudo-potential $V_P = V_C + V_R$. This result is known as the Phillips–Kleinman cancellation theorem [18] which provides justification why the electronic structure of strongly-bound valence electrons can be described using a nearly-free electron model and weak potentials.

To simplify the problem further, model pseudo-potentials are used in place of the actual pseudo-potential. Fig. 4 summarizes the various models employed. Note that the three-dimensional Fourier transforms (for bulk systems) of each of the described model potentials are of the following general form

$$V(q) \sim \frac{Ze^2}{\varepsilon_0 q^2} \cos(q r_c). \tag{9}$$

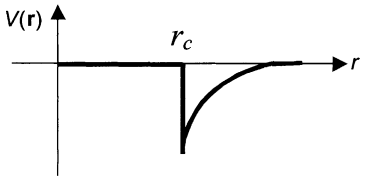
(a) Constant effective potential in the core region:

$$V(r) = \begin{cases} \dfrac{-Ze^2}{4\pi\varepsilon_0 r}; & r > r_C \\ \dfrac{-Ze^2}{4\pi\varepsilon_0 r_C}; & r \le r_C \end{cases}$$



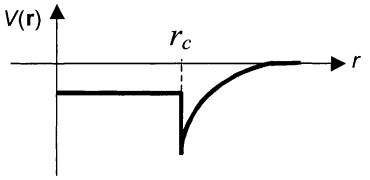
(b) Empty core model:

$$V(r) = \begin{cases} \dfrac{-Ze^2}{4\pi\varepsilon_0 r}; & r > r_C \\ 0; & r \le r_C \end{cases}$$



(c) Model potential due to Heine and Abarenkov:

$$V(r) = \begin{cases} \dfrac{-Ze^2}{4\pi\varepsilon_0 r}; & r > r_C \\ A; & r \le r_C \end{cases}$$



(d) Lin and Kleinman model potentials:

$$V(r) = \begin{cases} 2\dfrac{-Ze^2}{4\pi\varepsilon_0 r}\{1 - \exp[-\beta(r - r_C)]\}; & r > r_C \\ 0; & r \le r_C \end{cases}$$

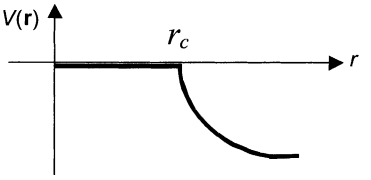


Fig. 4. Various model potentials.

This q-dependent pseudo-potential is then used to calculate the energy band structure along different crystallographic directions, using the procedure outlined in Section 2.1.2.

2.1.2. *Description of the empirical pseudo-potential method*

Recall from Section 2.1.1 that the Phillips–Kleinman cancellation theorem provides a means for the energy band problem to be simplified into a one-electron-like problem. For this purpose, Eq. (7) can be re-written as

$$\left(\frac{p^2}{2m} + V_{\mathrm{P}}\right)\varphi_k = E\varphi_k, \tag{10}$$

where V_{P} is the smoothly varying crystal pseudo-potential. In general, V_{P} is a linear combination of atomic potentials, V_{A}, which can be expressed as summation over lattice translation vectors $\boldsymbol{R}$ and atomic basis vectors $\boldsymbol{\tau}$ to arrive at the following expression

$$V_{\mathrm{P}}(\boldsymbol{r}) = \sum_{R}\sum_{\tau} V_{\mathrm{A}}(\boldsymbol{r} - \boldsymbol{R} - \boldsymbol{\tau}). \tag{11}$$

To simplify Eq. (11), the inner summation over $\boldsymbol{\tau}$ can be expressed as the total potential, V_0, in the unit cell located at $\boldsymbol{R}$. Eq. (11) then becomes

$$V_{\mathrm{P}}(\boldsymbol{r}) = \sum_{R} V_0(\boldsymbol{r} - \boldsymbol{R}). \tag{12}$$

Because the crystal potential is periodic, the pseudo-potential is also a periodic function and can be expanded into a Fourier series over the reciprocal lattice to obtain

$$V_{\mathrm{P}}(\boldsymbol{r}) = \sum_{G} V_0(\boldsymbol{G}) \mathrm{e}^{\mathrm{i}G r}, \tag{13}$$

where the expansion coefficient is given by

$$V_0(\boldsymbol{G}) = \frac{1}{\Omega} \int \mathrm{d}^3 r \, V_0(\boldsymbol{r}) \mathrm{e}^{-\mathrm{i}G r}, \tag{14}$$

and Ω is the volume of the unit cell.

To apply this formalism to the zinc-blende lattice, it is convenient to choose a two-atom basis centered at the origin ($\boldsymbol{R} = 0$). If the atomic basis vectors are given by $\boldsymbol{\tau}_1 = \boldsymbol{\tau} = -\boldsymbol{\tau}_2$, where $\boldsymbol{\tau}$, the atomic basis vector, is defined in terms of the lattice constant a_0 as $\boldsymbol{\tau} = a_0(1/8, 1/8, 1/8)$, $V_0(\boldsymbol{r})$ can be expressed as

$$V_0(\boldsymbol{r}) = V_1(\boldsymbol{r} - \boldsymbol{\tau}) + V_2(\boldsymbol{r} + \boldsymbol{\tau}), \tag{15}$$

where V_1 and V_2 are the atomic potentials of the cation and anion. Substituting Eq. (15) into Eq. (14) and using the displacement property of Fourier transforms, $V_0(\boldsymbol{r})$ can be recast as

$$V_0(\boldsymbol{G}) = \mathrm{e}^{\mathrm{i}G\tau} V_1(\boldsymbol{G}) + \mathrm{e}^{-\mathrm{i}G\tau} V_2(\boldsymbol{G}). \tag{16}$$

Writing the Fourier coefficients of the atomic potentials in terms of symmetric ($V_{\mathrm{S}}(\boldsymbol{G}) = V_1 + V_2$) and anti-symmetric ($V_{\mathrm{A}}(\boldsymbol{G}) = V_1 - V_2$) form factors, $V_0(\boldsymbol{G})$ is given by

$$V_0(\boldsymbol{G}) = \cos(\boldsymbol{G}\boldsymbol{\tau}) V_{\mathrm{S}}(\boldsymbol{G}) + \mathrm{i} \sin(\boldsymbol{G}\boldsymbol{\tau}) V_{\mathrm{A}}(\boldsymbol{G}), \tag{17}$$

where the pre-factors are referred to as the symmetric and anti-symmetric structure factors. The form factors are treated as adjustable parameters that can be fit to experimental data, hence the name EPM. For diamond lattice materials, with two identical atoms per unit cell, the $V_{\mathrm{A}} = 0$ and the structure factor is simply $\cos(\boldsymbol{G}\boldsymbol{\tau})$.

Now with the potential energy term specified, the next task is to recast the SWE in a matrix form. Recall that the solution to the SWE in a periodic lattice is a Bloch function, which is composed of a plane wave component and a cell-periodic part that has the periodicity of the lattice, i.e.

$$\varphi_k(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}kr} u_k(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}kr} \sum_{G'} U(\boldsymbol{G}') \mathrm{e}^{\mathrm{i}G' r}. \tag{18}$$

By expanding the cell-periodic part $u_k(\boldsymbol{r})$ of the Bloch function appearing in Eq. (18) into Fourier components and substituting the pseudo-wave function φ_k and potential V_0 into the SWE, the following matrix equation results

$$\sum_{G} \left\{ \left[\frac{\hbar^2 (\boldsymbol{k} + \boldsymbol{G})^2}{2m} - E \right] U(\boldsymbol{G}) + \sum_{G'} V_0(|\boldsymbol{G} - \boldsymbol{G}'|) U(\boldsymbol{G}') \right\} = 0. \tag{19}$$

The expression given in Eq. (19) is zero when each term in the sum is identically zero, which implies the following condition

$$\left[\frac{\hbar^2 (\boldsymbol{k} + \boldsymbol{G})^2}{2m} - E \right] U(\boldsymbol{G}) + \sum_{G'} V_0(|\boldsymbol{G} - \boldsymbol{G}'|) U(\boldsymbol{G}') = 0. \tag{20}$$

In this way, the band structure calculation is reduced to solving the eigen value problem specified by Eq. (20) for the energy E. As obvious from Eq. (18), $U(\boldsymbol{G}')$ is the Fourier component of the cell-periodic part of the Bloch function. The number of reciprocal lattice vectors used determines both the matrix size and calculation accuracy.

The eigenvalue problem of Eq. (20) can be written in the more familiar form $\boldsymbol{HU} = E\boldsymbol{U}$, where $\boldsymbol{H}$ is a matrix, $\boldsymbol{U}$ the column vector representing the eigenvectors and E the energy eigenvalue corresponding to its respective eigen vector. For the diamond lattice, the diagonal matrix elements of $\boldsymbol{H}$ are then given by

$$H_{i,j} = \frac{\hbar^2}{2m}|\boldsymbol{k} + \boldsymbol{G}_i|^2, \tag{21}$$

for $i = j$ and the off-diagonal matrix elements of $\boldsymbol{H}$ are given by

$$H_{i,j} = V_S(|\boldsymbol{G}_i - \boldsymbol{G}_j|)\cos[(\boldsymbol{G}_i - \boldsymbol{G}_j)\boldsymbol{\tau}] \tag{22}$$

for $i \neq j$. Note that the term $V_S(0)$ is neglected in arriving at Eq. (21) because it will only give a rigid shift in energy to the bands. The solution to the energy eigenvalues and corresponding eigenvectors can then be found by diagonalizing matrix $\boldsymbol{H}$.

2.1.3. Implementation of the empirical pseudo-potential method

For a typical semiconductor system, 137 plane waves are sufficient, each corresponding to vectors in the reciprocal lattice, to expand the pseudo-potential. The reciprocal lattice of a face-centered cubic (FCC), i.e. diamond or zinc-blende structure, is a body-centered cubic (BCC) structure. Reciprocal lattice vectors up to and including the 10th nearest neighbor from the origin are usually considered which results in 137 plane waves for the zinc-blende structure. The square of the distance from the origin to each equivalent set of reciprocal lattice sites is an integer in the set $|\boldsymbol{G}^2| = 0, 3, 4, 8, 11, 12$, etc. where $|\boldsymbol{G}^2|$ is expressed in units of $(2\pi/a_o)^2$. Note that the argument of the pseudo-potential term V_S in Eq. (22) is the difference between reciprocal lattice vectors. It can be shown that the square of the difference between reciprocal lattice vectors will also form the set of integers previously described. This means that V_S is only needed at discrete points corresponding to nearest-neighbor sites. The pseudo-potential, on the other hand, is a continuous quantity. Therefore, its Fourier transform $V_S(q)$ is also a continuous function that is shown in Fig. 5. The points corresponding to first three nearest neighbors are also indicated on this figure.

Recall that the pseudo-potential is only needed at a few discrete points along the $V(q)$ curve. The discrete points correspond to the q^2-values that match the integer set described previously. There is some controversy, however, regarding the value of V_S as q vanishes. There are two common values seen in the literature: $V_S(0) = -3/2E_F$ and $V_S(0) = 0$. In most cases, the term $V_S(0)$ is ignored because it only gives a rigid shift in energy to the bands. The remaining form factors needed to compute the band structure for non-polar materials correspond to $q^2 = 3$, 8 and 11. For $q^2 = 4$, the cosine term in Eq. (22) will always vanish. Furthermore, for values of q^2 greater than 11, $V(q)$ quickly approaches zero. This comes from the fact that the pseudo-potential is a smoothly varying function and only few plane waves are needed to represent it. If a function is rapidly varying in space, then many more plane waves would be required. Another advantage of the EPM is that only three parameters are needed to describe the band structure of non-polar materials.

Using the form factors listed in Table 1, where the Si form factors are taken from [19] and the Ge form factors are taken from [20], the band structures for Si and Ge are plotted in Fig. 6 [21]. Note that spin–orbit interaction is not included in these simulations. The lattice constants specified for Si

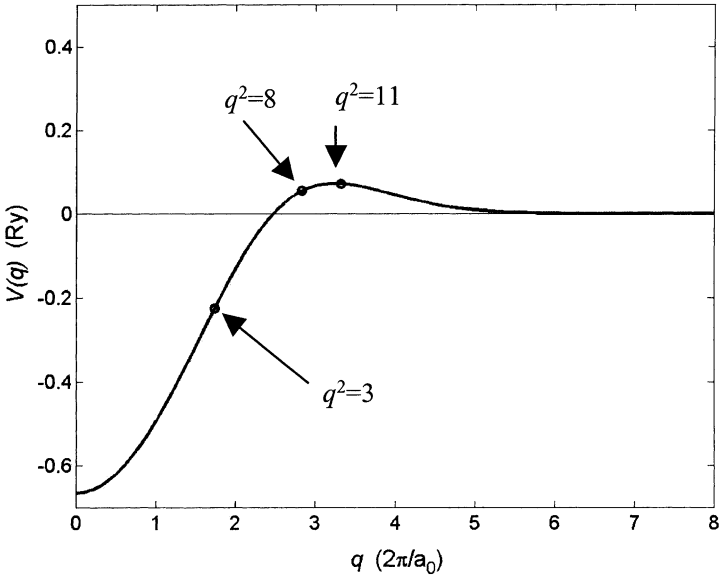


Fig. 5. Fourier transform of the pseudo-potential. (Note that $\boldsymbol{q} = |\boldsymbol{G} - \boldsymbol{G}'|$).

Table 1
Local pseudo-potential form factors

Form factor (Ry)	Si	Ge
V_3	−0.2241	−0.2768
V_s	0.055	0.0582
V_{11}	0.0724	0.052

and Ge are 5.43 and 5.65 Å, respectively. Si is an indirect band-gap semiconductor. Its primary gap, i.e. minimum gap, is calculated from the valence band maximum at the Γ-point to the conduction band minimum along the Δ direction, 85% of the distance from Γ to X. The band-gap of Si, using the parameters from Tables 1 and 2, is calculated to be $E_g^{\mathrm{Si}} = 1.08$ eV, in agreement with experimental

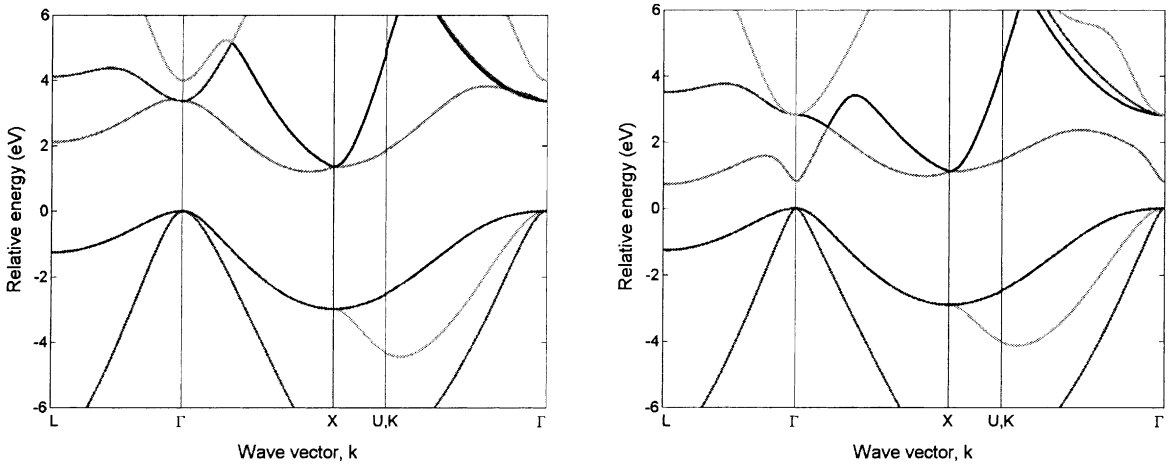


Fig. 6. Left panel: band structures of silicon. Right panel: band structure of germanium.

Table 2
Chadi and Cohen TB parameters [12]

	$E_p - E_s$	V_{ss}	V_{sp}	V_{xx}	V_{xy}
C	7.40	−15.2	10.25	3.0	8.3
Si	7.20	−8.13	5.88	1.71	7.51
Ge	8.41	−6.78	5.31	1.62	6.82

findings. Ge is also an indirect band-gap semiconductor. Its band-gap is defined from the top of the valence band at Γ to the conduction band minimum at L. The band-gap of Ge is calculated to be $E_g^{Ge} = 0.73$ eV. The direct gap, which is defined from the valence band maximum at Γ to the conduction band minimum at Γ, is calculated to be 3.27 eV and 0.82 eV for Si and Ge, respectively. Note that the curvature of the top valence band of Ge is larger than that of Si. This corresponds to the fact that the effective hole mass of Si is larger than that of Ge. Note that the inclusion of the spin–orbit interaction will lift the triple degeneracy of the bands at the Γ point, leaving doubly-degenerate heavy and light-hole bands and a split-off band moved downward in energy by few 10's of meV (depending upon the material under consideration).

In summary, the local EPM described in this section is rather good for an accurate description of the optical gaps. However, as noted by Chelikowsky and Cohen [22], when these local calculations are extended to yield the valence band electronic density of states, the results obtained are far from satisfactory. The reason for this discrepancy arises from the omission of the low cores in the derivation of the pseudo-potential in the previous section. This, as previously noted, allowed the usage of a simple plane wave basis. To correct for the errors introduced, an energy-dependent non-local correction term is added to the local atomic potential. This increases the number of parameters needed but leads to better convergence and more exact band structure results [23,24].

2.2. The tight-binding method

TB is a semi-empirical method for electronic structure calculations. While it retains the underlying quantum mechanics of the electrons, the Hamiltonian is parametrized and simplified before the calculation, rather than constructing it from first principles. The method is detailed by Slater and Coster [25], who laid the initial ground work. Conceptually, TB works by postulating a basis set which consists of atomic-like orbitals (i.e. they share the angular momentum components of the atomic orbitals and are easily split into radial and angular parts) for each atom in the system and the Hamiltonian is then parametrized in terms of various high symmetry interactions between these orbitals. For tetrahedral semiconductors, a conceptual basis set of 1s-like orbital and 3p-like orbitals has been used. In the most common form of TB (nearest neighbor, orthogonal TB), the orbitals are assumed to be orthogonal and interactions between different orbitals are only allowed to be non-zero within a certain distance, which is placed somewhere between the first and second nearest neighbors in the crystal structure. A further simplification which is made, is to neglect three-center integrals (i.e. an interaction between orbitals on atoms A and B which is mediated by the potential on atom C), meaning that each interaction is a function of the distance between the atoms only.

The quantitative description of the method presented below is due to Chadi and Cohen [12]. Let us denote the position of the atom in the primitive cell as

$$\boldsymbol{r}_{jl} = \boldsymbol{R}_j + \boldsymbol{r}_l, \tag{23}$$

where $\boldsymbol{R}_j$ is the position of the jth primitive cell and $\boldsymbol{r}_l$ is the position of the atom within the primitive cell. Let $h_l(\boldsymbol{r})$ be the Hamiltonian of the isolated atom, such that

$$h_l\phi_{ml}(\boldsymbol{r}-\boldsymbol{r}_{jl}) = E_{ml}\phi_{ml}(\boldsymbol{r}-\boldsymbol{r}_{jl}), \tag{24}$$

where E_{ml} and ϕ_{ml} are the eigenvalues and the eigenfunctions of the state indexed by m. The atomic orbitals ϕ_{ml} are called Löwdin orbitals [26] and they are different from the usual atomic wavefunctions in that they have been constructed in such a way that wavefunctions centered at different atomic sites are orthogonal to each other. The total Hamiltonian of the system is then

$$H_0 = \sum_{j,l} h_l(\boldsymbol{r}-\boldsymbol{r}_{jl}). \tag{25}$$

Note that the sum over l refers to a sum within the different atoms in the basis, therefore, $l = 1, 2$ for diamond and zinc-blende crystals. The unperturbed Bloch functions, that have the proper translational symmetry, are constructed to be of the following form

$$\Phi_{mlk} = \frac{1}{\sqrt{N}}\sum_j \mathrm{e}^{\mathrm{i}\boldsymbol{r}_{jl}\cdot\boldsymbol{k}}\phi_{ml}(\boldsymbol{r}-\boldsymbol{r}_{jl}). \tag{26}$$

The eigen values of the total Hamiltonian $H = H_0 + H_{\mathrm{int}}$ (where H_{int} is the interaction part of the Hamiltonian) are then represented as a linear combination of the Bloch functions

$$\Psi_k = \sum_{ml} c_{ml}\Phi_{\mathrm{mlk}}. \tag{27}$$

Operating with the total Hamiltonian of the system H on Ψ_k and using the orthogonality of the atomic wave functions, one arrives at the following matrix equation

$$\sum_{ml}[H_{m'l',ml} - E_k\delta_{mm'}\delta_{ll'}]c_{ml} = 0, \tag{28}$$

where the matrix element appearing in this expression is given by

$$H_{m'l',ml}(\boldsymbol{k}) = \sum_j \mathrm{e}^{\mathrm{i}(\boldsymbol{R}_j+\boldsymbol{r}_l-\boldsymbol{r}_{l'})\cdot\boldsymbol{k}}\langle\phi_{mlk}(\boldsymbol{r}-\boldsymbol{r}_{jl})|H|\phi_{mlk}(\boldsymbol{r}-\boldsymbol{r}_{jl'})\rangle. \tag{29}$$

Note that in the simplest implementation of this method, instead of summing over all the atoms, one sums over the nearest neighbor atoms only. Also note that the index m represents the s- and p-states of the outermost electrons ($|S\rangle$, $|X\rangle$, $|Y\rangle$ and $|Z\rangle$) and l is the number of distinct electrons in the basis. For the case of tetrahedrally coordinated semiconductors, the number of nearest neighbors is four and are located at

$$\begin{cases} d_1 = (1,1,1)a_0/4 \\ d_2 = (1,-1,-1)a_0/4 \\ d_3 = (-1,1,-1)a_0/4 \\ d_4 = (-1,-1,1)a_0/4 \end{cases} \tag{30}$$

For a diamond lattice, one also defines the following matrix elements

$$\begin{cases} V_{ss} = 4V_{ss\sigma} \\ V_{sp} = -4V_{sp\sigma}/\sqrt{3} \\ V_{xx} = 4[V_{pp\sigma}/3 + 2V_{pp\pi}/3] \\ V_{xy} = 4[V_{pp\sigma}/3 - V_{pp\pi}/3] \end{cases} \tag{31}$$

As an example, consider the matrix element between two s-states

$$H_{s_1,s_2} = [e^{i\boldsymbol{k}\cdot\boldsymbol{d}_1} + e^{i\boldsymbol{k}\cdot\boldsymbol{d}_2} + e^{i\boldsymbol{k}\cdot\boldsymbol{d}_3} + e^{i\boldsymbol{k}\cdot\boldsymbol{d}_4}]\langle s_1|H_{int}|s_2\rangle = g_1(\boldsymbol{k})V_{ss\sigma}. \tag{32}$$

Notice the appearance of the Bloch sum $g_1(\boldsymbol{k})$ in expression (32). This observation suggests that for different basis states, there will be four different Bloch sums g_1 through g_4, of the form

$$\begin{cases} g_1(\boldsymbol{k}) = [e^{i\boldsymbol{k}\cdot\boldsymbol{d}_1} + e^{i\boldsymbol{k}\cdot\boldsymbol{d}_2} + e^{i\boldsymbol{k}\cdot\boldsymbol{d}_3} + e^{i\boldsymbol{k}\cdot\boldsymbol{d}_4}] \\ g_2(\boldsymbol{k}) = [e^{i\boldsymbol{k}\cdot\boldsymbol{d}_1} + e^{i\boldsymbol{k}\cdot\boldsymbol{d}_2} - e^{i\boldsymbol{k}\cdot\boldsymbol{d}_3} - e^{i\boldsymbol{k}\cdot\boldsymbol{d}_4}] \\ g_3(\boldsymbol{k}) = [e^{i\boldsymbol{k}\cdot\boldsymbol{d}_1} - e^{i\boldsymbol{k}\cdot\boldsymbol{d}_2} + e^{i\boldsymbol{k}\cdot\boldsymbol{d}_3} - e^{i\boldsymbol{k}\cdot\boldsymbol{d}_4}] \\ g_4(\boldsymbol{k}) = [e^{i\boldsymbol{k}\cdot\boldsymbol{d}_1} - e^{i\boldsymbol{k}\cdot\boldsymbol{d}_2} - e^{i\boldsymbol{k}\cdot\boldsymbol{d}_3} + e^{i\boldsymbol{k}\cdot\boldsymbol{d}_4}] \end{cases} \tag{33}$$

It is also important to note that the Hamiltonian matrix elements between a s- and p-states on the same atom or two different p-states on the same atom, are zero because of symmetry in diamond and zinc-blende crystals. The 8×8 secular determinant representing all possible nearest neighbor interactions between the TB s- and p-orbitals centered on each atom in the crystal is

$$\begin{vmatrix} & S_1 & S_2 & X_1 & Y_1 & Z_1 & X_2 & Y_2 & Z_2 \\ S_1 & E_s - E_k & V_{ss}g_1 & 0 & 0 & 0 & V_{sp}g_2 & V_{sp}g_3 & V_{sp}g_4 \\ S_2 & V_{ss}g_1^* & E_s - E_k & -V_{sp}g_2^* & -V_{sp}g_3^* & -V_{sp}g_4^* & 0 & 0 & 0 \\ X_1 & 0 & -V_{sp}g_2 & E_p - E_k & 0 & 0 & V_{xx}g_1 & V_{xy}g_4 & V_{xy}g_3 \\ Y_1 & 0 & -V_{sp}g_3 & 0 & E_p - E_k & 0 & V_{xy}g_4 & V_{xx}g_1 & V_{xy}g_2 \\ Z_1 & 0 & -V_{sp}g_4 & 0 & 0 & E_p - E_k & V_{xy}g_3 & V_{xy}g_2 & V_{xx}g_1 \\ X_2 & V_{sp}g_2^* & 0 & V_{xx}g_1^* & V_{xy}g_4^* & V_{xy}g_3^* & E_p - E_k & 0 & 0 \\ Y_2 & V_{sp}g_3^* & 0 & V_{xy}g_4^* & V_{xx}g_1^* & V_{xy}g_2^* & 0 & E_p - E_k & 0 \\ Z_2 & V_{sp}g_4^* & 0 & V_{xy}g_3^* & V_{xy}g_2^* & V_{xx}g_1^* & 0 & 0 & E_p - E_k \end{vmatrix}. \tag{34}$$

The TB parameters appearing in Eqs. (32) and (34) are obtained by comparison with empirical pseudo-potential calculations, which are shown in [12].

Using the described method one can quite accurately describe the valence bands, whereas the conduction bands are not reproduced that well due to the omission of the interaction with the higher-lying bands. The accuracy of the conduction bands can be improved with the addition of more overlap parameters. However, there are only four conduction bands and the addition of more orbitals destroys the simplicity of the method.

Also, one has to take into consideration that this derivation applies only to the electronic energy calculation, which allows the band energy to be found together with the eigenvectors of the occupied states of the Hamiltonian. To obtain, for example, a total energy for the system being modeled, the effects of ion–ion repulsion, and those parts of the electron–electron interaction neglected

above, need to be accounted for. This is most often done via a pair potential, which is again fitted to ab initio data or experiment.

2.3. The $\boldsymbol{k}\cdot\boldsymbol{p}$ method

In contrast to the two previously described methods, the $\boldsymbol{k}\cdot\boldsymbol{p}$ method is based upon perturbation theory [27,28]. In this method, the energy is calculated near a band maximum or minimum by considering the wave number (measured from the extremum) as a perturbation. The description of this method starts with the one-electron wave function, which for the case of a periodic lattice, is of the form described in the introduction part of this section and repeated here for completeness. Namely,

$$\psi_k(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} u_k(\boldsymbol{r}), \tag{35}$$

where $u_k(r)$ is cell-periodic part of the Bloch function. The Schrödinger equation can then be written as

$$\left[\frac{p^2}{2m} + V(\boldsymbol{r}) + \frac{\hbar}{m}\boldsymbol{k}\cdot\boldsymbol{p}\right] u_k(\boldsymbol{r}) = \left[E_k - \frac{\hbar^2 k^2}{2m}\right] u_k(\boldsymbol{r}). \tag{36}$$

The term $(\hbar/m)\boldsymbol{k}\cdot\boldsymbol{p}$ is treated as a perturbation for determining $u_k(\boldsymbol{r})$ and E_k in the vicinity of $k = 0$ in terms of the complete set of cell-periodic wave functions and energy eigen values at $k = 0$, which are assumed known. To simplify the form of Eq. (36), it is convenient to define

$$E_k' = E_k - \frac{\hbar^2 k^2}{2m}. \tag{37}$$

When spin–orbit effects are included in the calculation, instead of the usual s- and p-orbitals discussed in conjunction with the TB method, one needs to take as basis set spinors that are represented as a linear combination of the s- and p-basis functions. Kane [27,28] takes as basis the following wave functions: $|\mathrm{i}S\downarrow\rangle$, $|(X - \mathrm{i}Y)\uparrow/\sqrt{2}\rangle$, $|Z\downarrow\rangle$, $|(X + \mathrm{i}Y)\uparrow/\sqrt{2}\rangle$, $|\mathrm{i}S\uparrow\rangle$, $|-(X + \mathrm{i}Y)\downarrow/\sqrt{2}\rangle$, $|Z\uparrow\rangle$ and $|(X - \mathrm{i}Y)\downarrow/\sqrt{2}\rangle$, where $\uparrow$ and $\downarrow$ represent the spin-up and the spin-down state. Taking the wave vector $\boldsymbol{k}$ in the z-direction and considering the total Hamiltonian corresponding to the first four terms, leads to the following 8×8 interaction matrix

$$\begin{bmatrix} H & 0 \\ 0 & H \end{bmatrix}, \tag{38}$$

where

$$H = \begin{bmatrix} E_\mathrm{s} & 0 & kP & 0 \\ 0 & E_\mathrm{p} - \Delta/3 & \sqrt{2}\Delta/3 & 0 \\ kP & \sqrt{2}\Delta/3 & E_\mathrm{p} & 0 \\ 0 & 0 & 0 & E_\mathrm{p} + \Delta/3 \end{bmatrix}. \tag{39}$$

The positive constant Δ is the spin–orbit splitting of the valence band. The real quantities P and Δ are defined by

$$P = -i\frac{\hbar}{m}\langle S|p_z|Z\rangle, \qquad \Delta = \frac{3\hbar i}{4m^2c^2}\left\langle X\left|\frac{\partial V}{\partial x}p_y - \frac{\partial V}{\partial y}p_x\right|Y\right\rangle. \tag{40}$$

In Eq. (39), E_s and E_p refer to the eigenvalues of the unperturbed Hamiltonian: E_s corresponds to the conduction band and E_p to the valence band. If the $\boldsymbol{k}$ vector is not in the z-direction, the Hamiltonian is more complicated, but it can be transformed to the form of Eq. (39) by a rotation of the basis functions. Kane [27,28] used this method to describe the energy band structure in a p-type germanium and silicon, and indium antimonide. This method has been very effective in describing the electronic states in quantum confined structures such as quantum wells and superlattices (see [29,30]).

2.4. Carrier dynamics

Under the influence of an external field, Bloch electrons in a crystal change their wave vector according to the acceleration theorem

$$\hbar \frac{\mathrm{d}\boldsymbol{k}(t)}{\mathrm{d}t} = \boldsymbol{F}, \tag{41}$$

where $\boldsymbol{F}$ is the external force acting on the particle. The effect on the actual velocity or momentum of the particle is, however, not straightforward as the velocity is related to the group velocity of the wave packet associated with the particle and is given by

$$v = \frac{1}{\hbar} \nabla_k E(\boldsymbol{k}), \tag{42}$$

where $E(\boldsymbol{k})$ is one of the dispersion relations from Fig. 6. As the particle moves through $\boldsymbol{k}$-space under the influence of an electric field, for example, its velocity can be positive or negative, eventually leading to Bloch oscillations if scattering did not limit the motion. Only near extremum of the bands, for example, at the Γ point in Fig. 6 for the valence band, or close to the L point in the conduction band, does the dispersion relation resemble that of free electrons, $E(\boldsymbol{k}) = \hbar^2 k^2 / 2m^*$, where m^* is the effective mass, which is different from the free electron mass. There, the electron velocity is simply given by $\boldsymbol{v} = \hbar \boldsymbol{k} / m^*$ and the momentum is $\boldsymbol{p} = \hbar \boldsymbol{k}$.

In the case of the valence band, the states are nearly full, and current can only be carried by the absence of electrons in a particular state, leading to the concept of holes, whose dynamics are identical to that of electrons except their motion is in the opposite direction of electrons, hence they behave as positively charged particles. In relation to transport and device behavior, these holes are then treated as positively charged particles in the presence of external fields, and one has to simulate the motion of both electrons and holes.

For device modeling and simulation, different approximate band models are employed. As long as carriers (electrons and holes) have relatively low energies, they may be treated using the so-called parabolic band approximation, where they simply behave as free particles having an effective mass. If more accuracy is desired, corrections due to deviation of the dispersion relation from a quadratic dependence on $\boldsymbol{k}$ may be incorporated in the non-parabolic band model. If more than one conduction band minimum is important, this model may be extended to a multi-valley model, where the term valley refers to different conduction band minima. Finally, if the entire energy dispersion is used, one usually refers to the model as full-band and some of the previously described methods is usually employed.

3. Semi-classical transport modeling

Fig. 3 illustrated various levels of approximation in describing charge transport within a hierarchical structure ranging from the exact quantum-mechanical solution of the n-particle problem

at the bottom, to analytical one-dimensional phenomenological modeling used in circuit simulation at the top. The exact quantum-mechanical solution of even a few particle system is a challenging computational task and clearly impossible for a semiconductor device with typical free-carrier electron densities that are on the order of 10^{17} cm^{-3} or more. Hence, simplifying approximations are necessary.

For conventional semiconductor devices, such as bipolar junction transistors (BJTs) and field effect transistors (FETs), the device behavior has been adequately described within the semi-classical model of charge transport, since the characteristic dimensions are typically at length scales much larger than those over which quantum-mechanical phase coherence is maintained. Hence a particle-based description is adequate as described within the Boltzmann equation framework and approximations thereof. As device dimensions continue to shrink, the channel lengths are now approaching the characteristic wavelength of particles (for example, the de Broglie wavelength at the Fermi energy) and quantum effects are expected to be increasingly important. It has in fact been well known for 30 years that quantum confinement effects occur for electrons in the inversion layers of Si MOSFET devices. However, at room temperature and under strong driving fields, such quantum effects have usually been found to be second-order at best in terms of the overall device behavior. However, as discussed in Section 1.1, it is not clear that this situation will persist as all spatial dimensions are reduced, and consideration of quantum effects, such as tunneling and interference, may in fact dominate.

As mentioned, the classical description of charge transport is given by the BTE in the hierarchy of Fig. 3. The BTE is an integral-differential kinetic equation of motion for the probability distribution function for particles in the six-dimensional phase space of position and (crystal) momentum

$$\frac{\partial f(\boldsymbol{r},\boldsymbol{k},t)}{\partial t}+\frac{1}{\hbar}\nabla_k E(\boldsymbol{k})\cdot\nabla_r f(\boldsymbol{r},\boldsymbol{k},t)+\frac{\boldsymbol{F}}{\hbar}\cdot\nabla_k f(\boldsymbol{r},\boldsymbol{k},t)=\left.\frac{\partial f(\boldsymbol{r},\boldsymbol{k},t)}{\partial t}\right|_{\text{Coll}}, \tag{43}$$

where $f(\boldsymbol{r}, \boldsymbol{k}, t)$ is the one-particle distribution function. The RHS is the rate of change of the distribution function due to randomizing collisions, and is an integral over the in-scattering and the out-scattering terms in momentum (wave vector) space. Once $f(\boldsymbol{r}, \boldsymbol{k}, t)$ is known, physical observables, such as average velocity or current, are found from averages of f. Eq. (43) is semi-classical in the sense that particles are treated as having distinct position and momentum in violation of the quantum uncertainty relations, yet their dynamics and scattering processes are treated quantum-mechanically through the electronic band structure (discussed in Section 2) and the use of time-dependent perturbation theory. Through moment expansion of the BTE, a set of approximate partial differential equations in position space, similar to those arising in the field of fluid dynamics, are obtained leading to the so-called hydrodynamic model for charge transport, discussed later in Section 3.2. The simplification of the hydrodynamic model to include just the continuity equation and the current density written in terms of the local electric field and concentration gradients leads to the so-called drift-diffusion model, also discussed in Section 3.2. Finally, the reduction of the drift-diffusion model to one-dimensional non-linear analytical expressions allows for the development of lumped parameter behavioral models suitable for circuit level simulation of many individual devices as well as passive elements.

The BTE itself is an approximation to the underlying many-body classical Liouville equation and quantum-mechanically by the Liouville–von Neumann equation of motion for the density matrix. The main approximations inherent in the BTE are the assumption of instantaneous scattering processes in space and time, the Markov nature of scattering processes (i.e. that they are uncorrelated

with prior scattering events) and the neglect of multi-particle correlations (i.e. that the system may be characterized by a single particle distribution function). In semi-classical simulation, some of these assumptions are relaxed through the use of molecular dynamics techniques discussed in Sections 3.1.5 and 5.1.2 (in the context of device simulations). However, the inclusion of quantum effects such as particle interference, tunneling, etc., which take one further down the hierarchy of Fig. 3 is more problematic in the semi-classical Ansatz and is an active area of research today as device dimensions approach the quantum regime.

3.1. Direct solution of Boltzmann transport equation: Monte Carlo method

The ensemble MC technique has been used now for over 30 years as a numerical method to simulate non-equilibrium transport in semiconductor materials and devices, and has been the subject of numerous books and reviews [31–33]. In application to transport problems, a random walk is generated to simulate the stochastic motion of particles subject to collision processes in some medium. This process of random walk generation may be used to evaluate integral equations and is connected to the general random sampling technique used in the evaluation of multi-dimensional integrals [34].

The basic technique is to simulate the free particle motion (referred to as the free flight) terminated by instantaneous random scattering events. The MC algorithm consists of generating random free flight times for each particle, choosing the type of scattering occurring at the end of the free flight, changing the final energy and momentum of the particle after scattering and then repeating the procedure for the next free flight. Sampling the particle motion at various times throughout the simulation allows for the statistical estimation of physically interesting quantities such as the single particle distribution function, the average drift velocity in the presence of an applied electric field, the average energy of the particles, etc. By simulating an ensemble of particles, representative of the physical system of interest, the non-stationary time-dependent evolution of the electron and hole distributions under the influence of a time-dependent driving force may be simulated.

The particle-based picture, in which the particle motion is decomposed into free flights terminated by instantaneous collisions, is basically the same picture underlying the derivation of the semi-classical BTE. In fact, it may be shown that the one-particle distribution function obtained from the random walk MC technique satisfies the BTE for a homogeneous system in the long-time limit [35].

3.1.1. Free flight generation

In the MC method, the dynamics of particle motion is assumed to consist of free flights terminated by instantaneous scattering events, which change the momentum and energy of the particle. To simulate this process, the probability density, $P(t)$, is required, in which $P(t)\,\mathrm{d}t$ is the joint probability that a particle will arrive at time, t, without scattering after the previous collision at $t = 0$ and then suffer a collision in a time interval $\mathrm{d}t$ around time t. The probability of scattering in the time interval $\mathrm{d}t$ around t may be written as $\Gamma[\boldsymbol{k}(t)]\,\mathrm{d}t$, where $\Gamma[\boldsymbol{k}(t)]$ is the scattering rate of an electron or hole of wave vector $\boldsymbol{k}$. The scattering rate, $\Gamma[\boldsymbol{k}(t)]$, represents the sum of the contributions from each individual scattering mechanism, which are usually calculated using perturbation theory, as described later. The implicit dependence of $\Gamma[\boldsymbol{k}(t)]$ on time reflects the change in $\boldsymbol{k}$ due to acceleration by internal and external fields. For electrons subject to time independent electric and magnetic fields, Eq. (41) may be integrated to give the time evolution of $\boldsymbol{k}$ between collisions as

$$\boldsymbol{k}(t) = \boldsymbol{k}(0) - \frac{e(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})t}{\hbar}, \tag{44}$$

where $\boldsymbol{E}$ is the electric field, $\boldsymbol{v}$ the electron velocity (given by Eq. (42) and $\boldsymbol{B}$ is the magnetic flux density. In terms of the scattering rate, $\Gamma[\boldsymbol{k}(t)]$, the probability that a particle has not suffered a collision after a time t is given by $\exp(-\int_0^t \Gamma[\boldsymbol{k}(t')]\,dt')$. Thus, the probability of scattering in the time interval dt after a free flight of time t may be written as the joint probability

$$P(t)\,dt = \Gamma[\boldsymbol{k}(t)]\exp\left[-\int_0^t \Gamma[\boldsymbol{k}(t')]\,dt'\right]dt. \tag{45}$$

Random flight times may be generated according to the probability density $P(t)$ using, for example, the pseudo-random number generator implicit on most modern computers, which generate uniformly distributed random numbers in the range [0, 1]. Using a direct method (see [31]), random flight times sampled from $P(t)$ may be generated according to

$$r = \int_0^{t_r} P(t)\,dt, \tag{46}$$

where r is a uniformly distributed random number and t_r is the desired free flight time. Integrating Eq. (46) with $P(t)$ given by Eq. (45) above yields

$$r = 1 - \exp\left[-\int_0^{t_r} \Gamma[\boldsymbol{k}(t')]\,dt'\right]. \tag{47}$$

Since $1 - r$ is statistically the same as r, Eq. (47) may be simplified to

$$-\ln r = \int_0^{t_r} \Gamma[\boldsymbol{k}(t')]\,dt'. \tag{48}$$

Eq. (48) is the fundamental equation used to generate the random free flight time after each scattering event, resulting in a random walk process related to the underlying particle distribution function. If there is no external driving field leading to a change of $\boldsymbol{k}$ between scattering events (for example, in ultra-fast photoexcitation experiments with no applied bias), the time dependence vanishes, and the integral is trivially evaluated. In the general case where this simplification is not possible, it is expedient to introduce the so-called self-scattering method [36], in which we introduce a fictitious scattering mechanism whose scattering rate always adjusts itself in such a way that the total (self-scattering plus real scattering) rate is a constant in time

$$\Gamma = \Gamma[\boldsymbol{k}(t')] + \Gamma_{\text{self}}[\boldsymbol{k}(t')], \tag{49}$$

where $\Gamma_{\text{self}}[\boldsymbol{k}(t')]$ is the self-scattering rate. The self-scattering mechanism itself is defined such that the final state before and after scattering is identical. Hence, it has no effect on the free flight trajectory of a particle when selected as the terminating scattering mechanism, yet results in the simplification of Eq. (48) such that the free flight is given by

$$t_r = -\frac{1}{\Gamma}\ln r. \tag{50}$$

The constant total rate (including self-scattering) Γ is chosen a priori so that it is larger than the maximum scattering encountered during the simulation interval. In the simplest case, a single value is chosen at the beginning of the entire simulation (constant-gamma method), checking to ensure that

the real rate never exceeds this value during the simulation. Other schemes may be chosen that are more computationally efficient, and which modify the choice of Γ at fixed time increments [37].

3.1.2. Final state after scattering

The algorithm described determines the random free flight times during which the particle dynamics is treated semi-classically according to Eq. (44). For the scattering process itself, we need the type of scattering (i.e. impurity, acoustic phonon, photon emission, etc.) which terminates the free flight, and the final energy and momentum of the particle(s) after scattering. The type of scattering which terminates the free flight is chosen using a uniform random number between 0 and Γ, and using this pointer to select among the relative total scattering rates of all processes including self-scattering at the final energy and momentum of the particle

$$\Gamma = \Gamma_{\text{self}}[n, \boldsymbol{k}] + \Gamma_1[n, \boldsymbol{k}] + \Gamma_2[n, \boldsymbol{k}] + \cdots + \Gamma_N[n, \boldsymbol{k}], \tag{51}$$

with n the band index of the particle (or subband in the case of reduced-dimensionality systems) and $\boldsymbol{k}$ the wave vector at the end of the free flight.

Once the type of scattering terminating the free flight is selected, the final energy and momentum (as well as band or subband) of the particle due to this type of scattering must be selected. For this selection, the scattering rate, $\Gamma_j[n, \boldsymbol{k}; m, \boldsymbol{k}']$, of the jth scattering mechanism is needed, where n and m are the initial and final band (subband) indices, and $\boldsymbol{k}$ and $\boldsymbol{k}'$ are the particle wave vectors before and after scattering. Defining a spherical coordinate system around the initial wave vector $\boldsymbol{k}$, the final wave vector $\boldsymbol{k}'$ is specified by $|\boldsymbol{k}'|$ (which depends on conservation of energy) as well as the zimuthal and polar angles, φ and θ around $\boldsymbol{k}$. Typically the scattering rate $\Gamma_j[n, \boldsymbol{k}; m, \boldsymbol{k}']$ only depends on the angle θ between $\boldsymbol{k}$ and $\boldsymbol{k}'$. Therefore, φ may be chosen using a uniform random number between 0 and 2π (i.e. $2\pi r$), while θ is chosen according to the cross-section for scattering arising from $\Gamma_j[n, \boldsymbol{k}; m, \boldsymbol{k}']$. If the probability for scattering into a certain angle $P(\theta)\,\mathrm{d}\theta$ is integrable, then random angles satisfying this probability density may be generated from a uniform distribution between 0 and 1 through inversion of Eq. (46). Otherwise, a rejection technique (for example, see [31,32]) may be used to select random angles according to $P(\theta)$.

3.1.3. Ensemble Monte Carlo simulation

The algorithm may be used to track a single particle over many scattering events in order to simulate the steady-state behavior of a system. Transient simulation requires the use of a synchronous ensemble of particles in which the algorithm is repeated for each particle in the ensemble representing the system of interest until the simulation is completed. Fig. 7 illustrates an

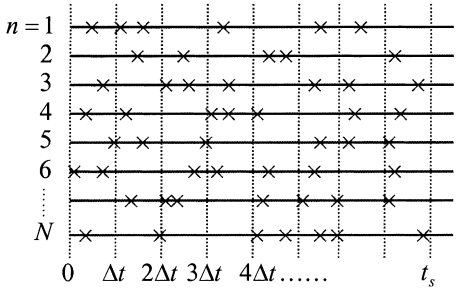


Fig. 7. Ensemble MC simulation in which a time step, Δt, is introduced over which the motion of particles is synchronized. The crosses (×) represent the scattering events.

ensemble MC simulation in which a fixed time step, Δt, is introduced to which the motion of all the carriers in the system is synchronized. The crosses ($\times$) illustrate random, instantaneous, scattering events, which may or may not occur during one time step. Basically, each carrier is simulated only up to the end of the time step and then the next particle in the ensemble is treated. Over each time step, the motion of each particle in the ensemble is simulated independent of the other particles. Non-linear effects such as carrier–carrier interactions or the Pauli exclusion principle are then updated at each times step, as discussed in more detail below.

The non-stationary one-particle distribution function and related quantities such as drift velocity, valley or subband population, etc. are then taken as averages over the ensemble at fixed time steps throughout the simulation. For example, the drift velocity in the presence of the field is given by the ensemble average of the component of the velocity at the nth time step as

$$\bar{v}_z(n\Delta t) \cong \frac{1}{N}\sum_{j=1}^{N} v_z^j(n\Delta t), \tag{52}$$

where N is the number of simulated particles and j labels the particles in the ensemble. This equation represents an estimator of the true velocity, which has a standard error given by

$$s = \frac{\sigma}{\sqrt{N}}, \tag{53}$$

where σ^2 is the variance which may be estimated from [34]

$$\sigma^2 \cong \frac{N}{N-1}\left\{\frac{1}{N}\sum_{j=1}^{N}(v_z^j)^2 - \bar{v}_z^2\right\}. \tag{54}$$

Similarly, the distribution functions for electrons and holes may be tabulated by counting the number of electrons in cells of $\boldsymbol{k}$-space. From Eq. (53), we see that the error in estimated average quantities decreases as the square root of the number of particles in the ensemble, which necessitates the simulation of many particles. Typical ensemble sizes for good statistics are in the range of 10^4–10^5 particles. Variance reduction techniques to decrease the standard error given by Eq. (53) may be applied to enhance statistically rare events, such as impact ionization or electron–hole recombination [32].

3.1.4. Scattering processes

Free carriers (electrons and holes) interact with the crystal and with each other through a variety of scattering processes which relax the energy and momentum of the particle. Based on first-order, time-dependent perturbation theory, the transition rate from an initial state $\boldsymbol{k}$ in band n to a final state $\boldsymbol{k}'$ in band m for the jth scattering mechanism is given by Fermi's Golden rule [38]

$$\Gamma_j[n,\boldsymbol{k};m,\boldsymbol{k}'] = \frac{2\pi}{\hbar}|\langle m,\boldsymbol{k}'|V_j(r)|n,\boldsymbol{k}\rangle|^2\delta(E_{k'} - E_k \mp \hbar\omega), \tag{55}$$

where $V_j(r)$ is the scattering potential of this process, E_k and $E_{k'}$ are the initial and final state energies of the particle. The delta function results in conservation of energy for long times after the collision is over, with $\hbar\omega$ the energy absorbed (upper sign) or emitted (lower sign) during the process. Scattering rates calculated by Fermi's Golden rule are typically used in MC device simulation as well as simulation of ultra-fast processes. The total rate used to generate the free

flight in Eq. (51), discussed in the previous section, is then given by

$$\Gamma_j[n,\boldsymbol{k}] = \frac{2\pi}{\hbar}\sum_{m,\boldsymbol{k}'}|\langle m,\boldsymbol{k}'|V_j(r)|n,\boldsymbol{k}\rangle|^2\delta(E_{k'} - E_k \mp \hbar\omega). \tag{56}$$

There are major limitations to the use of the Golden rule due to effects such as collision broadening and finite collision duration time [35]. The energy conserving delta function is only valid asymptotically for times long after the collision is complete. The broadening in the final state energy is given roughly by $\Delta E \approx \hbar/\tau$, where τ is the time after the collision, which implies that the normal $E(\boldsymbol{k})$ relation is only recovered at long times. Attempts to account for such collision broadening in MC simulation have been reported in the literature [39,40], although this is still an open subject of debate. Inclusion of the effects of finite collision duration in MC simulation have also been proposed [41,42]. Beyond this, there is still the problem of dealing with the quantum-mechanical phase coherence of carriers, which is neglected in the scatter free flight algorithm of the MC algorithm. This topic is discussed later in Section 6.

Fig. 8 lists the scattering mechanisms one should in principle consider in a typical MC simulation. They are roughly divided into scattering due to crystal defects, which is primarily elastic in nature, lattice scattering between electrons (holes) and lattice vibrations or phonons, which is inelastic and finally scattering between the particles themselves, including both single particle and collective type excitations. Phonon scattering involves different modes of vibration, either acoustic or optical, as well as both transverse and longitudinal modes. Carriers may either emit or absorb quanta of energy from the lattice, in the form of phonons, in individual scattering events. The designation of inter-valley scattering versus intra-valley scattering comes from the multi-valley band structure model mentioned in Section 2 and refers to whether the initial and final states are in the same valley or in different valleys. The scattering rates $\Gamma_j[n,\boldsymbol{k};m,\boldsymbol{k}']$ and $\Gamma_j[n,\boldsymbol{k}]$ are calculated using time-dependent perturbation theory using Fermi's rule, Eqs. (55) and (56), and the calculated rates are then tabulated in a scattering table in order to select the type of scattering and final state after scattering as discussed earlier.

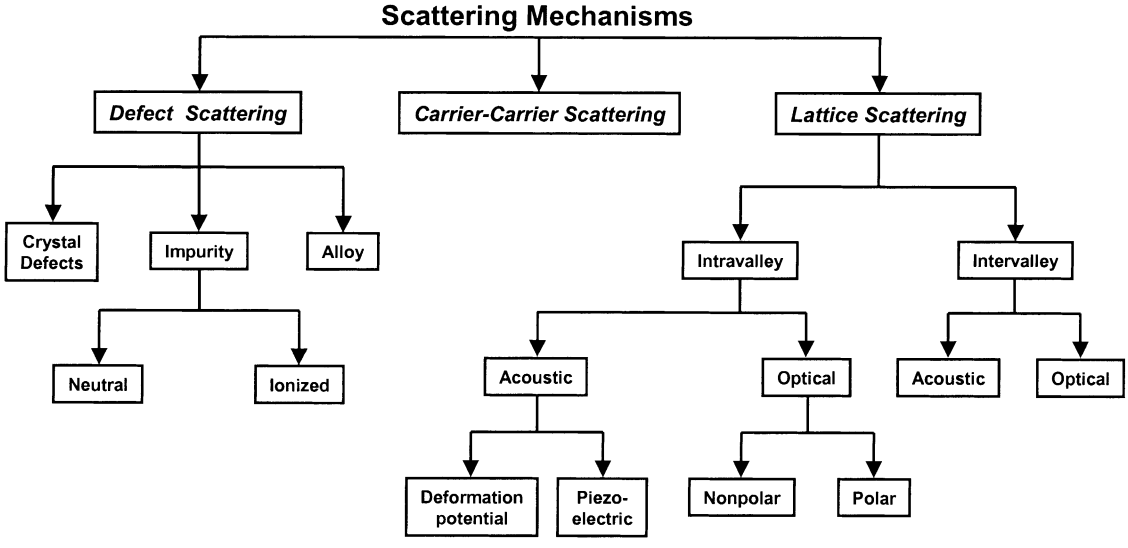


Fig. 8. Scattering mechanisms in a typical semiconductor.

3.1.5. Multi-carrier effects

Multi-particle effects relate to the interaction between particles in the system, which is a non-linear effect when viewed in the context of the BTE, due to the dependence of such effects on the single particle distribution function itself. Most algorithms developed to deal with such effects essentially linearize the BTE by using the previous value of the distribution function to determine the time evolution of a particle over the successive time step. Multi-carrier effects may range from simple consideration of the Pauli exclusion principle (which depends on the exact occupancy of states in the system), to single particle and collective excitations in the system. Inclusion of carrier–carrier interactions in MC simulation has been an active area of research for quite some time and is briefly discussed below. Another carrier–carrier effect, that is of considerable importance when estimating leakage currents in MOSFET devices, is impact ionization, which is a pure generation process involving three particles (two electrons and a hole or two holes and an electron). The latter is also discussed below.

3.1.5.1. Pauli exclusion principle. The Pauli exclusion principle requires that the bare scattering rate given by Eq. (56) be modified by a factor $1 - f_m(\boldsymbol{k}')$ in the collision integral of the BTE, where $f_m(\boldsymbol{k}')$ is the one-particle distribution function for the state $\boldsymbol{k}'$ in band (subband) m after scattering. Since the net scattering rate including the Pauli exclusion principle is always less than the bare scattering rate, a self-scattering rejection technique may be used in the MC simulation as proposed by Bosi and Jacoboni [43] for one particle simulation and extended by Lugli and Ferry [44] for EMC. In the self-scattering rejection algorithm, an additional random number r is generated (between 0 and 1), and this number is compared to $f_m(\boldsymbol{k}')$, the occupancy of the final state (which is also between 0 and 1 when properly normalized for the numerical $\boldsymbol{k}$-space discretization). If r is greater than $f_m(\boldsymbol{k}')$, the scattering is accepted and the particle's momentum and energy are changed. If this condition is not satisfied, the scattering is rejected, and the process is treated as a self-scattering event with no change of energy or momentum after scattering. Through this algorithm, no scattering to this state can occur if the state is completely full.

3.1.5.2. Carrier–carrier interactions. Carrier–carrier interactions, apart from degeneracy effects, may be treated as a scattering process within the MC algorithm on the same footing as other mechanisms. In the simplest case of bulk electrons in a single parabolic conduction band, the process may be treated as a binary collision where the scattering rate for a particle of wave vector $\boldsymbol{k}_0$ due to all the other particles in the ensemble is given by [45]

$$\Gamma_{\mathrm{ee}}(\boldsymbol{k}_0) = \frac{nm_n e^4}{4\pi\hbar^3\varepsilon^2\beta^2} \int \mathrm{d}\boldsymbol{k} f(\boldsymbol{k}) \frac{|\boldsymbol{k} - \boldsymbol{k}_0|}{|\boldsymbol{k} - \boldsymbol{k}_0|^2 + \beta^2}, \tag{57}$$

where $f(\boldsymbol{k})$ is the one-particle distribution function (normalized to unity), ε the permittivity, n the electron density and β the screening constant. In deriving Eq. (57), one assumes that the two particles interact through a statically screened Coulomb interaction, which ignores the energy exchange between particles in the screening which in itself is a dynamic, frequency-dependent effect. Similar forms have been derived for electrons in two-dimensional [46,47] and one-dimensional system [48], where carrier–carrier scattering leads to inter-subband as well as intra-subband transitions. Since the scattering rate in Eq. (57) depends on the distribution function of all the other particles in the system, this process represents a non-linear term as discussed earlier. One method is to tabulate $f(\boldsymbol{k})$ on a discrete grid, as is done for the Pauli principle, and then numerically integrate Eq. (57) at each time step. An alternate method is to use a self-scattering rejection technique [49], where the integrand excluding $f(\boldsymbol{k})$ is replaced by its maximum value and taken outside the integral over $\boldsymbol{k}$. The integral

over $f(\boldsymbol{k})$ is just unity, giving an analytic form used to generate the free flight. Then, the self-scattering rejection technique is used when the final state is chosen to correct for the exact scattering rate compared to this artificial maximum rate, similar to the algorithm used for the Pauli principle.

The treatment of inter-carrier interactions as binary collisions neglects scattering by collective excitations such as plasmons or coupled plasmon–phonon modes. These effects may have a strong influence on carrier relaxation, particularly at high carrier density. One approach is to make a separation of the collective and single particle spectrum of the interacting many-body Hamiltonian and treat them separately, i.e. as binary collisions for the single particle excitations and as electron–plasmon scattering for the collective modes [50]. Another approach is to calculate the dielectric response within the random phase approximation and associate the damping given by the imaginary part of the inverse dielectric function with the electron lifetime [51].

A semi-classical approach to carrier–carrier interaction, which is fully compatible with the MC algorithm, is the use of molecular dynamics [52], in which the carrier–carrier interaction is treated continuously in real space during the free flight phase through the Coulomb force of all the particles. A very small time step is required when using molecular dynamics to account for the dynamic distribution of the system. A time step on the order of 0.5 fs is often sufficiently small for this purpose. The small time step assures that the forces acting on the particles during the time of flight are essentially constant, that is $f(t) \cong f(t + \Delta t)$, where $f(t)$ is the single particle distribution function.

Using Newtonian kinematics, we can write the real space trajectories of each particle as

$$\boldsymbol{r}(t + \Delta t) = \boldsymbol{r}(t) + \boldsymbol{v}\Delta t + \frac{1}{2}\frac{\boldsymbol{F}(t)}{m}\Delta t^2, \tag{58}$$

and

$$\boldsymbol{v}(t + \Delta t) = \boldsymbol{v}(t) + \frac{\boldsymbol{F}(t)}{m}\Delta t. \tag{59}$$

Here, $\boldsymbol{F}(t)$ is the force arising from the applied field as well as that of the Coulomb interactions. We can write $\boldsymbol{F}(t)$ as

$$\boldsymbol{F}(t) = q\left[\boldsymbol{E} - \sum_i \nabla\varphi(\boldsymbol{r}_i(t))\right] \tag{60}$$

where $q\boldsymbol{E}$ is the force due to the applied field and the summation is the interactive force due to all particles separated by distance $\boldsymbol{r}_i$, with $\varphi(\boldsymbol{r}_i)$ the electrostatic potential. As in MC simulation, one has to simulate a finite number of particles due to practical computational limitations on execution time. In real space, this finite number of particles corresponds to a particular simulation volume given a certain density of carriers, $V = N/n$, where n is the density. Since the carriers can move in and out of this volume and since the Coulomb interaction is a long-range force, one must account for the region outside V by periodically replicating the simulated system. The contributions due to the periodic replication of the particles inside V in cells outside has a closed form solution in the form of an Ewald sum [53], which gives a linear as well as $1/r^2$ contribution to the force. The equation for the total force in the molecular dynamics technique then becomes

$$\boldsymbol{F} = \frac{-e^2}{4\pi\varepsilon}\sum_i^N \left(\frac{1}{{r_i}^2}\boldsymbol{a}_i + \frac{2\pi}{3V}\boldsymbol{r}_i\right). \tag{61}$$

This equation is easily incorporated in the standard MC simulation discussed up to this point. At every time step the forces on each particle due to all the other particles in the system are calculated from Eq. (61). From the forces, an interactive electric field is obtained which is added to the external electric field of the system to couple the molecular dynamics to the MC.

The inclusion of the carrier–carrier interactions in the context of particle-based device simulations is discussed in Section 5.1.2. The main difficulty in treating this interaction term in device simulations arises from the fact that the long-range portion of the carrier–carrier interaction is included via the numerical solution of the quasi-static Poisson equation (see Section 4.2). Under these circumstances, special care has to be taken when incorporating the short-range portion of this interaction term to prevent double counting of the force.

3.1.5.3. Band-to-band impact ionization. Another carrier–carrier scattering process is that of impact ionization, in which an energetic electron (or hole) has sufficient kinetic energy to create an electron–hole pair. Impact ionization therefore leads to the process of carrier multiplication. This process is critical, for example, in the avalanche breakdown of semiconductor junctions and is a detrimental effect in short channel MOS devices in terms of excess substrate current and decreased reliability.

The ionization rate of valence electrons by energetic conduction band electrons is usually described by Fermi's rule Eq. (55), in which a screened Coulomb interaction is assumed between the two particles, as was done in Section 3.1.5.3, where screening is described by an appropriate dielectric function such as that proposed by Levine and Louie [54]. In general, the impact ionization rate should be a function of the wave vector of the incident electron, hence of the direction of an electric field in the crystal, although there is still some debate as to the experimental and theoretical evidence. More simply, the energy-dependent rate (averaged over all wave vectors on a constant energy shell) may be expressed analytically in the power law form

$$\Gamma_{\mathrm{ii}}(E) = P[E - E_{\mathrm{th}}]^a, \tag{62}$$

where E_{th} is the threshold energy for the process to occur, which is determined by momentum and energy conservation considerations, but minimally is the band-gap of the material itself. P and a are parameters which may be fit to more sophisticated models. The Keldysh formula [55] is derived by expanding the matrix element for scattering close to threshold, which gives $a = 2$ and the constant $P = C/E_{\mathrm{th}}^2$, with $C = 1.19 \times 10^{14}\ \mathrm{s}^{-1}$ and assuming a parabolic band approximation,

$$E_{\mathrm{th}} = \frac{3 - 2m_{\mathrm{v}}/m_{\mathrm{c}}}{1 - m_{\mathrm{v}}/m_{\mathrm{c}}} E_g, \tag{63}$$

where m_{v} and m_{c} are the effective masses of the valence and conduction band, respectively, and E_g is the band-gap. More complete full-band structure calculations of the impact ionization rate have been reported for Si [56,57], GaAs [57,58] and wide band-gap materials [59], which are fairly well fit using power law model given in Eq. (62).

Within the ensemble MC method, the scattering rate given by Eq. (62) is used to generate the free flight time. The state after scattering of the initial electron plus the additional electron and hole must satisfy both energy and momentum conservation within the Fermi rule model, which is somewhat complicated unless simple parabolic band approximations are made.

3.1.6. Full-band particle-based simulation

The MC algorithm discussed in this section initially evolved during the 1970s and early 1980s using simplified representations of the electronic band structure in terms of a multi-valley parabolic

or non-parabolic approximation close to band minima and maxima. This simplifies the particle tracking in terms of the E–k relationship and particle motion in real space and greatly simplifies the calculated scattering rates such that analytical forms may be used. It soon became apparent that for devices where high field effects are important or for the correct simulation of high energy processes like impact ionization, the full-band structure of the material is required. Particle-based simulation which incorporates part or all of the band structure directly into the particle dynamics and scattering is commonly referred to as full-band MC simulation [33].

Typically, the EPM discussed in Section 2.1 has been utilized in full-band MC codes due to the relative simplicity of the calculation and the plane wave basis which facilitates calculation of some scattering processes. Early full-band codes developed at the University of Illinois utilized the full-band structure for the particle dynamics, but assumed isotropic energy-dependent scattering rates using the full-band density of states [33]. This is due to the computational difficulty and memory requirements to store the full k-dependent scattering rates throughout the whole Brillouin zone. Later simulators relaxed this restriction, although often assuming quasi-isotropic rates. Probably the most completely developed full-band code for full-band MC device simulation is the DAMOCLES code developed at IBM by Fischetti and Laux [60], which has been used extensively for simulation of a variety of device technologies [61].

These full-band codes are based on essentially the same algorithm as was discussed in Section 3.1, in which a particle scatters based on the total scattering rate, then the type of scattering and the final state after scattering are selected using the full k-dependent rates for each mechanism. An alternative approach referred to as cellular MC [62], stores the entire transition table for the total scattering rate for all mechanisms from every initial state k to every final state k'. Particle scattering is accomplished in a single step, at the expense of large memory consumption (on the order of 2 GB of RAM) necessary to store the necessary scattering tables.

Fig. 9 shows the calculated steady-state drift velocity and average energy for Si as a function of electric field for the CMC method and the earlier results from DAMOCLES which are essentially the

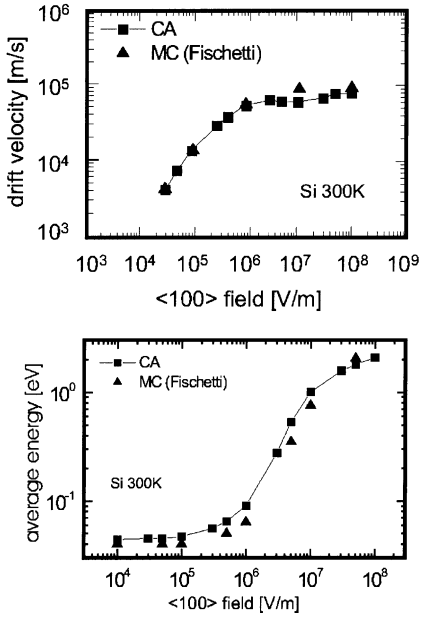


Fig. 9. Comparison of full-band MC simulation results using DAMOCLES [60] (triangles) to those using the CMC approach [62]. The upper plot is the steady-state drift velocity and the lower plot the average energy vs. electric field.

same. In such simulations, steady-state is typically reached after 2 ps of simulation time and then averages are calculated over the ensemble and in time for several picoseconds thereafter.

3.2. Hydrodynamic and drift-diffusion model

In a number of practical applications, it is not necessary to know the exact distribution function obtained by solving the BTE. Instead, it suffices to know only the lowest moments, like the mean and the variance, for example. For this purpose, semi-classical transport equations are derived based on either the first three or four moments of the distribution function that describe the carrier concentration, current, average carrier energy and energy flux variation (if four moments are retained). The various coefficients that appear in these equations may be assumed to be a function of the average carrier energy. The relationship between these coefficients and the energy is usually determined from steady-state MC calculations and experimental data for homogeneous samples.

The first three moments of the BTE for electrons, which describe conservation of particles, momentum (mass flow) and energy, are expressed by the following set of equations [63]

$$\frac{\partial n}{\partial t} = -\nabla(n\boldsymbol{v}), \tag{64}$$

$$\frac{\partial P_i}{\partial t} = -\sum_j 2\frac{\partial W_{ji}}{\partial x_j} - neE_i - P_i\left\langle\frac{1}{\tau_m}\right\rangle, \tag{65}$$

$$\frac{\partial W}{\partial t} = -\sum_i \frac{\partial J_{W_i}}{\partial x_i} + (\boldsymbol{J}\cdot\boldsymbol{E}) - (W - W_0)\left\langle\frac{1}{\tau_\varepsilon}\right\rangle, \tag{66}$$

where n is the electron density, $\boldsymbol{v}$ the average electron velocity, P_i and E_i ($i = 1, 2, 3$ for x, y and z-coordinates) are the ith components of the total momentum and the electric field, W_{ij} is a component of the total kinetic energy density tensor, W the total kinetic energy density (W_0 being the equilibrium electron energy corresponding to the lattice temperature T_L), $\boldsymbol{J} = -e\boldsymbol{P}/m^*$ the current density and $\boldsymbol{J}_W$ the kinetic energy flux. The momentum and energy relaxation rates that appear in Eqs. (65) and (66) are defined as

$$\left\langle\frac{1}{\tau_m}\right\rangle = \frac{\sum_k Pf(\boldsymbol{k})/\tau_m(\boldsymbol{k})}{P}, \quad \text{and} \quad \left\langle\frac{1}{\tau_\varepsilon}\right\rangle = \frac{\sum_k \varepsilon(\boldsymbol{k})f(\boldsymbol{k})/\tau_\varepsilon(\boldsymbol{k})}{W - W_0}. \tag{67}$$

Assuming that the carrier velocity $\boldsymbol{v}$ equals the sum of a drift ($\boldsymbol{v}_d$) and a thermal component ($\boldsymbol{c}$), i.e. $\boldsymbol{v} = \boldsymbol{v}_d + \boldsymbol{c}$, one can express the total kinetic energy as being equal the sum of a drift and a thermal energy component due to the random thermal motion of the carriers, i.e.

$$W = \tfrac{1}{2}nm^*v_d^2 + \tfrac{3}{2}k_BT_C = K + \tfrac{3}{2}k_BT_C, \tag{68}$$

where k_B is the Boltzmann constant and K is the drift component of the kinetic energy density. The kinetic energy flux term appearing in Eq. (66) then reduces to

$$\boldsymbol{J}_W = W\boldsymbol{v}_d + \boldsymbol{v}_d(nk_B\overleftrightarrow{T}_C) + \boldsymbol{Q} \tag{69}$$

where $\leftrightarrow T_C$ is the temperature tensor and $\boldsymbol{Q}$ is the heat flux vector.

Further simplifications to the momentum and energy balance equations are usually made assuming a displaced Maxwellian form for the distribution function, which leads to a diagonal

temperature tensor. This approximation is valid for systems in which the electron–electron interactions play a significant role. The use of a displaced Maxwellian distribution function leads to the following set of balance (also known as hydrodynamic) equations

$$\frac{\partial n}{\partial t} = -\nabla(n\boldsymbol{v}), \tag{70}$$

$$\frac{\partial v_d}{\partial t} + \boldsymbol{v}_\mathrm{d}\nabla\boldsymbol{v}_\mathrm{d} + \frac{1}{nm^*}\nabla(nk_\mathrm{B}T) = -\boldsymbol{v}_\mathrm{d}\left\langle\frac{1}{\tau_\mathrm{m}}\right\rangle, \tag{71}$$

$$\frac{\partial W}{\partial t} = -\nabla(W\boldsymbol{v}_\mathrm{d} + nk_\mathrm{B}T\boldsymbol{v}_\mathrm{d} - \kappa\nabla T) + \boldsymbol{J}\cdot\boldsymbol{E} - (W - W_0)\left\langle\frac{1}{\tau_\varepsilon}\right\rangle. \tag{72}$$

Note that the displaced Maxwellian distribution, which is symmetric in momentum space, will lead to zero heat flux, since it involves the third moment of the distribution function. However, Bløtekjær [64] has pointed out that this term may be significant for non-Maxwellian distributions, so that a phenomenological description for the heat flux $\boldsymbol{Q} = -\kappa\nabla T$ has been used in Eq. (72), where κ is the thermal conductivity. As already mentioned, the ensemble averaged energy-dependent momentum and energy relaxation rates that appear in Eqs. (71) and (72), are determined by steady-state MC simulation for bulk material under uniform electric fields.

For simulations where steady-state solutions are required, or transient events with relatively large time scales are being investigated, it is possible to neglect the terms $\partial\boldsymbol{v}_\mathrm{d}/\partial t$, $\partial W/\partial t$ and $\nabla(n\boldsymbol{v}_\mathrm{d})$ in the momentum and energy balance equations. Furthermore, if carrier heating effects are negligible, then the drift component of the kinetic energy density can be ignored. With these simplifications, and assuming that there are no temperature gradients in the system, the steady-state momentum balance equation leads to the following expression for the current density

$$\boldsymbol{J} = -en\boldsymbol{v}_\mathrm{d} = en\mu\boldsymbol{E} + eD\nabla n, \tag{73}$$

where the mobility and the diffusion coefficient are calculated using [65]

$$\mu_n = \frac{e}{m^*\langle 1/\tau_\mathrm{m}\rangle}, \quad \text{and} \quad D_n = \frac{1}{e}k_\mathrm{B}T\mu_n. \tag{74}$$

The result given in Eq. (73), together with the Eq. (70), constitutes the drift-diffusion model for electrons. A similar set of equations can be written for holes. Since a low-field limit has been assumed in order to arrive at the result given in Eq. (74), the mobility and the diffusion coefficients are energy independent quantities. To extend the validity of the drift-diffusion model to the high-field regime, ad hoc inclusion of field-dependent mobilities and diffusion coefficients is usually used in standard device simulators, such as Silvaco's ATLAS. However, the applicability of such an approach becomes questionable in nano-scale devices in which non-stationary and ballistic transport effects play significant role.

4. Field equations

In the previous sections, we have discussed transport models within the context of the semi-classical BTE. Of equal or greater importance in terms of the behavior of electronic devices are the

self-consistent fields inside the device associated with the external bias and internal charge and current distributions. In particular, the carriers within a semiconductor are accelerated by the electric and magnetic fields according to the Lorentz force equation

$$\boldsymbol{F} = \hbar \frac{\mathrm{d}\boldsymbol{k}}{\mathrm{d}t} = q(\boldsymbol{E} + \boldsymbol{v}\cdot\boldsymbol{B}) \tag{75}$$

where $\boldsymbol{B}$ and $\boldsymbol{E}$ are the magnetic flux density and electric field intensity, respectively. In general, these fields correspond to the solution of Maxwell's equations originating from the microscopic charges and currents in the device. For high frequency device modeling, in which the device dimensions are comparable to the wavelength, wave propagation effects may be important and full wave solutions of Maxwell's equations are necessary. In considering optoelectronic devices, direct solutions of either Maxwell's or the associated Helmholtz equation are necessary to represent optical field within the device. Such approaches has been taken, for example, in modeling microwave transistors under the context of 'global modeling' [66], as well as for the analysis of semiconductor device, in which optical cavity modes are coupled to semiconductor device simulation [67]. For most device modeling applications, the magnetic contribution to the Lorentz force is much smaller than the electric field contribution, and wave propagation effects are negligible, so that quasi-static representation of the fields in terms of the solution of Poisson's equation are sufficient.

In Section 4.1, we first discuss direct solution of Maxwell's equations using the finite-difference time-domain (FDTD) method, which has been used extensively in the electromagnetics community, and is employed in simulation of high frequency devices and circuits. The subsequent section (Section 4.2) is devoted to the description of efficient numerical solution methods for the Poisson's equation, as applied to semiconductor device simulation. The coupling of the various field solvers in semiconductor device simulation are then discussed in Section 5.

4.1. Finite-difference time-domain techniques

For electromagnetic solvers in general, there exist a number of general commercial packages for solving Maxwell's equations, such as ANSOFT's HFSS program [68]. For semiconductor device simulation, the FDTD method mentioned is convenient since time-domain methods are also used in the solution of the transport equations as discussed earlier. Commercial codes are available for FDTD electromagnetic simulation (see, for example, the XFDTD [69] and Fidelity [70] codes). ISE has recently released a commercial simulation tool combining their device simulation tool DESSIS with an FDTD-based simulator (EMLAB) for high frequency simulation of semiconductor devices [71]. In the following, we give a brief description of the FDTD method itself, and its coupling to particle-based simulators.

Maxwell's equations in SI units are written as

$$\begin{aligned} \nabla\cdot\boldsymbol{E} &= -\frac{\partial \boldsymbol{B}}{\partial t}, & \nabla\cdot\boldsymbol{D} &= \rho \\ \nabla\cdot\boldsymbol{H} &= \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}, & \nabla\cdot\boldsymbol{B} &= 0 \end{aligned} \tag{76}$$

Here ρ is the free charge density

$$\rho(\boldsymbol{r}) = q(N_{\mathrm{D}} - N_{\mathrm{A}} + p - n), \tag{77}$$

where N_D and N_A are the ionized donor and acceptor concentrations, while p and n are the hole and electron concentrations which are functions of position. For linear isotropic media, the relations for the various fields is simplified by the constitutive relationships

$$\boldsymbol{D} = \varepsilon \boldsymbol{E}, \quad \text{and} \quad \boldsymbol{B} = \mu \boldsymbol{H}, \tag{78}$$

where μ is the permeability ($\mu_0 = 4\pi \times 10^{-7}$ for non-magnetic semiconductors) and ε is the permittivity.

In Cartesian coordinates, the curl equations are expanded as

$$\begin{aligned}
\frac{\partial H_x}{\partial t} &= \frac{1}{\mu}\left(\frac{\partial E_y}{\partial z} - \frac{\partial E_z}{\partial y}\right) \\
\frac{\partial H_y}{\partial t} &= \frac{1}{\mu}\left(\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z}\right) \\
\frac{\partial H_z}{\partial t} &= \frac{1}{\mu}\left(\frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x}\right) \\
\frac{\partial E_x}{\partial t} &= \frac{1}{\varepsilon}\left(\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z}\right) + J_x \\
\frac{\partial E_y}{\partial t} &= \frac{1}{\varepsilon}\left(\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x}\right) + J_y \\
\frac{\partial E_z}{\partial t} &= \frac{1}{\varepsilon}\left(\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y}\right) + J_z
\end{aligned} \tag{79}$$

These may then be discretized [72] using the so-called Yee cell [73]. The Yee cell consists of a set of interpenetrating finite-difference grids, one representing the electric the other the magnetic fields, over which the derivatives in space are expanded. The electric fields are assumed to be updated at time step n while the magnetic fields are updated at time step $n + (1/2)$. Using central differences, the discretized Maxwell's equations are written as

$$\begin{aligned}
&E_z^{n+1}(i,j,k+0.5) \\
&\quad = E_z^n(i,j,k+0.5) + \frac{\Delta t}{\varepsilon}\left[\frac{H_y^{n+0.5}(i+0.5,j,k+0.5) - H_y^{n+0.5}(i-0.5,j,k+0.5)}{\Delta x}\right. \\
&\qquad \left. - \frac{H_x^{n+0.5}(i,j+0.5,k+0.5) - H_x^{n+0.5}(i,j-0.5,k+0.5)}{\Delta y} + J_z^{n+0.5}(i,j,k+0.5)\right],
\end{aligned} \tag{80}$$

for the electric field and

$$\begin{aligned}
&H_x^{n+0.5}(i,j+0.5,k+0.5) \\
&\quad = H_x^{n-0.5}(i,j+0.5,k+0.5) - \frac{\Delta t}{\mu}\left[\frac{E_z^n(i,j+1,k+0.5) - E_z^n(i,j,k+0.5)}{\Delta y}\right. \\
&\qquad \left. - \frac{E_y^n(i,j+0.5,k+1) - E_y^n(i,j+0.5,k)}{\Delta z}\right],
\end{aligned} \tag{81}$$

for the magnetic field strength. Similar equations hold for the other components of $\boldsymbol{E}$ and $\boldsymbol{H}$. As can be seen in Eqs. (80) and (81), the electric field at time step $n + 1$ is determined explicitly by the

electric field at time step n, and the magnetic fields in the adjoining Yee cell mesh points at the previous half time step, $n + (1/2)$. Likewise, the magnetic field at time step $n + (1/2)$ is calculated explicitly by the magnetic field at time step $n - (1/2)$ and the electric field in adjacent mesh points at time step n. Hence, the time evolution of the electric and magnetic fields is calculated non-iteratively in a time marching fashion in half time step intervals. The stability of this technique naturally depends on the time step and grid spacing.

The grid cell size is typically chosen to minimize the effects of numerical dispersion. If the highest frequency component in the simulation is characterized by a wavelength $\lambda = c/v$, then empirically, the cell size should be smaller than approximately $\lambda/10$ to avoid artificial dispersion effects. Other considerations in the grid size depend on the geometrical considerations of the structure being simulated. Once the spatial grid has been determined, the time step, Δt, used to propagate the electric and magnetic fields forward in time, has an upper bound determined by the furthest distance a signal can propagate over this time interval [74]. Setting this distance equal to the minimum grid spacing, this constraint yields

$$\Delta t \leq \frac{\mathrm{d}r\sqrt{\mu\varepsilon}}{\sqrt{3}} = \frac{\mathrm{d}r}{\sqrt{3}v_{\mathrm{p}}}, \tag{82}$$

where $\mathrm{d}r$ is the minimum space increment in any direction, and v_{p} is the phase velocity. More generally, the Courant stability condition [75] is of the form

$$\Delta t \leq \frac{1}{v_{\mathrm{p}}\sqrt{(1/(\Delta x)^2) + (1/(\Delta y)^2) + (1/(\Delta z)^2)}}, \tag{83}$$

where Δx, Δy and Δz are the minimum grid spacings in the three Cartesian coordinate directions.

In the FDTD method, the current density, $\boldsymbol{J}$, appearing earlier is the primary coupling between charge transport and the coupled electromagnetic fields. During the half time interval between the calculation of the electric or magnetic field components, charge transport is simulated over this time interval using the frozen field components of the previous time step. The updated current density at each grid point is then used in Eq. (80) to update the electric field in the FDTD algorithm. Within the drift-diffusion or hydrodynamic models, the current density is calculated directly from the continuity equation. Within a particle-based scheme, one has to map the continuous particle motion onto the discrete grid points. In a nearest grid point (NGP) scheme, the weighted velocities of the particle in the ensemble that lie within a unit cell volume around a given grid point, are summed according to

$$J(i,j,k,t) = \frac{1}{\Delta x \Delta y \Delta z} \sum_{n=1}^{N(i,j,k)} S_n v_n, \tag{84}$$

where S_n and v_n refer, respectively, to the charge and velocity of the nth particle associated with the grid point and $N(i, j, k)$ the total number of particles within a unit cell around the grid point (i, j, k).

The solution of Maxwell's equations using the FDTD technique requires the imposition of boundary conditions. The usual conditions for the continuity of the tangential and perpendicular components of the electric and magnetic fields are applied at dielectric and metallic boundaries. The simulation of open systems (e.g. an infinite domain) requires special care in that boundary conditions on the simulation domain must be specified to minimize the artificial reflection of the outgoing

wave. Various types of absorbing boundary conditions have been proposed in the literature [76,77]. Currently, the most popular technique that minimizes artificial reflection on open boundaries is the so-called perfectly matched layer (PML) method proposed by Berenger [78]. This method utilizes a fictitious magnetic loss for impedance matching of the outgoing wave to a highly lossy material. The PML has been found effective in attenuating outgoing waves for a wide range of frequencies and incident angles on the boundary surface, and is currently the state of the art in the FDTD method.

4.2. Poisson's equation

For most semiconductor device modeling, the device dimensions themselves are much smaller than the characteristic wavelengths associated with the maximum frequency of operation, so that quasi-static solutions of Maxwell's equations are sufficient. In this case, the electric field may be expressed as the gradient of a scalar potential. The divergence equation for the electric displacement in Eq. (74) then yields Poisson's equation, which in two- or three-dimensions is written as

$$\nabla^2 V(\boldsymbol{r}) = -\frac{q(N_{\mathrm{D}} - N_{\mathrm{A}} + p - n)}{\varepsilon_{\mathrm{sc}}} = -\frac{\rho(\boldsymbol{r})}{\varepsilon_{\mathrm{sc}}}, \tag{85}$$

where, as previously noted, N_{D} and N_{A} are the ionized donor and acceptor concentrations, while p and n the hole and electron concentrations, which are functions of position. Poisson's equation assumes that the field may be described as the gradient of a scalar potential V, valid in the quasi-static limit for the associated temporal variation.

In the numerical solution of the two- or three-dimensional Poisson equation, the application of a conventional finite-difference or finite elements scheme leads to algebraic equations having a well-defined structure, defined over a finite mesh or grid. For example, using central differences in two-dimensions to write the Laplacian appearing in Eq. (85), the discretized form becomes a system of linear equations

$$a_{\mathrm{N}} V_{i,j+1} + a_{\mathrm{S}} V_{i,j-1} + a_{\mathrm{E}} V_{i+1,j} + a_{\mathrm{W}} V_{i-1,j} + a_{\mathrm{C}} V_{i,j} = -\frac{\rho_{i,j}}{\varepsilon_{\mathrm{sc}}}, \tag{86}$$

where i and j label the two dimensional coordinates of a particular grid point, and the coefficients representing the grid cells to the north (N), south (S), etc. are determined from the grid spacing in the usual way.

In general, the resulting system of equations can be represented by the matrix equation $\boldsymbol{Ax} = \boldsymbol{b}$ [79]. The most suitable methods for the solution of this matrix equation are direct methods, but the computational cost becomes prohibitive as the number of equations increases, which is normally the case in two- and three-dimensional device simulations. This has led to the development of iterative procedures that utilize the well-defined structure of the coefficient matrix. The simplest and most commonly used iterative procedures are the successive over-relaxation (SOR) and the alternating direction implicit (ADI) methods [80]. Both methods lose their effectiveness when complex problems are encountered and when the equation set becomes large, as it is usually the case in three-dimensional problems. One alternative to avoiding this problem is the incomplete lower–upper (ILU) decomposition method described in Section 4.2.1, which provides a significant increase in the power of iterative methods and is, therefore, more suitable for solving three-dimensional problems. Other alternatives for solving large-scale three-dimensional problems are the multi-grid and the conjugate gradient methods that are discussed in Sections 4.2.2 and 4.2.3, respectively.

4.2.1. Incomplete lower–upper decomposition method

Within incomplete factorization schemes [81] for two-dimensional problems, the matrix $\boldsymbol{A}$ is decomposed into a product of lower ($\boldsymbol{L}$) and upper ($\boldsymbol{U}$) triangular matrices, each of which has four non-zero diagonals in the same locations as the ones of the original matrix $\boldsymbol{A}$. The unknown elements of the $\boldsymbol{L}$ and $\boldsymbol{U}$ matrices are selected in such a way that the five diagonals common to both $\boldsymbol{A}$ and $\boldsymbol{A}'=\boldsymbol{LU}$ are identical and the four superfluous diagonals represent the matrix $\boldsymbol{N}$, i.e. $\boldsymbol{A}'=\boldsymbol{A}+\boldsymbol{N}$. Thus, rather than solving the original system of equations $\boldsymbol{Ax}=\boldsymbol{b}$, one solves the modified system $\boldsymbol{LUx}=\boldsymbol{b}+\boldsymbol{Nx}$, by solving successively the matrix equations $\boldsymbol{LV}=\boldsymbol{b}+\boldsymbol{Nx}$ and $\boldsymbol{V}=\boldsymbol{Ux}$, where $\boldsymbol{V}$ is an auxiliary vector. It is important to note that the four superfluous terms of $\boldsymbol{N}$ affect the rate of convergence of the ILU method. Stone [82] suggested the introduction of partial cancellation, which minimizes the influence of these additional terms and accelerates the rate of convergence of the ILU method. By using a Taylor series expansion, the superfluous terms appearing in $\boldsymbol{A}'$ are partially balanced by subtracting approximately equal terms.

4.2.2. Multi-grid methods

The multi-grid method represents an improvement over the SOR and ILU methods in terms of iterative techniques available for solving large systems of equations [83]. The basic principle behind the multi-grid method is to reduce different Fourier components of the error on grids with different mesh sizes. Most iterative techniques work by quickly eliminating the high frequency Fourier components, while the low frequency ones are left virtually unchanged. The result is a convergence rate that is initially fast, but slows down dramatically as the high frequency components disappear. The multi-grid method utilizes several grids, each with consecutively coarser mesh sizes. Each of these grids acts to reduce a different Fourier component of the error, therefore, increasing the rate of convergence with respect to single grid based methods, such as an SOR.

The initial setup for the multi-grid solver is to create a sequence of grids. The finest grid is generated according to the device structure, and each consecutive grid is obtained by doubling the spacing of the previous one. This is repeated until the final, coarsest grid contains 3×3 points. The coarsening process must ensure propagation of the boundary conditions to all grids in order to obtain a unique solution. The Poisson equation is solved on the finest grid, and the residual, which is a computational measure of the error, is passed down or restricted to the next, coarser grid. The next grid solves the error equation. This results in a reduction of the relatively lower-frequency error components, as compared with the initial error on the previous grid. This process is continued through all subsequent grids. At the coarsest grid, the error equation is solved exactly and the error is prolonged up through the finer grids, adding its correction at each grid level. At the finest grid, the correction is used to update the final solution. In this way, the multiple error components are reduced simultaneously and the procedure is repeated until convergence of a solution is obtained.

4.2.3. Conjugate gradient methods with pre-conditioning

The basic conjugate gradient (CG) algorithm is one of the best known iterative techniques for solving sparse symmetric positive definite (SPD) systems, but it loses its applicability when the resulting system of equations is not SPD. In such circumstances, the best alternative are the Lanczos-type algorithms, which solve not only the original system $\boldsymbol{Ax}=\boldsymbol{b}$ but also solve the dual linear system $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}^{*}=\boldsymbol{b}^{*}$. In recent years, the CG-squared (CGS) method due to Sonneveld [84] has been recognized as an attractive transpose-free variant of the Bi-CG iterative method [85]. This method works quite well in many cases, but the very high variations in the residual vectors often cause the

residual norms to become inaccurate, which can lead to substantial buildup of rounding errors and overflow. The Bi-CGSTAB method due to Van der Vorst [86] is a variant of the CGS algorithm, which avoids squaring of the residual polynomial. It has been demonstrated that the convergence behavior of this method is smoother because it produces more accurate residual vectors and, therefore, more accurate solutions. In conjunction with the Bi-CGSTAB method, a successful pre-conditioning matrix can be obtained by using ILU factorization [87]. If $\boldsymbol{L}$ and $\boldsymbol{U}$ are the strictly lower and the strictly upper triangular parts of $\boldsymbol{A}$, then the pre-conditioning matrix is

$$\boldsymbol{K}_{\mathrm{ILU}(k)} = (\boldsymbol{L} + \tilde{\boldsymbol{D}})\tilde{\boldsymbol{D}}^{-1}(\boldsymbol{U} + \tilde{\boldsymbol{D}}), \tag{87}$$

where $\mathrm{diag}(\boldsymbol{K}_{\mathrm{ILU}(k)}) = \mathrm{diag}(\boldsymbol{A})$. The case $k = 0$ is used here, and means no fill-ins are allowed. Once the diagonal $\tilde{\boldsymbol{D}}$ is computed, scaling of the original matrix $\boldsymbol{A}$ is performed using

$$\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{D}}^{-1/2}\boldsymbol{A}\tilde{\boldsymbol{D}}^{-1/2} = \mathrm{diag}(\tilde{\boldsymbol{A}}) + \tilde{\boldsymbol{L}} + \tilde{\boldsymbol{U}}. \tag{88}$$

The pre-conditioning matrix for this symmetrically scaled matrix is of the form

$$\tilde{\boldsymbol{K}} = (\tilde{\boldsymbol{L}} + \boldsymbol{I})(\boldsymbol{I} + \tilde{\boldsymbol{U}}), \tag{89}$$

where $\boldsymbol{I}$ is the identity matrix. The Bi-CGSTAB method is now applied to the pre-conditioned system

$$(\tilde{\boldsymbol{L}} + \boldsymbol{I})^{-1}\tilde{\boldsymbol{A}}(\boldsymbol{I} + \tilde{\boldsymbol{U}})^{-1}\tilde{\boldsymbol{x}} = (\tilde{\boldsymbol{L}} + \boldsymbol{I})^{-1}\tilde{\boldsymbol{b}}, \tag{90}$$

where $\tilde{\boldsymbol{b}} = \tilde{\boldsymbol{D}}^{-1/2}\boldsymbol{b}$ and the solution of the original system of equations is obtained via $\boldsymbol{x} = \tilde{\boldsymbol{D}}^{-1/2}(\boldsymbol{I} + \tilde{\boldsymbol{U}})^{-1}\tilde{\boldsymbol{x}}$. In the calculation of the product $(\tilde{\boldsymbol{L}} + \boldsymbol{I})^{-1}\tilde{\boldsymbol{A}}(\boldsymbol{I} + \tilde{\boldsymbol{U}})^{-1}\tilde{\boldsymbol{x}}$, appearing in Eq. (90), extra work is avoided by using the Eisenstat's trick [88]

$$(\tilde{\boldsymbol{L}} + \boldsymbol{I})^{-1}\tilde{\boldsymbol{A}}(\boldsymbol{I} + \tilde{\boldsymbol{U}})^{-1}\tilde{\boldsymbol{p}} = \tilde{\boldsymbol{t}} + (\tilde{\boldsymbol{L}} + \boldsymbol{I})\{\tilde{\boldsymbol{p}} + [\mathrm{diag}(\tilde{\boldsymbol{A}}) - 2\boldsymbol{I}]\tilde{\boldsymbol{t}}\}, \tag{91}$$

where $\tilde{\boldsymbol{t}} = (\boldsymbol{I} + \tilde{\boldsymbol{U}})^{-1}\tilde{\boldsymbol{p}}$.

5. Device simulations

In previous sections, we introduced the numerical solution of the BTE using MC methods (Section 3.1), the approximate solutions of the BTE using either hydrodynamic or drift-diffusion model (Section 3.2), and the solution of Maxwell's equations (Section 4.1) and Poisson's equation (Section 4.2) over a finite mesh. Within a device, both the transport kernel and the field solver are coupled to each other. The field associated with the potential coming from Poisson's equation is the driving force accelerating particles in the MC phase, for example, while the distribution of mobile (both electrons and holes) and fixed charges (e.g. donors and acceptors) provides the source of the electric field in Poisson's equation corresponding to the right-hand side of Eq. (85). In the later sections we give an extensive description of the MC particle-based device simulators with emphasis on the particle–mesh (PM) coupling and the inclusion of the short-range Coulomb interaction (Section 5.1). This discussion is followed by a brief summary of hydrodynamic/drift-diffusion device simulators with reference to commercially available simulation software (Section 5.2). We finish this section with an application of the FDTD methods coupled with a MC transport kernel on the example of a co-planar strip on a GaAs substrate (Section 5.3).

5.1. Particle-based device simulations

Within the particle-based EMC method with its time-marching algorithm, Poisson's equation may be decoupled from the BTE over a suitably small time step (typically less than the inverse plasma frequency corresponding to the highest carrier density in the device). Over this time interval, carriers accelerate according to the frozen field profile from the previous time step solution of Poisson's equation, and then Poisson's equation is solved at the end of the time interval with the frozen configuration of charges arising from the MC phase (see discussion in [52]). Note that Poisson's equation is solved on a mesh, whereas the solution of charge motion using EMC occurs over a continuous range of coordinate space in terms of the particle position. Therefore, a PM coupling is needed for both the charge assignment and the force interpolation. The PM coupling is broken into four steps: (1) assign particle charge to the mesh; (2) solve the Poisson equation on the mesh; (3) calculate the mesh-defined forces; and (4) interpolate to find forces on the particle. There are a variety of schemes that can be used for the PM coupling and these are discussed in Section 5.1.1.

Another issue that has to be addressed in particle-based simulations is the real space boundary conditions for the particle part of the simulation. Reflecting boundary conditions are usually imposed at the artificial boundaries. As far as the Ohmic contacts are concerned, they require more careful consideration because electrons crossing the source and drain contact regions contribute to the corresponding terminal current. In order to conserve the charge in the device, the electrons exiting the contact regions must be re-injected. Commonly employed models for the contacts include [89]:

- Electrons are injected at the opposite contact with the same energy and wave vector $\boldsymbol{k}$. If the source and drain contacts are in the same plane, as in the case of MOSFET simulations, the sign of $\boldsymbol{k}$ normal to the contact will change. This is an unphysical model, however [90].
- Electrons are injected at the opposite contact with a wave vector randomly selected based upon a thermal distribution. This is also an unphysical model.
- Contact regions are considered to be in thermal equilibrium. The total number of electrons in a small region near the contact are kept constant, with the number of electrons equal to the number of dopant ions in the region. This is a very good model most commonly employed in actual device simulations.
- Another method uses 'reservoirs' of electrons adjacent to the contacts. Electrons naturally diffuse into the contacts from the reservoirs, which are not treated as part of the device during the solution of Poisson's equation. This approach gives results similar to the velocity-weighted Maxwellian [89], but at the expense of increased computational time due to the extra electrons simulated. It is an excellent model employed in few most sophisticated particle-based simulators.

There are also several possibilities for the choice of the distribution function: Maxwellian, displaced Maxwellian, and velocity-weighted Maxwellian [91].

To simulate the steady-state behavior of a device, the system is started in some initial condition, with the desired potentials applied to the contacts, and then the simulation proceeds in a time stepping manner until steady-state is reached. This typically takes several picoseconds of simulation time, and consequently several thousand time steps based on the usual time increments required for stability. Fig. 10 shows the particle distribution in a three-dimensional metal semiconductor field effect transistor (MESFET) structure, where the dots indicate the individual simulated particles [92]. In these simulations, the charge-neutral method, discussed earlier, is used.

After sufficient time has elapsed, so that the system is driven into a steady-state regime, one can calculate the steady-state current through a specified terminal. The device current can be determined

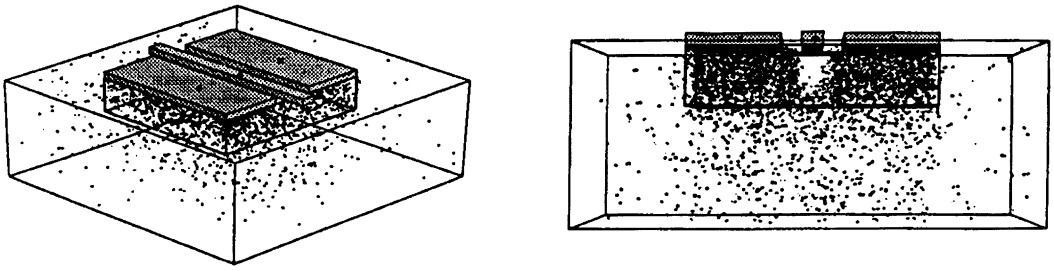


Fig. 10. Example of the particle distribution in a MESFET structure simulated in three-dimensional using an EMC approach.

via two different, but consistent methods. First, by keeping track of the charges entering and exiting each terminal, the net number of charges over a period of the simulation can be used to calculate the terminal current. The method is quite noisy due to the discrete nature of the carriers. In a second method, the sum of the carrier velocities in a portion of the device are used to calculate the current. For this purpose, the device is divided into several sections along, for example, the x-axis (from source to drain for the case of a MOSFET or MESFET simulation). The number of carriers and their corresponding velocity is added for each section after each free flight time step. The total x-velocity in each section is then averaged over several time steps to determine the current for that section. The total device current can be determined from the average of several sections, which gives a much smoother result compared to counting the terminal charges. By breaking the device into sections, individual section currents can be compared to verify that the currents are uniform. In addition, sections near the source and drain regions of a MOSFET or a MESFET may have a high y-component in their velocity and should be excluded from the current calculations. Finally, by using several sections in the channel, the average energy and velocity of electrons along the channel is checked to ensure proper physical characteristics.

As in the case of solving the full Maxwell's equations, for a stable MC device simulation, one has to choose the appropriate time step, Δt, and the spatial mesh size (Δx, Δy, and/or Δz). The time step and the mesh size may correlate to each other in connection with the numerical stability. For example, the time step Δt must be related to the plasma frequency

$$\omega_{\mathrm{p}} = \sqrt{\frac{e^2 n}{\varepsilon_{\mathrm{s}} m^*}}, \tag{92}$$

where n is the carrier density. From the viewpoint of stability criterion, Δt must be much smaller than the inverse plasma frequency. The highest carrier density specified in the device model is used to estimate Δt. If the material is a multi-valley semiconductor, the smallest effective mass to be experienced by the carriers must be used in Eq. (92) as well. In the case of GaAs, with the doping of 5×10^{17} cm^{-3}, $\omega_{\mathrm{p}} \cong 5 \times 10^{13}$; hence, Δt must be smaller than 0.02 ps.

The mesh size for the spatial resolution of the potential is dictated by the charge variations. Hence, one has to choose the mesh size to be smaller than the smallest wavelength of the charge variations. The smallest wavelength is approximately equal to the Debye length (for degenerate semiconductors the relevant length is the Thomas–Fermi wavelength), given as

$$\lambda_{\mathrm{D}} = \sqrt{\frac{\varepsilon_{\mathrm{s}} k_{\mathrm{B}} T}{e^2 n}}. \tag{93}$$

The highest carrier density specified in the model should be used to estimate λ_D from the stability criterion. The mesh size must be chosen to be smaller than the value given by Eq. (93). In the case of GaAs, with the doping density of 5×10^{17} cm^{-3}, $\lambda_D \cong 6$ nm.

Based on this discussion, the time step (Δt), and the mesh size (Δx, Δy, and/or Δz) are specified independently based on physical arguments. However, there are numerical constraints as well. This means that Δt chosen must be checked again by calculating the distance l_{max}, defined as

$$l_{max} = \boldsymbol{v}_{max} \Delta t, \tag{94}$$

where $\boldsymbol{v}_{max}$ is the maximum carrier velocity that can be approximated by the maximum group velocity of the electrons in the semiconductor (on the order of 10^8 cm/s). The distance l_{max} is the maximum distance the carriers can propagate during Δt. The time step is, therefore, chosen to be small enough so that l_{max} is smaller than the spatial mesh size chosen using Eq. (93). This constraint is too large because of a time step (Δt) may result in substantial change in the charge distribution, while the field distribution in the simulation is only updated every Δt, leading to unacceptable errors in the carrier force.

5.1.1. Particle–mesh (PM) coupling

The charge assignment and force interpolation schemes usually employed in self-consistent MC device simulations are the nearest-grid-point (NGP) and the cloud-in-cell (CIC) schemes [93]. In the NGP scheme, the particle position is mapped into the charge density at the closest grid point to a given particle. This has the advantage of simplicity, but leads to a noisy charge distribution, which may exacerbate numerical instability. Alternatively, within the CIC scheme a finite volume is associated with each particle spanning several cells in the mesh, and a fractional portion of the charge per particle is assigned to grid points according to the relative volume of the 'cloud' occupying the cell corresponding to the grid point. This method has the advantage of smoothing the charge distribution due to the discrete charges of the particle-based method, but may result in an artificial 'self-force' acting on the particle, particularly if an inhomogeneous mesh is used.

To better understand the NGP and the CIC scheme, consider a tensor product mesh with mesh lines $x_i, i = 1, \ldots, N_x$ and $y_j, j = 1, \ldots, N_y$. If the mesh is uniformly spaced in each axis direction, then $x_{l+1} - x_l = x_{l+2} - x_{l+1}$. The permittivities are considered constant within each mesh element and are denoted by $\varepsilon_{kl}, k = 1, \ldots, N_x - 1$ and $l = 1, \ldots, N_y - 1$. Define centered finite-differences of the potential ψ in the x- and y-axis at the midpoints of element edges as follows

$$\begin{cases} \Delta^x_{k+(1/2),l} = -\dfrac{\psi_{k+1,l} - \psi_{k,l}}{x_{k+1} - x_k} \\ \Delta^y_{k,l+(1/2)} = -\dfrac{\psi_{k,l+1} - \psi_{k,l}}{y_{l+1} - y_l} \end{cases}, \tag{95}$$

where the minus sign is included for convenience because the electric field is negative of the gradient of the potential. Consider now, a point charge in two-dimensional located at (x, y) within an element $\langle i, j \rangle$. If the restrictions for the permittivity (P) and the tensor-product meshes with uniform spacing in each direction (M) apply, the standard NGP/CIC schemes in two dimensions can be summarized by the following four steps.

1. *Charge assignment to the mesh*: the portion of the charge ρ_L assigned to the element nodes (k, l) is $w_{kl}\rho_L$, $k = i, i + 1$ and $l = j, j + 1$, where w_{kl} are the four charge weights which sum to unity by charge conservation. For the NGP scheme, the node closest to (x, y) receives a weight $w_{kl} = 1$,

with the remaining three weights set to zero. For the CIC scheme, the weights are $w_{ij} = w_x w_y$, $w_{i+1,j} = (1 - w_x)w_y$, $w_{i,j+1} = w_x(1 - w_y)$, and $w_{i+1,j+1} = (1 - w_x)(1 - w_y)$, $w_x = (x_{i+1} - x)/(x_{i+1} - x_i)$ and $w_y = (y_{j+1} - y)/(y_{j+1} - y_j)$.

2. *Solve the Poisson equation*: the Poisson equation is solved by some of the numerical techniques discussed in Section 4.2.
3. *Compute forces on the mesh*: the electric field at mesh nodes (k, l) is computed as: $E^x_{kl} = (\Delta^x_{k-1/2,l} + \Delta^x_{k+1/2,l})/2$ and $E^y_{kl} = (\Delta^y_{k,l-1/2} + \Delta^y_{k,l+1/2})/2$, for $k = i, i+1$ and $l = j, j+1$.
4. *Interpolate to find forces on the charge*: interpolate the field to position (x, y) according to $E^x = \sum_{kl} w_{kl} E^x_{kl}$ and $E^y = \sum_{kl} w_{kl} E^y_{kl}$, where $k = i, i+1$, $l = j, j+1$ and the w_{ij} are the NGP or CIC weights from step 1.

The requirements (*P*) and (*M*) severely limit the scope of devices that may be considered in device simulations using the NGP and the CIC schemes. Laux [94] proposed a new PM coupling scheme, namely, the nearest-element-center (NEC) scheme, which relaxes the restrictions (*P*) and (*M*). The NEC charge assignment/force interpolation scheme attempts to reduce the self-forces and increase the spatial accuracy in the presence of non-uniformly spaced tensor-product meshes and/or spatially-dependent permittivity. In addition, the NEC scheme can be utilized in one axis direction (where local mesh spacing is non-uniform) and the CIC scheme can be utilized in the other (where local mesh spacing is uniform). Such hybrid schemes offer smoother assignment/interpolation on the mesh compared to the pure NEC. The new steps of the pure NEC PM scheme are the following:

1. *Charge assignment to the mesh*: divide the line charge ρ_L equally to the four mesh points of the element $\langle i, j \rangle$.
2. *Compute forces on the mesh*: calculate the fields $\Delta^x_{i+1/2,l}$, $l = j, j+1$, and $\Delta^y_{k,j+1/2}$, $k = i, i+1$.
3. *Interpolate to find force on the charge*: interpolate the field according to the following $E^x = (\Delta^x_{i+1/2,j} + \Delta^x_{i+1/2,j+1})/2$ and $E^y = (\Delta^x_{i,j+1/2} + \Delta^x_{i+1,j+1/2})/2$.

The NEC designation derives from the appearance, in step (1) of moving the charge to the center of its element and applying a CIC-like assignment scheme. The NEC scheme involves only one mesh element and its four nodal values of potential. This locality makes the method well-suited to non-uniform mesh spacing and spatially-varying permittivity. The interpolation and error properties of the NEC scheme are similar to the NGP scheme.

5.1.2. The short-range force

In modern deep-submicrometer devices, for achieving optimum device performance and eliminating the so-called punch-through effect, the doping densities must be quite high. This necessitates a careful treatment of the electron–electron (e–e) and electron–impurity (e–i) interactions, an issue that has been a major problem for quite some time. Many the approaches used in the past have included the short-range portions of the e–e and e–i interactions in the $\boldsymbol{k}$-space portion of the MC transport kernel, as discussed in Section 3.1.5, thus neglecting the important inelastic properties of these two interaction terms [95,96]. An additional problem with this screened scattering approach is that, unlike the other scattering processes, e–e and e–i scattering rates need to be re-evaluated frequently during the simulation process to take into account the changes in the distribution function and the screening length. The calculation of the distribution function is highly CPU intensive, and it cannot account for local variations of the electron density in real space. Furthermore, the ionized impurity scattering is usually treated as a simple two-body event, thus ignoring the multi-ion contributions to the overall scattering potential. A simple screening model is usually used that ignores the dynamical perturbations to the Coulomb fields caused by the movement

of the free carriers. To overcome these difficulties, several authors have advocated the use of the coupled ensemble MC molecular dynamics approach [97,98,99], that gives simulation mobility results in excellent agreement with the experimental data for high substrate doping levels [99]. However, it has proven to be quite difficult to incorporate this coupled ensemble MC molecular dynamics approach when inhomogeneous charge densities, characteristic of semiconductor devices, are encountered [96,100]. An additional problem with this approach in a typical particle-based device simulation arises from the fact that both the e–e and e–i interactions are already included, at least within the Hartree approximation (long-range carrier–carrier interaction), through the self-consistent solution of the three-dimensional Poisson equation via the PM coupling discussed in Section 5.1.1. The magnitude of the resulting so-called mesh force, that arises from the force interpolation scheme, depends upon the volume of the cell and, for commonly employed mesh sizes in device simulations, usually leads to double-counting of the force.

To overcome the described difficulties of incorporation of the short-range e–e and e–i force into the problem, one can follow two different paths. One way is to use the particle–particle–particle–mesh (P^3M) scheme introduced by Hockney and Eastwood [93]. An alternative to this scheme is to use the corrected-Coulomb approach due to Gross et al. [101,102,104] and Vasileska et al. [103].

5.1.2.1. The P^3M method. The P^3M algorithms are a class of hybrid algorithms developed by Hockney and Eastwood [93]. These algorithms enable correlated systems with long-range forces to be simulated for a large ensemble of particles. The essence of the method is to express the inter-particle forces as a sum of two component parts: the short-range part $\boldsymbol{F}_{\rm sr}$, which is non-zero only for particle separations less than some cutoff radius $\boldsymbol{r}_{\rm e}$, and the smoothly varying part $\boldsymbol{F}$, which has a transform that is approximately band-limited. The total short-range force on a particle $\boldsymbol{F}_{\rm sr}$ is computed by direct particle–particle (PP) pair force summation, and the smoothly varying part is approximated by the PM force calculation.

Two meshes are employed in the P^3M algorithms: the charge–potential mesh and a coarser mesh, the so-called chaining mesh. The charge–potential mesh is used at different stages of the PM calculation to store, in turn, charge density values, charge harmonics, potential harmonics and potential values. The chaining mesh is a regular array of cells whose sides have lengths greater than or equal to the cutoff radius $\boldsymbol{r}_{\rm e}$ of the short-range force. Associated with each cell of this mesh is an entry in the head-of-chain array: This addressing array is used in conjunction with an extra particle coordinate, the linked-list coordinate, to locate pairs of neighboring particles in the short-range calculation.

The particle orbits are integrated forward in time using the leapfrog scheme

$$\boldsymbol{x}_i^{n+1} = \boldsymbol{x}_i^n + \frac{\boldsymbol{p}^{n+1/2}}{m}\Delta t, \tag{96}$$

$$\boldsymbol{p}_i^{n+1/2} = \boldsymbol{p}_i^{n-1/2} + (\boldsymbol{F}_i + \boldsymbol{F}_i^{\rm sr})\Delta t. \tag{97}$$

The positions $\{\boldsymbol{x}_i\}$ are defined at integral time levels and momenta $\{\boldsymbol{p}_i\}$ are defined at half-integral time levels. Momenta $\{\boldsymbol{p}_i\}$ are used rather than velocities for reasons of computational economy.

To summarize, the change in momentum of particle i at each time step is determined by the total force on that particle. Thus, one is free to choose how to partition the total force between the short-range and the smoothly varying part. The reference force $\boldsymbol{F}$ is the inter-particle force that the mesh calculation represents. For reasons of optimization, the cutoff radius of $\boldsymbol{F}_{\rm sr}$ has to be as small as possible and, therefore, $\boldsymbol{F}$ to be equal to the total inter-particle force down to as small a particle separation as possible. However, this is not possible due to the limited memory storage and the required CPU time even in the state-of-the-art computers.

The harmonic content of the reference force is reduced by smoothing. A suitable form of reference force for a Coulombic long-range force is one which follows the point particle force law beyond the cutoff radius $\boldsymbol{r}_e$, and goes smoothly to zero within that radius. The smoother the decay of $\boldsymbol{F}(\boldsymbol{x})$ and the large $\boldsymbol{r}_e$ becomes, the more rapidly the harmonics $\boldsymbol{R}(\boldsymbol{k})$ decay with increasing $\boldsymbol{k}$. Such smoothing procedure is equivalent to ascribing a finite size to the charged particle. As a result, a straightforward method of including smoothing is to ascribe some simple density profile $S(\boldsymbol{x})$ to the reference inter-particle force. Examples of shapes, which are used in practice, and give comparable total force accuracy are the uniformly charged sphere, the sphere with uniformly decreasing density, of the form

$$S(\boldsymbol{r}) = \begin{cases} \dfrac{48}{\pi a^4}\left(\dfrac{a}{2} - r\right), & r < a/2 \\ 0, & \text{otherwise} \end{cases}, \tag{98}$$

and the Gaussian distribution of density. The second scheme gives marginally better accuracies in three-dimensional simulations. Note that the cutoff radius of the short-range force implied by Eq. (98) is a rather than r_e. In practice, one can make r significantly smaller than a, because continuity of derivatives at $r = a$ causes the reference force to closely follow the point particle force for radii somewhat less than a. It has been found empirically that a good measure of the lower bound of r_e is given by the cube root of the auto-correlation volume of the charge shapes, which for the case of uniformly decreasing density gives

$$r_e \geq \left(\frac{5\pi}{48}\right)^{1/3} a \approx 0.7a. \tag{99}$$

Once the reference inter-particle force $\boldsymbol{F}$ for the PM part of the calculation is chosen, the short-range part $\boldsymbol{F}_{sr}$ is found by subtracting $\boldsymbol{F}$ from the total inter-particle force, i.e.

$$\boldsymbol{F}_{sr} = \boldsymbol{F}_{tot} - \boldsymbol{F}. \tag{100}$$

5.1.2.2. The corrected Coulomb approach. This second approach is a purely numerical scheme that generates a corrected Coulomb force look-up table for the individual e–e and e–i interaction terms. To calculate the proper short-range force, one has to define a three-dimensional box with uniform mesh spacing in each direction. A single (fixed) electron is then placed at a known position within a three-dimensional domain, while a second (target) electron is swept along the 'device' in, for example, 0.2 nm increments so that it passes through the fixed electron. The three-dimensional box is usually made sufficiently large so that the boundary conditions do not influence the potential solution. The electron charges are assigned to the nodes using one of the charge-assignment schemes discussed previously [94]. A three-dimensional Poisson equation solver is then used to solve for the node or mesh potentials. At self-consistency, the force on the swept electron $\boldsymbol{F} = \boldsymbol{F}_{mesh}$ is interpolated from the mesh or node potential. In a separate experiment, the Coulomb force $\boldsymbol{F}_{tot} = \boldsymbol{F}_{coul}$ is calculated using standard Coulomb law. For each electron separation, one then tabulates $\boldsymbol{F}_{mesh}$, $\boldsymbol{F}_{coul}$ and the difference between the two $\boldsymbol{F}' = \boldsymbol{F}_{coul} - \boldsymbol{F}_{mesh} = \boldsymbol{F}_{sr}$, which is called the corrected Coulomb force or a short-range force. The later is stored in a separate look-up table.

As an example, the corresponding fields to these three forces for a simulation experiment with mesh spacing of 10 nm in each direction are shown in Fig. 11. It is clear that the mesh force and the Coulomb force are identical when the two electrons are separated several mesh points (30–50 nm) apart.

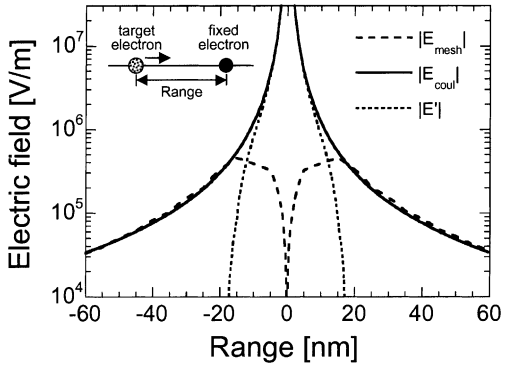


Fig. 11. Mesh, Coulomb and corrected Coulomb field vs. the distance between the two electrons. Note: $\boldsymbol{F} = -e\boldsymbol{E}$.

Therefore, adding the two forces in this region would result in double-counting of the force. Within three to five mesh points, $\boldsymbol{F}_{\text{mesh}}$ starts to deviate from $\boldsymbol{F}_{\text{coul}}$. When the electrons are within the same mesh cell, the mesh force approaches zero, due to the smoothing of the electron charge when divided amongst the nearest node points. The generated look-up table for $\boldsymbol{F}'$ also provides important information concerning the determination of the minimum cutoff range based upon the point where $\boldsymbol{F}_{\text{coul}}$ and $\boldsymbol{F}_{\text{mesh}}$ begin to intersect, i.e. $\boldsymbol{F}'$ goes to zero.

Fig. 12 shows the simulated doping dependence of the low-field mobility, derived from three-dimensional resistor simulations, which is a clear example demonstrating the importance of the proper inclusion of the short-range electron–ion interactions. For comparison, also shown in this figure are the simulated mobility results reported in [105], calculated with a bulk EMC technique using the Brooks approach [106] for the e–i interaction, and finally the measured data [107] for the case when the applied electric field is parallel to the $\langle 1\,0\,0 \rangle$ crystallographic direction. From the results shown, it is obvious that adding the corrected Coulomb force to the mesh force leads to mobility values that are in very good agreement with the experimental data. It is also important to note that, if only the mesh force is used in the free flight portion of the simulator, the simulation mobility data points are significantly higher than the experimental ones due to the omission of the short-range portion of the force.

The short-range e–e and e–i interactions also play a significant role in the operation of semiconductor devices. For example, carriers thermalization at the drain end of the MOSFET channel is significantly affected by the short-range e–e and e–i interactions. This is illustrated in

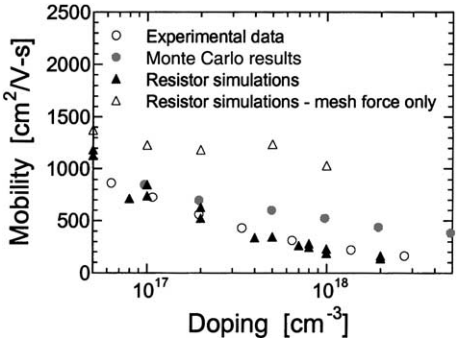


Fig. 12. Low-field electron mobility derived from three-dimensional resistor simulations vs. doping. Also shown on this figure are the ensemble MC results and the appropriate experimental data.

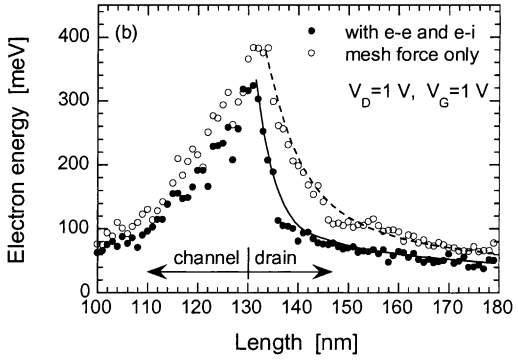


Fig. 13. Average energy of the electrons coming to the drain from the channel. Filled (open) circles correspond to the case when the short-range e–e and e–i interactions are included (omitted) in the simulations. The channel length extends from 50 to 130 nm.

Fig. 13 on the example of a 80 nm channel length *n*-MOSFET. Carrier thermalization occurs over distances that are on the order of few nanometers when the e–e and e–i interactions are included in the problem. Using the mesh force alone does not lead to complete thermalization of the carriers along the whole length of the drain extension, and this can lead to inaccuracies when estimating the device on-state current.

5.2. *Hydrodynamic/drift-diffusion device simulations*

As already discussed in Section 3.2, the balance equations are a set of coupled conservation laws in the form of differential equations that can be easily derived from the BTE. In fact, BTE can be fully represented by an infinite set of such conservation laws, starting with particle density conservation, followed by momentum conservation, energy conservation, etc. This set can be regarded as equivalent to a series expansion of the BTE. However, in order to be of any use, the expansion has to be truncated after a suitable number of terms. Hence, in practice, only a limited number of the most important conservation laws are retained, which may suffice for a satisfactory analysis of most devices. In fact, the much used drift-diffusion formalism, discussed at the end of Section 3.2, is based on the first two conservation laws, only the particle density and the momentum balance equations. But in order to describe important effects in modern-day devices related to non-stationary electron transport and heating of the carrier gas, the third conservation law—the energy balance equation—is also needed, which leads to what is known as the hydrodynamic formulation, described in details in Section 3.2. With this model, phenomena such as velocity overshoot and thermalization of energetic carriers via collisions are described. When the corresponding hydrodynamic or drift-diffusion equations are coupled with a field solver, one has a hydrodynamic or a drift-diffusion device simulator.

Commercial two-dimensional and even three-dimensional simulators based on either drift-diffusion or the hydrodynamic formalism, such as PISCES [108], MEDICI [109], MINIMOS [110], DESSIS from ISE [111], SILVACO [112], etc. are quite popular, especially for analyzing silicon devices, since the effects of electron heating are not as pronounced in silicon as in compound semiconductors. But these simulators are usually not accurate enough for a quantitative description of short-channel compound semiconductor devices. Still, the drift-diffusion formalism may be quite useful for numerically challenging tasks, such as three-dimensional FET modeling. Another example may be the analysis of the field distribution near the pinch-off for estimates of the breakdown voltage

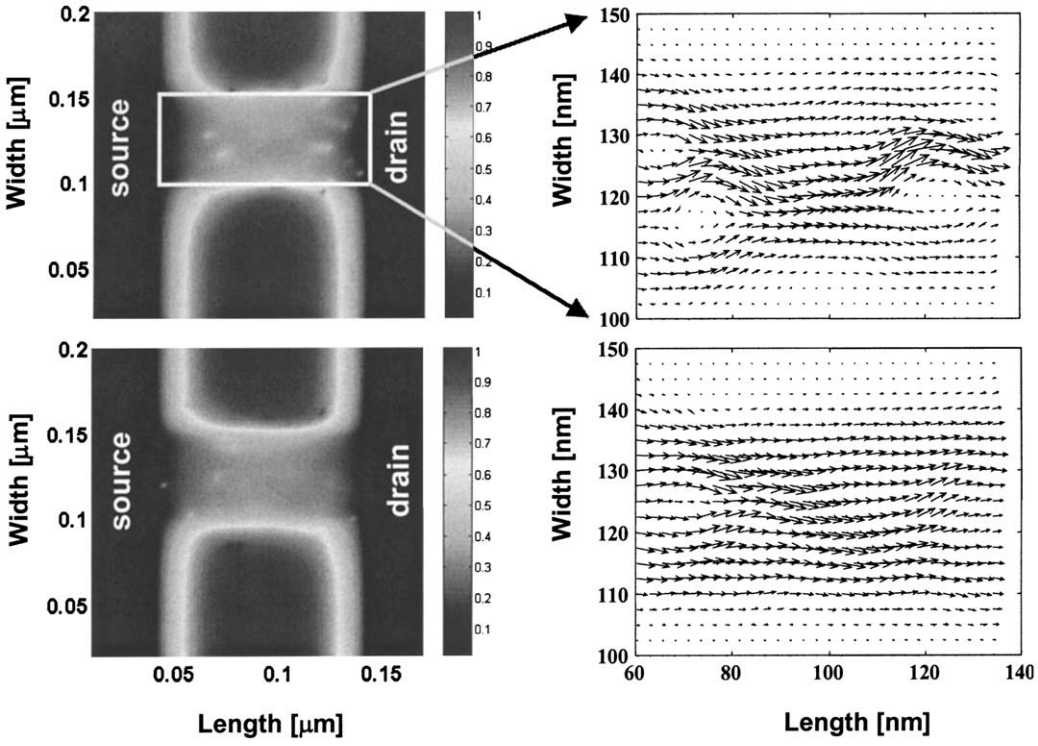


Fig. 14. Conduction band edge (left panel) and current stream lines in *n*-channel ultra-small MOSFET devices in which atomistic description of the impurity atoms is used in the device active region.

in MESFETs and HFETs. They also have been very successfully used in investigating fluctuations in the threshold voltage and the off-state power dissipation in nano-scale MOSFETs, in which there are very few impurity atoms in the device active region, and the position of each impurity will have significant influence on the actual device performance [113].

An example for the role of the atomistic description of the impurity atoms on the potential profile and current stream lines is given in Fig. 14 for a device with gate length equal to 0.1 μm and gate width equal to 0.05 μm. The effect of the randomly sited impurities can be seen on the potential plots on the left for two devices with different number and different impurity distribution. The potential fluctuations force the current to divert around the potential peak of the random impurity. This may be seen in the figures on the right where the current flow vectors avoid several regions, which represent the role of the impurities. Such non-uniform current flow leads to threshold voltage and off-state current fluctuations amongst devices fabricated on the same chip. The effect becomes more prominent as channel lengths are scaled into the nanometer range, thus, giving rise to device reliability concerns.

5.3. Electromagnetic device simulation

Simulation of optoelectronic and high frequency devices requires the solution of the full set of Maxwell's equations rather than just the Poisson's equation. We previously discussed the FDTD method for solving Maxwell's equations in Section 4.1. Numerous other solution techniques are of course available, such as finite element methods (FEMs), moment methods, frequency domain techniques, etc. [114]. For semiconductor lasers, drift-diffusion and hydrodynamic models have been

coupled to solutions of, for example, the Helmholtz wave equation in the optical cavity to simulate laser performance in MINILASE-II [115,116].

The modeling of optoelectronic devices such as semiconductor lasers and light-emitting diodes using particle-based device simulation is computationally difficult due to the characteristic time scale for spontaneous emission, which is on the order of nanoseconds. In contrast, the time step in a MC simulation is typically a femtosecond, which requires an enormous number of time steps just to characterize a few radiative transitions. MC is still used to calibrate the moment method models used in MINILASE. For example, Rota et al. [117] used MC simulation to investigate the capture process for electrons into a quantum well laser for calibrating the rate models used in MINILASE. However, the issue of dealing with vastly different time scale phenomena within a time-domain algorithm, such as the EMC method, is challenging.

For high frequency devices and circuits, the characteristic time scales of scattering events and the inverse frequency are somewhat commensurate, and FDTD methods, as discussed in Section 4.1, have been successfully applied. For time-domain techniques, such as the FDTD method, the same coupled algorithm for device simulation is used, as already discussed in this section, in which the field equations are decoupled from the transport equations over a small time interval, and the solution of one is used as the input for the other. For coupled MC/FDTD simulation, the carriers are accelerated by the full Lorentz force given by Eq. (75), while the source for Maxwell's equations is provide by the current density on the FDTD mesh, calculated by projecting the carrier velocities onto the nearest mesh point, as given by Eq. (84).

As an example of the application of the coupled MC/FDTD simulation method, a structure which has been utilized to study carrier dynamics under high field conditions is the biased co-planar strip configuration shown in Fig. 15 [118,119]. In this structure, an ultra-short optical pulse (switching beam) is used to excite electron-hole pairs between the co-planar strips. Due to the dc bias on the strips, the electrons and holes are excited in opposite directions giving rise to a time-dependent photocurrent induced in the waveguide structure. The pulse propagates down the waveguide, where it is detected by a time-delayed pulse (sampling beam). The change in electric field due to the propagating pulse is detected (for example) by the shift in an excitonic resonance due to the Stark effect [119]. Measurement of the time-dependent photocurrent detects the time-dependent velocity of the electrons and holes as they accelerate in the electric field from an initial state of essentially zero velocity.

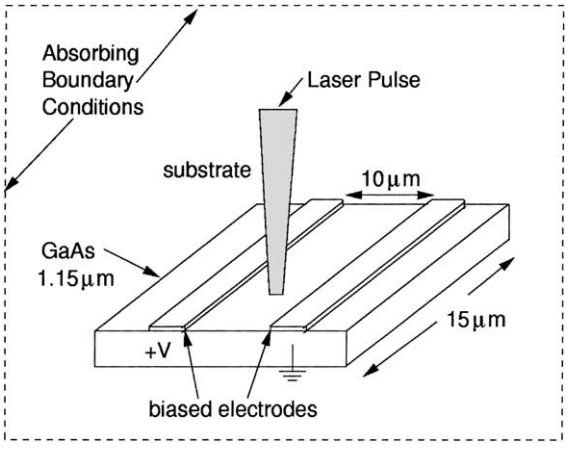


Fig. 15. Illustration of the experimental configuration for electro-optic sampling of a biased co-planar strip on a GaAs substrate.

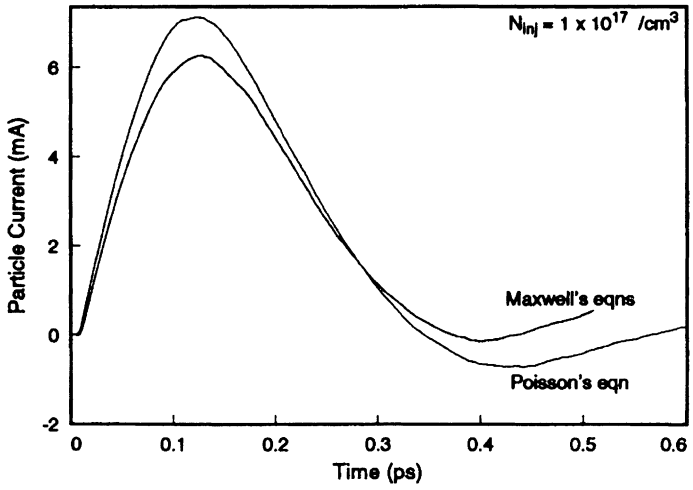


Fig. 16. Calculated particle current vs. time for the solution based on Poisson's equation only, and solutions considering FDTD solutions to Maxwell's equations for the structure shown in Fig. 15 and an average field of 40 kV/cm.

MC simulation has been employed in the interpretation of these results using FDTD solutions of Maxwell's equations, which are solved self-consistently with the particle dynamics [120,121]. In these simulations, Maxwell's equations for the electric and magnetic fields are discretized onto a three-dimensional Yee cell grid using the FDTD method described in Section 4.1, and solved at each time step using as a source term the current density calculated from the previous MC phase of the simulation during the previous time step. Using the solutions for $\boldsymbol{E}$ and $\boldsymbol{B}$ from this step, the particles then accelerate under the influence of the Lorentz force during the next MC phase. Assuming the time step is properly chosen, this method allows the evolution of the system to be modeled during and after photoexcitation by the switching beam. Fig. 16 shows a typical result of the calculated particle current induced by the switching beam for an average field of 40 kV/cm in a GaAs co-planar strip structure with an excited carrier population of 1×10^{17} cm^{-3} in a 1 μm spot diameter between the strip lines. The particle current shows an overshoot behavior, which is expected for the short-time dynamics of carriers accelerated in an electric field [32]. However, the decay of the current back to zero, and the undershoot is not expected from simple carrier dynamics in a constant field, and arises due to the self-consistent field of the electrons and holes themselves which collapses the dc field existing in the gap. Whether one uses full solutions to Maxwell's equations, or simply Poisson's equation (quasi-static solutions), the result is fairly similar, and this is clearly seen from the results shown in Fig. 16.

The time-dependent separation of electrons and holes after photoexcitation in the experiments resembles a Hertzian dipole, which has a characteristic frequency in the terahertz range (based upon the time scale shown in Fig. 16). The emission of terahertz radiation in such structures has long been recognized and has potential applications in sources and detectors in this frequency range [122]. Son et al. [123] have used the measured terahertz radiation in a similar experimental structure to Fig. 15 in order to study the particle dynamics after photoexcitation. Modeling of such terahertz radiation using the coupled MC/FDTD simulation described earlier is accomplished by modeling a much larger domain than the co-planar strip structure of Fig. 15 to include the free space outside. Attention must be paid to the proper boundary conditions on the larger domain, which should be purely absorbing to first-order. Recent improvements in the FDTD method to approximate absorbing boundary conditions have been developed which result in very little reflected radiation [72].

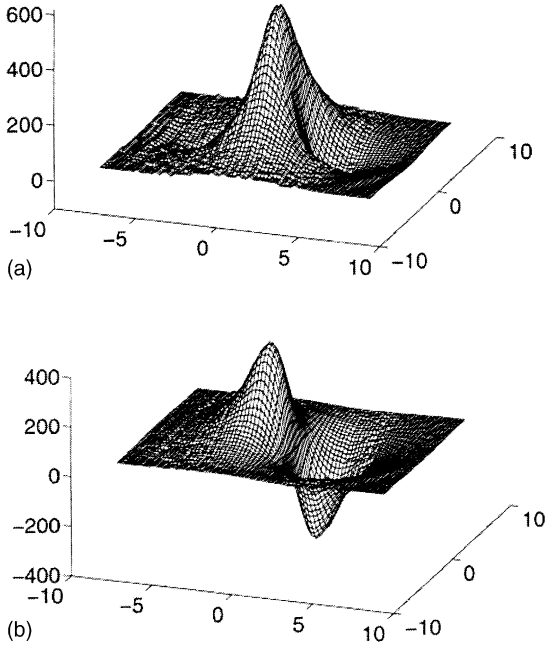


Fig. 17. Calculated near-field electric field outside of the microstrip. (a) The tangential component and (b) the radial component [124]. The x and y scales are in microns.

Fig. 17 illustrates the calculated FDTD results for the normal and tangential electric fields in the near-field regime at an observation point directly above the co-planar strips [124]. The results are shown for a time corresponding to the delay time for the electromagnetic pulse to arrive at the observation point. The results bear a close resemblance to the expected waveforms due to radiation from an ideal Hertzian dipole.

6. Quantum corrections to semi-classical approaches

In the past, quantum effects have been known to dominate the operation of resonant tunneling diodes [125], quantum cascade lasers [126], etc. As discussed in Section 1.1, tunneling through the gate oxide [127], source to drain tunneling and space-quantization effects are expected to be important in nano-scale MOSFETs and will require solution of the one-dimensional Schrödinger–Poisson problem. Solutions of the two-dimensional Schrödinger–Poisson problem are needed, for example, for describing the channel charge in narrow-width MOSFETs. With regard to gate oxide tunneling, the one-electron effective mass approximation may not be sufficiently accurate and ab initio calculations will most probably be needed [128,129].

It is also relevant to recall from the discussion given in the Section 1.1 that discrete impurity effects in nano-scale MOSFETs will lead to potential fluctuations which, in turn, will affect the magnitude of the device terminal characteristics (threshold voltage, off-state current, off-state power dissipation, etc.). In the nano-scale MOSFETs, these potential fluctuations will eventually lead to further electron confinement into small boxes containing only a few electrons as device dimensions scale even further. This implies that understanding transport in future ultra-small devices will also require understanding transport in single and coupled quantum dots. The quantum dots by themselves have been the focus of numerous studies (see [130]). For example, controllable loading

of these dots with few electrons has already been achieved, thus allowing one to speak of artificial quantum-dot hydrogen atoms and quantum-dot helium atoms [131]. Computing architectures for quantum devices, so-called quantum cellular automata, which consist of cells of coupled quantum dots occupied by only a few electrons, have also been proposed [132,133] and realized experimentally [134].

In these nano-scale MOSFETs, the time scale for the carrier transport is relatively short, as they leave the source with a memory of its distribution, traverse the channel under high fields, and enter the drain. Questions that arise in this context are, for example, what are the requirements for proper transport equations, and how can these be incorporated into existing simulators? Understanding transport in quantum dot structures is yet another challenging problem that needs further consideration. For example, one of the main difficulties in explaining transport in open quantum dots is the determination of the exact energy spectrum and how the dot states couple with the leads (the quasi-two-dimensional electron gas) via the quantum point contacts (QPCs). Fluctuations in the confining potential, due to the atomistic nature of the impurity atoms and how they affect the energy level spectrum in the dot, is yet another issue that prevents one in establishing a one to one correspondence between experiments and device simulations.

Due to the complexity of dealing with quantum transport at the lowest level of the hierarchy of Fig. 3 (Green's function method or direct solution of the n-body Schrödinger equation), and due to the desire to have device simulation tools which are able to deal with multiple levels of length scales and complexity, from the quantum regime to the classical regime, increasing interest is being focused at present on the use of quantum-mechanically derived potentials that may be added as 'corrections' to the semi-classical simulation tools. A way to include quantum effects into classical simulation tools is to add such quantum potentials to the mean field potential computed from Poisson's equation. In the past, such potential corrections have been employed mostly in the context of fluid approximations leading to the so-called QHD equations [135], where the corresponding equations are usually derived under the assumption that the electron gas is near thermal equilibrium. Even so, they were expected to be more generally valid and allow one to simulate quantum effects in ultra-small scale semiconductor and nano-electronic devices. More recently, these quantum corrections are introduced as modifications of the Hartree potential obtained from solving the Poisson's equation. Also note that this concept of multi-scale simulation, currently in the focus of the scientific research, is particularly critical in semiconductor devices where much of the device domain is in quasi-equilibrium and behaves classically (e.g. substrate, source-drain contacts, gate, etc.), whereas the critical regions governing the current are spatially small, subject to high fields and high degrees of potential confinement leading to quantum effects. It is, therefore, expected to be quite successful in overcoming some of the limitations of the semi-classical transport approaches discussed in Sections 3 and 5 of this review article. An in-depth description of the effective potential approach, utilized in particle-based simulations, is given in Section 6.1. In Section 6.2 we give brief description of the quantum hydrodynamic model and its use in device simulations on the example of a high electron mobility modulation-doped SiGe device structure.

6.1. The effective potential approach

From a circuit modeling point of view, even the one-dimensional solution of the Schrödinger–Poisson problem is an burdensome approach in terms of both complexity and computational cost. Because of this, it is common practice in industry to use analytical and macroscopic (in the sense of retaining the classical transport framework by adding correction terms to account for the quantum-mechanical effects) models that have provided some practical solutions. However, there are a

number of problems associated with these approaches and all of them are directly related to the non-stationary nature of carrier transport (velocity overshoot) in deep submicrometer devices. Hence, more sophisticated models are needed that are able to capture the appropriate transport physics of the processes occurring in the smallest device sizes.

A solution to this dilemma might be the use of quantum-mechanically derived potentials that are added as 'corrections' to semi-classical simulation tools. The idea of quantum potentials derives from the hydrodynamic formulation of quantum mechanics first introduced by de Broglie [136,137] and Madelung [138], and later developed by Bohm [139,140]. In this picture, the wave function is written in complex form in terms of its amplitude and phase, $\psi(r,t) = R(r,t)\exp[\mathrm{i}S(r,t)/\hbar]$, and when substituted back into the Schrödinger equation, it leads to coupled equations of motion for the density and phase, of the form

$$\frac{\partial \rho(r,t)}{\partial t} + \nabla\left(\rho\frac{1}{m}\nabla S\right) = 0, \tag{101}$$

$$-\frac{\partial S(r,t)}{\partial t} = \frac{1}{2m}(\nabla S)^2 + V(r,t) + Q(\rho,r,t), \tag{102}$$

where the probability density is $\rho(r,t) = R(r,t)^2$. With identification of the velocity as $\boldsymbol{v} = \nabla S/m$, and the flux as $\boldsymbol{j} = \rho\boldsymbol{v}$, Eq. (101) is the continuity equation. Hence, Eqs. (101) and (102), arising from this so-called Madelung transformation to the Schrödinger equation, have the form of classical hydrodynamic equations with the addition of an extra potential, often referred to as the quantum or Bohm potential, written as

$$Q = -\frac{\hbar^2}{2mR}\nabla^2 R \rightarrow -\frac{\hbar^2}{2m\sqrt{n}}\frac{\partial^2\sqrt{n}}{\partial x^2}, \tag{103}$$

where the square root of the density n, represents the magnitude of the wave function R. The Bohm potential essentially represents a field through which the particle interacts with itself. It has been used, for example, in the study of wave packet tunneling through barriers [141], where the effect of the quantum potential is seen to lower or smooth barriers, and hence allows particles to leak through.

An alternate form of the quantum potential has been derived by Iafrate et al. [142] and arises from taking moments of the Wigner–Boltzmann equation, the kinetic equation describing the time evolution of the Wigner distribution function [143]. This later quantum potential takes the form

$$V_{\mathrm{Q}} = -\frac{\hbar^2}{8m}\frac{\partial^2(\ln n)}{\partial x^2}, \tag{104}$$

and is sometimes referred to as the Wigner potential, or as the density gradient correction. Zhou and Ferry [144] derived a form for a smooth quantum potential based on the effective classical partition function of Feynman and Kleinert [145]. More recently, Gardner and Ringhofer derived a smooth quantum potential for hydrodynamic modeling valid to all orders of $\hbar^2$, which involves a smoothing integration of the classical potential over space and temperature [146].

A standard way to include quantum effects into classical simulation tools is to add such quantum potentials to the mean-field potential computed from solving the Poisson's equation. Such potential corrections have been employed mostly in the context of fluid approximations leading to the so-called QHD equations [147]. The Bohm potential of Eq. (103) has also been used in quantum particle-based simulations for quantum molecular dynamics calculations in quantum chemistry [148].

Again, the result differs from the classical molecular dynamics picture only by an additional quantum force term arising from the Bohm potential. These potential terms may lead to problems for particle simulators since they involve higher derivatives of the electron density, which are almost impossible to compute from particle-based methods due to statistical fluctuations. Moreover, for thermalized systems, their derivation relies on small potential variations and, therefore, these approximations are not valid close to the barriers where they are actually needed most.

Analogous to the smoothed potential representations discussed for the QHD model, it is desirable to define a smooth quantum potential for use in quantum particle-based simulation. Ferry [149] suggested an 'effective potential' that is derived from a wave packet description of particle motion, where the extent of the wave packet is defined from the range of wave vectors established by the thermalized distribution function (characterized by an electron temperature). The effective potential seen by electrons is given by

$$V_{\text{eff}}(x) = \frac{1}{\sqrt{2\pi}a_0}\int_{-\infty}^{\infty} V(x')\exp\left(-\frac{(x-x')^2}{2a_0^2}\right)\mathrm{d}x', \tag{105}$$

where $V(x')$ is the actual potential, and a_0 is the spatial spread of the wave packet. The effective potential accounts for the 'size of the electron' and its associated wavepacket, which feels the presence of barriers, etc. at a distance. From this Ansatz, the actual particle is treated as point-like in the presence of the effective potential associated with its wave-like nature, leading back to a classical particle simulation scheme. Representative simulation results for assymmetric MOSFET structures (focused ion beam MOSFET (FIBMOS)), which utilize this effective potential approach, are shown in Fig. 18 [150,151]. The inclusion of quantum-mechanical space-quantization effects leads to threshold voltage shift of about 150 mV and drain current reduction between 30 and 40%, depending upon the gate bias. These results are in agreement with experimental findings, thus demonstrating the applicability of such scheme in accurately representing the quantum-mechanical effects in the device channel region.

6.2. Quantum hydrodynamic model

The equations, which explicitly include quantum corrections and describe the particle conservation, momentum conservation and energy conservation, discussed in detail in [152], are

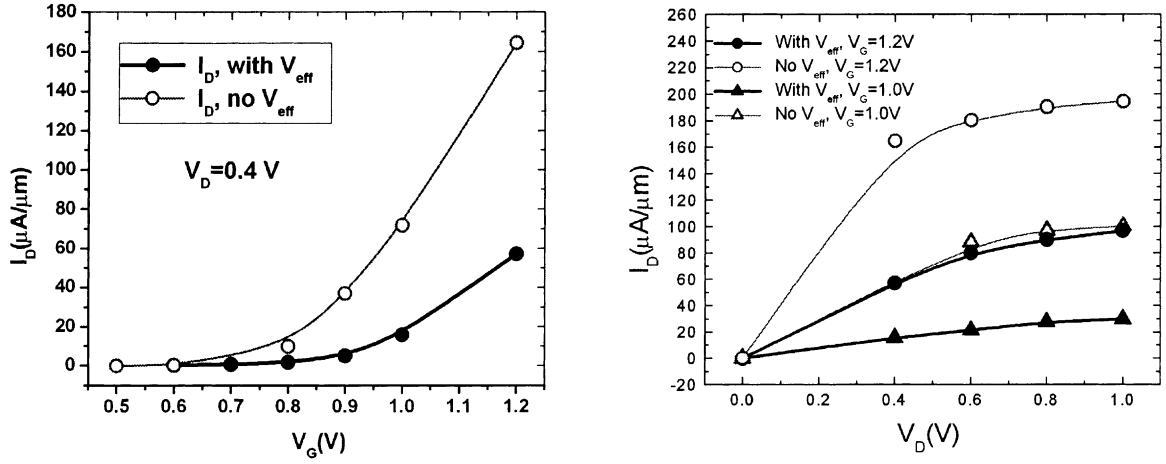


Fig. 18. Left panel: transfer characteristics of a FIBMOS device. Right panel: device output characteristics.

the following

$$\frac{\partial n}{\partial t} + \nabla \cdot (n\boldsymbol{v}) = 0, \tag{106}$$

$$\frac{\partial v}{\partial t} + \boldsymbol{v} \cdot \nabla v = -\frac{q\boldsymbol{E}}{m^*} - \frac{1}{nm^*}\nabla(nk_{\mathrm{B}}T_{\mathrm{q}}) - \frac{\boldsymbol{v}}{\tau_{\mathrm{m}}}, \tag{107}$$

$$\frac{\partial T}{\partial t} + \frac{1}{3\gamma}\boldsymbol{v} \cdot \nabla T_{\mathrm{q}} = -\frac{2}{3\gamma}\nabla \cdot (vT_{\mathrm{q}}) + \frac{m^* v^2}{3\gamma k_{\mathrm{B}}}\left(\frac{2}{\tau_{\mathrm{m}}} - \frac{1}{\tau_{\mathrm{w}}}\right) - \frac{T - T_0}{\tau_{\mathrm{w}}}, \tag{108}$$

where n is the average electron density, v is the average electron velocity, T is the effective electron temperature, m^* is the effective electron mass, $\boldsymbol{E}$ is the electric field, τ_{m} is the momentum relaxation time, τ_{w} is the energy relaxation time, and T_{q} is given by

$$T_{\mathrm{q}} = \gamma T + \frac{2}{3k_{\mathrm{B}}}U_{\mathrm{q}}, \tag{109}$$

with

$$U_{\mathrm{q}} = -\frac{\hbar^2}{8m^*}\nabla^2 \ln n, \tag{110}$$

where U_{q} is the quantum correction. The explicit quantum correction, as already discussed in Section 6.1, involves the second-order space derivative of the log of the density. Hence, it tends to smoothen the electron distribution, especially where the electron density has sharp changes. The factor, γ, is the degeneracy factor [153], given by

$$\gamma = \frac{F_{3/2}(\mu_{\mathrm{f}}/k_{\mathrm{B}}T)}{F_{1/2}(\mu_{\mathrm{f}}/k_{\mathrm{B}}T)}, \tag{111}$$

where μ_{f} is the Fermi energy measured from the conduction band edge, and is introduced as a correction to the total average electron kinetic energy

$$w = \tfrac{1}{2}m^* v^2 + \tfrac{3}{2}\gamma k_{\mathrm{B}}T + U_{\mathrm{q}}. \tag{112}$$

The relaxation times τ_{m} and τ_{w} are functions of energy, and, as discussed in Section 3.2, are determined by fitting the homogeneous hydrodynamic equations to the velocity-field and energy-field relations from MC simulations described in Section 3.1.

This quantum hydrodynamic model has been successfully applied to a variety of device structures realized in different material systems. As an example, the investigation of transport in a 0.18 μm gate length, modulation-doped structure, shown schematically in the right panel of Fig. 19, is discussed here. The doping of the top $Si_{0.7}Ge_{0.3}$ layer is 3.5×10^{18} cm^{-3}, and a doping of 1×10^{14} cm^{-3} is used in the $Si_{0.7}Ge_{0.3}$ substrate. The lattice temperature in the simulation is taken to be 300 K. The typical simulation domain is 1 μm × 0.09 μm. The thickness of the top $Si_{0.7}Ge_{0.3}$ layer is 19 nm, and the strained-Si channel is 18 nm thick.

The simulated I–V characteristics for gate biases 0.7, 0.5, 0.2 and 0 V, respectively, are shown on the right panel of Fig. 19. The small thickness of the top $Si_{0.7}Ge_{0.3}$ layer provides a normally off device, since the Schottky barrier height of 0.9 eV leads to an estimated depletion width of 18.4 nm.

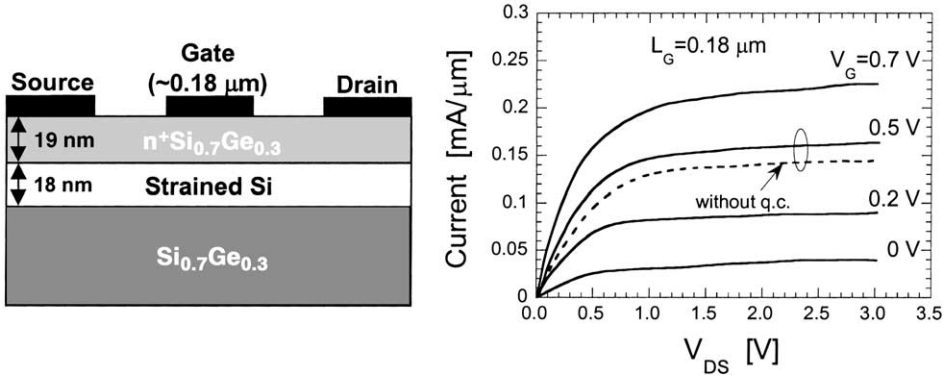


Fig. 19. Left panel: schematic description of the device structure under investigation. Right panel: simulated *I–V* characteristics.

The peak transconductance is about 300 mS/mm, and a good saturation with a drain conductance of 4.6 mS/mm is obtained for 0.5 V on the gate. Approximately the same current level and transconductance is found in a 0.25 μm device. These simulation results are comparable with corresponding experimental measurements. The relatively larger current level (0.3 μA/mm) and transconductance (330 mS/mm) found in the experiment is thought to be due to a higher sheet-charge density (2.5×10^{12} cm^{-2} compared with 1×10^{12} cm^{-2} in this simulation) in the quantum well for their particular modulation-doped structure. It is interesting to note that the transconductance of this device approaches the same order of magnitude as that of the AlGaAs/GaAs device with the same geometry, although the transconductance of the SiGe device is about three times smaller. The inclusion of quantum corrections leads to about 15% current increase for gate voltage of 0.5 V. By inspecting the electron density distribution along the channel region of the device (not shown here), one can see that this is due to the rapid change in the electron density at the gate end close to the drain contact within a region that is much shorter than the gate length. The inclusion of the quantum potential also leads to increase of the electron density in the channel.

7. Summary

To summarize this review, we have attempted to overview the field of computational electronics with particular emphasis on the main numerical methods used in semiconductor device simulation, and some examples taken primarily from our own research. The review is by no means inclusive of the extensive research in this field since the mid-1970s, which is continuing at an uninterrupted pace. An interested reader is referred to several books and review articles referenced in the present review.

In this review, we have tried to emphasize contemporary issues in simulation of leading edge semiconductor device technologies such as nanometer scale devices. The challenges of simulating increasingly smaller devices with more complicated geometries include the necessity of full three-dimensional modeling, inclusion of atomistic effects in terms of discrete dopant profiles and other device inhomogeneities, non-stationary/ballistic transport with proper treatment of both the long-range and the short-range particle–particle interactions, full electromagnetic simulation for very high frequency and optoelectronic devices, and the considerations of quantum-mechanical effects such as interference and tunneling. The inclusion of all these effects comes at the cost of vastly increased computational burden, as one would expect. Fortunately, there has been a concurrent improvement in

the performance in terms of speed and memory of the computational platforms available, based on the same technologies this field is attempting to ameliorate. The desktop computer system today has essentially the same computational performance of the world's fastest supercomputer of only a decade ago. Over the same time-frame, there has also been significant algorithmic speedup of many of the techniques discussed herein, which has greatly aided the development of this field. The prospect of having a fully 'first-principles' approach to semiconductor device modeling in which the input is ultimately the position of atoms themselves, seems less of a fantasy today with this continued exponential growth in computation.

The last point concerning atomic positions leads to one final observation concerning the field of computational electronics. Herein, we are concerned, almost exclusively, with techniques in the simulation of electronic structure and transport in semiconductor materials and devices, with relatively little mention of the equally (if not more) important field of semiconductor process simulation. As device technologies become increasingly smaller and more three-dimensional, our ability to know what we actually simulate is becoming increasingly uncertain. The number of process steps (as well as the cost of manufacturing) in a typical device technology is increasing along a similar 'Moore's law' growth curve as that of the device and circuit performance. Device and process simulation are becoming inexorably linked, and one of the great challenges in this field will be the ability to provide tools which accurately simulate the entire manufacture processes from the bare semiconductor to circuit level electronic performance based only on the specification of the process flow.

Acknowledgements

The authors would like to thank the numerous individuals whose work, comments or discussion contributed to this review. These include D.K. Ferry, S. El-Ghazaly, S. Gonzalez, W. Gross, S. Pennathur, K. Remely and M. Saraniti. The authors would also like to acknowledge support from the National Science Foundation DESCARTES Center, as well as individual NSF grants.

References

[1] P. Antognetti, G. Massobrio, Semiconductor Device Modeling with SPICE, McGraw-Hill, New York, 1988.
[2] 2001 International Technology Road Map of Semiconductors, http://public.itrs.net.
[3] R. Chau, J. Kavalieros, B. Roberds, A. Murthy, B. Doyle, D. Barlage, M. Doczy, R. Arghavani, http://www.intel.com/research/silicon/iedm.htm.
[4] D.K. Ferry, S.M. Goodnick, Transport in Nanostructures, Cambridge University Press, Cambridge, 1997.
[5] T.H. Ning, in: Proceedings of the IEEE 2000 Custom Circuits Conference, 2000, p. 49.
[6] M. Depas, B. Vermeire, P.W. Mertens, R.L. Van Meirhaegne, M.M. Heyns, Solid State Electron. 38 (1995) 14657.
[7] R. Lake, G. Klimeck, R.C. Bowen, D. Jovanovic, J. Appl. Phys. 81 (1997) 7845.
[8] G. Baccarani, M. Wordeman, Solid State Electron. 28 (1985) 407.
[9] S. Cordier, Math. Mod. Methods Appl. Sci. 4 (1994) 625.
[10] P.Y. Yu, M. Cardona, Fundamentals of Semiconductors, Springer, Berlin, 1999.
[11] C. Herring, Phys. Rev. 57 (1940) 1169.
[12] D.J. Chadi, M.L. Cohen, Phys. State Solids B 68 (1975) 405.
[13] J. Luttinger, W. Kohn, Phys. Rev. 97 (1955) 869.
[14] M.L. Cohen, T.K. Bergstresser, Phys. Rev. 141 (1966) 789.
[15] J.R. Chelikowsky, M.L. Cohen, Phys. Rev. B 14 (1976) 556.
[16] E. Fermi, Nuovo Cimento 11 (1934) 157.
[17] H.J. Hellman, J. Chem. Phys. 3 (1935) 61.
[18] J.C. Phillips, L. Kleinman, Phys. Rev. 116 (1959) 287.

[19] J.R. Chelikowsky, M.L. Cohen, Phys Rev. B 10 (1974) 12.
[20] L.R. Saravia, D. Brust, Phys. Rev. 176 (1968) 915.
[21] S. Gonzalez, Masters Thesis, Arizona State University, 2001.
[22] J.R. Chelikowsky, M.L. Cohen, Phys. Rev. B 10 (1974) 5059.
[23] K.C. Padney, J.C. Phillips, Phys. Rev. B 9 (1974) 1552.
[24] D. Brust, Phys. Rev. B 4 (1971) 3497.
[25] J.C. Slater, G.F. Coster, Phys. Rev. 94 (1954) 1498.
[26] P.-O. Löwdin, J. Chem. Phys. 19 (1951) 1396.
[27] E.O. Kane, J. Phys. Chem. Solids 1 (1956) 82.
[28] E.O. Kane, J. Phys. Chem. Solids 1 (1957) 249.
[29] G. Bastard, Wave Mechanics Applied to Semiconductor Heterostructures, Halsted Press, New York, 1988.
[30] D.L. Smith, C. Mailhiot, Phys. Rev. B 33 (1986) 8345.
[31] C. Jacoboni, L. Reggiani, Rev. Mod. Phys. 55 (1983) 645.
[32] C. Jacoboni, P. Lugli, The Monte Carlo Method for Semiconductor Device Simulation, Springer, Vienna, 1989.
[33] K. Hess, Monte Carlo Device Simulation: Full-Band and Beyond, Kluwer Academic Publishing, Boston, 1991.
[34] M.H. Kalos, P.A. Whitlock, Monte Carlo Methods, Wiley, New York, 1986.
[35] D.K. Ferry, Semiconductors, Macmillan, New York, 1991.
[36] H.D. Rees, J. Phys. Chem. Solids 30 (1969) 643.
[37] R.M. Yorston, J. Comp. Phys. 64 (1986) 177.
[38] L.I. Schiff, Quantum Mechanics, McGraw-Hill, New York, 1955.
[39] Y.-C. Chang, D.Z.-Y. Ting, J.Y. Tang, K. Hess, Appl. Phys. Lett. 42 (1983) 76.
[40] L. Reggiani, P. Lugli, A.P. Jauho, Phys. Rev. B 36 (1987) 6602.
[41] D.K. Ferry, A.M. Kriman, H. Hida, S. Yamaguchi, Phys. Rev. Lett. 67 (1991) 633.
[42] P. Bordone, D. Vasileska, D.K. Ferry, Phys. Rev. B 53 (1996) 3846.
[43] S. Bosi, C. Jacoboni, J. Phys. C 9 (1976) 315.
[44] P. Lugli, D.K. Ferry, IEEE Trans. Electron. Devices 32 (1985) 2431.
[45] N. Takenaka, M. Inoue, Y. Inuishi, J. Phys. Soc. Jpn. 47 (1979) 861.
[46] S.M. Goodnick, P. Lugli, Phys. Rev. B 37 (1988) 2578.
[47] M. Moško, A. Mošková, V. Cambel, Phys. Rev. B 51 (1995) 16860.
[48] L. Rota, F. Rossi, S.M. Goodnick, P. Lugli, E. Molinari, W. Porod, Phys. Rev. B 47 (1993) 1632.
[49] R. Brunetti, C. Jacoboni, A. Matulionis, V. Dienys, Physica B & C 134 (1985) 369.
[50] P. Lugli, D.K. Ferry, Phys. Rev. Lett. 56 (1986) 1295.
[51] J.F. Young, P.J. Kelly, Phys. Rev. B 47 (1993) 6316.
[52] R.W. Hockney, J.W. Eastwood, Computer Simulation Using Particles, Institute of Physics Publishing, Bristol, 1988.
[53] D.J. Adams, G.S. Dubey, J. Comp. Phys. 72 (1987) 156.
[54] Z.H. Levine, S.G. Louie, Phys. Rev. B 25 (1982) 6310.
[55] L.V. Keldysh, Zh. Eksp. Teor. Fiz. 37 (1959) 713.
[56] N. Sano, A. Yoshii, Phys. Rev. B 45 (1992) 4171.
[57] M. Stobbe, R. Redmer, W. Schattke, Phys. Rev. B 47 (1994) 4494.
[58] Y. Wang, K. Brennan, J. Appl. Phys. 71 (1992) 2736.
[59] M. Reigrotzki, R. Redmer, N. Fitzer, S.M. Goodnick, M. Dür, W. Schattke, J. Appl. Phys. 86 (1999) 4458.
[60] M.V. Fischetti, S.E. Laux, Phys. Rev. B 38 (1988) 9721.
[61] For a complete overview, see www.research.ibm.com/DAMOCLES.
[62] M. Saraniti, S.M. Goodnick, IEEE Trans. Electron. Devices 47 (2000) 1909.
[63] C.M. Snowden, Introduction to Semiconductor Device Modeling, World Scientific, Singapore, 1986.
[64] K. Bløtekjær, IEEE Trans. Electron. Devices 17 (1970) 38.
[65] K. Tomizawa, Numerical Simulation of Submicron Semiconductor Devices, Artech House, Boston, 1993.
[66] R.O. Grondin, S.M. El-Ghazaly, S.M. Goodnick, IEEE Trans. MTT 47 (1999) 817.
[67] M. Grupen, K. Hess, IEEE J. Quantum Electron. 34 (1998) 120.
[68] See www.ansoft.com/products/hf/hfss for details.
[69] www.remcom.com.
[70] www.reland.com.
[71] See www.ise.ch/news/release_7.html.
[72] A. Taflove, Computational Electrodynamics: The Finite-Difference Time-Domain Method, Artech House, Boston, 1995.
[73] Y.S. Yee, IEEE Trans. Antenn. Propagat. 14 (1966) 302.
[74] W.L. Stutzman, G.A. Thiele, Antenna Theory and Design, Wiley, New York, 1998, p. 493.
[75] A. Taflove, M.E. Brodwin, IEEE Trans. Microw. Theory Tech. 23 (1975) 623.
[76] G. Mur, IEEE Trans. Electromagn. Comp. 23 (1981) 1073.
[77] T.G. Moore, F.G. Blaschak, A. Taflove, G.A. Kriegsmann, IEEE Trans. Antenn. Propagat. 36 (1988) 1797.
[78] J.-P. Berenger, J. Comp. Phys. 114 (1994) 185.

[79] S. Selberherr, Analysis and Simulation of Semiconductor Devices, Springer, New York, 1984.
[80] G. Dahlqüist, Å. Björck, Numerical Methods, Prentice-Hall, Englewood Cliffs, NJ, 1974.
[81] G.V. Gadiyak, M.S. Obrecht, in: Proceedings of the Second International Conference on Simulation of Semiconductor Devices and Processes, 1986, p. 147.
[82] H.L. Stone, SIAM J. Num. Anal. 5 (1968) 536.
[83] W. Hackbush, Multi-Grid Methods and Applications, Springer, Berlin, 1985.
[84] P. Sonneveld, SIAM J. Sci. Stat. Comput. 10 (1989) 36.
[85] Y. Saad, Iterative Methods for Sparse Linear Systems, PWS Publishing Company, Boston, 1996.
[86] H.A. Van der Vorst, SIAM J. Sci. Stat. Comput. 13 (1992) 631.
[87] H.A. Van der Vorst, SIAM J. Sci. Stat. Comput. 10 (1989) 1174.
[88] S.C. Eisenstat, SIAM J. Sci. Stat. Comput. 2 (1981) 1.
[89] T. Gonzalez, D. Pardo, Solid State Electron. 39 (1996) 555.
[90] P.A. Blakey, S.S. Cherensky, P. Sumer, Physics of Submicron Structures, Plenum Press, New York, 1984.
[91] T. Gonzalez, D. Pardo, Solid State Electron. 39 (1996) 555.
[92] S.S. Pennathur, S.M. Goodnick, Inst. Phys. Conf. Ser. 141 (1995) 793.
[93] R.W. Hockney, J.W. Eastwood, Computer Simulation Using Particles, Institute of Physics Publishing, Bristol, 1988.
[94] S.E. Laux, IEEE Trans. Comp. Aided Des. Int. Circ. Syst. 15 (1996) 1266.
[95] M.E. Kim, A. Das, S.D. Senturia, Phys. Rev. B 18 (1978) 6890.
[96] M.V. Fischetti, S.E. Laux, Phys. Rev. B 38 (1988) 9721.
[97] P. Lugli, D.K. Ferry, Phys. Rev. Lett. 56 (1986) 1295.
[98] A.M. Kriman, M.J. Kann, D.K. Ferry, R. Joshi, Phys. Rev. Lett. 65 (1990) 1619.
[99] R.P. Joshi, D.K. Ferry, Phys. Rev. B 43 (1991) 9734.
[100] M.V. Fischetti, S.E. Laux, J. Appl. Phys. 78 (1995) 1058.
[101] W.J. Gross, D. Vasileska, D.K. Ferry, IEEE Electron. Device Lett. 20 (1999) 463.
[102] W.J. Gross, D. Vasileska, D.K. Ferry, VLSI Des. 10 (2000) 437.
[103] D. Vasileska, W.J. Gross, D.K. Ferry, Superlattices Microstructures 27 (2000) 147.
[104] W.J. Gross, D. Vasileska, D.K. Ferry, IEEE Trans. Electron. Devices 47 (2000) 1831.
[105] K. Tomizawa, Numerical Simulation of Submicron Semiconductor Devices, Artech House, Norwood, 1993.
[106] H. Brooks, Phys. Rev. 83 (1951) 879.
[107] C. Canali, G. Ottaviani, A. Alberigi-Quaranta, J. Phys. Chem. Solids 32 (1971) 1707.
[108] S. Beebe, F. Rotella, Z. Sahul, D. Yergeau, G. McKenna, L. So, Z. Yu, K. Wu, E. Kan, J. McVittie, R. Dutton, in: Proceedings of the International Electron Devices Meeting, 1994, p. 213.
[109] MEDICI, Two-Dimensional Device Simulation Program, Version 1999.2, Avant! Corporation, Fremont, CA, 1999.
[110] S. Selberherr, A. Schütz, H. Pötzl, IEEE Trans. Electron. Devices 27 (1980) 1540.
[111] DESSIS-ISE, ISE TCAD Release 6.0, ISE Integrated Systems Engineering AG, Zürich, Switzerland, 1999.
[112] ATLAS User's Manual, 6th Edition, Silvaco International, Santa Clara, CA, 1998.
[113] A. Asenov, IEEE Trans. Electron. Devices 45 (1998) 2505.
[114] M.N.O. Sadiku, Numerical Techniques in Electromagnetics, CRC Press, Boca Raton, FL, 1992.
[115] M. Grupen, K. Hess, L. Rota, in: W.W. Chow, M.A. Osinski (Eds.), Physics and Simulation of Optoelectronic Devices III, Vol. 2399, SPIE, 1995, p. 292.
[116] M. Grupen, K. Hess, Appl. Phys. Lett. 65 (1994) 2454.
[117] L. Rota, M. Grupen, K. Hess. in: K. Hess, J.P. Leburton, U. Ravaioli (Eds.), Hot Carriers in Semiconductors, Plenum Press, New York, 1996, p. 563.
[118] K.E. Meyer, M. Pessot, G. Mourou, R.O. Grondin, S.N. Chamoun, Appl. Phys. Lett. 53 (1988) 2254.
[119] W.H. Knox, J.E. Henry, K.W. Goossen, K.D. Li, B. Tell, D.A.B. Miller, D.S. Chemla, A.C. Gossard, J. English, S. Schmitt-Rink, IEEE J. Quantum Electron. 25 (1989) 2586.
[120] S.M. El-Ghazaly, R.P. Joshi, R.O. Grondin, IEEE Trans. Microw. Theory Tech. 38 (1990) 629.
[121] S.M. Goodnick, S. Pennathur, U. Ranawake, P. Lenders, V. Tripathi, Int. J. Num. Model. 8 (1995) 205.
[122] P.R. Smith, D.H. Auston, M. Nuss, IEEE J. Quantum Electron. 24 (1988) 255.
[123] J. Son, W. Sha, T. Norris, J. Whitaker, G. Mourou, Appl. Phys. Lett. 63 (1993) 923.
[124] K.A. Remley, A. Weisshaar, V.K. Tripathi, S.M. Goodnick, VLSI Des. 8 (1998) 407.
[125] R. Tsu, L. Esaki, Appl. Phys. Lett. 22 (1973) 562.
[126] J. Faist, F. Capasso, D.L. Sivco, C. Sirtori, A.L. Hutchinson, A.Y. Cho, Science 264 (1994) 553.
[127] C.-J. Sheu, S.-L. Jang, Solid State Electron. 44 (2000) 1819.
[128] M. Städele, B.R. Tuttle, K. Hess, J. Appl. Phys. 89 (2001) 348.
[129] A.A. Demkov, X. Zhang, D.A. Brabold, Phys. Rev. B 64 (1–4) (2001) 125306.
[130] J.P. Bird, R. Akis, D.K. Ferry, Y. Aoyagi, T. Sugano, J. Phys. Condens. Matter 9 (1997) 5935.
[131] D. Pfannkuche, R.R. Gerhardts, Phys. Rev. B 44 (1991) 13132.
[132] C.S. Lent, P.D. Tougaw, W. Porod, G.H. Bernstein, Nanotechnology 4 (1993) 49.
[133] C.S. Lent, P.D. Tougaw, W. Porod, Appl. Phys. Lett. 62 (1993) 714.
[134] A.O. Orlov, I. Amlani, G.H. Bernstein, C.S. Lent, G.L. Snider, Science 277 (1997) 928.

[135] C.L. Gardner, SIAM J. Appl. Math. 54 (1994) 409.
[136] L. de Broglie, C.R. Acad. Sci. Paris 183 (1926) 447.
[137] L. de Broglie, C.R. Acad. Sci. Paris 184 (1927) 273.
[138] E. Madelung, Z. Phys. 40 (1926) 322.
[139] D. Bohm, Phys. Rev. 85 (1952) 166.
[140] D. Bohm, Phys. Rev. 85 (1952) 180.
[141] C. Dewdney, B.J. Hiley, Found. Phys. 12 (1982) 27.
[142] G.J. Iafrate, H.L. Grubin, D.K. Ferry, J. Physique 42 (1981) 307.
[143] E. Wigner, Phys. Rev. 40 (1932) 749.
[144] J.-R. Zhou, D.K. Ferry, IEEE Trans. Electron. Devices 39 (1992) 473.
[145] R.P. Feynman, H. Kleinert, Phys. Rev. A 34 (1986) 5080.
[146] C. Gardner, C. Ringhofer, Phys. Rev. E 53 (1996) 157.
[147] C.L. Gardner, SIAM J. Appl. Math. 54 (1994) 409.
[148] C. Lopreore, R. Wyatt, Phys. Rev. Lett. 82 (1999) 5190.
[149] D.K. Ferry, Superlattices Microstructures 27 (2000) 61.
[150] D. Vasileska, X. He, I. Knezevic, D.K. Schroder, J. Comput. Electron., in press.
[151] D. Vasileska, R. Akis, I. Knezevic, S.N. Milicic, A.S. Ahmed, D.K. Ferry, Microelectron. Eng., in press (Special issue).
[152] J.-R. Zhou, D.K. Ferry, IEEE Trans. Electron. Devices 39 (1992) 473.
[153] J.-R. Zhou, D.K. Ferry, IEEE Trans. Electron. Devices 40 (1993) 421.