A Review of Global Modeling of Charge Transport in Semiconductors and Full-Wave Electromagnetics

Robert O. Grondin, Senior Member, IEEE, Samir M. El-Ghazaly, Senior Member, IEEE, and Stephen Goodnick, Senior Member, IEEE

Abstract—Models of semiconductor devices that combine complex transport models with full-wave solutions of Maxwell’s equations are reviewed. Both hydrodynamic and ensemble Monte Carlo transport models are studied. The coupling of such transport models with electrodynamics is discussed and sample results are presented. The physical consistency of the models is considered and some open computational challenges are reviewed.

I. INTRODUCTION

C ONTEMPORARY high-performance electronics are based on technologies such as monolithic microwave integrated circuits (MMIC’s), with large numbers of closely packed active and passive structures, many levels of transmission lines and discontinuities, all functioning at high speeds, frequencies, and sometimes over very broad bandwidths. The highly varied physical dimensions, physical processes, and time and frequency scales require us to address topics such as electromagnetics (EM’s), charge transport, and thermal properties on more than an individual basis. In addition, we must worry about how the details of one such solution affect the details of the others. This is not how we have normally proceeded in the past. Instead of a global approach, we used independent simulators that dealt in detail with one part of the problem while making simplistic assumptions about other aspects of the problem. For example, we would worry about all of the details of the EM fields, but willingly use simple models of materials and electron devices. Alternatively, we would worry about all the details of the materials and electron devices, but willingly use simple models of the EM fields inside the system and represent external systems as equivalent circuits. In both of these cases, we would assume that thermal effects and heat extraction are simple. Examples of situations where the problem involves both a complicated field solution and a complicated materials/electron-device solution include situations where the active electron device will be coupled to a passive environment that cannot be accurately described by a simple equivalent circuit [1]–[4] and structures such as wide-gate field-effect transistors (FET’s) inside of which wave propagation cannot be ignored [5].

Here, we will review progress at trying to incorporate both full-wave electrodynamics and complex charge transport into a single combined simulation. We will begin in Section II by reviewing the charge transport models. In Section III, we will then discuss EM models. Following these discussions of the individual parts of our more global simulation, we will discuss the coupling of the two in Sections IV and V. This paper will close with a discussion of some open issues in Section VI and computational challenges in Section VII. Throughout this paper, we assume a target audience of workers familiar with either transport models or full-wave EM’s, but not both. Thousands of papers have been published in either of these two areas and we make no pretense at providing a comprehensive literature review of either area. Our focus will be on combining the two. We will assume that the reader is familiar with finite-difference (FD) techniques, but not necessarily familiar with certain forms of statistical simulation.

II. CHARGE TRANSPORT IN SEMICONDUCTORS

A. General Introduction

Since most active electron devices are built using semiconductors, several decades of work have been extended into the study and modeling of charge transport in a semiconductor. The general problem is illustrated in Fig. 1. Energy is exchanged between some external power supply and the internal electric field of the device. This field then accelerates the mobile charge carriers, the electrons and holes. Energy and momentum transfers between the field and these charges. As the charge carriers move though, they also interact with the media through which they move. This is usually modeled by having the charge carrier “emit” or “absorb” a phonon, a quanta of lattice vibrational energy. The heat energy stored in these phonons then flows out to a heat sink. Effectively, in most semiconductor device models, the processes by which the topmost and bottommost arrows occur in Fig. 1 are modeled in a simple way. Great attention is then paid to the details of how the other processes or arrows work. Here, we will be interested in approaches that supplement this model by building, in accurate versions of the top arrow, the processes by which energy enters and leaves the device electromagnetically. We will retain the hope that the bottom arrow (heat extraction) can be viewed as a simple process.

Up until the 1970’s, semiconductor devices were well modeled with a drift–diffusion transport approach. The drift–diffusion approach includes a drift current along the
electric field and a diffusion current that flows along carrier density gradients. In a spatially homogenous system, it reduces to Ohm’s law for low electric fields. However, such models assume that the time and space scales of the problem are such that there is, at all times, a balance of sorts between the arrows on Fig. 1. It assumes that the microscopic distribution of momentum and energy over the charge carriers at any location and time inside the device is equal to that which one would find in a large sample with a dc field equal to the local instantaneous field. As length scales shrink into the submicrometer range and time scales shrink into the picosecond range, this assumption breaks down [6]–[8].

In the following several sections, we review the basic models used to replace the drift–diffusion model. All of the models that we review are semiclassical. In semiclassical models, we view electrons and holes as charged points that move in accordance with Newton’s laws. However, they also are constrained by an energy band structure that complicates our description of how the kinetic energy of a particle is related to its velocity and momentum. In addition, as they move, they experience various scattering events in which they suddenly exchange energy and momentum with the media through which they move. The models are referred to as “semiclassical” because the descriptions used for the band structure and scattering processes are generated by quantum mechanical calculations.

B. Particle-Based Models

The first category of transport models, which we review, works by tracking trajectories of individual particles. Usually, it is assumed that these particles do not interact with each other. There are no particle–particle scattering processes included. The workhouse model for such cases is the Ensemble Monte Carlo (EMC) method [6]–[8]. The EMC technique applies a statistical or Monte Carlo approach to modeling a random walk of an individual particle subject to both a continuous acceleration by an applied EM field and collisions.

The basic technique is to simulate the motion of each particle in the ensemble as a sequence of free flights during which the particle is accelerated by EM fields. Each free flight is terminated by a scattering event. Although many possible scattering processes exist, we assume that only one process occurs and that the scattering is instantaneous. The algorithm proceeds as follows:

1) generate a statistically distributed flight time for the particle;
2) deterministically compute the resulting flight;
3) statistically choose the type of terminating scattering event from a menu of scattering processes;
4) statistically determine the final energy and momentum of the particle after scattering.

The procedure is then repeated over and over. Sampling the particle motion at various times throughout the simulation allows for the statistical estimation of physically interesting parameters. This particle-based picture in which we view the particle motion as free flights terminated by instantaneous randomizing collisions also underlies the derivation of the Boltzmann transport equation (BTE). In fact, the one-particle distribution function obtained from the random-walk Monte Carlo technique satisfies the time-dependent BTE for a homogeneous system [9]. A more detailed description of this Monte Carlo process follows.

The process starts by generating a flight time. To generate the flight time, we need to know the probability that a particle will suffer a collision in a time interval $dt$ around time $t$, after the previous collision at $t = 0$. This probability may be written as the product of the probability that the particle did not scatter in the interval $(0, t)$ and the probability that it does a scattering in the time interval $dt$ around $t$. Both of these probabilities depend on the scattering rate, $\Gamma(k(t))$, which is the sum of the contributions from each individual scattering. The implicit dependence of $\Gamma(k(t))$ on the particle wave vector $k$ complicates the free-flight generation because $k$ changes during the flight due to acceleration by an applied EM field. For example, in the simple case of an electric field $E$ that does not vary in time and space, the time dependence of $k$ between collisions is

$$k(t) = k(0) = \frac{eEt}{\hbar}. \quad (1)$$

Now, since $k$ depends on time, the probability that the free flight starts at time zero and ends at time $t$ is

$$P(t) dt = \Gamma(k(t)) \exp \left[ - \int_0^t \Gamma(k(t')) dt' \right] dt \quad (2)$$

where the exponential term on the right-hand side is the probability of arriving at time $t$ without scattering, and the term $\Gamma(k(t)) dt$ is the probability that the particle scatters in $dt$ around time $t$. Note that (2) does not depend on (1), but may be applied to any time-varying $k$. The fact that $k$ varies over the flight here makes $\Gamma(k(t))$ vary during the flight and, thus, gives us a nontrivial integral in (2).

Random flight times may be generated according to the probability density $P(t)$ above by using a random number generator that generates random numbers uniformly distributed in the range $(0, 1)$. Using a direct method (see, e.g., [6]), random flight times sampled from $P(t)$ may be generated according to

$$r = \int_0^\tau P(t) dt \quad (3)$$

![Fig. 1. The basic energy flows involved in the use of a semiconductor device.](image-url)
where \( r \) is a uniformly distributed random number, and \( t_{fr} \) is the desired free flight time. Integrating (3) with \( P(t) \) given by (2) yields

\[
    r = 1 - \exp \left( - \int_0^{t_{fr}} \Gamma(k(t')) \, dt' \right). \tag{4}
\]

Since \( r \) is uniform, \( 1 - r \) is statistically the same as \( r \) and (4) may be simplified to read

\[
    -\ln r = \int_0^{t_{fr}} \Gamma(k(t')) \, dt'. \tag{5}
\]

Equation (5) is the fundamental equation used to generate the random free-flight duration.

If there is no external driving field leading to a change of \( k \) between scattering, the time dependence vanishes and the relationship between the random number and the corresponding flight time is

\[
    t_{fr} = -\frac{1}{\Gamma} \ln r. \tag{6}
\]

In the general case, and essentially any device modeling case, this simplification is not possible. Fortunately, there is a trick. A fictitious scattering mechanism called self-scattering [10] is introduced. Since it is an artificial mechanism, we are free to set its parameters in any convenient way. We select the self-scattering rate in such a way that the sum of the self-scattering rates and the real scattering rate is constant over the free flight

\[
    \Gamma = \Gamma_{\text{self}}[k(t')] + \Gamma_{\text{self}}[k(t')] \tag{7}
\]

where \( \Gamma_{\text{self}}[k(t')] \) is the self-scattering rate. To avoid introducing an effect on the free-flight trajectory of a particle, the self-scattering mechanism is defined such that the states before and after a self-scattering event are identical. Since we now have a scattering rate that is constant during the flight, we can use (6) instead of (5) to generate our flight time.

Having selected a flight time, we then deterministically compute the trajectory followed by the particle as it moves under the influence of the applied EM fields. Equation (1) is a particularly simple example of how we would determine the wave vector \( k \) at the end of the flight. In addition, we would also compute the corresponding velocities of the particle and follow its trajectory through our system. In cases where the field is varying in space and time, one needs to carefully choose the fields used in the acceleration. Having performed the flight, we need to decide which type of scattering process ends the flight. The type of scattering terminating the free flight is chosen using a second random number, uniformly distributed over a different interval, the interval (0, \( \Gamma \)), to generate a pointer that selects among the relative total scattering rates of all processes including self-scattering at the final energy and momentum of the particle

\[
    \Gamma = \Gamma_{\text{self}}[n, k] + \Gamma_1[n, k] + \Gamma_2[n, k] + \cdots + \Gamma_N[n, k] \tag{8}
\]

where \( n \) is the band index of the particle (or subband in the case of reduced-dimensionality systems), and \( k \) is the wave vector at the end of the free flight.

Once the type of scattering 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 according to energy conservation and the scattering rate, \( \Gamma_j(n, k; m, k') \), of the \( j \)th scattering mechanism. Here, \( n \) and \( m \) are the initial and final band (subband) indexes and \( k \) and \( k' \) are the particle wave vectors before and after scattering. This involves generating and using one or more additional random numbers, depending on the scattering mechanism. Details on how such decisions are made are found in [6]–[8] and references therein.

EMC calculations attempt an “exact” stochastic simulation of the microscopic processes of the semiclassical transport picture. Their accuracy depends on the accuracy of the model (there are significant questions concerning this for very short times, spaces, and very high fields [7], [8]) and the number of particles in the ensemble. This number may range from \( 10^3 \)–\( 10^7 \). As a result, these simulations are compute intensive. A feature often not recognized though is that EMC calculations are naturally three-dimensional (3-D) and as one moves into the modeling of a 3-D system, the costs of the field computation become comparable to and may even exceed the costs associated with the EMC. Additionally, an EMC model does not involve the same concerns with numerical stability that occur in other models.

In the above discussion, we assumed that one could track each particle’s trajectory independently of all other particles. Since the particles in an electron device are charged, they affect each other’s motion when they are close. For a sufficiently “low” density, such close encounters are statistically rare. However, at densities above a few times \( 10^{18} \) cm\(^{-3} \), many electron effects due to particle–particle interactions will be important. One cannot merely incorporate a carrier–carrier scattering mechanism into the Monte Carlo model of the trajectory of a single particle and fully model the carrier–carrier interactions. Consequently, a real-space molecular dynamics approach is needed to accurately treat the carrier–carrier scattering [11]. In these calculations, a scalar potential seen by each particle due to all other particles is computed by summing all of the individual pairwise potentials, and the gradient of this total potential is used as the force seen by the particle at this instant of time. No vector potential contributions are generally included. Lugli and Ferry [11] developed an approach that incorporated the exchange interaction into the semiclassical EMC procedure and led to a distribution of carriers that satisfies both the exclusion principle and the uncertainty principle, while obeying a set of equations of motion which resemble that of a classical gas [12]. This model has been successfully applied to several important high-density situations [7, Ch. 4], [13]–[16]. Such calculations are computationally expensive and are generally restricted to rather small spatial regions.

\section*{C. Energy-Momentum Conserving or Hydrodynamic Models}

EMC calculations are numerically costly. Due to this, we often use models in which we view the particles as a fluid. The standard development of such “hydrodynamic” models is based on obtaining the moments of Boltzmann’s transport
equation by integration over momentum space. The integration results in a strongly coupled highly nonlinear set of partial differential equations that conserve average carrier density, energy, and momentum [17]–[20]. These equations are as follows.

1) The Current-Continuity Equation:
\[ \frac{\partial n}{\partial t} + \nabla \cdot (n \mathbf{v}) = 0. \quad (9) \]

2) The Energy-Conservation Equation:
\[ \frac{\partial (n e)}{\partial t} + \nabla \cdot (n \mathbf{v} e) = q n \mathbf{v} \cdot (\mathbf{E} + \mathbf{v} \times \mathbf{B}) - \nabla \cdot (n k_B T \mathbf{v}) - \frac{n(e - \epsilon_0)}{\tau_e}, \quad (10) \]

3) The Momentum-Conservation Equation:
\[ \frac{\partial (n p_x \mathbf{v})}{\partial t} + \nabla \cdot (n p_x \mathbf{v}) = q n \mathbf{v} \cdot (\mathbf{E}_x + \mathbf{v} \times \mathbf{B})_x - \nabla (n k_B T) - \frac{n p_x}{\tau_m}. \quad (11) \]

where \( n \) is the electron density, \( t \) is time, \( \mathbf{v} \) is the average electron velocity, \( e \) is the average electron energy, \( q \) is the electronic charge, \( k_B \) is the Boltzmann’s constant, \( T \) is the electron temperature, \( \epsilon_0 \) is the equilibrium thermal energy, \( p_x \) is the \( x \)-component of electron momentum, and \( \tau_e \) and \( \tau_m \) are the energy and momentum relaxation times, respectively. The relaxation times, in general, are nonlinear functions of energy. Similar equations are written for the momentum in the other directions. The electronic current density distribution \( \mathbf{J} \) inside the active device at any time \( t \) is given by
\[ \mathbf{J}(t) = -q n(t) \mathbf{v}(t). \quad (12) \]

This approach requires less computational time and effort than does the EMC. However, it is still not a trivial task. Various parameters such as the relaxation times, need to be obtained somehow, probably through comparison with EMC results. Additionally, these partial differential equations pose challenges in terms of numerical stability and cost. For example, in many simulations, the momentum-conservation equation is greatly simplified, often to the level of an assumption that the electron momentum is able to adjust instantaneously to a change in the electric field. While this assumption is justified for long-gate devices, it leads to inaccurate estimations of device internal distributions and microwave characteristics for submicrometer gate devices [21].

Equations (9)–(11) are coupled highly nonlinear partial differential equations. To decouple these equations in time, a FD-based scheme is used. The scalar variables are defined at the basic nodes, whereas the vector quantities are defined at the complementary nodes (i.e., between two basic nodes). This type of staggered mesh arrangement separates the component of a vector variable and facilitates a convenient method for decoupling the variables and expressing the approximations of the space derivatives using the FD method. The solution is obtained in a self-consistent evaluation of the three equations in conjunction with a field solution. The order in which these equations are solved is critical to the stability of the solution. The change in carrier momentum is computed first. The carrier density and energy equations are then solved using the updated distribution of carrier velocity. The reason is that the relaxation time of the electron momentum is about one order of magnitude shorter than dielectric and energy relaxation times. If the order of the solution is reversed, large changes in momentum will be introduced leading to potential instabilities. As we will note later, the EM solutions have stability criteria which, when taken in conjunction with the small space steps desired in semiconductor device models, provide the most stringent limits on time-step selection. Therefore, in the combined models we discuss here, all the equations are represented by explicit FD schemes as using other FD formulations (e.g., implicit or semi-implicit) provides little advantage. Best accuracy and stability in solving the continuity and energy equations is obtained using the up-wind scheme, where the direction of electron velocity dictates the correct differencing. Since all the inertia terms in the momentum equation are taken into account, extra attention is given to its solution method. The velocity gradient terms in the momentum equation make direct FD solution sensitive to small fluctuations and prone to instability. To overcome this problem, a scheme called the Lax method [22], which utilizes average values over two time and space steps, is used.

III. EM MODELING

A. Quasi-Electrostatic Models for the Fields Inside the Device

We start our discussion of EM’s by briefly reviewing the method generally used in modeling semiconductor devices. This is a quasi-electrostatic approach. It is not truly electrostatic because we allow the field to vary in time. However, we assume that it always has a spatial behavior that is electrostatic in nature. That is, we assume that
\[ \nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon} \quad (13) \]
and
\[ \nabla \times \mathbf{E} = 0. \quad (14) \]

The assumption that the electric field is irrotational is the essential approximation that is made. It of course allows us to introduce a scalar potential \( V \) where
\[ \mathbf{E} = -\nabla V \quad (15) \]
and \( V \) is obtained by solving Poisson’s equation
\[ \nabla^2 V = -\frac{\rho}{\varepsilon}. \quad (16) \]

A typical semiconductor device model solves Poisson’s equation subject to boundary conditions at each time step. A great deal of sophistication in fact has been developed in solving
this particular problem [e.g., (24)–(27)]. Then, when a force or field is needed, one computationally takes the gradient of $V$.

In discussing the limitations of this approach, someone who approaches the question from the point-of-view of a traditional semiconductor device modeler must remember that key issue is the validity of (15) and not the validity of Poisson’s equation, (16). Unfortunately, the boundary value problem associated with Poisson’s equation is computationally challenging, whereas (15) is computationally trivial and, as a result, most device modelers are not used to viewing (15) as being more important than (16). This discussion can also be complicated by the fact that there exist gauges in which Poisson’s equation is electrodynamically exact \[27, 28\]. However, (15) is never electrodynamically exact. In these gauges where Poisson’s equation is exact, to compute the electric field, one must solve an additional vector equation for the vector potential $\mathbf{A}$ and then use

$$
\mathbf{E} = -\nabla V - \frac{\partial \mathbf{A}}{\partial t},
$$

(17)

There exists a rich literature in which the transport models reviewed here (and others) are solved self-consistently with a quasi-electrostatic field model \[24\]–\[26\].

**B. Equivalent Circuits and Their Limitations**

Part of the justification for the quasi-electrostatic approach in semiconductor device modeling is an assumption that an equivalent circuit will represent the universe outside of the device. Geometrically dependent quantities such as fields, current densities, and charge densities are tracked inside the device only. The device model is usually used to relate these internal geometric dependencies to voltages and currents at the device terminals. To complete the description, one then needs to include a description of the external circuit connected to these terminals. A common problem faced in such situations is the nonlinearity of the device. Even if we could assume a single-frequency sinusoidal input, we get harmonic generation. To then properly model the system performance, one needs a good representation of the circuit as seen by the semiconductor element at a large number of harmonics. This can be particularly troublesome in the genuinely transient case where the circuit needs to be described over a broad band of frequencies.

Getting an accurate equivalent-circuit representation at even a single frequency can be difficult in millimeter-wave integrated circuits due to the EM discontinuities associated with real physical structures. A discontinuity in a transmission line is a variation from a straight and longitudinally homogeneous circuit. The dimensions of the variation are generally small compared to the wavelength. These variations distort field and current lines, lead to radiation and can excite propagating surface modes. To account for these effects, in a circuit model, the discontinuities are often analyzed as individual circuit elements, separate from the transmission line itself. The circuit elements are combined to give the final analysis \[29\]. Inclusion of discontinuities in the computer-aided-design tools for circuits leads to a remarkable improvement in accuracy of the predicted circuit performance \[30\]. However, the discontinuities associated with active circuits cannot be analyzed separately. The energy scattered from the discontinuity is coupled to the active device and may severely affect the device performance.

To further complicate matters, in conventional high-frequency GaAs transistors, the electrode structures of the transistor itself form an EM discontinuity at the input and output ports of the device. A common practice in analyzing the discontinuity effects on active devices is to analyze the passive structure separately from the behavior of the real active device \[31\]. This approach, however, does not take into account the active nonlinear device characteristics. In addition, it is very difficult to determine the discontinuity effects from measurement. This situation prevents, in many cases, the possibility of matching the FET. Indeed, we almost see a shift from regarding the FET as a nonlinear active circuit element into a picture of an FET as a nonlinear, active EM discontinuity.

**C. EM’s—Full-Wave Approaches**

The point of Section III-B was that accurate equivalent-circuit representations of the EM response of the system external to the device are often nontrivial. One alternative to an equivalent-circuit representation is a full-wave EM solution of the system external to the active device. This then dramatically alters the computational environment inside of which the device model is to be inserted. For example, we probably no longer need to know the voltages at the device terminals. Additionally, we are solving field equations outside of the device. It, therefore, may be sensible for us to include the device itself inside the same EM model as one is using for the external system. It certainly is sensible to ask for the fields internal and external to the device to match at the device boundaries. In the following sections, we discuss such approaches, starting with a brief review of finite-difference time-domain (FDTD) approaches to full-wave EM’s.

EM-wave propagation can be completely characterized by solving Maxwell’s equations. Assuming uniform linear isotropic media for the dielectric and magnetic relations, Maxwell’s equations are given as

$$
\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t},
$$

(18)

and

$$
\nabla \times \mathbf{H} = \varepsilon \frac{\partial \mathbf{E}}{\partial t} + \mathbf{J}
$$

(19)

where $\mathbf{E}$ is the electric field, $\mathbf{H}$ is the magnetic field, $\mathbf{J}$ is the conduction current density, and $\varepsilon$ and $\mu$ are the permittivity and the permeability of the medium, respectively. The conductivity of the medium is obtained using the transport model.

A common formulation is to view these two curl equations and Gauss’s law as independent equations. From these three, one can then derive the law of conservation of magnetic flux and the current continuity equation. In the models we are about to describe though, a second choice is made. The transport
model directly implements the continuity equation and we will supplement this with Maxwell’s curl equations. With the exception of the process of identifying an initial state, we will not solve Gauss’s law. This approach can work because it is possible to view Gauss’s law as an equation that is dependent upon Maxwell’s curl equations and the continuity equation [32].

The time-domain solution of Maxwell’s equations is obtained using a 3-D mesh where field components are arranged following Yee’s method [33]. The mesh scheme is shown in Fig. 2. Using first-order differencing, Maxwell’s curl equations can then be decoupled over a small time interval $dt$. To reduce the computational requirements, a procedure similar to the one described in [34] is used. The electrodes are taken as perfect conductors, which leads to zero tangential electric fields. We cannot use an infinite mesh to simulate the infinite space surrounding the structure. Therefore, we need to implement some form of absorbing boundaries when we terminate the mesh in an open region. The choice of absorbing boundaries is very critical to the overall stability of the FD scheme. Several different absorbing boundary conditions have been used in the work reviewed here. In one set of calculations, we use a second-order boundary condition. In the other computations, we use a PML approach.

The reason for selecting second-order boundary conditions is that for long time simulation or for boundaries that are placed where the field strength is high, instability is likely to occur if the absorption is not perfect [35] when first-order absorbing boundaries are used. A discussion of some of the most common second-order absorbing boundary conditions can be found in [36] where it was concluded that Higdon’s boundary conditions [37] gave the best overall efficiency for a waveguided structure. Higdon’s boundary conditions are based on the fact that an arbitrary wave can be decomposed into a summation of plane waves with different angles of incidence $\theta$. For second-order boundaries, two velocities have to be chosen according to the relation $v_{iz} = c/\cos\theta z$, where $c$ is the propagation velocity of the wave inside the medium. Here, when this approach is used, one of these velocities is related to the effective dielectric constant of the medium, which corresponds to zero angle of incidence [38]. For the other velocity, however, numerical experimentation was performed to obtain the optimum absorption for the given geometry and simulation parameters of the problem. It has been observed that the reflections by the Higdon’s boundary conditions are an order of magnitude less than those from the first-order absorbing boundary conditions. Other of the calculations reviewed here use a perfectly matched layer (PML) or an EM absorbing boundary condition [39]–[41]. The essential idea of a PML is to terminate the grid in a fictitious layer of absorbing material. The properties of this artificial layer are picked such that there is no wave reflection at the interface between the region of interest and fictitious material. Since our fictitious material is absorbing, by adjusting the width of the fictitious layer, we can minimize the return of a reflected component off of the eventual grid termination back into the region of interest.

IV. COUPLING FDTD EM’S WITH HYDRODYNAMIC TRANSPORT

A. Linking the Two Models

The link between these two models is established by properly transforming the physical parameters (e.g., fields and current densities) from one model to the other. The simulation starts by obtaining the steady-state dc solution, using Poisson’s equation and the transport model. The dc solution is used in the ac analysis as initial values. The ac excitation is then applied. Maxwell’s equations are solved for the electric- and magnetic-field distributions. The new fields are used in the semiconductor model to find the current density. This process is repeated for each time interval.

We now have two different sets of partial differential equations: the hydrodynamic equations [(9)–(11)] and Maxwell’s curl equations [(18) and (19)]. Both FD and finite-element schemes can be used to solve such equations and both have been used in semiconductor device simulations. In our work, we typically use FD techniques. They are easily formulated and considerable information is available on their stability and convergence properties [42]. In cases where we couple Maxwell’s equations with the hydrodynamic equations, several FD techniques, such as the upwind and the Lax methods, are used in conjunction with the basic FD formulation of Maxwell’s equations to achieve stable and accurate solutions. The space increments are adjusted to satisfy Debye length criteria for semiconductor simulation and wavelength criteria for EM simulation for a stable solution simultaneously. The same space increments, used in the two models, are identical in all three dimensions.

The initial state used for the time-dependent calculations is obtained from a steady-state dc solutions for electric fields, current densities, carrier density, carrier velocities, carrier energy, and the transport parameters. These are obtained from the semiconductor model by solving Poisson’s, continuity, energy balance, and momentum balance equations. As the dc solution is performed in a two-dimensional (2-D) plane $(x-y)$ and the ac analysis is performed in 3-D, all the sections of the semiconductor device along the third dimension (the $z$-direction) are initialized with the same dc values for the above-mentioned quantities. In the dc steady-state, Maxwell’s equations become

$$\nabla \times \mathbf{E} = 0$$

(20)
and

$$\nabla \times \mathbf{H}^{dc} = J^{dc}$$  \hspace{1cm} (21)$$

where $\mathbf{H}^{dc}$ is the dc magnetic field, and $J^{dc}$ is the dc current density. Equation (21) suggests that it is not necessary to find the dc distribution of the magnetic field, since the dc current density carries the information about the dc magnetic-field distribution. This approach can be implemented as one of the measures taken to reduce the computation time when the effects of $\mathbf{H}^{dc}$ are not of interest in the simulation.

After completing the above initializations, the ac excitation is applied. The mesh for the EM fields are extended along the $z$-axis before and after the semiconductor device. The extended regions are made of a passive transmission line. At the device input, this passive section allows the correct mode of the excitation wave to develop before it is fed to the active device. At the device output, it is used for absorbing the EM wave.

The time-domain distribution of the EM fields is obtained using Maxwell’s equations. Using (21), the curl of $\mathbf{H}$ can be written as

$$\nabla \times \mathbf{H} = \nabla \times \mathbf{H}^{dc} + J^{dc}. \hspace{1cm} (22)$$

The electric field is then specified as

$$\frac{\partial \mathbf{E}}{\partial t} = \frac{1}{\varepsilon} (\nabla \times \mathbf{H}^{dc} + J^{dc} - J^{tot}) \hspace{1cm} (23)$$

where $J^{tot}$ is the total current density obtained from the active device. The magnetic field is given by

$$\frac{\partial \mathbf{H}}{\partial t} = \frac{1}{\mu} \nabla \times \mathbf{E}. \hspace{1cm} (24)$$

Equations (23) and (24) give the electric- and magnetic-field distributions at each time step. These new values are used by the semiconductor model to update the current density at the same time step. These current densities are fed back to the Maxwell’s equations again to calculate the electric and magnetic fields in the following time step. This process of updating the fields and current densities progresses in time in response to applied excitation and moves on for an appreciable amount of time to study the device behavior.

### B. Some Sample Results

In this section, we present some examples of the type of result obtained from a model in which hydrodynamic transport equations are coupled with a full-wave Maxwell solver [50]. Other results are available in the references [4], [43]–[50]. We will compare the physically based model described here with a more conventional circuit-based model. In both cases in this comparison, techniques were used to incorporate EM-wave effects into the semiconductor device simulation. We apply both models to the same device.

The effect of using the EM model in device simulation versus using just a quasi-static semiconductor model is demonstrated in Figs. 3 and 4. A modulation-doped FET (MODFET) and metal–semiconductor FET (MESFET) are simulated using the approaches. In the quasi-static model, Poisson’s equation is solved to get the electric fields. In the combined EM solid-state simulator (CESS), Maxwell’s equations are solved to obtain the electric and magnetic fields. In both cases, a hydrodynamic transport model was used and these results, therefore, reflect the difference due to EM’s. The approach used in the quasi-static model was to extract small-signal parameters from the dc solution. In Fig. 3, the output voltage wave is shown for different device width using these two models. The MODFET has higher output voltage than the MESFET, which is evident from the models. The output voltage wave monotonously increases along the device width in the quasi-static model. On the other hand, in the EM model, the output voltage wave nonlinearly increases with the device width. This phenomenon is expected due to the device EM-wave interaction. Up to $100 \mu m$, the EM wave receives energy from the device, then from $100$ to $150 \mu m$, the EM wave gives back some of the energy to the device due to phase cancellation (i.e., the phase velocity mismatches between the waves on the different electrodes). Again, it gets energy from the device after $150 \mu m$. This behavior is absent when the output is obtained from
the quasi-static analysis. Thus, the full-wave model provides the opportunity to choose the optimized device width for maximum output voltage.

A comparison between circuit-based and full-wave models is also accomplished by simulating a MESFET structure similar to the one presented in [43]. The gate-source spacing is 1.5 μm, the gate length is 0.5 μm, the gate width is 1 mm, the active-layer thickness = 0.2 μm and the active layer doping is $10^{17}$ cm$^{-3}$. The amplifier gain characteristics resulting from the two simulation techniques are shown in Fig. 4. Two input voltages were used in the full-wave simulation: 0.1 and 0.3 V. The results exhibit the same trend; however, as expected, there are quantitative differences. The sources of such differences can be identified by examining the nature of the two simulators. The circuit-based simulator is strictly small signal, while the full-wave model retains the large-signal aspects of the device. To demonstrate the large-signal potential of the full-wave simulator, the input signal level was increased to 0.3 V. Fig. 3 shows that the gain becomes lower as the signal amplitude increases. The strength of this approach is demonstrated not only by predicting the amount of gain degradation at higher input levels, but also by calculating the harmonics generated by the device nonlinearity. The harmonic content of the output is analyzed by taking its Fourier transform, as will be shown in Section V. To the best of our knowledge, this is the only direct approach that can perform such a study.

V. COMBINED EMC AND FULL-WAVE SIMULATIONS

In this section, we will review efforts at combining EMC transport calculations with full-wave EM’s. To do this, one needs to be able to relate the fields on the EM mesh to forces in the EMC model, and then one will need to be able to extract from the ensemble of particles a set of currents to serve as source terms in the EM model. EMC calculations are essentially untroubled by spurious numerical instabilities and the overall time step, therefore, will again be controlled by the EM solver.

To date, all such efforts have coupled the EMC results with a FDTD full-wave solver built using Yee’s mesh [33], as shown in Fig. 2. To go from the field solution to the EMC, we need to relate the force seen by the particle to the field result. We assume that a particle lies inside such a “bin” and that it sees a constant field. The particle location is tracked during its free flight and when a particle moves from one bin to another, the field seen by this particle is appropriately modified. The fields used as forces in the EMC calculation are obtained by simply averaging the values on the surrounding surface. To go from the EMC to the field solver, we need to obtain a current density from the ensemble. There are several means of obtaining a current at a time step from the EMC results [51]. One approach is to simply associate each particle with the closest associated current mesh point of the full-wave solver. Alternatively, one could also average all the particles found inside a given bin and then use the average of the associated bins to obtain a current. Lim [51] used a variation of this second approach, in which the instantaneous particle currents are replaced by the time average of a single particle’s current over a single time step. This helps ensure that the product $\mathbf{E} \cdot \mathbf{J}$ during a time step is the same as the energy absorbed by the particles through field acceleration during the time step [51].

We illustrate the application of such models by discussing the simulation [51]–[55] of ultrapower photoconductive switching experiments [1]–[3] carried out in microstrip systems. The only difference between this system and the earlier FET’s modeled hydrodynamically is the presence of a gate electrode. In these experiments, laser pulses are used to photo-excite gaps in microstrip lines. The optical pulses are very short, with pulse widths in the few tens of femtoseconds. They are focused on a region where a dc bias drops across a gap between two metal electrodes. The photoexcited carriers initially have an average velocity of zero. This velocity then builds as the carriers are accelerated by the field. As it builds, a photocurrent flows between the two electrodes, which excites an EM pulse that propagates down a transmission-line structure. This guided-wave response is observed using electrooptic sampling [56] with subpicosecond resolution. In addition, as the photocarriers accelerate, they radiate a freely propagating pulse of EM radiation of picosecond duration and terahertz bandwidth that can be observed using other photoexcited gaps as detecting antennas [2]. A typical feature of these experiments is that the bias field in the system collapses as energy is transferred from the field into the carriers. The energy cannot be resupplied during the experiment as the energy flowing into the gap region from the power supply during the actual transient itself is just the flow needed to maintain the dark state leakage. This particular effect can be very clearly seen in the experimental data on quantum-well structures where it is possible to use optical techniques to perform an in situ observation of the actual electric field in the well itself [57]. This effect is also clearly manifest in the simulation results of Lim [51].

These experiments are of interest because they represent one of the most direct experimental observations of a phenomena called velocity overshoot, which has long been predicted using the transport models reviewed here [58]. In Fig. 5, we show such an overshoot. Here, a simulation was done [59] of photoexcitation of electron–hole pairs in GaAs by an 80-fs-long optical pulse at a wavelength of 856 nm. The sample was assumed to have a constant field of 40 kV/cm. The velocity plotted is the average electron velocity. As can be seen, this quantity undergoes a transient from initial to final state in which there is a “velocity overshoot.” The

Fig. 5. Results of an EMC study of photoexcitation into GaAs. A 856-nm optical pulse (80-fs duration) was used to excite a sample biased at 40 kV/cm.
The issue confronted directly by Remley et al. [55] was that the simulation produces a near-field result. It would be extremely onerous in terms of memory and simulation time to extend the grid to include the far field; yet the experimental data on the freely propagating radiation is obtained in the far field. The earliest and conceptually simplest model for these experiments is that we form a Hertzian dipole in the photoexcited region. In Fig. 6(a) and (b), we compare the actual fields found on the surface of the simulation domain with those expected from an ideal Hertzian dipole driven by a time-varying current comparable to the computed photocurrent.

To model the far-field case, a near- to far-field transformation was employed. This is illustrated in Fig. 7. Various methods for performing a near- to far-field transformation exist and Remley et al. concluded that the Kirchhoff surface integral representation [27] was the most appropriate to use here. In Fig. 8, we show the results. The existence of the GaAs substrate in the experiment has a visible complicating effect. These results are generally consistent with those seen experimentally.

Another feature of the experiments is the guided wave, which also has a terahertz bandwidth. In Fig. 9, we show how the wave propagates down the line. This figure shows...
Let us now compare this with the simulations just described. We have proceeded by using an average or macroscopic charge and current as sources in Maxwell’s equations, and then computing the resulting macroscopic fields. In other words, we related the macroscopic fields to the microscopic particles by first performing an ensemble average to obtain a macroscopic field source, the ensemble average current. We then computed the fields. This can be contrasted with models using molecular dynamics, where we instead compute potentials associated with single particles and then sum the single particle results at the field point. It should be possible to proceed in a similar fashion for the photoconductive simulations. We would compute the radiation fields of the individual carriers in the ensemble and then take an ensemble average of these fields [59], [60] at the field point to obtain a macroscopic field. We do not suggest this as an effective simulation method, but rather as a thought experiment that probes the relationship between semiclassical transport models and classical electrodynamics.

To carry out the above “thought computation,” we must be able to connect all the accelerative phenomena implicit to Fig. 10 with EM radiation. There are difficulties with the collisional accelerations. First, as we assume the collisions to be instantaneous, each collision actually has an impulsive acceleration in our microscopic framework. It will have a radiation spectra that would be white (this is an extension of the general point of view described in Jackson [27]). Secondly, the scattering rates are calculated neglecting these radiation fields. Since EM radiation in the low terahertz range involves photons in the low meaelectronvolt range, we would envision a study of inelastic scattering events that alter energy thresholds. In performing such a calculation, one would have to reopen an old issue inside the field of electron transport theory, that being the existence and role of finite-collision thresholds. In performing such a calculation, one would have to reopen an old issue inside the field of electron transport theory, that being the existence and role of finite-collision thresholds.

VI. IS SEMICLASSICAL TRANSPORT THEORY CONSISTENT WITH CLASSICAL ELECTRODYNAMICS?

Throughout this paper, we have addressed the self-consistent modeling of electrodynamics with semiclassical charge transport models. Numerical consistency was the goal. However, are these two pictures, as usually developed, actually physically consistent? The potential that they are not truly consistent is illustrated in Fig. 10, where we show the accelerative processes present in the study leading to Fig. 5. Here, in each time step of the EMC, a sum was made of the various accelerative events. As can be seen early in the transient, acceleration by the field dominates. In this particular case, even though the field is constant, there is a fall off in the field acceleration due to a transferring of carriers to higher mass valleys. More importantly, there is a counterbalancing collisional acceleration that builds during the transient. In the eventual steady state, this collisional acceleration counter balances the field acceleration. Note that when one integrates the total acceleration in Fig. 10, the result is Fig. 5. Also note that with the exception of a sign flip, the total acceleration in Fig. 10 resembles the general shape of the radiated pulse of Fig. 8.

Fig. 9. Simulation results showing how the guided-wave signal degrades as one moves away from the photoexcited region (from Lim [51]).

Fig. 10. Comparison of acceleration due to field, collisions, and the total acceleration in the photoexcitation study. The squares are the acceleration due to collisions, the diamonds the acceleration by the field, and the triangles are the total acceleration.

an important detail, which is that as one moves down the line away from the photoexcited region, that the perceived overshoot will decrease [51].
exist. The difficulty is this. The field solutions tend to assign a specific region of space to each processor in the array. However, typically, the complex transport occurs in only a small portion of the total system volume. Therefore, if we use the same assignment of regions of space to individual processors when we parallelize the transport problem, we find that most of the processor array is idle during the transport calculation. One can, of course, use two different mappings, one for the field computation and one for the transport computation, but then one encounters communication overhead as one shuffles data in and out of the array every time one swaps between the field and particle solutions. Lim [51], [54] demonstrated that it is possible to find schemes where increasing the communications overhead in exchange for effective load balance, which visibly improved the cost effectiveness of the solution.

It also should be noted that when one examines the costs associated with a quasi-electrostatic approach in 3-D, full-wave solutions may be cost competitive [54]. This may surprise some, but the fact that an approximation helps purely analytical solutions of textbook problems does not mean that it simplifies real computation. In particular, the Poisson solvers used in conventional simulations require multiple iterations per time step, whereas an explicit FDTD requires no iteration inside a time step. In fact, if one is principally interested in choosing between coupling a 2-D or 3-D quasi-electrostatic approach with either a drift–diffusion transport model or a hydrodynamic model, usually there is little difference in computational effort between these two. The Poisson solver is likely to dominate the costs in either instance. For example, Alsunaidi et al. [49] found that when running a hydrodynamic model in conjunction with Poisson’s equation, the Poisson solution consumed 94% of the computational run time.

B. Hybridization

From the previous discussion, it is obvious that some techniques are well suited to handle certain problems or part of problems, more than the others. When we consider developing global simulators that can handle the diverse regions of electric circuits with various material characteristics and, maybe, drastic variations in accuracy and bandwidth requirements, the best approach is to use more than one analysis scheme. Each technique is efficient in solving certain aspects of the problem. Hybrid numerical techniques, therefore, are quite useful [45]. The real challenge is in interfacing the different techniques. Exchanging the data accurately and efficiently between simulator segments employing different techniques is certainly a major concern, especially when this process has to be repeated many times, as it is the case in iterative solvers. One should notice that most semiconductor-device physical simulators employ iterative schemes. Therefore, hybridized techniques should be selected such that the data exchange process should be simple enough to not overshadow any gain in computation speed obtained from the hybridization.

The hybridization technique in MMIC’s was previously used in coupling the lumped elements and the distribution network of the transmission lines in many different forms [62]–[66]. Some works were done for hybrid systems with active and passive lumped elements by the FDTD [67] and transmission-line matrix (TLM) methods [68], [69]. The technique was applied to simulate the picosecond pulse generation by an optically excited GaAs devices using 3-D charge transport models [52], [70]. Some researchers modeled the three-terminal active devices by incorporating two-terminal lumped elements into several FDTD subgrids [71] and by using the SPICE lumped circuits in the FDTD solution of Maxwell’s equation [72]. The microwave amplifiers were simulated using the equivalent voltage and current source approaches for the transistor [73]. The characterization of microwave circuits containing various lumped components was performed using a global EM analysis [74]. In [67]–[74], the active devices were not physically simulated. Instead, equivalent-circuit models were used. This process of replacing the active device with the equivalent sources or the equivalent-circuit parameters may not represent all aspects of device nonlinearity and wave–device interactions at the millimeter-wave frequencies. The alternative approach is to simulate the integrated circuit by coupling the physical equations representing the semiconductor device with the EM fields in the transmission lines with an accurate FDTD algorithm. However, the FDTD method requires intensive computer memory and consumes a considerable amount of time for its operation. We used several hybridization techniques to simulate millimeter-wave circuits with active devices. Details of these techniques can be found in [50] and [75].

VIII. Conclusion

This paper reviewed the basic semiconductor transport models that are used to model high-speed and high-frequency semiconductor devices. The basic concept of particle-based models, hydrodynamic models, and the factors affecting their accuracy are discussed. The limitations of quasi-electrostatic and circuit-based models are pointed out. This paper showed that using full-wave EM models considerably improves the accuracy of the simulator and facilitates interfacing semiconductor devices with the circuit without significantly increasing the computer time over the quasi-static models. Several examples were presented and the results show both that quasi-static models can be inaccurate and that the combined models can address concerns that simply cannot be well formulated at all using other approaches.

ACKNOWLEDGMENT

The authors thank D. Ferry, R. Joshi, T. Norris, W. Knox, K. Meyer, G. Mourou, M. Alsunaidi, S. Hammadi, S. M. Sohel Imtiaz, I.-S. Lim, M. Megahed, K. Connolly, K. Remley, S. Pennathur, and V. K. Tripathi for their thoughts, work, and suggestions with regard to these matters. Many of the specific computational result were generated by some of these individuals, as noted in the references and figure captions.

REFERENCES


Robert O. Grondin (S’72–M’81–SM’88), photograph and biography not available at the time of publication.

Samir M. El-Ghazaly (S’84–M’86–SM’91), for photograph and biography, see this issue, p. 815.

Stephen Goodnick (M’88–SM’91), photograph and biography not available at the time of publication.