Highly Efficient Full Band Monte Carlo Simulations

R. Kent Smith and Jeff Bude
AT&T Bell Labs, Murray Hill NJ, 07974-0636

Abstract

We present a full band monte carlo algorithm based on phase-space simplexes which has all of the advantages of analytical band monte carlo while preserving the accuracy of a full band structure. An adaptive, contour-aligned grid algorithm is used to represent the energy band structure within the irreducible wedge and is calculated once for each semiconductor band. The complexity of generating the adaptive grids in phase-space is more than compensated for by the simplicity and efficiency introduced into the equations of motion and final state selection, which are treated exactly within the physics of the model. Consequently, our method confines simulation error to statistical error and a consistently bounded discretization error determined only by the choice of phase-space grid and can be set to a user-defined tolerance. Results using this method show at least an order of magnitude improvement in performance over previous full band codes.

Monte Carlo simulation has earned an important place in semiconductor transport simulation because it offers a practical way of solving the full Boltzmann equation – essential for a rigorous understanding of non-stationary/high-field transport which includes impact ionization, hot carrier dynamics, and velocity overshoot. Naturally, the more rigorously based the physics in the Monte Carlo simulation, the more accurate and detailed the solution. The numerical solution of the Boltzman equation is obtained by integrating particles along trajectories in phase space. The momemtum of each particle is abruptly changed when scattering events occur.

One of the most important elements in establishing an accurate physical model is the inclusion of a full band structure (FB) which, for example, can be calculated from empirical pseudo-potential or local density functional theory. Unfortunately the inclusion of a full band structure considerably complicates the MC simulation. First, the particle velocity is not simply related to $\vec{k}$ but is determined by the gradient of the energy band. Failure to integrate the equations of motion exactly will result is unphysical gains or loses in the particle energy. Secondly, the scattering rates conserve energy which, for simple dispersionless phonon models, restricts the final

state momentum, $\vec{k}'$, to lie on energy isosurfaces. For a general band structure, these surfaces may be extremely complicated requiring elaborate searches of the Brilluoin Zone for the final state momentum, $\vec{k}'$.

The computation complexity of MC simulations may be significantly reduced if an analytic expression is used for the band structure. Isotropic polynomial approximations to the energy bands enable efficient and accurate calculations of the free flight trajectories and final states momentum selection. Analytical bands (AB) are rigorously justified for low field transport since the carriers remain close to the band minimum, where the bands are parabolic. Although AB models are computationally efficient, accuracy is sacrificed, at high energies, where simple band approximations are difficult to construct.

We propose a FB MC algorithm [2, 3], based on simplexes, which has the computational advantages of the faster AB simulators, while preserving the physics contained in the full band structure. In k-space, the irreducible wedge is represented by a contour-aligned tetrahedral grid. An adaptive algorithm is developed to approximate each energy band by a piecewise linear polynomial to a pre-defined accuracy. Within each element, the equations of motion may therefore be integrated exactly, thus conserving energy at each time step. Furthermore, the energy isosurfaces for the entire Brilluoin Zone, may be readily calculated providing an accurate final k-state selection. As a result, the simulation error to the statistical error inherent in MC, and a consistently bounded discretization error deriving from the choice of grid in phase-space.

Our algorithm represents phase space as a cartesian product of two simplex grids, $T_k \times T_x$. In k-space, only the irreducible wedge, defined by the 48 elements of the point group pertaining to cubic semiconductors is triangulated. A rotation matrix, Q, is assigned to each particle rotates the position in the irreducible wedge to its proper place in the full Brilluoin zone. In both spaces, piecewise linear polynomial approximations are used to represent the electrostatic potential and each energy band.

Since the equation of motions are linear, particle trajectories may be computed exactly. Pointers to adjacent elements are used to facilitate motion through the grid. Neighbor pointers are also constructed at the phase space boundaries, where particles may be either transmitted or reflected. Element pointers at reflected surfaces point to themselves, where as transmitted particles reenter the phase space grid at some other point. In either case, the simulation particles remain confined to the solution domain and enough information is retained to reproduce the entire phase space.

The selection of the final state momentum is also exactly computed for piecewise linear polynomial approximations. Since ∇E is a constant, the constant energy surfaces are defined by the intersection of a plane with the tetrahedron. An efficient, single parameter, quadratic interpolation procedure may be developed when the vertices of the tetrahedron are constrained to lie between two constant energy surfaces. Each tetrahedra classified by the number of vertices lying on an given energy contour, E_i. The density of states is then characterized by the single interpolation parameter,

$\phi(E)$, and given by

$$\mathcal{D}^{(I)}(E) = \mathcal{D}_T^{(I)}(1-\phi)^2 \tag{1}$$

$$\mathcal{D}^{(II)}(E) = \mathcal{D}_T^{(II)}\phi(1-\phi) \tag{2}$$

$$\mathcal{D}^{(III)}(E) = \mathcal{D}_T^{(III)}\phi^2 \tag{3}$$

where $\phi(E) = (E - E_i)/(E_{i+1} - E_i)$ and $E_i \leq E \leq E_{i+1}$. The major advantage of this decomposition is that global quantities, obtained by summing over tetrahedra, may be interpolated in the same manner. The selection of the final state momentum is readily computed by proceeding down a k-space tree. The rotation matrix, Q, is first determined, then the tetrahedra type, and finally a single tetrahedron containing the final state. In addition to ϕ, two uniformly distributed random variables are sufficient to determine the barycentric coordinates within the tetrahedron. The final state momentum is then computed by a linear interpolation of the tetrahedron vertices, $\vec{k}_i$.

Since the the trajectory calculations and the selection of the final state momentum are performed exactly, the only numerical error encountered in the solution of the discretized BTE is the statistical error associated with sampling the phase space. This error is bounded during each monte carlo run by monitoring the fluctuations in computed quantities. However, a second source of numerical error arises from approximating the continuous BTE with a discrete set of equations. Complete control of the numerical error is obtained by constructing meshes that provide a consistent bound on the discretization error.

We use an adaptive grid algorithm to construct meshes that represent functions in both k-space and x-space to a desired accuracy. These algorithms are based upon estimating the discretization error, ϵ, in approximating continuous functions by piecewise linear polynomials. In x-space, simple, inexpensive error estimates are constructed directly from Poisson's equation, [1]. These estimates provide both lower and upper bounds to the true discretization error, $0 \leq C_1|||\epsilon||| \leq |||\psi - \psi_h||| \leq C_2|||\epsilon|||$ For k-space grids, the discretization error is computed as deviations of the true band energies from piecewise linear approximations. The meshes are continually refined or unrefined until $\|\epsilon\| \leq \delta$ where δ is a user specified error tolerance. By controlling the discretization error of these phase space grids, we can insure that numerical error associated with our MC simulations is bounded.

Several considerations are necessary to produce an acceptable, contour-aligned, tetrahedral grid for our monte carlo simulations. Unlike conventional adaptive refinement algorithms, each grid edge must lie either on an contour or connect two adjacent contours. Mesh refinement is accomplished by adding new energies contours rather than simply refining element edges. Sharp cusps are formed in the higher energy bands, due to band crossings, which produce discontinuous changes in $\nabla E(\vec{k})$. These features significantly effect the accuracy of any polynominal approximation to the energy bands. As with any mesh generation code, the grid should consist of reasonably shaped tetrahedra.

Table 1: Tetrahedra Grid Summary

Band	Vertices	Tetrahedra	Contours	Quality
1	349	1198	24	0.567
2	530	2268	19	0.611
3	414	1617	18	0.590
4	1343	6143	16	0.529
5	1658	8077	17	0.564
6	613	2431	13	0.456
7	419	1542	11	0.421

Critical points and ridge lines play an essential role the production of accurate contour-aligned grids. The role of critical points, $\vec{k}$ such that$\nabla E(\vec{k}) = 0$, is twofold. Local extrema make excellent vertices, especially for coarse piecewise linear approximations. Also, saddle points describe how components of contours split and connect [4, 9, 8, 10]. The role of ridge lines, extrema in the curvature of the isosurface, is used to decompose the space between two contour surfaces. When a ridge line connects, say, a minimum to an adjacent saddle point, the "watershed" regions between ridge lines are empirically observed to provide a good starting decomposition.

The initial grid is generated from a skeleton of each energy band. The skeleton consists of a set of ridge lines and critical point energies that divide the wedge into several regions. Within each region, vertices are generated along contours that adhere to both the wedge boundaries and ridge lines. Curved region boundaries are represented by piecewise linear line segments with sufficient resolution to capture the geometry of the boundary. Contour-aligned grids are produced by refining grid edges. Edges that cross more then one contour are divided by adding a point at the midpoint contour energy. Edges that lie on a contour are refined either to reduce the discretization error or to improve the quality of the resultant tetrahedra. After all edges are divided at most once, the grid is then retriangled producing a new set of tetrahedra. The overall accuracy of the contour-aligned grids is controlled by selectively adding new energy contours when necessary.

The lowest seven silicon energy bands were computed using the nonlocal pseudopotential of Chelikowsky and Cohen [5] . The first 3 bands correspond to holes while the last 4 bands correspond to electrons. To easily accommodate interband transitions, a single set of energy contours was calculated for all bands. A error tolerance of $\delta_k = 0.02E(\vec{k})$ was used in generating all grids. The low energy spectrum was further selectively refined to produce an error if $\delta = 0.002ev$ near the minimum energy. As shown in Table 1 , accurate, high quality, contour-aligned tetrahedral grids and be obtained with relatively few grid points. The grid for the lowest electron band in the x-y plane of the irreducible wedge is shown in figure 1.

The performance of our algorithm is compared with several existing monte carlo

Table 2: Performance Comparison

Code	Band Structure	CPU Time (YMP)	
		Time Step	Total
IBM	Full	5.0*	100.0*
Illinois	Full	1.5	25.0
BEBOP	Analytic	1.0	2.0
MMC	Full	1.0	1.0

* estimated

codes in Table 2. The test example consisted of simulating 10000 electrons in a constant electric field of $10^5 V/cm$ for a simulation time of 1 picosecond. To obtain a fair comparison, the full band Illinois code [11], the analytic band BEBOP code [12] , and our full band code, MMC, were run on the same machine, a Cray YMP. The CPU times for IBM monte carlo code [7] were taken from published data [6]. These timings are qualitative since a different problem was solved on a different machine. Large time steps may be taken for our simulations and BEBOP, since the equations of motions and the final state energy selection are computed to the machine precision. Small time steps for the other full band codes are necessary to integrate the equations of motion to a reasonable numerical accuracy due to the general band structure. The significant increase in CPU time of the IBM code as compared to the Illinois code is largely attributed to the exhaustive time spent in the final state momentum selection. By any measure, our code combines the rigorous physics of full band simulations with the computational efficiency of analytical band models.

A computationally efficient, full band structure, monte carlo algorithm has been developed. The algorithm is based on a simplex decomposition of phase space, and has all of the advantages of analytic band simulators, while preserving the physics contained in the full band structure. The work required for free-flight computations as well as final state selection is considerably reduced through the use of piecewise linear approximations to both the energy and the electrostatic potential. This method treats motion in k-space and x-space symmetrically allowing exact integrations of the equations of motion and final state selection.

In k-space, the irreducible wedge is represented by a contour-aligned tetrahedral grid. An adaptive grid algorithm is used to generate a mesh that approximates the energy bands to a pre-defined accuracy. Furthermore, the equal-energy surfaces may be readily calculated providing an accurate final k-state selection. As a result, the simulation error is reduced to the monte carlo statistical error and a consistently bounded discretization error deriving from the choice of grid in phase-space.

References

[1] R. E. BANK AND R. K. SMITH, *A posteriori error estimates based on hiearchical bases*, SIAM J. Numer. Anal., (1993).

[2] J. BUDE, E. GROSSE, AND R. K. SMITH, *Highly efficient phase-space simplex full band monte carlo simulations.* to be published.

[3] J. BUDE AND R. K. SMITH, *Phase-space simplex monte carlo for semiconductor transport*, Proceedings of HCIS-8, (1993).

[4] A. CAYLEY, *On contour and slope lines*, Philosophical Magazine, XVIII (1859), pp. 264–268.

[5] J. R. CHELIKOWSKY AND M. L. COHEN, *Nonlocal psuedopotential calculations for the electronic structure of eleven diamond and zinc-blende semiconductors*, Phys. Rev. B, 14 (1976), pp. 556–582.

[6] M. V. FISCHETTI AND S. E. LAUX, *Full band monte carlo program for electrons in silicon*, Phys. Rev B, 38 (1988), pp. 9721–9745.

[7] ——, *Monte carlo analysis in semiconductor devices: the DAMOCLES program*, IBM J. Res. Develop., 34 (1990), pp. 444–494.

[8] C. JOHNSON AND E. H. GROSSE, *Interpolation polynomials, minimal spanning trees and ridge-line analysis in density map interpretation*, American Crystallographic Association Program and Abstracts, 4:2 (1976), p. 48.

[9] S. P. MORSE, *Concepts of use in computer map processing*, Comm. ACM, 12 (1969), pp. 145–152.

[10] J. L. PFALTZ, *Surface networks*, Geographical Analysis, 8 (1976), pp. 77–93.

[11] H. SHICHIJO, J. Y. TANG, J. BUDE, AND D. YODER, *Full band monte carlo program for electrons in silicon*, in Monte Carlo Device Simulation: Full Band and Beyond, K. Hess, ed., 1991.

[12] F. VENTURI, R. K. SMITH, E. SANGIORGI, M. R. PINTO, AND B. RICCO, *A generalized purpose device simulator coupling poisson and monte carlo transport with applications to deep submicron mosfets*, IEEE Trans. CAD, 8 (1989), pp. 360–369.

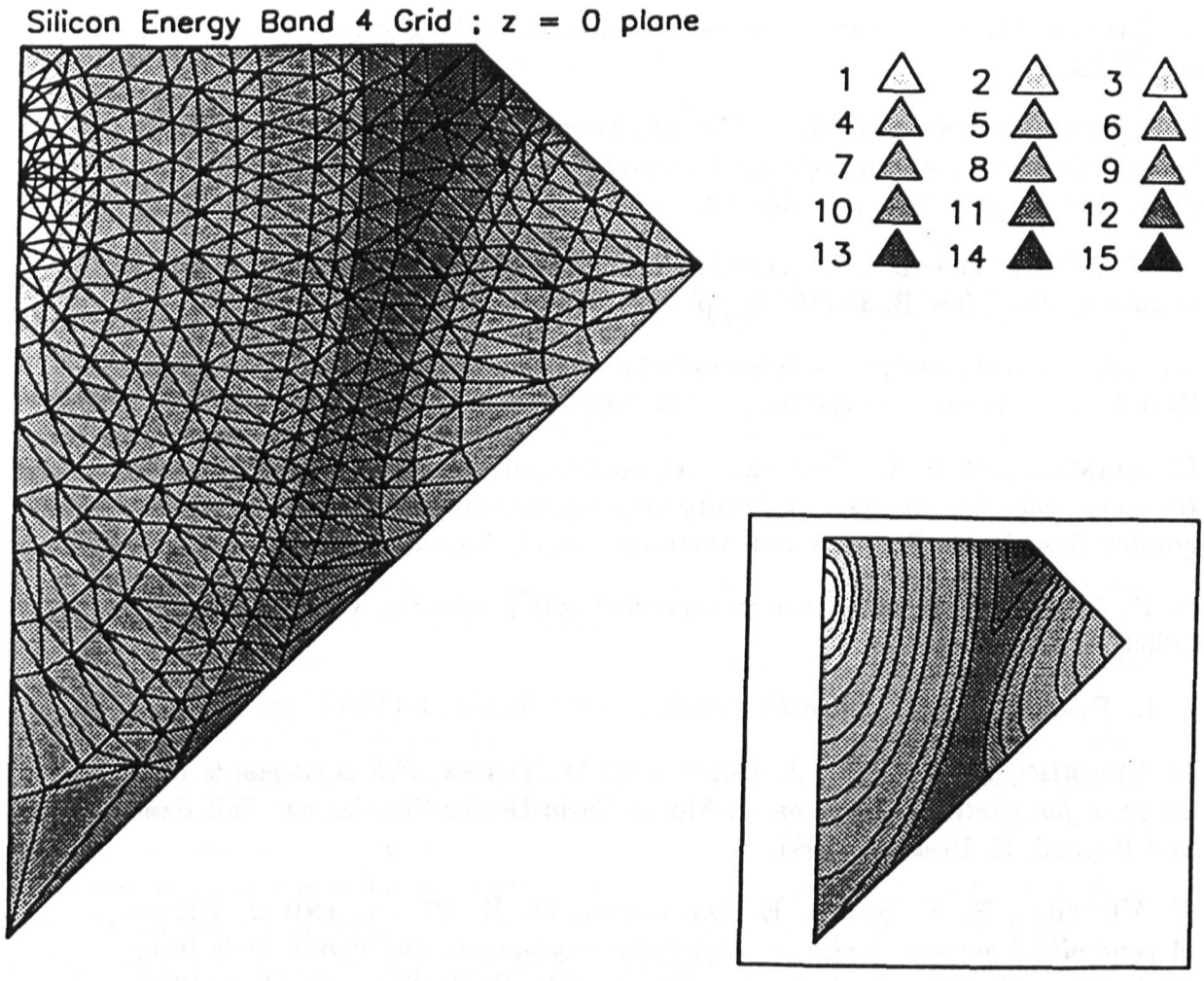

Figure 1: Contour-Aligned Tetrahedral Grid