Parallel Preconditioning for Spherical Harmonics Expansions of the Boltzmann Transport Equation

Karl Rupp*†, Tibor Grasser* and Ansgar Jüngel†
*Institute for Microelectronics, TU Wien. Gußhausstraße 27-29/E360, A-1040 Wien, Austria
†Institute for Analysis and Scientific Computing, TU Wien. Wiedner Hauptstraße 8-10/E101, A-1040 Wien, Austria
Email: {rupp,grasser}@iue.tuwien.ac.at, juengel@asc.tuwien.ac.at

***Abstract*—While the Monte Carlo method for the Boltzmann transport equation for semiconductors has already been parallelized, this is much more difficult to accomplish for the deterministic spherical harmonics expansion method which requires the solution of a linear system of equations. For the typically employed iterative solvers, preconditioners are required to obtain good convergence rates. These preconditioners are serial in nature and cannot be applied efficiently in a black-box manner to arbitrary systems.**

Motivated by the underlying physical processes, we present a parallel block-preconditioning scheme that allows us to use existing serial preconditioners in a parallel setting. A reduction of execution times by up to one order of magnitude on current multi-core processors as well as graphics processing units is observed.

I. Introduction

Since its introduction in the early 1990s, the spherical harmonics expansion (SHE) method has become an attractive alternative to the stochastic Monte Carlo method for the numerical solution of the Boltzmann transport equation (BTE). While the application of the SHE method has long been restricted to one-dimensional device simulations due to high memory requirements, enough memory is available on modern computers to allow for two-dimensional device simulations [1].

With the emergence of parallel computing architectures in desktop computers, parallel algorithms increase in attractiveness. Recently, massively parallel computing architectures in the form of graphics processing units (GPUs) for the use as accelerators have gained a lot of popularity. While fully parallel implementations of the Monte Carlo method have already been reported [2], an artificial restriction of the SHE method to a single CPU core would be detrimental to the attractiveness of the method.

The SHE method ultimately leads to the solution of large systems of linear equations, typically employed within a nonlinear iteration scheme to ensure self-consistency of the BTE with the Poisson equation. Due to the large number of unknowns, iterative solution methods have to be used for the solution of these systems. The convergence rate of such iterative methods can be substantially improved by the use of preconditioners. As discussed by Jungemann *et al.* [3], the system of linear equations resulting from the SHE equations requires a good preconditioner in order to obtain reasonable rates. In recent publications on the SHE method [1], [3], an incomplete LU factorization (ILU) preconditioner was used for that purpose. ILU is a widely accepted black-box preconditioner [4], but it is in its pure form restricted to single-threaded execution. Even though parallel block-variants of ILU as well other parallel preconditioning techniques such as sparse approximate inverses [5] or polynomial preconditioners have been developed, their convergence enhancement can be typically considerably lower than single-threaded variants [4], [6].

The purpose of this work is to show that the preconditioner for the SHE method can be well parallelized. We study the structure of the linear system resulting from the SHE method and propose a general block-preconditioning scheme which can also be used with serial preconditioners. We demonstrate a considerable reduction of execution times on multi-core CPUs as well as on GPUs by employing the initially serial ILU preconditioner within the proposed parallel framework.

II. Physics-Based Block-Preconditioning

In operator form, the SHE equations in steady state can be written as

$$L_{l,m}\{f\} = Q_{l,m}\{f\}\ , \quad l = 0, \dots, L, \quad m = -l, \dots, l\ ,$$

where $L_{l,m}$ and $Q_{l,m}$ denote the projections of the streaming operator and the scattering operator onto the spherical harmonics $Y_{l,m}$, respectively. Employing the H-transform [1], [7], carrier trajectories in free flight are given by hyperplanes of constant total energy H in the simulation domain $(\boldsymbol{x}, H)$, cf. Fig. 1. This is reflected in the model by the fact that $L_{l,m}$ does not couple any of the, say, N_{H} different energy levels in the simulation domain.

Carriers within the device can change their total energy only by inelastic scattering events, thus the scattering operator $Q_{l,m}\{f\}$ is responsible for coupling different energy levels. However, if only elastic scattering processes are considered, the total energy of the involved particles remains unchanged and the different energy levels do not couple. Therefore, in a SHE simulation using only elastic scattering and N_{H} different energy levels, the resulting system of linear equations is consequently decoupled into N_{H} independent problems. Such a decomposition has been observed already in early publications on SHE [8], but it has been of no practical relevance since inelastic scattering processes are essential for predictive device simulation.

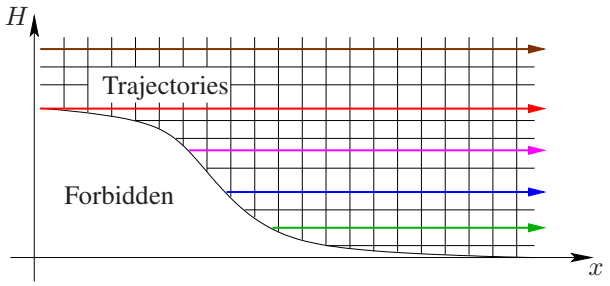


Fig. 1. Trajectories of carriers in free flight within the device are given by constant total energy H.

Inelastic scattering processes like optical phonon scattering couple different energy levels. As devices are scaled down, the average number of scattering events of a carrier while moving through the device decreases. As a consequence, the coupling between different energy levels gets weaker. At the algebraic level this can be reasoned as follows: Using a box integration scheme as proposed by Hong *et al.* [1], the volume integral over the free streaming operator $L_{l,m}$ is transformed to a surface integral due to the divergence operator with respect to the spatial variable $\boldsymbol{x}$. Therefore, if the typical device length d is scaled to $d' := \alpha d$ with $0 < \alpha < 1$, the contributions from the free streaming operator scale as α^{n-1}, where n denotes the spatial dimension considered in the simulation. However, the scattering terms are obtained by an integration over the control volume, which scales as α^n. Therefore, in the limit of extremely scaled devices, the coupling between different energy levels is negligible.

We propose a construction of a preconditioner based on the decoupled problem and using it as an approximation for the coupled problem including inelastic scattering. More precisely, let $\boldsymbol{S}_{\text{full}}$ denote the system matrix of the coupled problem after elimination of the odd order unknowns (cf. [3]) and $\boldsymbol{S}_{\text{elastic}}$ the system matrix of the decoupled problem. Then we construct the preconditioner $\boldsymbol{P}_{\text{full}}$ for $\boldsymbol{S}_{\text{full}}$ as

$$\boldsymbol{P}_{\text{full}} \approx (\boldsymbol{S}_{\text{full}})^{-1} \approx (\boldsymbol{S}_{\text{elastic}})^{-1} \approx \boldsymbol{P}_{\text{elastic}} \ . \qquad (1)$$

Since the elastic problem is decoupled into N_{H} subproblems, $\boldsymbol{S}^{\text{elastic}}$ decomposes into N_{H} independent blocks. For each of these blocks, a (possibly serial) preconditioner can be efficiently set up as well as applied to the residual vector in parallel.

III. Symmetrization of the Scattering Processes

Naturally the scattering rate from higher energy to lower energy is much higher than vice versa. This asymmetry of inelastic scattering processes for energies H_i and H_j, $i < j$, with respect to energy manifests itself in the system matrix in the form of large values in the block with energy index (H_i, H_j), and small entries in the block (H_j, H_i), cf. Fig. 2. Therefore, the upper triangular part of the system matrix is populated with much larger values than the lower triangular part. It should be noted that this asymmetry ensures that the equilibrium solution is a Maxwell (or more generally, a Fermi-Dirac) distribution.

The large values in the upper triangular part of the matrix are a hindrance for the construction of the preconditioner by neglecting off-diagonal blocks. We reduce this asymmetry by rescaling the unknowns of the discrete system according to the expected exponential decay. The new discrete unknowns $f'_{l,m}(\boldsymbol{x}_i, H_i, t)$ are obtained from the old discrete unknowns $f_{l,m}(\boldsymbol{x}_i, H_i, t)$ by

$$f'_{l,m}(\boldsymbol{x}_i, H_i, t) := \exp\left(\frac{\varepsilon_i}{k_{\text{B}}T}\right) f_{l,m}(\boldsymbol{x}_i, H_i, t) \ , \qquad (2)$$

where ε_i denotes the kinetic energy at point $(\boldsymbol{x}_i, H_i)$, k_{B} is the Boltzmann constant and T denotes a scaling temperature which can be seen as a numerical parameter. The benefit of this rescaling is that in equilibrium the primed unknowns are then of similar order and show little to no exponential behavior. We note that this rescaling can be written in matrix form as

$$\boldsymbol{S}\boldsymbol{f} = \boldsymbol{b} \qquad \Leftrightarrow \qquad \boldsymbol{S}\boldsymbol{D}\boldsymbol{f}' = \boldsymbol{b} \ ,$$

where $\boldsymbol{D}$ is a diagonal matrix with the diagonal terms given by the reciprocals of the exponentials in (2). The matrix $\boldsymbol{S}' := \boldsymbol{S}\boldsymbol{D}$ represents the system matrix with symmetrized scattering entries. Here, symmetrization refers to rescaling the unknowns such that the entries in the off-diagonal blocks (H_i, H_j) and (H_j, H_i) are of similar magnitude – it does not mean symmetry of the system matrix in the strict mathematical sense.

IV. Practical Considerations

For the construction of the preconditioner it is not necessary to set up another system matrix $\boldsymbol{S}_{\text{elastic}}$ explicitly. Since the contribution of inelastic scattering operators to the diagonal blocks is positive, it is of advantage to use the block diagonal of $\boldsymbol{S}_{\text{full}}$ for setting up the preconditioner. Thereby, extra memory for a second system matrix is avoided.

It has been observed in numerical experiments that the rescaling of unknowns leads to better results if the temperature T in (2) is set above room temperature. The physical interpretation is that carriers are heated in areas of large electric fields, thus having a lower exponential decay rate, which relates to a higher temperature. Good results are obtained with $T = 400\,\text{K}$ and only a low sensitivity of the number of iterations on the parameter T is observed.

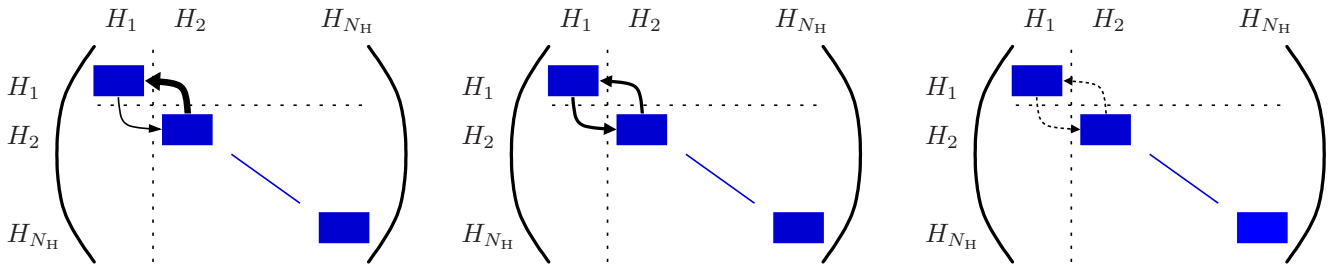
The rows of the system matrix $\boldsymbol{S}'$ can be normalized prior to the block-factorization. This leads to a matrix $\boldsymbol{S}''$ given as

$$\boldsymbol{S}'' = \boldsymbol{E}\boldsymbol{S}' = \boldsymbol{E}\boldsymbol{S}\boldsymbol{D} \ ,$$

where the diagonal matrix $\boldsymbol{E}$ consists of the inverses of the row norms. Thus, a two-sided diagonal preconditioner is applied to the initial system matrix $\boldsymbol{S}$ before launching the block-preconditioning scheme.

V. Results

As a benchmark for the proposed block preconditioner we consider the spatially two-dimensional simulation of an n^+nn^+ diode with different lengths of the intrinsic region. The parallel preconditioning scheme is implemented in our



(a) System matrix structure without any modifications. Scattering from higher to lower energies leads to non-negligible off-diagonal coupling blocks.

(b) System matrix structure after rescaling the unknowns such that solution variables are of similar magnitude.

(c) As devices are scaled down, the entries in off-diagonal blocks become small compared to the entries in the diagonal blocks.

Fig. 2. Structure of the system matrix for total energy levels $H_1 < H_2 < \ldots < H_{N_H}$. Unknowns at the same total energy H_i are enumerated consecutively, leading to a block-structure of the system matrix. For simplicity, scattering only between energy levels H_1 and H_2 is depicted using arrows with thickness proportional to the magnitude of the entries.

simulator ViennaSHE. As a preconditioner for each block, we consider an incomplete LU factorization with threshold (ILUT). The same preconditioner is used as a single-threaded preconditioner, since ILU preconditioners have been employed in other recent works. The stabilized bi-conjugate gradient algorithm (BiCGStab) is used as linear solver, since it provides a lower memory footprint than the GMRES method used in [3].

Execution times of the iterative BiCGStab solver are compared for a single CPU core and for multiple CPU cores of a quad-core Intel Core i7 960 CPU with eight logical cores. In addition, comparisons for an NVIDIA Geforce GTX 580 GPU are found in Figs. 3 and 4. The GPU is programmed and accessed using our OpenCL-based library ViennaCL [9], [10], which also provides the iterative BiCGStab solver.

As can be seen in Figs. 3 and 4, the performance increase for each linear solver step is more than one order of magnitude compared to the single-core implementation. This super-linear scaling with respect to the number of cores on the CPU is due to the better caching possibilities obtained by the higher data locality within the block-preconditioner.

The required number of iterations using the block-preconditioner decreases with the device size. For a 25 nm intrinsic region, the number of iterations is only twice than that of an ILUT preconditioner for the full system. At an intrinsic region of 200 nm, four times the number of iterations are required. This is a very small price to pay for the excellent parallelization possibilities.

Overall, the multi-core implementation is by a factor of three to ten faster than the single core-implementation even though a slightly larger number of solver iterations is required. The purely GPU-based solver with hundreds of simultaneous lightweight threads is by up to one order of magnitude faster than the single-core CPU implementation.

A comparison of Figs. 3 and 4 further shows that the SHE order does not have a notable influence on the block-preconditioner efficiency compared to the full preconditioner. The slightly larger number of solver iterations for third order expansions is due to the higher number of unknowns in the linear system. The performance gain is almost uniform over the length of the intrinsic region and slightly favors shorter devices, thus making the scheme an ideal candidate for current and future scaled-down devices.

VI. Conclusions

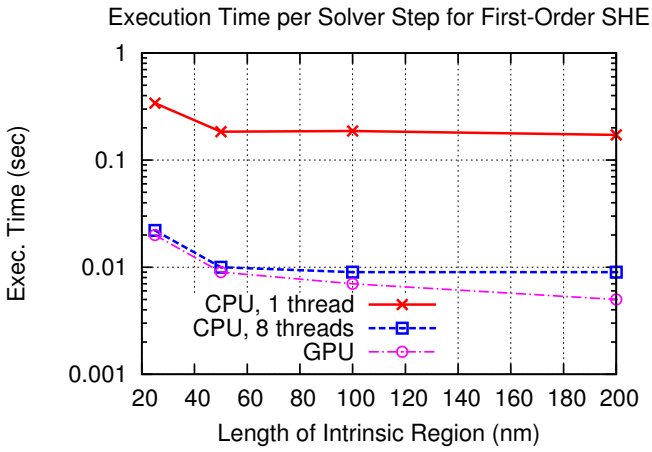
A parallel block-preconditioning scheme is proposed and demonstrated to be very efficient especially for scaled-down devices. In contrast to black-box block preconditioners, the proposed scheme is based on a sound physical principle. The number of iterations compared to a single-threaded ILUT preconditioner for the full system matrix is two to three times as large, but this is only a minor price to pay for the huge degree of parallelism provided for the crucial preconditioning step. On the whole, an overall performance improvement of one order of magnitude is obtained.

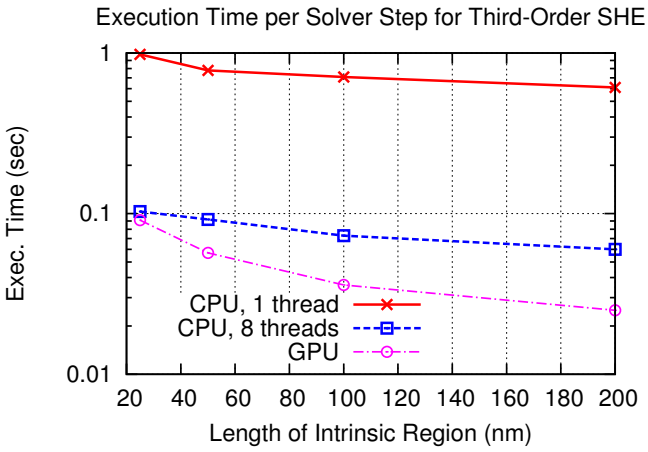
Acknowledgment

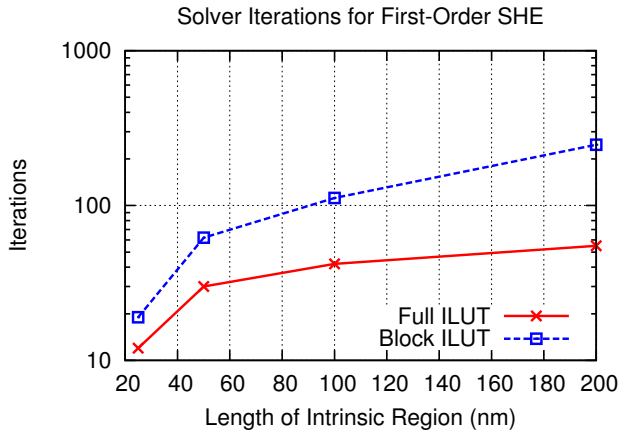
Karl Rupp and Ansgar Jüngel acknowledge support from the Austrian Science Fund (FWF), grants I395 and P20214. The authors gratefully acknowledge support by the Graduate School PDEtech at the TU Wien.

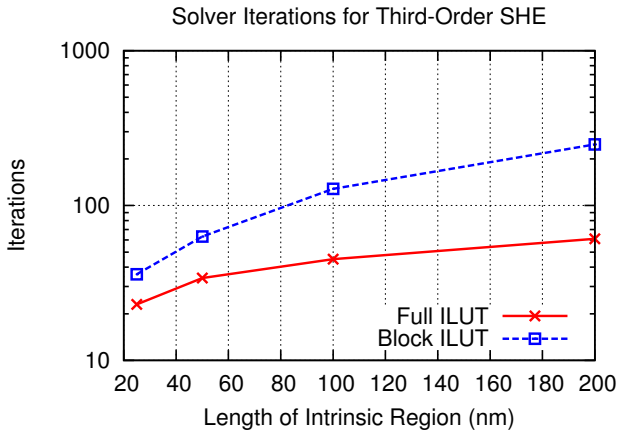
References

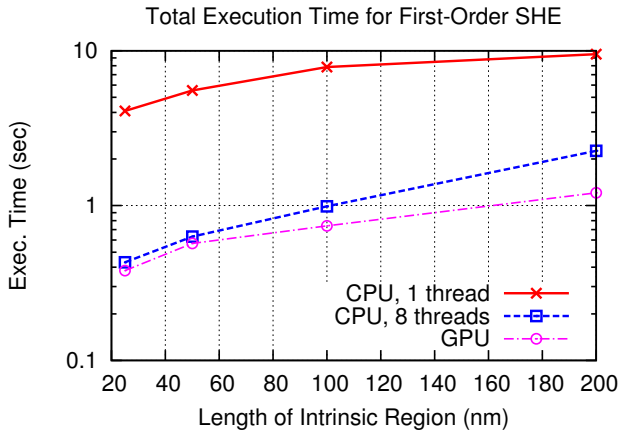
[1] S. M. Hong and C. Jungemann, A Fully Coupled Scheme for a Boltzmann-Poisson Equation Solver based on a Spherical Harmonics Expansion. *J. Comp. Elec.*, vol. 8, p. 225–241 (2009).
[2] W. Zhang et. al., A 3D Parallel Monte Carlo Simulator for Semiconductor Devices. *Proc. IWCE 2009*, p. 1–4 (2009).
[3] C. Jungemann et. al., Stable Discretization of the Boltzmann Equation based on Spherical Harmonics, Box Integration, and a Maximum Entropy Dissipation Principle. *J. Appl. Phys.*, vol. 100, no. 2, p. 024502-+ (2006).
[4] Y. Saad, *Iterative Methods for Sparse Linear Systems, Second Edition*, SIAM (2003).
[5] M. J. Grote and T. Huckle, Parallel Preconditioning with Sparse Approximate Inverses. *SIAM J. Sci. Comp.*, vol. 18, no. 3, p. 838–853 (1997).
[6] P. S. Vassilevski, *Multilevel Block Factorization Preconditioners*, Springer (2008).
[7] A. Gnudi *et al.*, Two-Dimensional MOSFET Simulation by Means of a Multidimensional Spherical Harmonics Expansion of the Boltzmann Transport Equation. *S. S. Electr.*, vol. 36, no. 4 p. 575–581 (1993).
[8] Gnudi, A. *et al.*, One-Dimensional Simulation of a Bipolar Transistor by means of Spherical Harmonics Expansion of the Boltzmann Transport Equation. *Proc. SISDEP*, vol. 4, p. 205–213 (1991).
[9] Khronos Group. OpenCL. `http://www.khronos.org/opencl/`.
[10] ViennaCL. `http://viennacl.sourceforge.net/`.











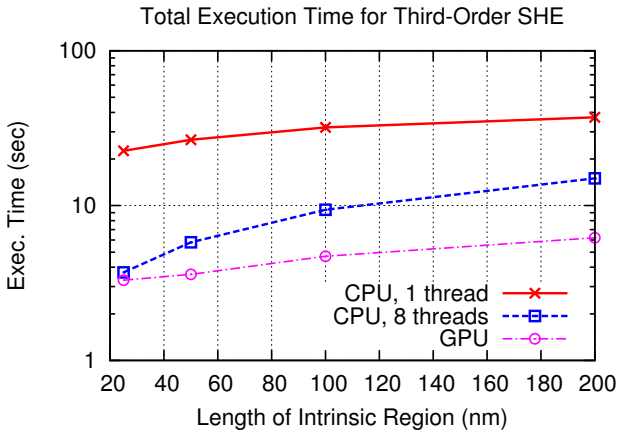


Fig. 3. Execution times per solver iteration, number of solver iterations and total solver execution time for a first-order SHE simulation of n^+nn^+ diodes with different lengths of the intrinsic region. A reduction of total execution times compared to a single-threaded implementation by one order of magnitude is obtained.

Fig. 4. Execution times per solver iteration, number of solver iterations and total solver execution time for a third-order SHE simulation of n^+nn^+ diodes with different lengths of the intrinsic region. Similar to first-order expansions, a reduction of execution times up to one order of magnitude with respect to a single-threaded implementation is obtained.