System Matrix Compression for Spherical Harmonics Expansions of the Boltzmann Transport Equation

Karl Rupp[1], Tibor Grasser[2] and Ansgar Jüngel[3]
[1] Christian Doppler Laboratory for Reliability Issues in Microelectronics
at the Institute for Microelectronics, TU Wien, Vienna, Austria
[2] Institute for Microelectronics, TU Wien, Vienna, Austria
[3] Institute for Analysis and Scientific Computing, TU Wien, Vienna, Austria
Email: {rupp,grasser}@iue.tuwien.ac.at, juengel@asc.tuwien.ac.at

Abstract

Due to its deterministic nature, the spherical harmonics expansion of the Boltzmann transport equation is an attractive alternative to the Monte Carlo method for the purpose of electronic device simulation. The major drawback when using higher order expansions is the huge memory requirement, especially for two- and three-dimensional simulations. We propose a method to compress the resulting system of linear equations, such that memory requirements are reduced by up to two orders of magnitude. In that context we discuss criteria for the selection of an appropriate linear equation solver and show that execution times for matrix-vector multiplications using the compressed matrix scheme on a single CPU core are comparable to that of an uncompressed system matrix. Numerical results demonstrate the applicability of our method and confirm our theoretical results.

Introduction

As long as quantum mechanical effects are negligible, the microscopic behavior of electrons is very well described by a distribution function $f(\boldsymbol{x}, \boldsymbol{k}, t)$ that depends on the spatial coordinate $\boldsymbol{x} = (x, y, z)$, the momentum $\hbar\boldsymbol{k} = (\hbar k_x, \hbar k_y, \hbar k_z)$ and time t, and fulfills the Boltzmann transport equation (BTE). The most commonly used method to solve the BTE is the Monte Carlo method, with the main disadvantage of its computational expense, especially when attempting to reduce the statistical noise in the low density tails of the distribution function. The most prominent alternative to the stochastic Monte Carlo (MC) method is the deterministic spherical harmonics expansion (SHE) method [1]. Traditionally, only expansion up to first order have been employed [1], [2], which gives reasonable but not perfect agreement with MC simulations. Recent results demonstrate that higher order expansions, e.g. of order nine, result in excellent agreement with MC simulations, while maintaining the performance benefit [3], [4].

The major challenge of the SHE method is the huge memory consumption reported even for two-dimensional devices [3] at moderate expansion orders, which has so far prohibited an application of the SHE method to three-dimensional simulations. To overcome these limitations, we present a new system matrix compression scheme that reduces the memory requirements by up to two orders of magnitude and paves the way for three-dimensional device simulations using the SHE method.

The Projected Equations

After a truncated expansion of the distribution function into real valued, orthonormal spherical harmonics $Y_{l,m}(\theta, \varphi)$ up to order L,

$$f(\boldsymbol{x}, \varepsilon, \theta, \varphi, t) \approx \sum_{l=0}^{L} \sum_{m=-l}^{l} f_{l,m}(\boldsymbol{x}, \varepsilon, t) Y_{l,m}(\theta, \varphi) \ , \quad (1)$$

a spherical projection of the BTE and application of the H-transform [2], one obtains with Einstein's summation convention a system of coupled partial differential equations with shifted arguments [3]

$$\begin{aligned} &\frac{\partial f_{l,m} Z}{\partial t} + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{v}_{l,m}^{l',m'} f_{l',m'} Z - \boldsymbol{F} \cdot \boldsymbol{\Gamma}_{l,m}^{l',m'} f_{l',m'} Z \\ &= \sum_{\eta} s_{l,m}^{l',m';\mathrm{in}} f_{l',m'}(\boldsymbol{x}, \varepsilon \mp \hbar\omega_\eta, t) Z(\boldsymbol{x}, \varepsilon \mp \hbar\omega_\eta, t) \\ &\quad - s_{l,m}^{l',m';\mathrm{out}} f_{l',m'} Z \end{aligned}$$

for all $l \in \{0, \ldots, L\}$, $m \in \{-l, \ldots, l\}$. The generalized density of states Z depends on the band structure, $\boldsymbol{F}$ is the force and $s_{l,m}^{l',m';\mathrm{in}}$ and $s_{l,m}^{l',m';\mathrm{out}}$ denote the in- and out-scattering coefficients.

If all coupling coefficients $\boldsymbol{v}_{l,m}^{l',m'}$, $\boldsymbol{\Gamma}_{l,m}^{l',m'}$, $s_{l,m;\eta}^{l',m';\mathrm{in}}$ and $s_{l,m}^{l',m';\mathrm{out}}$ were multiples of the Kronecker delta $\delta_{l,l'}\delta_{m,m'}$, all equations would be decoupled and could be solved individually. Conversely, nonzero coupling coefficients for all quadruples (l, m, l', m') indicate a tight coupling, which usually complicates the solution process. This is in analogy to systems of linear equations: If the system matrix is diagonal, the solution is found immediately, but if the matrix is dense, typically a lot of computational effort is required to solve the system.

It has been shown in [4] that the scattering terms $s_{l,m}^{l',m';\mathrm{in}}$ and $s_{l,m}^{l',m';\mathrm{out}}$ do not couple different expansion coefficients in the case of spherical energy bands. For non-spherical energy bands each expansion coefficient $f_{l,m}$ is coupled with $f_{0,0}$ only, thus the coupling among the equations induced by the scattering operator is very weak.

978-1-4244-7700-5/10/$26.00 ©2010 IEEE

For general band structures, the symmetry of the underlying processes yields that

$$\boldsymbol{v}_{2i,m}^{2i',m'} = \boldsymbol{v}_{2i+1,m}^{2i'+1,m'} = \boldsymbol{0}, \quad \boldsymbol{\Gamma}_{2i,m}^{2i',m'} = \boldsymbol{\Gamma}_{2i+1,m}^{2i'+1,m'} = \boldsymbol{0} \ .$$

for all permissible integers i, i' and m, m' holds [4]. Therefore, all nonzero coupling coefficients possess different parities in the leading indices. This structural information about the coupling was already used in a preprocessing step for the solution of the discretized equations in [4].

Under the assumption of spherical energy bands, i.e. $\varepsilon(\boldsymbol{k}) = \tilde{\varepsilon}(|\boldsymbol{k}|)$, the velocity $\boldsymbol{v}$, the modulus of the wave vector $|\boldsymbol{k}|$ and the generalized density of states only depend on the energy ε, but not on the angles θ, φ. Consequently, we rewrite

$$\begin{aligned} \boldsymbol{v}_{l,m}^{l',m'}(\varepsilon) &= v(\varepsilon) \int Y_{l,m} \boldsymbol{e}_\varepsilon Y_{l',m'} \, \mathrm{d}\Omega \\ &=: v(\varepsilon) \boldsymbol{a}_{l,m}^{l',m'} \ , \end{aligned} \tag{2}$$

$$\begin{aligned} \boldsymbol{\Gamma}_{l,m}^{l',m'}(\varepsilon) &= \frac{1}{\hbar|\boldsymbol{k}|} \int \Big(\frac{\partial Y_{l,m}}{\partial \theta} \boldsymbol{e}_\theta \\ &\qquad + \frac{1}{\sin\theta} \frac{\partial Y_{l,m}}{\partial \varphi} \boldsymbol{e}_\varphi \Big) Y_{l',m'} \, \mathrm{d}\Omega \\ &=: \frac{1}{\hbar|\boldsymbol{k}|} \boldsymbol{b}_{l,m}^{l',m'} \ , \end{aligned} \tag{3}$$

where $\boldsymbol{e}_\theta$ and $\boldsymbol{e}_\varphi$ denote the unit vectors in angular direction. The coupling between index pairs (l,m) and (l',m') is determined by the integral terms $\boldsymbol{a}_{l,m}^{l',m'}$ and $\boldsymbol{b}_{l,m}^{l',m'}$ only. With this it can now be shown that the coupling induced by $\boldsymbol{v}_{l,m}^{l',m'}$ and $\boldsymbol{\Gamma}_{l,m}^{l',m'}$ is still weak:

Theorem 1. *Under the assumption of spherical energy bands, the following holds true for indices $l, l' \in \{0, \dots, L\}$, $m \in \{-l, \dots, l\}$ and $m' \in \{-l', \dots, l'\}$:*

1) *If $\boldsymbol{v}_{l,m}^{l',m'}$ is nonzero, then $l \in \{l' \pm 1\}$ and $m \in \{\pm|m'| \pm 1, m'\}$.*
2) *If $\boldsymbol{\Gamma}_{l,m}^{l',m'}$ is nonzero, then $l \in \{l' \pm 1\}$ and $m \in \{\pm|m'| \pm 1, m'\}$.*

A proof is given in [5]. The theorem allows one to better eliminate those coefficients $\boldsymbol{v}_{l,m}^{l',m'}$ and $\boldsymbol{\Gamma}_{l,m}^{l',m'}$, which may not vanish in simulations due to numerical noise, even though they are analytically zero.

Discretization and System Matrix Compression

In steady state, a discretization of the expansion coefficients on a staggered grid (cf. [6]) is obtained by a Galerkin method

$$f_{l,m}^{\mathrm{even}} = \sum_{i=1}^{N^{\mathrm{even}}} \alpha_{i;l,m}^{\mathrm{even}}(t) \varphi_i^{\mathrm{even}}(\boldsymbol{x}, H) \ , \tag{4}$$

$$f_{l,m}^{\mathrm{odd}} = \sum_{i=1}^{N^{\mathrm{odd}}} \alpha_{i;l,m}^{\mathrm{odd}}(t) \varphi_i^{\mathrm{odd}}(\boldsymbol{x}, H) \ , \tag{5}$$

where H denotes the total energy. This leads to a system of linear equations represented by a system matrix $\boldsymbol{S}$ of size $(N^{\mathrm{even}} + N^{\mathrm{odd}})(L+1)^2 \times (N^{\mathrm{even}} + N^{\mathrm{odd}})(L+1)^2$. The sparsity of $\boldsymbol{S}$ is clear due to the local support of the basis functions $\varphi_i^{\mathrm{even}}(\boldsymbol{x}, H)$ and $\varphi_i^{\mathrm{odd}}(\boldsymbol{x}, H)$.

Using the results of Theorem 1, it can be shown [5] that the number of entries in each row of $\boldsymbol{S}$ is at most $11 C_{\mathrm{sparse}}$, where C_{sparse} is a constant that depends only on the regularity of the underlying mesh. Hence, with the typical values $L = 9$ and $C_{\mathrm{sparse}} = 10$, there are at most $11000(N^{\mathrm{even}} + N^{\mathrm{odd}})$ entries in $\boldsymbol{S}$, which still prohibits sufficiently fine grids for three-dimensional device simulations. If there were a dense coupling among the spherical harmonics expansion equations, the number of entries per row in $\boldsymbol{S}$ would be $\mathcal{O}(L^2 C_{\mathrm{sparse}}) = \mathcal{O}(L^2)$, so the memory requirements in this example would have been two orders of magnitude larger.

For spherical energy bands, the decoupling (2) and (3) allows to separate the spatial coupling due to the overlap of basis functions $\varphi_i^{\mathrm{even}}(\boldsymbol{x}, H)$ and $\varphi_i^{\mathrm{odd}}(\boldsymbol{x}, H)$ from the coupling of different expansion coefficients. This allows one to write the resulting system matrix as

$$\begin{aligned} \boldsymbol{S} &= \begin{pmatrix} \boldsymbol{S}^{\mathrm{ee}} & \boldsymbol{S}^{\mathrm{eo}} \\ \boldsymbol{S}^{\mathrm{oe}} & \boldsymbol{S}^{\mathrm{oo}} \end{pmatrix} \\ &= \sum_{i=1}^{8} \begin{pmatrix} \boldsymbol{Q}_i^{\mathrm{ee}} \otimes \boldsymbol{R}_i^{\mathrm{ee}} & \boldsymbol{Q}_i^{\mathrm{eo}} \otimes \boldsymbol{R}_i^{\mathrm{eo}} \\ \boldsymbol{Q}_i^{\mathrm{oe}} \otimes \boldsymbol{R}_i^{\mathrm{oe}} & \boldsymbol{Q}_i^{\mathrm{oo}} \otimes \boldsymbol{R}_i^{\mathrm{oo}} \end{pmatrix}, \end{aligned} \tag{6}$$

where $\otimes$ denotes the Kronecker product. The dimensions of the matrices $\boldsymbol{Q}_i^{\mathrm{ee}}$, $\boldsymbol{Q}_i^{\mathrm{eo}}$, $\boldsymbol{Q}_i^{\mathrm{oe}}$, $\boldsymbol{Q}_i^{\mathrm{oo}}$ only depend on the number of degrees of freedom in $(\boldsymbol{x}, H)$-space, while the dimensions of the matrices $\boldsymbol{R}_i^{\mathrm{ee}}$, $\boldsymbol{R}_i^{\mathrm{eo}}$, $\boldsymbol{R}_i^{\mathrm{oe}}$, $\boldsymbol{R}_i^{\mathrm{oo}}$ are determined by the spherical harmonics expansion order L only.

This allows for a representation of $\boldsymbol{S}$ using only $\mathcal{O}((L+1)^2 + C_{\mathrm{sparse}} N)$ numbers. Since N is typically much larger than $(L+1)^2$, the full system matrix can be stored for $C_{\mathrm{sparse}} = 10$ with roughly $80N$ numbers, which means a reduction by a factor 137 compared to the uncompressed case.

In the case of non-spherical bands, the velocity and the modulus of the wave vector depend on the energy *and* on the angles. In order to decouple the radial energy contributions from the angular ones, we perform a spherical expansion up to order L' of the coupled terms in the integrands by approximating [4]

$$\boldsymbol{v}(\varepsilon, \theta, \varphi) \approx \sum_{l''=0}^{L'} \sum_{m''=-l''}^{l''} \boldsymbol{v}^{l'',m''}(\varepsilon) Y_{l'',m''}(\theta, \varphi) \ , \tag{7}$$

$$\frac{1}{\hbar|\boldsymbol{k}(\varepsilon, \theta, \varphi)|} \approx \sum_{l''=0}^{L'} \sum_{m''=-l''}^{l''} \Gamma^{l'',m''}(\varepsilon) Y_{l'',m''}(\theta, \varphi) \ , \tag{8}$$

where the expansion coefficients are given by

$$\begin{aligned} \boldsymbol{v}^{l'',m''}(\varepsilon) &= \int \boldsymbol{v}(\varepsilon, \theta, \varphi) Y_{l'',m''}(\theta, \varphi) \, \mathrm{d}\Omega \ , \\ \Gamma^{l'',m''}(\varepsilon) &= \int \frac{1}{\hbar|\boldsymbol{k}(\varepsilon, \theta, \varphi)|} Y_{l'',m''}(\theta, \varphi) \, \mathrm{d}\Omega \ . \end{aligned}$$

For simplicity, the expansion order L' is the same for both $\boldsymbol{v}_{l,m}^{l',m'}$ and $\boldsymbol{\Gamma}_{l,m}^{l',m'}$ and has to be chosen such that the complexity of the band structure is captured. Values of L' in the range five to ten are expected to be sufficient to obtain a good approximation of the non-spherical bands of interest. In this way, certain full-band effects can also be considered in the deterministic SHE approach [7].

Similar to the case of spherical energy bands, the system matrix $\boldsymbol{S}$ can then be written in the form

$$\boldsymbol{S} = \sum_{i=1}^{3+6(L'+1)^2} \begin{pmatrix} \boldsymbol{Q}_i^{\mathrm{ee}} \otimes \boldsymbol{R}_i^{\mathrm{ee}} & \boldsymbol{Q}_i^{\mathrm{eo}} \otimes \boldsymbol{R}_i^{\mathrm{eo}} \\ \boldsymbol{Q}_i^{\mathrm{oe}} \otimes \boldsymbol{R}_i^{\mathrm{oe}} & \boldsymbol{Q}_i^{\mathrm{oo}} \otimes \boldsymbol{R}_i^{\mathrm{oo}} \end{pmatrix} , \quad (9)$$

which allows one to store the system matrix using only $\mathcal{O}(L'^2(L^4+N))$ matrix entries. The term L^4 stems from the conservative assumption that the coupling among the SHE equations might be dense for non-spherical energy bands. However, the number of grid points N is typically larger than $L^4 \approx 10\,000$ for $L \approx 10$, so even a dense coupling does not influence the total memory requirements considerably, since the number of grid points N is typically still larger than $10\,000$ and thus the dominant term.

Solution of the Linear System

The matrix compression scheme is of use only if the resulting system of linear equations can be solved without the need to recover the full matrix again. Such a reconstruction is, in principle, necessary if direct solvers such as the Gauss algorithm are used, because during the solution process the matrix structure is altered in a way that destroys the block structure. For many popular iterative solvers from the family of Krylov methods, it is usually sufficient to provide matrix-vector multiplications [8], [9].

Matrix-vector products for a matrix given as a Kronecker product of two smaller matrices can be carried out in a straightforward manner by decomposing the vector into blocks of suitable size. This allows for the realization of a very memory efficient matrix-vector multiplication if the system matrix is given in the form (6) or (9).

However, the full system matrix for the even and odd order expansions coefficients was found by numerical experiments to be ill-conditioned. A substantial improvement of the system matrix condition number can be obtained if the unknowns for the odd order expansion coefficients are eliminated in a preprocessing step. This is possible since $\boldsymbol{S}^{\mathrm{oo}}$ as in (6) is a diagonal matrix [5]. However, a direct elimination by altering the system matrix is not possible without destroying the Kronecker product structure. This can be avoided by using the Schur complement. Writing the system as

$$\boldsymbol{S}\boldsymbol{f} = \begin{pmatrix} \boldsymbol{S}^{\mathrm{ee}} & \boldsymbol{S}^{\mathrm{eo}} \\ \boldsymbol{S}^{\mathrm{oe}} & \boldsymbol{S}^{\mathrm{oo}} \end{pmatrix} \begin{pmatrix} \boldsymbol{f}^{\mathrm{e}} \\ \boldsymbol{f}^{\mathrm{o}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{r}^{\mathrm{e}} \\ \boldsymbol{r}^{\mathrm{o}} \end{pmatrix} \quad (10)$$

with the vector of unknowns $\boldsymbol{f}$ split into $\boldsymbol{f}^{\mathrm{e}}$ and $\boldsymbol{f}^{\mathrm{o}}$ as unknowns associated with even and odd order harmonics respectively and analogously for the right hand side vector $\boldsymbol{r}$, the elimination of odd order unknowns using the Schur complement leads to

$$(\boldsymbol{S}^{\mathrm{ee}} - \boldsymbol{S}^{\mathrm{eo}}(\boldsymbol{S}^{\mathrm{oo}})^{-1}\boldsymbol{S}^{\mathrm{oe}})\boldsymbol{f}^{\mathrm{e}} = \boldsymbol{r}^{\mathrm{e}} - \boldsymbol{S}^{\mathrm{eo}}(\boldsymbol{S}^{\mathrm{oo}})^{-1}\boldsymbol{r}^{\mathrm{o}} . \quad (11)$$

Therefore, the system matrix compression scheme can also benefit from the improved system matrix condition number after elimination of the odd order unknowns.

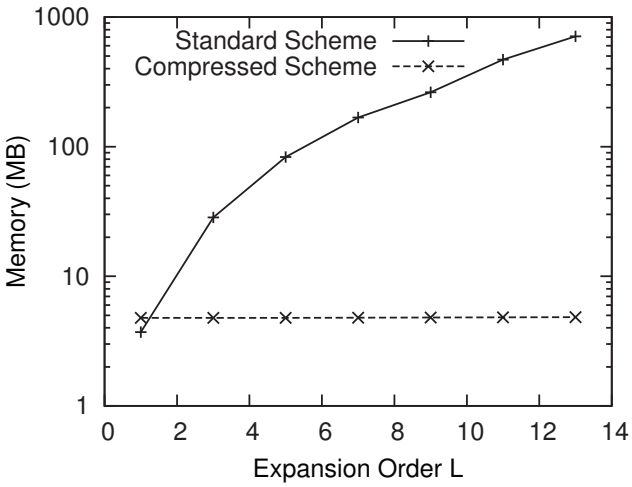

Fig. 1. Memory used for the uncompressed and the compressed system matrix for different expansion orders L on a three-dimensional $(\boldsymbol{x}, H)$-grid with $12\,500$ nodes.

The total memory needed for the SHE equations is essentially given by the memory required for the unknowns, which adds another perspective on the selection of the iterative solver. Recent publications on SHE simulations [3], [4] used GMRES [8], which is typically restarted after, say, s steps, denoted by GMRES(s). If the system matrix is uncompressed, the additional memory needed for GMRES(s) is approximately the amount of memory needed for the storage of the system matrix; thus, it is not a major concern. However, using the proposed matrix compression scheme, the memory of order $\mathcal{O}(NL^2)$ needed for the unknowns is dominant, so the additional memory for GMRES(s) of order $\mathcal{O}(sNL^2)$ makes the method very unattractive from the memory consumption point of view when used with the typical values of s in the range 20 to 30. Therefore we conclude that either GMRES(s) with small values of s or other iterative solvers with smaller memory consumption such as BiCGStab [9] should be used for SHE simulations that rely on a compressed matrix scheme.

Results

We have compared memory requirements for the storage of the system matrix at several expansion orders in a two-dimensional device simulation. The results in Fig. 1 clearly demonstrate the asymptotic advantage of our approach: Already at an expansion order of $L = 5$, memory savings by a factor of 18 are observed, which increases to 146 at $L = 13$. With the compressed scheme, the memory required

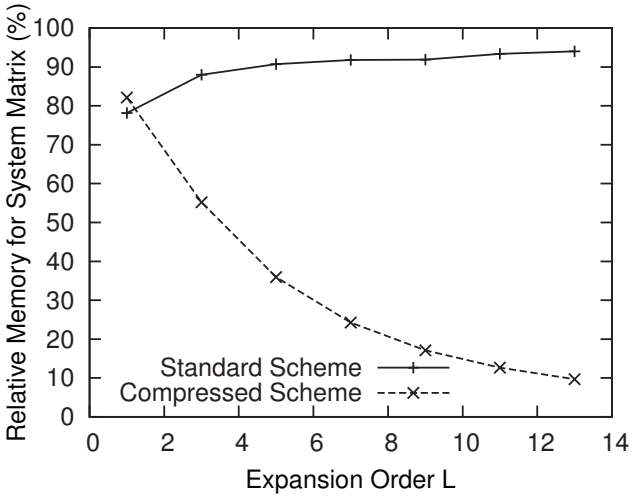

Fig. 2. Memory used for the system matrix in relation to the total amount of memory used (i.e. system matrix, unknowns and right hand side).

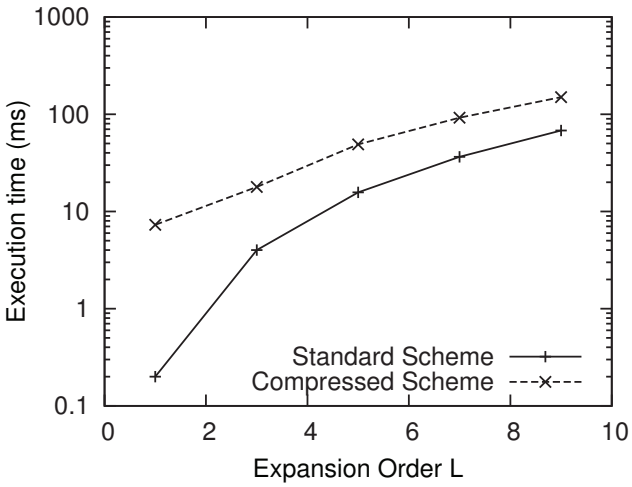

Fig. 3. Execution times for matrix-vector multiplication with the Schur complement $(\boldsymbol{S}^{\mathrm{ee}} - \boldsymbol{S}^{\mathrm{eo}}(\boldsymbol{S}^{\mathrm{oo}})^{-1}\boldsymbol{S}^{\mathrm{oe}})$ on a single core of an Intel Core 2 Quad Q9550 CPU.

for the system matrix increases only by a few kilobytes as L increases, which is negligible.

Since the memory required by the system matrix is of order $\mathcal{O}(N + L^2)$ and the memory for the unknowns is of order $\mathcal{O}(NL^2)$, the memory required for the unknowns is much larger than the memory required for the representation of the system matrix for large values of L, cf. Fig. 2. Therefore, the asymptotic memory requirements for the full simulation is still $\mathcal{O}(NL^2)$, but the constant of proportionality is of order one, while for the full system matrix it is around $11C_{\mathrm{sparse}} \approx 100$, so a reduction of memory requirements by two orders of magnitude is obtained.

On a single CPU core, the minor price to pay for the dramatic reduction in memory consumption is that the execution times of matrix-vector products with the Schur complement $(\boldsymbol{S}^{\mathrm{ee}} - \boldsymbol{S}^{\mathrm{eo}}(\boldsymbol{S}^{\mathrm{oo}})^{-1}\boldsymbol{S}^{\mathrm{oe}})$ increase by a factor of about two, cf. Fig. 3. However, the proposed matrix compression scheme is very well suited for parallel architectures, because the data required for the system matrix may even fit into the CPU caches, allowing for a very high performance. Moreover, since the blocks $\boldsymbol{S}^{\mathrm{ee}}$, $\boldsymbol{S}^{\mathrm{eo}}$, $\boldsymbol{S}^{\mathrm{oe}}$ and $\boldsymbol{S}^{\mathrm{oo}}$ are given by sums of Kronecker products, each summand can be computed on a separate core.

Conclusion

The matrix compression scheme presented in this work reduces the memory requirements for the system matrix arising from a SHE of the BTE from order $\mathcal{O}(NL^2)$ to $\mathcal{O}(N + L^2)$, which results in total memory savings for the full simulation run by up to two orders of magnitude. Therefore, our scheme paves the way for three-dimensional device simulations especially for larger expansion orders L. On a single CPU core, the small price to pay is a runtime penalty on matrix-vector multiplication of about a factor of two. However, the proposed method is especially attractive for parallel architectures where it is expected to outperform the traditional storage scheme.

Acknowledgments

Support by the Graduate School PDEtech at the Vienna University of Technology is gratefully acknowledged. A. Jüngel acknowledges partial support from the Austrian Science Fund (FWF), grants P20214, I395, and WK "Differential Equations", and the Austrian Exchange Service (ÖAD), grants ES 08/2010, FR 07/2010, HR 10/2010.

References

[1] N. Goldsman et. al., A Physics-based Analytical/Numerical Solution to the Boltzmann Transport Equation for the Use in Device Simulation. *Solid State Electronics*, vol. 34, p. 389–396 (1991).

[2] A. Gnudi et. al., Two-dimensional MOSFET Simulation by Means of a Multidimensional Spherical Harmonics Expansion of the Boltzmann Transport Equation. *Solid State Electronics*, vol. 36, no. 4 p. 575–581 (1993).

[3] 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).

[4] 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).

[5] K. Rupp et. al., Matrix Compression for Spherical Harmonics Expansions of the Boltzmann Transport Equation for Semiconductors. *submitted*.

[6] C. Ringhofer, Numerical Methods for the Semiconductor Boltzmann Equation Based on Spherical Harmonics Expansions and Entropy Discretizations. *Transport Theory and Statistical Physics*, vol. 31, p. 431–452 (2002).

[7] M. C. Vecchi and M. Rudan, Modeling electron and hole transport with full-band structure effects by means of the Spherical-Harmonics Expansion of the BTE. *IEEE Transactions on Electron Devices*, vol. 45, p. 230–238 (1998).

[8] Y. Saad and M. H. Schultz, GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems. *SIAM J. Sci. Stat. Comput.*, vol. 7, no. 3, p. 856–869 (1986).

[9] H. A. van der Vorst, Bi-CGSTAB: A Fast and Smoothly Converging Variant of Bi-CG for the Solution of Non-Symmetric Linear Systems. *SIAM Journal on Scientific and Statistical Computing*, vol. 12, p. 631–644 (1992).