

A Wavelet Method to Solve High-dimensional Transport Equations in Semiconductor Devices

Vincent Peikert* and Andreas Schenk*

*Integrated Systems Laboratory, ETH Zurich, CH-8092 Zürich, Switzerland

Email: vpeikert@iis.ee.ethz.ch

Abstract—This paper reports the first numerical solver for the Boltzmann transport equation (BTE) that uses wavelets as basis functions. The main advantage of wavelets is that they offer modern compression and adaptation techniques that could cope with the "curse of dimensionality" of the 6-dimensional phase space. An adequate numerical method for the BTE has been developed which combines a conservative discontinuous Galerkin (DG) formulation with a Multi-Wavelets (MW) basis. NIN device simulations in a 3-dimensional phase space prove that the DG formulation performs well together with MWs. On the other hand, it shows that MWs provide a very efficient basis for the BTE. The number of degrees of freedom can be compressed to about 1-10% in comparison to other modern solvers. Even greater advantages are expected for higher-dimensional phase spaces.

I. INTRODUCTION

Transport equations (Boltzmann, Wigner) on 6-dimensional phase spaces describe the behavior of charge carriers in semiconductors and form the basis for Technology Computer Aided Design (TCAD) on device level. However, the numerical solution of these high-dimensional equations involves huge matrices which practically limits contemporary direct solvers to problems with only one real space dimension. Already the simulation of 2-dimensional devices calls for very coarse meshes and geometrical simplicity which sets limitations to the accuracy and the practical application [1], [2]. Therefore, moments of the BTE and the Monte Carlo method are still widely used in commercial TCAD.

To overcome this long-standing obstacle, the techniques of wavelet compression and adaptivity (having increasing advantages in higher dimensions) are applied to the BTE in this work for the first time.

Recently, a mesh generation tool based on a Discrete Wavelet Transformation was successfully used for 3-dimensional semiconductor devices in the drift-diffusion model [3]. Pioneering work in the field of numerical solution of high-dimensional transport equations with wavelets was presented in Ref. [4], where the authors solved the radiative transfer equation by using a L^2 least-square formulation. Such a L^2 ansatz, however, could be expected to perform badly in the case of the BTE, since all regions of low density would be unattended by the L^2 optimization. On the other hand, a local control of the current conservation is essential. For this reason, a numerical method for the BTE was developed in this paper which combines a conservative DG formulation with MWs.

II. MULTI-WAVELET ADAPTIVITY

MWs were constructed for the first time in Ref. [5]. MWs build an orthonormal hierarchical basis for piecewise polynomials. Their most important advantage for the approximation of high-dimensional functions is their compressibility as demonstrated in Fig. 1. The first figure in Fig. 1 shows a function Φ consisting of 16 piecewise constant functions. The number of required coefficients is 16. This function can be uniquely transformed to 16 Haar-Wavelets (zeroth order MWs) Ψ_i ($\Phi = \sum_{i=1}^{16} a_i \Psi_i$, $a_i = \int \Phi \Psi_i$).

Due to the orthonormality, the norm equivalence $|\Phi|_{L^2} = \sum |a_i|^2$ holds so that the smallest Haar coefficients contribute least to the L^2 norm of Φ . The second figure in Fig. 1 shows the 12 Haar-Wavelets with the largest coefficients only. The first important note is, that the compressed function still shows 16 piecewise constants although only 12 coefficients are used. The second note is that the absolute error is small, i.e. the other 4 coefficients could be easily omitted without losing much of the L^2 norm (see Fig. 2).

The reason why MWs need less coefficients to represent Φ is that higher-order MWs only correct, what MWs of lower order already represent. Therefore, the coefficients decay very quickly with the hierarchical order. The size of the coefficients not only determines which coefficients to cut (wavelet compression), but also where additional higher-order wavelets are needed (wavelet refinement). To summarize, MW (hp-)adaptivity offers a lot of potential for the solution of high-dimensional semiconductor transport equations.

On the other hand, Fig. 2 also shows that the relative error is large in regions where Φ is small. Consequently, regions in semiconductor devices with a low density would be much less resolved than regions with a high density. As a first approach to solve that problem, a density-normalized norm as wavelet adaptivity criterion has been applied in this work instead of the usual L^2 norm.

III. DISCONTINUOUS GALERKIN METHOD

Recently, the well-known DG method has been applied to the BTE for the first time in [1]. The DG formulation is based on multi-dimensional integration by parts of the free-streaming operator of the BTE and is particle conserving, if a piecewise polynomial basis is chosen. The stabilization of the hyperbolic free-streaming operator naturally takes place in case of DG by applying a well-defined upwinding scalar flux. This is in

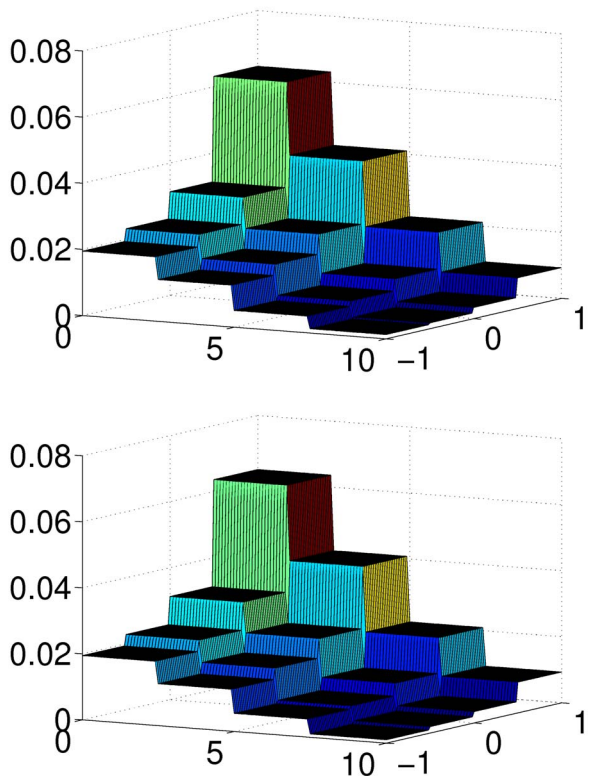


Fig. 1. Multidimensional wavelet compression. From top to bottom: 16 coefficients, 12 coefficients.

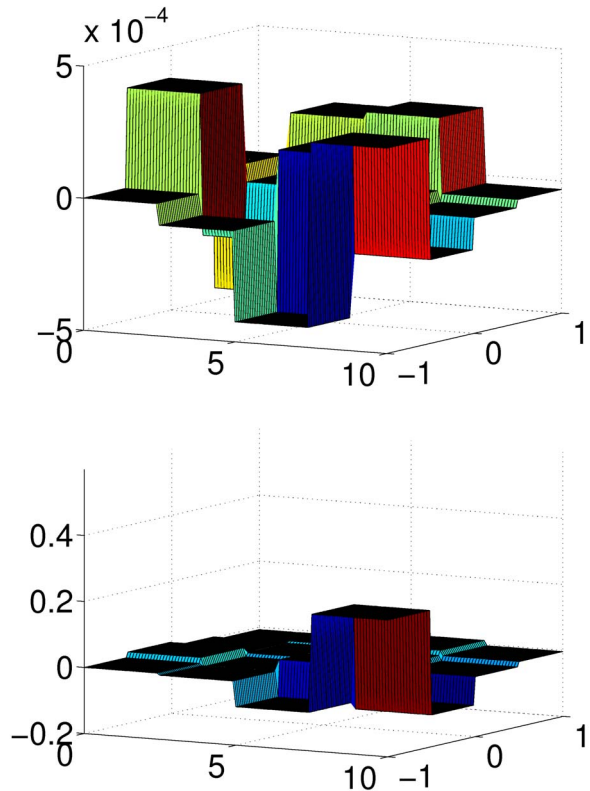


Fig. 2. Multidimensional wavelet compression. From top to bottom: absolute error, relative error.

contrast to the popular spherical harmonics expansion (SHE) where stabilization has to be forced by maximum entropy principles or the upwinding of complicated vector valued fluxes (see e.g. [6] and references there in). Whereas for SHE the boundary conditions are non-trivial, boundary conditions in the case of DG can be easily found by defining the in- and out-fluxes.

Due to numerical and physical reasons, it is common for the direct solution of the BTE to transform the momentum space into spherical coordinates on equi-energy surfaces. Following e.g. the derivations, assumptions and notations in [1] the 2-dimensional BTE can be transformed into the form

$$\frac{\partial}{\partial x}(g_1\Phi) + \frac{\partial}{\partial y}(g_2\Phi) + \frac{\partial}{\partial \omega}(g_3\Phi) + \frac{\partial}{\partial \mu}(g_4\Phi) + \frac{\partial}{\partial \phi}(g_5\Phi) = C(\Phi) \quad (1)$$

, where ω is the energy, μ is the cosine of the polar angle and ϕ is the azimuth angle. This transformation allows for the usage of easy-to-handle tensor grids, although the DG method would also enable the application of unstructured grids.

IV. MULTI-WAVELET DISCONTINUOUS GALERKIN METHOD

The idea of this work is to replace the commonly used piecewise polynomials in a nodal DG (NDG) method by

MWs to obtain a conservative multiscale solver for the BTE. This method is new not only in the TCAD community, but also in the applied mathematics community. Hence, no theoretical framework exists so far and the behavior of the scheme had been vague at the outset. Another application study with a more formal description of the same method (called the Multi-Wavelet discontinuous Galerkin method) has recently been accepted for publication [7] but the idea goes back to a similar method proposed in [8].

Although it is clear that a simple MW transformation of piecewise polynomials leads to an equivalence of the MWDG and the NDG methods, it remained unclear what will happen when wavelets are removed from or added to that multidimensional expansion: Does the MWDG problem keep uniquely solvable? How efficiently does MWDG approximate the coefficients? How many coefficients can be finally saved? It also has to be pointed out that the general local conservation property of the DG method is lost in the adaptive MWDG method. Nonetheless, this violation remains highly controllable and could be the basis for a new adaptive method which only becomes possible by a combination of wavelets and a conservative weak formulation. To the best of the authors knowledge this insight has never been mentioned before.

In one-dimensional devices, Eq. (1) reduces to

$$\frac{\partial}{\partial x}(g_1\Phi) + \frac{\partial}{\partial \omega}(g_3\Phi) + \frac{\partial}{\partial \mu}(g_4\Phi) = C(\Phi). \quad (2)$$

The discretization of the scattering operator is straightforward and produces very sparse matrices in the MW basis. However, as a possible drawback of using MWs in real and energy space instead of piecewise polynomials, the higher coupling between MWs in the free-streaming operator should be mentioned. On the other hand, the resulting matrices still show a quasi-sparse structure due to the vanishing moments of the MWs so that this drawback can be overcome to a certain degree by utilizing wavelet matrix compression techniques (see e.g. [9]).

For Eq. (2), the SHE reduces to the Legendre expansion (LE). It is worth to mention, that MWs can be constructed from Legendre polynomials in their lowest hierarchy so that the LE can be envisaged as a special case of the MW expansion corresponding to a pure p-refinement.

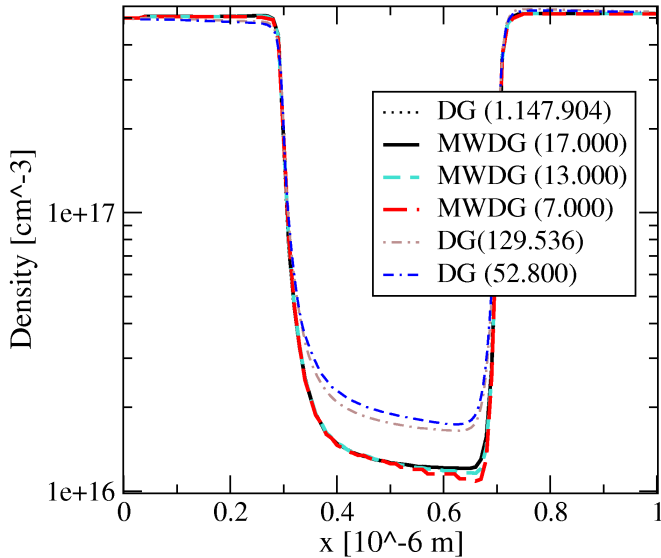


Fig. 3. Density in a 400 nm channel NIN device with 1 V bias. Comparison between NDG and MWDG.

V. SIMULATION RESULTS

For a proof of concept, stationary MWDG frozen-field simulations of a NIN structure at 1 V bias are shown in Figs. 3 - 7. An additional NDG solver had been implemented to allow comparisons with the MWDG method. For all simulations exactly the same formulation of the BTE (variable transformations, physical models, DG weak formulation) as well as the same 400 nm NIN structure as in [1] have been used. The only difference to [1] is that piecewise constant functions were applied instead of piecewise linear functions. Hence, to adapt the well-tested mesh in [1] (which leads to 691.200 degrees of freedom) for piecewise constant NDG, an additional grid point was placed into every mesh gap. Then the mesh was adaptively coarsened by hand until the moments

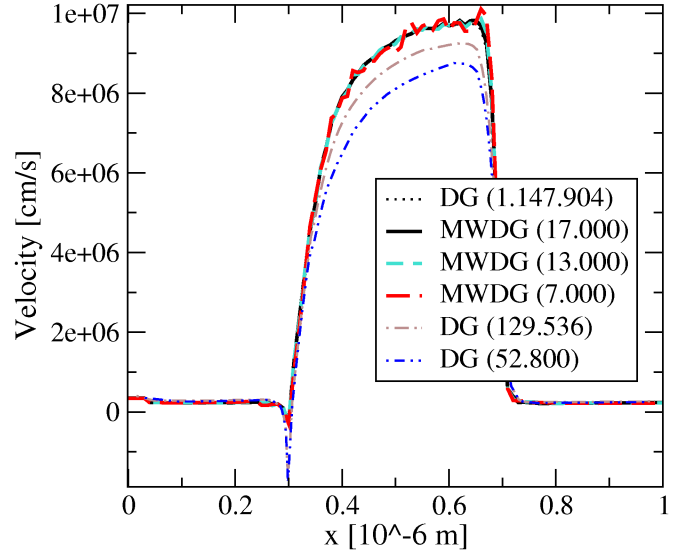


Fig. 4. Velocity in a 400 nm channel NIN device with 1 V bias. Comparison between NDG and MWDG.

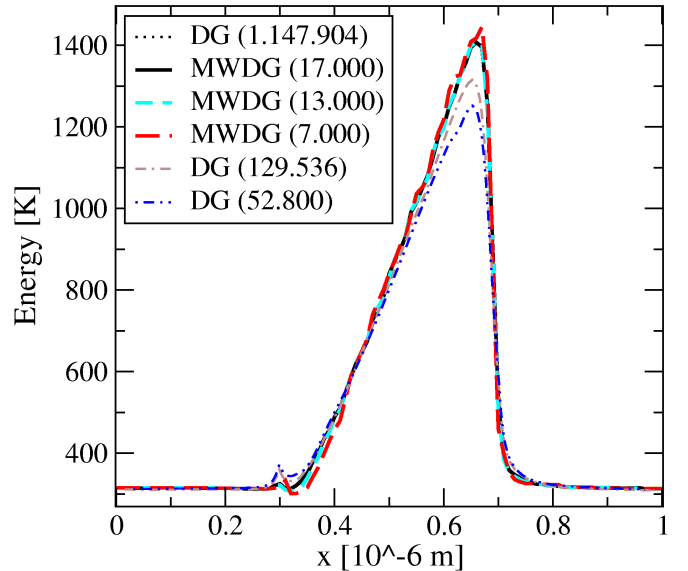


Fig. 5. Energy in a 400 nm channel NIN device with 1 V bias. Comparison between NDG and MWDG.

(density, velocity and energy) deviated by more than 1% from their original values. This led to 1.147.904 degrees of freedom (DoF). Furthermore, the mesh was coarsened two times more which yielded 129.536 and 52.800 DoF, respectively. However, the mesh in the vicinity of the n^+-n junctions was not coarsened in the last step.

In the following, the performance of the MWDG compression is presented which reveals the most critical uncertainties. To do so, the solution of the NDG (1.147.904 DoF) was first transformed to MWs, then the coefficients with the smallest contribution to the density-normalized L^2 norm were skipped and finally, the MWDG was solved with the remaining coefficients.

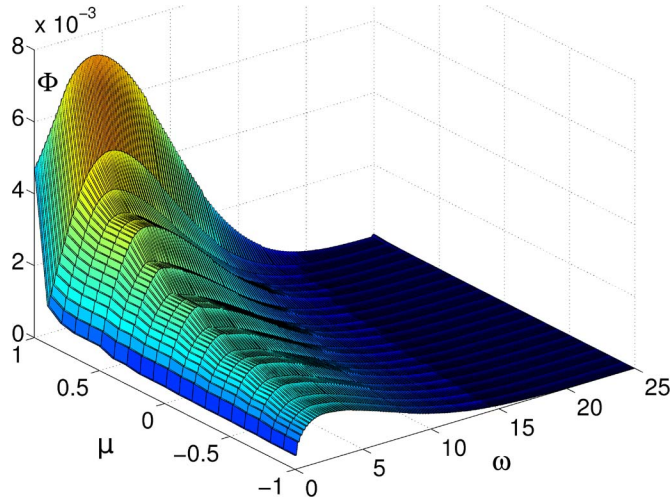


Fig. 6. NDG solution of the BTE with 1.147.904 coefficients in a 400 nm channel NIN device with 1 V bias at $x = 0.6$.

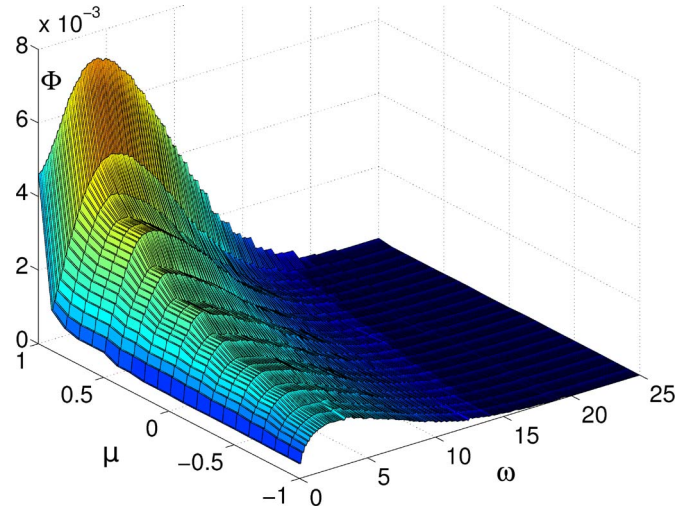


Fig. 7. MWDG solution of the BTE with 17.000 coefficients in a 400 nm channel NIN device with 1 V bias at $x = 0.6$.

Figs. 3 - 5 present the moments of the solution of the BTE at different compression stages.

- The MWDG is uniquely solvable no matter which and how many coefficient are removed from the hierarchical order.
- The figures show that the number of coefficients may be reduced from 1.147.904 DoF to about 13.000 DoF which corresponds to a compression rate of nearly 99%. They also prove that the remaining coefficients do not change much in the MWDG method when other wavelets are removed (or added) which is essential for wavelet adaptivity.
- The solution starts to fluctuate around the true value, if the number of wavelets becomes too small (for 7000 DoF in this case). However, one could still smooth this solution to obtain an acceptable approximation. This is in contrast to the mesh coarsening of the NDG method, where the approximate solution departs from the correct solution.

This outperforming reflects the multiscale nature of MWs.

Figs. 6 and 7 compare the NDG solution (1.147.904 DoF) with the compressed MWDG solution (17.000 DoF) in the wave vector space at $x = 0.6 \mu\text{m}$ (where one of the biggest errors occurs). One can see that the solution becomes rougher, but that it does not drift. Therefore, a smoothing algorithm could help to even enhance the performance of the MWDG method.

VI. CONCLUSION

A new Multi-Wavelet discontinuous Galerkin method for the Boltzmann transport equation has been developed. The method combines the advantages of multiscale wavelet bases with the advantages of local conservative weak formulations. Haar-Wavelet simulations show that the number of coefficients can be compressed to just 1% compared to the standard discontinuous Galerkin method, if a density-normed wavelet adaptivity scheme is applied. Higher-order polynomial Multi-

Wavelets are expected to further increase the compression rate especially in higher dimensions.

ACKNOWLEDGMENT

Partial funding of this project by Toshiba Corp. is gratefully acknowledged.

REFERENCES

- [1] Y. Cheng, I.M. Gamba, A. Majorana, C. Shu, *Comput. Methods Appl. Mech. Engrg.* 198, 3130 (2009).
- [2] S.M. Hong, C. Jungemann, C., M. Bollhöfer, *Proceedings 13th International Conference on Simulation of Semiconductor Processes and Devices, Hakone, Japan (2008)*, pp.293–296.
- [3] L. De Marchi, E. Baravelli, F. Franzè, N. Speciale, *IEEE Trans. Comp-Aid. Design Integr. Circ. A. Systems.* 26, 1967 (2007).
- [4] G. Widmer, R. Hiptmair, C. Schwab, *Journ. Appl. Phys.* 227, 6071 (2008).
- [5] B. Alpert, *SIAM Journ. Math. Anal.* 24, 246 (1993).
- [6] C. Jungemann, A.T. Pham, B. Meinerzhagen, C. Ringhofer, M. Bollhöfer, *Journ. Appl. Phys.* 100, 024502 (2006)
- [7] R. Archibald, G. Fann, W. Shelton, *Adaptive Discontinuous Galerkin Methods with Multiwavelet Bases*, to appear in *Journal of Scientific Computing*, 2011.
- [8] B. Alpert, G. Beylkin, D. Gines, L. Vozovoi, *Journ. Comp. Phys.* 182, 149 (2002).
- [9] A. Cohen, W. Dahmen, R. DeVore, 'Adaptive wavelet techniques in numerical simulation' in *Encyclopedia of Computational Mechanics*, Wiley (2004)