Is there anything left to do in TCAD?

Z. Stanojević, F. Schanovsky, G. Rzepa, X. Klemenschits, H. Demel, O. Baumgartner,
C. Kernstock, and M. Karner
Global TCAD Solutions GmbH., Bösendorferstraße 1/12, 1010 Vienna, Austria
Email: {z.stanojevic|…|m.karner}@globaltcad.com

***Abstract*—Over the past decade, the development of commercial technology computer-aided design (TCAD) software has followed an evolutionary rather than revolutionary path. Alongside established continuum and particle-based approaches in both process and device simulation, advanced carrier transport models – such as deterministic bulk and subband Boltzmann transport equation (BTE) solvers and non-equilibrium Green's functions (NEGF) – have been incorporated into the TCAD toolkit for single-device simulation. At the system level, the field of design-technology co-optimization (DTCO) has expanded to encompass variability, reliability, and the extension of TCAD methodologies from devices to circuits. However, most of these innovations were introduced over a decade ago, prompting the question: What remains to be developed in TCAD? We address this question by analyzing current limitations and potential future directions in TCAD development across three key dimensions: (1) fidelity, (2) integration, and (3) efficiency – each with particular relevance in commercial and industrial contexts. We examine ongoing challenges in classical TCAD, advanced transport modeling, and DTCO flows, and point to potential directions for future developments. Among these, we include various methodologies related to machine learning and hardware accelerators, particularly within the efficiency dimension.**

I. Introduction

Technology computer-aided design (TCAD) is partially responsible for the semiconductor revolution of the past decades. As process technology evolves to become more and more expensive, TCAD gains in importance, allowing limited process and device engineering to be performed outside the fab by means of process and device simulation. In this paper we will specifically look at three areas of major interest to the TCAD community in the recent years, which also have major implications in the industrial use of TCAD: (i) advanced carrier transport modeling, (ii) design-technology co-optimization, (iii) accelerators in TCAD.

We make a distinction between TCAD and computational material science (CMS); thus, we do not consider ab-initio methods such as density functional theory (DFT) or molecular dynamics (MD) as part of TCAD, although these play an important supporting role in modeling and simulation with TCAD. Material science provides the necessary material parameters that make meaningful TCAD simulations possible. These parameters are either sourced directly from experiments or by fitting simulations to experimental data, or they are obtained by the aforementioned CMS methods.

In both process and device simulation, the methodologies employed in TCAD can be broadly categorized in continuum models (usually Poisson combined with drift-diffusion) and particle-based models (Monte Carlo for electrons, ions, and atoms for device, ion implantation, and deposition simulations, respectively), and these two have been the work-horses of TCAD since its beginnings. The development of advanced transport models has introduced methods that do not fit these categories, most notably direct solutions of partial integro-differential equations (PIDE) for Boltzmann transport and non-equilibrium Green's functions for quantum transport, requiring novel algorithms to solve the associated problems.

Based on this observation, we introduce a classification of problems that will help understand what is needed to advance the state of the art in a particular topic:

Algorithm-limited problems are currently computationally intractable or impractically expensive and a break-through could be achieved either by finding a faster alternative algorithm, or by employing parallelization or accelerators to brute-force a solution. Also, an algorithm-limited problem might effectively be solved by finding an alternative formulation with minimal accuracy loss but allowing solution by a simpler algorithm. Typical examples of algorithm-limited problems are found in advanced transport simulations in silicon devices, where a certain well-observed phenomenon requires a detailed microscopic (possibly quantum-mechanical) and thus computationally expensive model.

Data-limited problems could potentially be solved by the currently available methods but material and model parameters are either not known or have only been characterized with a high degree of uncertainty. Typical examples of data-limited problems are device simulations of new channel materials, such as transition-metal dichalcogenide (TMD) mono-layers. While algorithm-limited problems can be solved by advancing the state of the art in theory alone, data-limited problems cannot; solving them requires collaboration between TCAD and material science, of which at least a part must be based in experimental work.

II. Advanced Transport

A. Silicon

Historically, the first venture into advanced transport modeling in silicon devices was through bulk and device Monte Carlo (MC) techniques [1], which solve the bulk Boltzmann transport equation (BTE) stochastically. Starting with quasi-parabolic band approximations, the method has evolved to include numerical dispersion relations (on $\mathbf{k}$-meshes), mechanical stress, and transport in alloys. However, the main weakness of MC is its statistical nature, making rare events (e.g. off-state current) difficult to simulate and adding random noise, which limits self-consistent convergence with electrostatics. Deterministic methods for solving the bulk BTE also exist, the most prominent being the spherical harmonics expansion (SHE) of the angular dependence of functions in $\mathbf{k}$-space [2]. However, even with analytical band structures, SHE is still computationally expensive. Bulk BTE methods are ideally suited for path-finding in short-channel devices that do not exhibit quantum confinement effects, such as high-speed bipolar transistors [3].

When considering path-finding in nano-scale MOSFETS, we need to address both the shortness and the narrowness of their channels, which are typically fins or gate-all-around (GAA) nanosheets. Here, quantum confinement is of major importance. Simulation methodologies that incorporate a Schrödinger-Poisson solution represent the confined states in the channel as wave modes, which leads to a new transport formalism: the subband Boltzmann transport equation (SBTE) [4]. SBTE becomes more efficient for narrower channels, since the computational burden grows with the number of modes in the channel, which scales with the square of its cross-section area. Non-parabolic band models with numerical dispersion relations can be readily used [5] and with that SBTE incorporates several additional effects: ground-state energy and effective-mass change induced by confinement, mechanical stress and alloy effects, density-of-states (DoS) and scattering in a low-dimensional gas, and performance-limiting physical mechanisms at very small device dimensions, such as roughness scattering and source-drain tunneling [6]. The SBTE has been demonstrated to work with realistically large device dimensions across a range of technology nodes, from bulk MOSFETs and FinFETs [7] to A14 nanosheets [8].

Following this trend, one may be led to believe that non-equilibrium Green's functions (NEGF) are the next logical advancement for nano-scale MOSFETs, as they replace semi-classical by coherent quantum transport. NEGF massively expands computational effort, even when ballistic, compared to SBTE, as instead of low-dimensional rate equations, wave equations need to be solved for each energy grid-point. Since inverse-matrix elements of indefinite matrices are needed, direct methods must be used, such as recursive Green's functions (RGF) or SelInv [9]. When modeling scattering in NEGF, local approximations are used for phonon scattering self-energies [10], which are derived from Fermi's golden rule and are thus not different from the semi-classical scattering models in SBTE. Spatially-correlated scattering self-energies, such as Coulomb and roughness scattering, result in non-local operators that need to be represented by full matrices, thus eliminating efficiency gains from operator sparsity. A common workaround is to simulate devices with random distributions of charges and randomly rough Si/SiO$_2$ interfaces [11]. However, this requires ensembles or hundreds to thousands of devices to be simulated with NEGF to obtain average characteristics.

Advanced transport in silicon is an algorithm-limited problem. Its recent success was enabled by the availability of band and scattering parameters that were investigated as early as the 1950s [12] and 1970s [13], respectively. With computing power becoming abundant and devices shrinking to reduce the number of conducting modes, methods such as SBTE have become feasible. The fidelity of advanced transport models in silicon is derived from the experimental characterization of band gaps, effective masses, and mobilities in bulk Si and thin films, and has been repeatedly confirmed by comparing device simulations to experiments. All methods (MC, SHE, SBTE, NEGF) are or have the potential to be well-integrated in TCAD-software, allowing input of process-simulated geometries [7] and embedding of advanced transport domains in a drift-diffusion (DD) simulation [14]. In the case of our GTS Nano Device Simulator (NDS), the SBTE solver is presented as an add-on to the DD-based device simulator Minimos-NT, simplifying usage and facilitating integration with other components. All presented methods can take advantage of modern parallel hardware, and SBTE in particular offers high-level parallelization opportunities and thus scales very well on parallel CPUs; however, the computational burden of SBTE scales $\approx \mathcal{O}(A^4)$ with the channel cross-sectional area and while 5 nm thin NWFETs can be simulated within 30 min, practical-sized FinFETs have a turn-around-time (TAT) of 1-2 days. This underlines the necessity for automated calibration of simpler DD-based [8] or even compact models against advanced transport models in order to leverage them in practice.

B. Novel Channel Materials

While advanced transport modeling in silicon is a algorithm-limited problem, it's quite the opposite for channel materials beyond silicon, of which most pose data-limited problems. One of the most prominent post-Si material candidates are transition metal dichalcogenides (TMDs). Advanced transport methods can be readily applied to materials such as MoS$_2$ with some modifications [15]. However, there is currently no consensus on the model parameters for MoS$_2$, especially for the complex electron-phonon interactions expected to occur within the material based on first-principles calculations [16]. While initially there was a large gap between experimentally observed and simulated mobilities, the gap has recently narrowed from the experimental side [17], thanks to improved process control and dielectrics.

III. Design-Technology Co-Optimization

TCAD plays an important role in Design-Technology Co-Optimization (DTCO): In principle, TCAD predicts device characteristics from process assumptions through process and device simulation, from which compact models can be extracted to predict the impact of technology parameters and circuit design on a design's performance. In this paper, we use DTCO to highlight the importance of tool integration in TCAD. One of the key components of our GTS Cell Designer (CD) DTCO flow is the parasitics extraction (PEX). Rather than implementing PEX as a stand-alone tool, it was implemented as add-on to our device simulator Minimos-NT. The integration has several immediate benefits: (i) as a device simulator Minimos-NT already provides a field solver core, (ii) PEX and Minimos-NT share the same library of material parameters, which ensures consistency between the two, and (iii) being built on Minimos-NT, PEX can include semiconductor regions in its R and C-extraction.

One guiding design principle of PEX was automation; the goal was that PEX can process a large number of similar logic cells that are generated in 3D from layouts either through process emulation or constructive solid geometry. PEX extracts and annotates a cell's netlist by pruning and simplifying its segment-adjacency graph according to a list of rules; tagging layers and materials in the process flow provides the necessary information to PEX for netlist extraction [18]. When the netlist graph is found, the R and C values are determined by probing the individual resistance and capacitance paths in the 3D model of the cell. This process is made efficient by probing disjoint resistance branches simultaneously and re-using LU-factors of the Poisson equation for multiple capacitance probings in parallel.

In a practical DTCO flow, NMOS and PMOS transistor compact models are extracted along with the cell's R/C-network to complete the netlist. To extract the compact models, C/V and I/V characteristics of the transistors are simulated in Minimos-NT and PEX is performed on the single transistors as well to extract a single-device netlist containing the local R/C-parasitics. This way the local parasitics are already accounted for when optimizing the compact model parameters, which would otherwis be double-counted if a bare transistor model was optimized [19].

Fidelity of this process is bounded by the fidelity of the models in Minimos-NT, which have been calibrated for advanced logic nodes – in part through NDS. The integration minimizes the losses between TCAD and extracted compact models and permits the verification of extracted circuits through comparison with full-cell TCAD [20]. While mainly algorithm-limited, introduction of novel interconnect materials, such as ruthenium or nano-structured carbon, add a data-limited problem to DTCO.

IV. Accelerators in TCAD

Classical TCAD is often time-consuming, especially in 3D. This is an algorithm-limited problem. Recently, methods have been discussed to accelerate TCAD simulations by means of dedicated hardware and through machine-learning methods.

A. Algebraic accelerators

Algebraic accelerators focus on accelerating the solution of the linear system in the solution phase of a simulator, which might be part of a Newton-Raphson scheme, which itself is inherently serial. Especially in 3D, the accumulated time in the linear solver makes up the majority of the simulator's runtime.

For 3D Poisson-like and convection-diffusion problems the baseline solution strategy is to use an iterative solver like GMRES or BiCGStab in combination with an incomplete sparse LU (ILU) preconditioner. On a single thread this is also the most efficient option. However, ILU algoritmhs based on sparse Gauss elimination are not suitable for acceleration by vectorization or multi-threading. Unlike complete sparse LU [21], ILU algorithms maintain a low fill-in, making supernodal optimizations ineffective.

The ILUPACK library parallelizes ILU using a multi-level approach, that resembles a multi-grid approach [22]. Another promising approach is using a fixed-point iteration to compute the ILU elements, where the inner loop within each iteration can be easily parallelized [23, 24]. However, during each iteration, the ILU sparsity pattern is fixed, and adjusting it requires additional steps, which are not easily parallelized, making the resulting preconditioner somwhat worse than serial ILU.

Finally, algebraic multi-grid (AMG) methods have seen resurgence in the form of the AMGX library [25], which has native support for GPU computing. While the results look impressive, it must be noted that the effectiveness of AMG is problem-specific; AMG works incredibly well for Poisson and diffusion-type problems but less so for convective and highly non-linear problems.

B. Development accelerators

Apart from the purely performance-focused algebraic acceleration, there are methods that can shorten the time-to-market of TCAD software by automating the model development in simulators. The use of expression templates [26, 27] can help speed up the development of models based on coupled PDE systems with many variables (e.g. dopant diffusion and stress evolution in process simulation). Static expressions in C++ allow for optimization and syntax checking of the expressions by the compiler. For non-linear systems, automatic differentiation (AD) and dual numbers can be used to automate Jacobian evaluation [28, 29], which would otherwise be done by hand. Both methods put model development on a foundation that can be developed and tested separately from the models, thereby making development less error-prone. Automated assembly is also more conducive to parallelization than hand-written code, thereby benefiting overall simulator performance.

We have used this approach to derive a generic non-linear coupled PDE solver platform from our Schrödinger-Poisson solver (VSP) [30]. The platform has been successfully used to develop a model for phase-change memories [31] as well as a full process simulator.

C. Machine-learning accelerators

A general approach to applying machine-learning (ML) in TCAD remains elusive and it is unclear if prediction of TCAD results can be achieved in a generic way with the currently available ML methods. Process and device simulation generally produce large amounts of data based on comparatively few input fields, therefore a corresponding ML model would need to "fill-in" a considerable amount of "gaps", not unlike image generation from a prompt. This raises major questions about the fidelity of such ML models.

Successful applications of ML have been problem-specific, where TCAD has been used to generate data to train an ML model to predict a few KPIs in a specific application; the trained model would then be be used in combination with a optimizer to find optimal values for one or several KPIs under varying constraints [32, 33]. This approach trades up-front simulation cost to generate training data for a much faster optimization loop. In some settings in 2D, the approach is close to a break-even in runtime [33] but might still be prohibitive in 3D, highlighting the need for effective acceleration of TCAD software. It also demonstrates that currently the most feasible way for engagement of ML with TCAD is to provide toolboxes for surrogate-model building that are well-integrated with the TCAD software.

V. Conclusions

In conclusion, we can say that there is plenty left to do in TCAD. While major advancements have been made in advanced transport models, the TAT for silicon devices makes these tools less-than-practical for daily use. Algorithmic improvements and automatic calibration of DD models would help bring advanced transport modeling further into mainstream TCAD. Collaborative efforts between computation and experimental material science and TCAD would significantly help unlocking the benefits of novel channel materials in the semiconductor industry. Scaling-up TCAD and DTCO to interface with circuit design would help accelerate chip development but requires close integration of the tools to address the varied challenges DTCO is facing. Finally, modernizing TCAD software would bring efficiency benefits that would enable novel ML-based optimization schemes.

References

[1] C. Jacoboni and L. Reggiani, "The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials," *Rev. Mod. Phys.*, vol. 55, pp. 645–705, Jul 1983. [Online]. Available: http://link.aps.org/doi/10.1103/RevModPhys.55.645

[2] S.-M. Hong, A.-T. Pham, and C. Jungemann, *Deterministic Solvers for the Boltzmann Transport Equation*. Springer Vienna.

[3] H. Leenders, M. Müller, C. Jungemann, and M. Schröter, "Physical modeling of inp/ingaas dhbts with augmented drift-diffusion and boltzmann transport equation solvers–part i: Simulation tools and application to sample structures," vol. 70, no. 10, pp. 5065–5072.

[4] S. Jin, M. V. Fischetti, and T.-W. Tang, "Theoretical Study of Carrier Transport in Silicon Nanowire Transistors Based on the Multisubband Boltzmann Transport Equation," *Electron Devices, IEEE Transactions on*, vol. 55, no. 11, pp. 2886 –2897, nov. 2008.

[5] Z. Stanojević, M. Karner, O. Baumgartner, H. W. Karner, C. Kernstock, H. Demel, and F. Mitterbauer, "Phase-space solution of the subband Boltzmann transport equation for nano-scale TCAD," in *SISPAD*, Sept 2016, pp. 65–67.

[6] Z. Stanojević, G. Strof, O. Baumgartner, G. Rzepa, and M. Karner, "Performance and Leakage Analysis of Si and Ge NWFETs Using a Combined Subband BTE and WKB Approach," in *SISPAD*, 2020, pp. 63–66.

[7] Z. Stanojević, C.-M. Tsai, G. Strof, F. Mitterbauer, O. Baumgartner, C. Kernstock, and M. Karner, "Nano device simulator–a practical subband-bte solver for path-finding and dtco," *IEEE TED*, vol. 68, no. 11, pp. 5400–5406, 2021.

[8] L.-C. Hung, G. Rzepa, M. Kampl, C.-M. Tsai, F. Schanovsky, O. Baumgartner, Z. Stanojević, and M. Karner, "Hierarchical transport modeling for path-finding dtco," in *2024 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*. IEEE, pp. 1–4.

[9] L. Lin, C. Yang, J. C. Meza, J. Lu, L. Ying, and W. E, "Selinv—an algorithm for selected inversion of a sparse symmetric matrix," vol. 37, no. 4, pp. 1–19.

[10] A. Martinez, A. Price, R. Valin, M. Aldegunde, and J. Barker, "Impact of phonon scattering in si/gaas/ingaas nanowires and finfets: a negf perspective," vol. 15, no. 4, pp. 1130–1147.

[11] H.-H. Park, Y. Lu, W. Choi, Y.-T. Kim, K.-H. Lee, and Y. Park, "Atomistic simulations of phonon- and alloy-scattering-limited mobility in SiGe nFinFETs," in *2014 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*. IEEE, 2014, pp. 257–260.

[12] G. Dresselhaus, A. F. Kip, and C. Kittel, "Cyclotron Resonance of Electrons and Holes in Silicon and Germanium Crystals," *Phys. Rev.*, vol. 98, pp. 368–384, Apr 1955. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRev.98.368

[13] C. Jacoboni, C. Canali, G. Ottaviani, and A. Alberigi Quaranta, "A review of some charge transport properties of silicon," *Solid-State Electronics*, vol. 20, no. 2, pp. 77 – 89, 1977. [Online]. Available: http://dx.doi.org/10.1016/0038-1101(77)90054-5

[14] S. Jin, S.-M. Hong, W. Choi, K.-H. Lee, and Y. Park, "Coupled drift-diffusion (dd) and multi-subband boltzmann transport equation (msbte) solver for 3d multi-gate transistors," in *2013 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*. IEEE, pp. 348–351.

[15] Z. Stanojević, C.-M. Tsai, J. M. Gonzalez Medina, L.-C. Hung, and M. Karner, "From si to mos2 – device simulation based on the direct solution of the boltzmann transport equation," in *2023 IEEE Nanotechnology Materials and Devices Conference (NMDC)*. IEEE, pp. 896–898.

[16] X. Li, J. T. Mullen, Z. Jin, K. M. Borysenko, M. B. Nardelli, and K. W. Kim, "Intrinsic electrical transport properties of monolayer silicene and MoS$_2$ from first principles," *Physical Review B*, vol. 87, no. 11, mar 2013.

[17] F. Zhuo, J. Wu, B. Li, M. Li, C. L. Tan, Z. Luo, H. Sun, Y. Xu, and Z. Yu, "Modifying the power and performance of 2-dimensional mos2field effect transistors," vol. 6.

[18] Z. Stanojević, X. Klemenschits, G. Rzepa, F. Mitterbauer, C. Schleich, F. Schanovsky, O. Baumgartner, and M. Karner, "Tcad for circuits and systems: Process emulation, parasitics extraction, self-heating," in *2024 IEEE BiCMOS and Compound Semiconductor Integrated Circuits and Technology Symposium (BCICTS)*. IEEE, pp. 294–297.

[19] G. Rzepa, K. K. Bhuwalka, O. Baumgartner, D. Leonelli, H.-W. Karner, F. Schanovsky, C. Kernstock, Z. Stanojevic, H. Wu, F. Benistant, C. Liu, and M. Karner, "Performance and variability-aware sram design for gate-all-around nanosheets and benchmark with finfets at 3nm technology node," in *2022 International Electron Devices Meeting (IEDM)*, 2022, pp. 15.1.1–15.1.4.

[20] G. Rzepa, M. Karner, O. Baumgartner, G. Strof, F. Schanovsky, F. Mitterbauer, C. Kernstock, H. Karner, P. Weckx, G. Hellings, D. Claes, Z. Wu, Y. Xiang, T. Chiarella, B. Parvais, J. Mitard, J. Franco, B. Kaczer, D. Linten, and Z. Stanojevic, "Reliability and variability-aware dtco flow: Demonstration of projections to n3 finfet and nanosheet technologies," in *2021 IEEE International Reliability Physics Symposium (IRPS)*, 2021, pp. 1–6.

[21] O. Schenk and K. Gärtner, "Solving unsymmetric sparse systems of linear equations with pardiso," vol. 20, no. 3, pp. 475–487.

[22] J. I. Aliaga, M. Bollhöfer, A. F. Martín, and E. S. Quintana-Ortí, *Parallelization of Multilevel ILU Preconditioners on Distributed-Memory Multiprocessors*. Springer Berlin Heidelberg, pp. 162–172.

[23] E. Chow and A. Patel, "Fine-grained parallel incomplete lu factorization," vol. 37, no. 2, pp. C169–C193.

[24] H. Anzt, E. Chow, and J. Dongarra, "Parilut—a new parallel threshold ilu factorization," vol. 40, no. 4, pp. C503–C519.

[25] M. Naumov, M. Arsaev, P. Castonguay, J. Cohen, J. Demouth, J. Eaton, S. Layton, N. Markovskiy, I. Reguly, N. Sakharnykh, V. Sellappan, and R. Strzodka, "Amgx: A library for gpu accelerated algebraic multigrid and preconditioned iterative methods," vol. 37, no. 5, pp. S602–S626.

[26] C. Pflaum, "Expression templates for partial differential equations," vol. 4, no. 1, pp. 1–8.

[27] A. Singh, P. Incardona, and I. F. Sbalzarini, "A c++ expression system for partial differential equations enables generic simulations of biological hydrodynamics," vol. 44, no. 9.

[28] E. Tijskens, D. Roose, H. Ramon, and J. De Baerdemaeker, "Fastder++, efficient automatic differentiation for non-linear pde solvers," vol. 65, no. 1–2, pp. 177–190.

[29] R. Anand Krishna, R. V. S. Krishna Dutt, and P. Premchand, *Automatic Differentiation Using Dual Numbers - Use Case*. Springer Nature Switzerland, pp. 68–78.

[30] O. Baumgartner, Z. Stanojevic, K. Schnass, M. Karner, and H. Kosina, "VSP–a quantum-electronic simulation framework," *J. Comput. Electron.*, vol. 12, pp. 701–721, 2013. [Online]. Available: http://dx.doi.org/10.1007/s10825-013-0535-y

[31] M. Thesberg, Z. Stanojevic, O. Baumgartner, C. Kernstock, D. Leonelli, M. Barci, X. Wang, X. Zhou, H. Jiao, G. Donadio, D. Garbin, T. Witters, S. Kundu, H. Hody, R. Delhougne, G. Kar, and M. Karner, "Monolithic tcad simulation of phase-change memory (pcm/pram) + ovonic threshold switch (ots) selector device," vol. 199, p. 108504.

[32] T. Herrmann, P. Jungmann, E. Silva, N. Mika, and A. Zaka, "Ring oscillator dtco using machine learning approach based on tcad," in *2024 IEEE European Solid-State Electronics Research Conference (ESSERC)*. IEEE, pp. 225–228.

[33] N. Ripamonti, J. Chakravorty, A. Sapozhnik, G. Gupta, E. Kuk, and D. M. Luca, "Novel neural-network-based method for advanced termination design in power semiconductor devices," in *2025 51st IEEE European Solid-State Electronics Research Conference (ESSERC)*.