Nano Device Simulator—A Practical Subband-BTE Solver for Path-Finding and DTCO

Zlatan Stanojević, *Member, IEEE*, Chen-Ming Tsai, Georg Strof, Ferdinand Mitterbauer, Oskar Baumgartner, *Member, IEEE*, Christian Kernstock, and Markus Karner, *Member, IEEE*

***Abstract*— We present an in-depth discussion on the subband Boltzmann transport (SBTE) methodology, its evolution, and its application to the simulation of nanoscale MOSFETs. The evolution of the method is presented from the point of view of developing a commercial general-purpose SBTE solver, the GTS nano device simulator (NDS). We show a wide range of applications SBTE is suited for, including state-of-the-art nonplanar and well-established planar technologies. It is demonstrated how SBTE can be employed both as a path-finding tool and a fundamental component in a DTCO-flow.**

***Index Terms*— Design-technology co-optimization, device simulation, path-finding, subband Boltzmann transport, TCAD.**

I. Introduction

GTS NANO device simulator (NDS) [1] is a device simulator centered around the *deterministic* solution of the coupled multisubband Boltzmann transport equation (SBTE). The NDS methodology was first presented in 2015 [2] as a versatile and effective SBTE-based framework that could be applied in almost the same way as a conventional drift-diffusion (DD)-based device simulator when high physical fidelity of transport modeling was required. Since then the methodology has evolved under that premise and has found application in academia and industry for CMOS-single-device path-finding [3]–[5], compact-model extraction from physics-based TCAD and DTCO [6], [7], and reliability modeling with an emphasis on hot-carrier degradation [8].

As an SBTE solver, NDS accurately models the effects of quantum confinement, crystal orientation, high values of mechanical stress, carrier scattering, and short-channel effects, such as velocity-overshoot [see Fig. 5 (right)], on device performance. It does that based on physics rather than fitted empirical models. Furthermore, NDS provides a platform for TCAD model building and a reference for the calibration of empirical and compact models.

NDS has been applied across multiple technologies including FinFET, stacked horizontal nanosheets and nanowires [3], vertical nanosheets and nanowires [7], as well as planar technologies, such as FDSOI [9] and bulk MOSFETs. In this article, we will give an overview of the model details of the NDS methodology and its extended use cases that go beyond the scope of previous publications.

We present several key developments that extend the state of the art in SBTE simulation: 1) the use of a trajectory-based formulation of the SBTE/Liouville equation that allows discretization on arbitrary numerical subband structures; 2) the incorporation of tunneling transport into the SBTE framework using WKB; 3) a Poisson-DD-SBTE iteration scheme that provides stable convergence and allows us to *embed* SBTE domains within Poisson-DD simulations; 4) a procedure for making process-simulated/emulated device topographies available for SBTE device simulation; and 5) a procedure for calibrating SBTE to hardware data.

II. Methods

In this section, we discuss the models and methods used in NDS. First, the SBTE-based transport core of NDS is presented, followed by our approach of self-consistently coupling the SBTE to electrostatics. Finally, procedures are shown that allow us to use topography process-simulated device structures including imported devices from third-party tools.

A. Transport Core

NDS solves the SBTE either in 2-D or 3-D *phase-space*. Phase-space is the tensor product of the real-space and **k**-space coordinates. For 3-D devices, such as FinFET or NWFET/NSFET, with 2-D confinement, both **r** and **k**-space are 1-D, hence phase-space is 2-D. For 2-D (planar) devices, **k**-space is 2-D (longitudinal and lateral), while **r**-space remains 1-D, thus phase-space becomes 3-D. The SBTE reads

$$\left[v_\nu(x,k_x)\frac{\partial}{\partial x}+\frac{F_\nu(x,k_x)}{\hbar}\frac{\partial}{\partial k_x}\right]f_\nu(x,k_x)=\mathbb{S}f \qquad (1)$$

with x being the direction of carrier transport, i.e., the channel direction, where v_ν and F_ν are velocity and force, respectively,

Manuscript received March 1, 2021; revised April 10, 2021 and May 6, 2021; accepted May 6, 2021. The review of this article was arranged by Editor S. Jin. *(Corresponding author: Zlatan Stanojević.)*

The authors are with Global TCAD Solutions GmbH, 1010 Vienna, Austria (e-mail: z.stanojevic@globaltcad.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2021.3079884.

Digital Object Identifier 10.1109/TED.2021.3079884

$f_\nu(x,k_x)$ is the subband distribution function (DF), and $\mathbb{S}$ is the scattering operator

$$\mathbb{S}f = \sum_{\nu',k_x'} \big\{ S_{\nu',\nu}(k_x',k_x)[1-f_\nu(k_x)]f_{\nu'}(k_x') - S_{\nu,\nu'}(k_x,k_x')[1-f_{\nu'}(k_x')]f_\nu(k_x) \big\} \quad (2)$$

where the index ν denotes the subband index n in 2-D phase-space (3-D device) and the combined subband and lateral **k**-vector $\nu=(n,k_\perp)$ in 3-D phase-space (2-D device). The SBTE can be directly discretized on a phase-space-grid, i.e., a tensor-product of the **r**-space and **k**-space grids. This approach is comparably straightforward in its implementation and enables the use of perturbative methods, such as small-signal ac-analysis, on the SBTE. However, its major drawback is the *artificial carrier heating* (ACH) phenomenon, arising due to numerical diffusion in phase-space [10], which broadens the distribution function on the source side causing an increase in subthreshold-slope as if the device was heated. Highly scaled FinFETs and GAA-devices have a near-ideal subthreshold-slope, thus ACH prohibits the use of phase-space discretization in determining their OFF-currents.

Alternatively, (1) can be transformed into a trajectory-based picture [11]–[13]. Setting $\mathbb{S}=0$, one obtains the (subband) Liouville equation

$$\left[\frac{\partial H_\nu(x,k_x)}{\partial k_x}\frac{\partial}{\partial x} - \frac{\partial H_\nu(x,k_x)}{\partial x}\frac{\partial}{\partial k_x}\right] f_\nu(x,k_x) = 0 \quad (3)$$

which was rewritten using the Hamiltonian $H_\nu = E_\nu^{\mathrm{kin}} + V_\nu$, with $v_\nu = \partial H_\nu/\partial k_x$ and $F_\nu = -\partial H_\nu/\partial x$. An example of a Hamiltonian in a transistor is shown in Fig. 1. Equation (3) can be solved through the *method of characteristics* using

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial H_\nu(x,k_x)}{\partial k_x}, \quad \frac{\mathrm{d}k_x}{\mathrm{d}t} = -\frac{\partial H_\nu(x,k_x)}{\partial x} \quad (4)$$

$$\frac{\mathrm{d}f_\nu}{\mathrm{d}t} = 0. \quad (5)$$

The variable t is the particle flight-time along each characteristic which can be viewed as a particle's (ballistic) trajectory through the device. The equations in (4) need not be solved numerically, as it follows from (3) and (4) that $\mathrm{d}H_\nu(x,k_x)/\mathrm{d}t = 0$, which means that a trajectory at energy E_0 can be conveniently extracted from the contours of the Hamiltonian at E_0 as shown in Fig. 1 (purple lines).

The Hamiltonian $H_\nu(x,k_x)$ is generated by solving the Schrödinger equation

$$\big[\mathbf{H}_{\mathrm{kin}} + V\big]\psi = E\psi \quad (6)$$

where $\mathbf{H}_{\mathrm{kin}}$ can be an effective-mass, $\mathbf{k}\cdot\mathbf{p}$, or other Hamiltonian. The Hamiltonian is decoupled along the transport direction x, assuming the wave function to be confined perpendicularly to x and a plane wave along x, resulting in a set of decoupled Schödinger equations at different positions x along the channel

$$\big[\mathbf{H}_{\mathrm{kin}}(k_x) + V(x,y,z)\big]\psi(x,k_x) = E(x,k_x)\psi(x,k_x). \quad (7)$$

By solving the eigenproblem in (7) for each (x,k_x), we obtain the phase-space Hamiltonian $H_\nu(x,k_x) = E_\nu(x,k_x)$; such a function is plotted for the first subband in Fig. 1. The subband index ν is consistent with the ordering of the eigenstates by the numerical eigensolver. In simulation, an (x,k_x) grid is used in the phase space, where the grid points along the x direction coincide with cuts of the SBTE domain (see Section II-C); on each cut, (7) is solved numerically. The contour-method extracts trajectories from the Hamiltonian for each point of an energy-grid (Fig. 2) and works with any numerical representation of $H_\nu(x,k_x)$, regardless of whether the subbands are parabolic or based on a higher order model ($\mathbf{k}\cdot\mathbf{p}$, TB, DFT) and regardless of whether $E_n(\mathbf{k})$ is monotonous with respect to $\|\mathbf{k}\|$.

The trajectories are 1-D curves in the phase-space. For each energy E and subband ν, we can have one or more trajectories: the trajectories highlighted in Fig. 1 are transmitting trajectories that follow carriers entering the device through the source and exiting at the drain, and vice-versa; other trajectories are reflecting and circular ones as shown in Fig. 2. We can denote each trajectory at E in subband ν using a trajectory index η. The coordinate along each trajectory is the *time-of-flight* t,

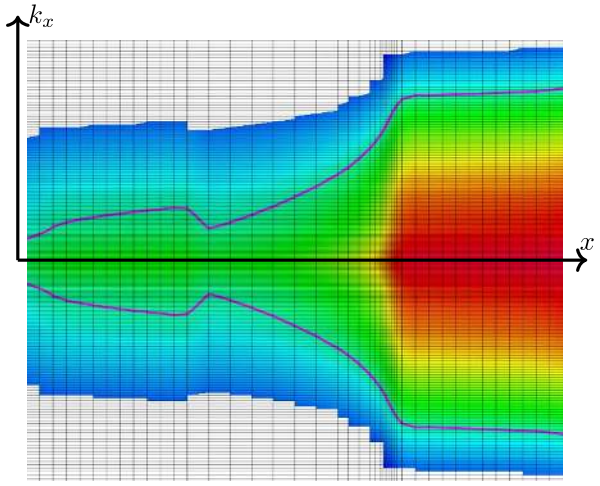

Fig. 1. Phase-space plot of the Hamiltonian of the lowest subband in a NWFET; the vertical grid lines correspond to the k-grid, while the horizontal grid lines correspond to cuts (and their refinements); colors indicate the total particle energy (potential + kinetic) of each state, which is obtained by the numerical solution of the $\mathbf{k}\cdot\mathbf{p}$ Schrödinger equation for each cut position x and wavevector k_x; the purple lines highlight two possible ballistic trajectories at a given energy E_0.

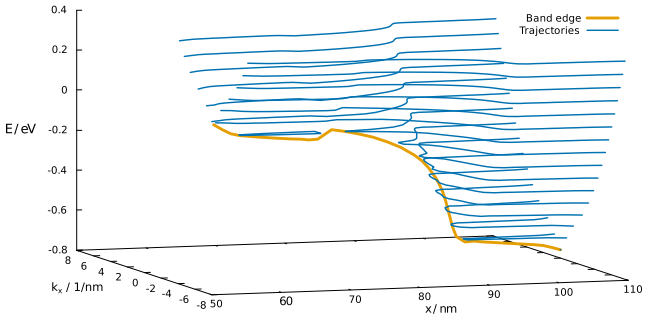

Fig. 2. Set of trajectories extracted at different energies from the energy-scape in Fig. 1; the subband edge is added in orange; trajectories below the top-of-the-barrier (ToB) are reflected off the subband edge, while the trajectories above the ToB allow unrestricted transmission between the left and right contacts; closed-loop trajectories correspond to carriers trapped in local potential wells and can only be accessed through inelastic scattering (and tunneling).

which by convention we define as 0 at the starting point of each trajectory. For every Hamiltonian H_ν, we can now map every point (x, k_x) in phase space to a point (E, η, t) in *trajectory-space* and vice-versa.

Equation (5) is the transformed Liouville equation in trajectory-space and from it follows that the DF is constant along a ballistic trajectory. Thus, ACH is eliminated from the left-hand side of the SBTE, i.e., the free-streaming operator, because the trajectories are inherently energy-conserving. In trajectory-space, the transformed SBTE from (1) takes the form of

$$\frac{\mathrm{d}f_{\nu,\eta}(E,t)}{\mathrm{d}t} = -\Gamma_{\nu,\eta}(E,t)f_{\nu,\eta}(E,t) + g_{\nu,\eta}(E,t) \tag{8}$$

where $\Gamma_{\nu,\eta}(E,t)$ is the out-scattering rate and $g_{\nu,\eta}(E,t)$ is the particle influx along the trajectory. NDS provides three methods to calculate Γ and g.

1) DF is interpolated from trajectory-space to phase-space, $f_{\nu,\eta}(E,t) \mapsto f_\nu(x,k_x)$, (2) is evaluated in (ν, k_x)-space for each cut x to obtain $\Gamma_\nu(x,k_x)$, $g_\nu(x,k_x)$, which are then interpolated back to trajectory-space, $\Gamma_\nu(x,k_x) \mapsto \Gamma_{\nu,\eta}(E,t)$, $g_\nu(x,k_x) \mapsto g_{\nu,\eta}(E,t)$. Evaluation of the scattering operator in phase-space is generally faster than in trajectory-space and this approach results in faster simulation times. However, the interpolations between the energy-grid and phase-space-grid are not perfectly energy-conserving and thus a residual ACH-effect persists.
2) The elements of (2) are transformed into trajectory-space, $S_{\nu,\nu'}(k_x,k_x';x) \mapsto S_{\nu,\nu'}(E,E';\eta,\eta';t,t')$, and evaluated by summation/integration over (E',η',t') to obtain $\Gamma_{\nu,\eta}(E,t)$ and $g_{\nu,\eta}(E,t)$ directly; for elastic scattering, $E' = E$. This method completely eliminates ACH, but at an increased computational cost.
3) A hybrid method, where all elastic scattering (ac phonons, Coulomb, and roughness) is evaluated in trajectory-space while inelastic processes are evaluated by interpolating back and forth between the two spaces. This method eliminates almost all ACH with minimal impact on simulation time.

While all three methods are implemented in NDS, method №3 is applied most often as it provides the best compromise in terms of runtime cost and ACH suppression.

Tunneling along the transport directions (S/D-tunneling, Schottky-barrier-tunneling, and band-to-band-tunneling) can be directly included in the SBTE as shown in Fig. 3. For intra-band tunneling, the transmission coefficients are calculated from the *complex subband structure* using the WKB-formula

$$T(E_0) = \exp\left[-2\int_{x_1}^{x_2} \kappa_x(E_0,x)\mathrm{d}x\right]. \tag{9}$$

The transmission coefficients are integrated into (8) as effective rates Γ_{tunn} and influxes into g_{tunn} [14].

Finally, the assembled system of equations is solved numerically. For Fermi–Dirac statistics, the scattering operator (2) is nonlinear and the SBTE is solved using the Newton method, with the linearized SBTE being solved in every iteration. Damping is applied to ensure the values of the DF are confined to the range [0,1], which also ensures the positive-definiteness of the linearized SBTE, which is important since the linearized SBTE itself is being solved iteratively.

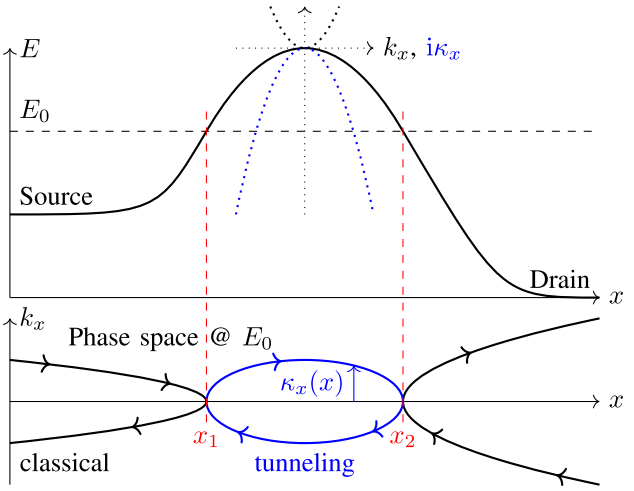

Fig. 3. Typical n-MOSFET potential in off-state along with the local real $E(x, k_x)$ and complex dispersion relation $E(x, \mathrm{i}\kappa_x)$; a cut through the $E(x, k_x)$ landscape at energy E_0 reveals the classical and tunneling trajectories an electron can take; the classical turning points are marked as x_1 and x_2.

After the SBTE is solved numerically, the DF is interpolated one more time to phase-space and macroscopic quantities are extracted from it, most notably the carrier concentration and current density

$$n(\mathbf{r}) = \sum_\nu \int \rho_{\nu,\mathbf{k}}(x,\mathbf{r}_\perp) f_\nu(x,k_x) N_{\mathbf{k}} \mathrm{d}k_x, \tag{10}$$

$$J(\mathbf{r}) = \pm q_0 \sum_\nu \int \rho_{\nu,\mathbf{k}}(x,\mathbf{r}_\perp) v_\nu(x,k_x) f_\nu(x,k_x) N_{\mathbf{k}} \mathrm{d}k_x \tag{11}$$

where $\rho_{\nu,\mathbf{k}}(x,\mathbf{r}_\perp) = \|\psi_{\nu,\mathbf{k}}(x,\mathbf{r}_\perp)\|^2$ is the probability density of state $\nu,\mathbf{k}$ in the cut cross section at position x and $N_{\mathbf{k}}$ is the $\mathbf{k}$-space density of states.

B. Scattering Models

The scattering processes modeled in this work include: 1) acoustic phonon scattering; 2) optical and inter-valley scattering; 3) local and remote Coulomb scattering caused by dopants, interface traps, and oxide charges; 4) Si/SiO_2-interface-roughness scattering; and 5) polar-optical scattering caused by remote optical phonons in high-k materials. The details of the scattering rate evaluation are discussed in [15], [16]. The parameters for optical and inter-valley scattering are taken from literature, see [17], while the parameters for acoustic phonons and roughness are calibrated to the universal low-field mobility data from Takagi *et al.* [18], [19].

C. Self-Consistency

To obtain the electrical characteristics of a semiconductor device at varying voltage biases, the SBTE has to be coupled with electrostatics and solved self-consistently. NDS provides two approaches to achieve self-consistency: 1) using a Poisson-DD-SBTE-loop and 2) through a direct Poisson-SBTE-loop.

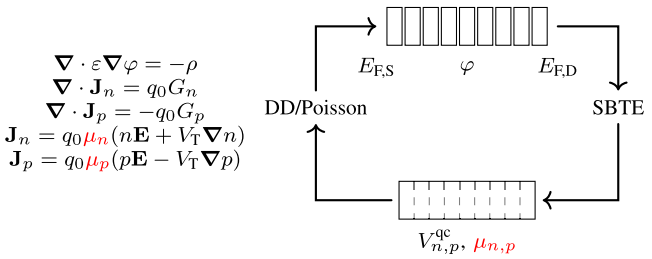

Fig. 4. Principle of operation of the Poisson-DD-SBTE loop.

1) Poisson-DD-SBTE-Loop: This approach jumps back and forth between solving the SBTE and a coupled set of Poisson-DD equations, where the DD equations use a correction potential and an *effective mobility* extracted from the SBTE solution. The correction potential is computed using

$$V_{\mathrm{qc}} = \pm \frac{k_\mathrm{B} T}{q_0} \ln\left(\frac{n_{\mathrm{SBTE}}}{n_{\mathrm{DD}}}\right) \tag{12}$$

where n_{SBTE} and n_{DD} represent the electron/hole concentrations from the SBTE solution (10) and the guess from DD. The effective mobility is obtained from the carrier concentration and current density (10) and (11) by applying an inverted Scharfetter–Gummel DD-discretization formula

$$\mu = \pm J[i \to j] \frac{\delta_{\mathrm{edge}}}{k_\mathrm{B} T} \Big/ \{n[j]\mathcal{B}(\pm\Delta) - n[i]\mathcal{B}(\mp\Delta)\}, \tag{13}$$

$$\Delta = \frac{1}{k_\mathrm{B} T}\big\{E_{\mathrm{cv}}[i] - E_{\mathrm{cv}}[j] - q0\big(V_{\mathrm{qc}}[i] - V_{\mathrm{qc}}[j]\big)\big\} \tag{14}$$

for each edge (i, j) of the mesh; here, $J[i \to j]$ represents the edge current density, δ_{edge} the edge length, and $\mathcal{B}$ the Bernoulli-function $\mathcal{B}(x) = x/(e^x - 1)$. Through this *effective mobility*, the DD equation is "forced" to reproduce the average carrier drift velocity and concentration resulting from the SBTE solution.

Fig. 4 illustrates the self-consistency loop: 1) an initial Poisson/DD/density gradient (DG) solution is used as an initial guess; 2) the electrostatic potential and contact Fermi energies are passed to the SBTE solver; 3) the SBTE is solved to obtain the DF and the macroscopic carrier concentrations and current densities; 4) correction potentials and effective mobilities are computed using (12)–(14) that account for the short-channel behavior of the simulated device; 5) the effective mobility and correction potentials are plugged into the DD equations and the resulting Poisson-DD system is solved to refine the potential guess. Upon convergence, the carrier concentrations, currents, and drift velocities produced by the Poisson-DD-system and the SBTE coincide.

This approach has two major benefits: 1) it provides a good initial guess for the Poisson-SBTE problem and ensures stable convergence due to the quality of solution guesses given by the Poisson-DD-system, so only a few solutions of the computationally expensive SBTE-solution are required; 2) it allows one to *embed* the SBTE domain into a larger device, where semiconductor segments outside the SBTE domain are simulated using Poisson-DD; these might be a bulk semiconductor substrate, a poly-gate, or parts of the contact or interconnect structure.

2) Poisson-SBTE-Loop: This approach allows one to directly iterate the solution of the Poisson and SBTE equations. The linearized Poisson equation

$$\nabla \cdot \varepsilon \nabla \varphi + \varrho + \frac{\mathrm{d}\varrho}{\mathrm{d}\varphi} = 0 \tag{15}$$

which is solved in every iteration, contains an additional term $\mathrm{d}\varrho/\mathrm{d}\varphi$ for the (approximate) derivative of the charge density with respect to the potential, which stabilizes convergence.

Since the Poisson-DD-SBTE approach is faster and more stable, the direct Poisson-SBTE approach is only used in cases where coupling between SBTE and Poisson-DD is not feasible. Two such cases are: 1) Schottky-FETs, where Schottky-barrier tunneling in the S/D-contacts must be included but cannot be coupled to the Poisson-DD system through correction potentials and effective mobilities alone; and 2) MOSFETs with semi-infinite leads for S/D contacts, a type of boundary condition not supported by the DD simulator. Neither of these two applications is the focus of this work and we simulated all applications in Section III using the Poisson-DD-SBTE approach.

D. Device Preparation

In a TCAD environment, device simulators must be capable of handling fairly complex device geometries based either on construction or topography simulation. This requirement applies to SBTE-based simulators as well. Fig. 5 shows an example FinFET device that was generated through the level-set method using a simple process flow. The shapes in the device's geometry (left) are clearly nonideal, posing a general challenge to meshing and device simulation.

For SBTE simulation, the device must be preprocessed. A portion of the semiconductor in the device, i.e., the channel, is designated as the *SBTE domain* and cut out to become its own segment. The SBTE domain needs to be a straight portion of the device's fin, wire, or sheet, although the orientation is arbitrary and need not be aligned with any coordinate axis. The SBTE domain's cross section perpendicular to the transport direction is meshed in 2-D, and then the 2-D grid is replicated at every cut position and extruded to fill the whole SBTE domain. The extruded grid is then passed on to a tetrahedral mesher [20] along with the rest of the device, to generate the complete device mesh. Finally, 2-D cuts are extracted at each plane of the extruded grid. In each 2-D cut, the parts not intersecting with the SBTE domain are remeshed since the 3-D mesh in these regions is not aligned with the cut position and thus would otherwise result in non-Delaunay 2-D cut meshes (Fig. 5, middle).

The tetrahedral mesher may add points between the cuts of the SBTE domain, especially at its surface. Quantities in these points cannot be extracted from the SBTE solution but are instead interpolated between the nearest two cuts. This interpolation mainly concerns the quantities needed for the self-consistent Poisson-DD-SBTE (12)–(14) and Poisson-SBTE loops (15), i.e., current densities, carrier concentrations, and their derivatives with respect to the potential. Despite the nonidealities and interpolations, a self-consistent Poisson-DD-SBTE simulation with robust convergence is achieved.

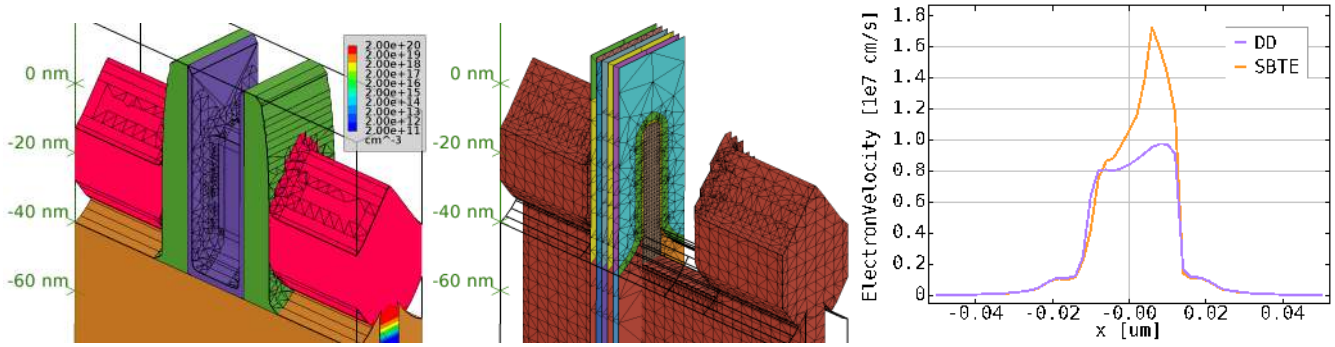

Fig. 5. NDS simulation of a device generated using level-set topography simulation; left: level-set generated FinFET with complex warped surfaces, typical of topography simulation; the analytical doping is shown; middle: the SBTE domain is cut out of the device and meshed using an extruded grid, and mixed with the mesh of the rest of the device; cuts are then extracted from the SBTE domain and remeshed; right: electron drift velocity in the FinFET, DD versus SBTE; the SBTE result clearly shows the velocity overshoot effect not seen in the DD solution.

E. Simulator Performance

Several measures have been implemented to enhance NDS' performance and reduce memory consumption and turn-around-time (TAT), especially for simulations using the $\mathbf{k}\cdot\mathbf{p}$ band structure model. Most of the simulation time is spent in calculating the subband structure and scattering transition rates, thus the focus lies on reducing the number of states and transitions.

1) In $\mathbf{k}\cdot\mathbf{p}$-based simulations, only one half of the $\mathbf{k}$-space is computed. The other half is constructed by exploiting the $\mathbf{k}\cdot\mathbf{p}$-Hamiltonian's property $\mathbf{H}(-\mathbf{k}) = \mathbf{H}^*(\mathbf{k})$ and converting states from $\mathbf{k}$ to $-\mathbf{k}$ through conjugation.
2) In $\mathbf{k}\cdot\mathbf{p}$-based simulation with spin-orbit coupling, each degenerate spin-up and down subband is joined into a single subband, reducing the number of states by two. The corresponding wave functions $\psi_{\nu,\mathbf{k},\uparrow}$, $\psi_{\nu,\mathbf{k},\downarrow}$ are preserved as two-state blocks for scattering evaluation.
3) Due to the effective suppression of ACH, the $\mathbf{k}$-grid density can be relaxed without perceptible changes in the results.
4) The $\mathbf{k}$-dependence of the wave functions can be neglected in scattering, so only the states in the subband minima, $\psi_{\nu,\mathbf{k}} \approx \psi_{\nu,\mathbf{k}_0}$ are used to evaluate scattering transitions.

Measures 1 and 2 reduce the TAT by a factor of 8 without *any loss* of accuracy. Measures 3 and 4 do impact the result and can be used to trade TAT for accuracy. Fig. 6 shows an example of the impact of these measures on TAT and simulation results. In total, we achieve a 30- to 40-fold TAT-reduction compared to a naïve SBTE implementation.

III. Applications

The NDS methodology can be applied to a wide range of technologies and applications. In this work, we have picked three very different use cases to demonstrate its versatility. The first case is focused on the iN14 technology, which we use to showcase the calibration process of physical parameters of the SBTE. The second case uses the NDS methodology to predict S/D-leakage in Ge-channel p-NWFETs and compares it to their Si-based counterparts. The third case demonstrates that the NDS methodology is also suited for the simulation of bulk-technology, namely, a 28-nm high-k metal-gate device.

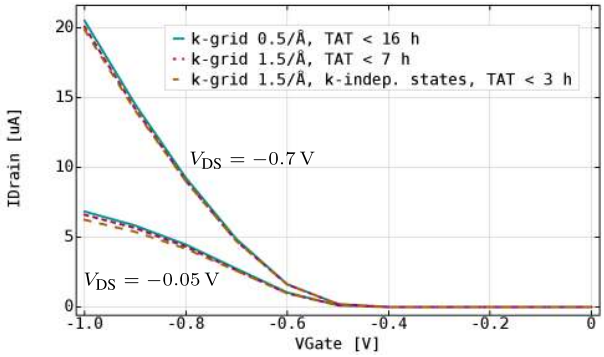

Fig. 6. Transfer characteristics from an NDS simulation of 6 nm wide 20 nm high pMOS FinFET showing the effects of performance measures 3 and 4; the device was first simulated with the default 5×10^7 m^{-1} k-grid to establish the baseline; the device was then simulated with a relaxed 1.5×10^8 m^{-1} k-grid (measure 3) showing negligible deviation from the baseline but halving TAT; finally, the device was simulated with the relaxed k-grid and neglecting the k-dependence of the states (measure 4) showing slightly more deviation but halving TAT once more; TAT was measured for the slowest bias point ($V_G = -1.0$ V, $V_{DS} = -0.7$ V) using ten cores on an Intel Xeon E5-2660 v2 CPU at 2.2 GHz.

A. Calibrating NDS on iN14 Hardware

In this example, we calibrated the NDS model parameters using iN14 bulk FinFET hardware data from IMEC [22], [23]. This is done in three steps:

In the first step, the electrostatics were calibrated. While the device geometry was recreated based on TEM-images, the active doping distribution was unknown. Thus, analytical doping profiles were calibrated by matching the simulated subthreshold slope and DIBL to the measurements. This step does not require SBTE and is done using DD as the transport model.

In the second step, scattering parameters for surface roughness and acoustic phonons were recalibrated by matching low-field mobility calculations with long-channel device measurements, as shown in Fig. 7 (left). Typically, only scattering parameters dependent on the fabrication process need to be adjusted, these being the Si/SiO_2-interface's roughness and trap density.

In the third step, short-channel simulations are performed with SBTE using the calibrated parameters from the first two steps. Now, by comparing the simulated and measured

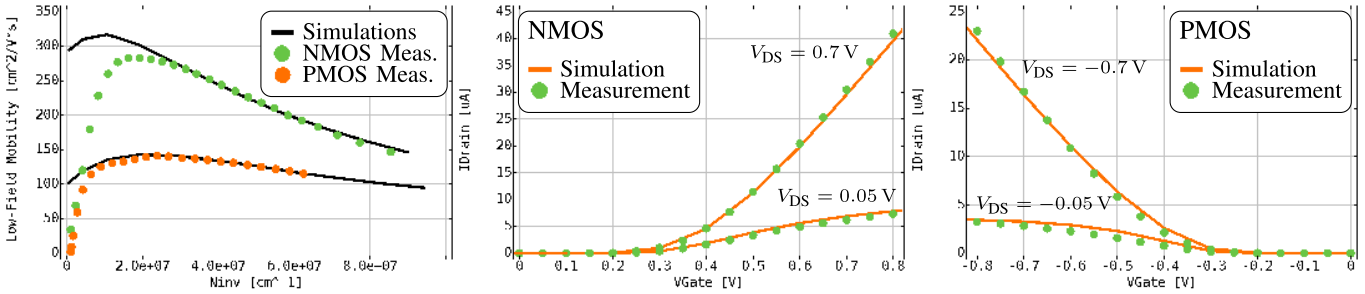

Fig. 7. Left: calibrated low-field mobility of the N14 bulk FinFET to nMOS and pMOS hardware data. Mobility calculations are done on 2-D cross sections using GTS VSP [21]. Middle and right: device characteristics of the IMEC N14 bulk FinFET technology for nMOS and pMOS and the corresponding simulation results from NDS; using the input of the low-field mobility calibration (left) and matching the electrostatics, NDS gives excellent agreement with measurements; nMOS simulations used an effective mass Hamiltonian with nonparabolicity correction and pMOS simulations a six-band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian as electronic structure models.

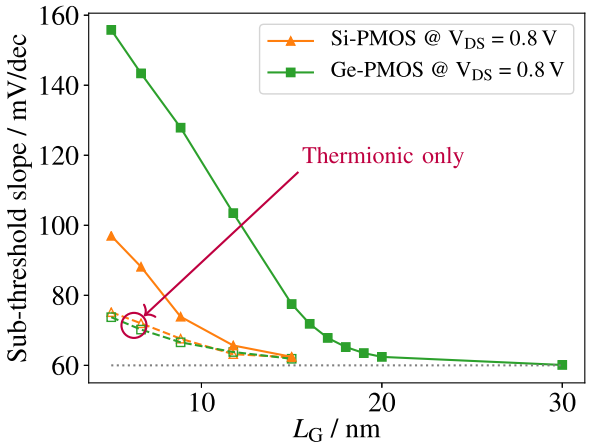

Fig. 8. SS-roll-off for Si- and Ge-pMOS, where the Ge-pMOS shows severe SS-degradation for $L_{gate} < 20$ nm.

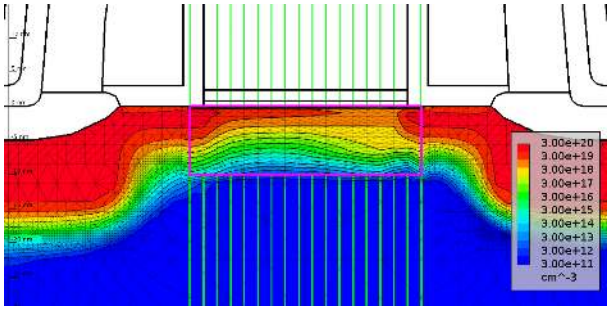

Fig. 9. NDS simulation of a 28-nm high-k metal-gate planar MOSFET in the saturated on-state; the electron concentration is shown; the pink box indicates the SBTE domain and the green lines show the placement of the cuts; the hybrid-extruded/unstructured mesh is clearly visible.

I_D/V_G-curves in the linear and saturated regimes, the gate work function is calibrated by matching the OFF-currents, and the external resistance is calibrated using the ON-currents. The final result in Fig. 7 (middle and right) shows excellent agreement between the experimental data and the model. The fitted simulation setups are transferable to other, more advanced, nodes assuming that similar process conditions apply.

B. Leakage in Ge-Channel p-NWFETs

With this example, we demonstrate the adequacy of the NDS methodology for predicting critical device performance. Ge-channel p-type NWFETs have been demonstrated recently [24], with gate lengths as low as 30 nm. The lower hole effective mass in Ge increases the hole velocity and thus the on-current. However, it also increases intra-band S/D-tunneling. In our study [14], we investigated the subthreshold roll-off of 6-nm-diameter Si- and Ge-channel p-type NWFETs down to a gate length of 5 nm.

The accuracy of the intraband tunneling current prediction in both the Si- and Ge p-type NWFETs strongly depends on the valence band structure model used. Here, we used a six-band $\mathbf{k}\cdot\mathbf{p}$ model, which includes spin-orbit coupling, for both Si- and Ge-channel devices.

The results in Fig. 8 show that tunneling leakage becomes the dominant source of subthreshold slope degradation in the Ge-channel devices as early as a 20 nm gate length, a value that is well within reach of fabrication technology. To rule out electrostatic effects, the same simulations were run with thermionic injection only, i.e., with the tunneling model disabled and those showed only moderate degradation. In Si-channel devices, this effect only becomes significant below 15 nm gate length.

C. Planar 28-nm HKMG Simulation

In our last example, we take a look at planar devices. While those are not found at the forefront of technology development, they are found in a wide range of applications, be it due to cost and volume reasons, or a better analog performance than their FinFET counterparts. Contemporary planar devices are usually made highly optimized for their specific application. The relaxed design rules for these technologies allow designers to co-optimize their devices and circuits on an application-to-application basis. In this context, NDS becomes an optimization rather than a path-finding tool. Hardware data on planar technology is readily available making NDS a very accurate device simulation tool that can enhance or replace DD-based device simulation flows.

We demonstrate this on a 28-nm high-k metal-gate n-type MOSFET shown in Fig. 9. The SBTE-domain consists of a box below the gate. The box needs to be deep enough to capture all of the channel electrons and the required box depth is dependent on the device's doping and operating voltage range. As discussed in Section II-C1, the Poisson-DD-solver *sees* the embedded SBTE domain as a region with a predefined mobility and correction potential while the remaining silicon

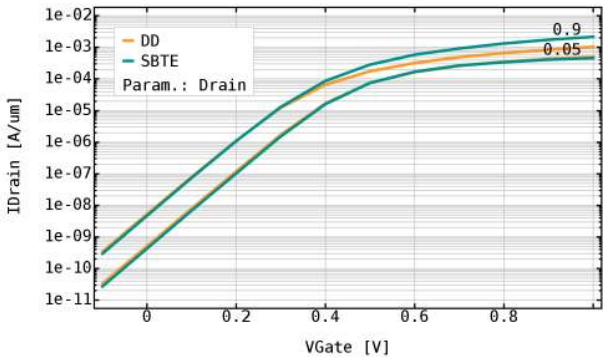

Fig. 10. Simulated transfer characteristics for the 28-nm high-k metal-gate planar MOSFET in Fig. 9, SBTE versus DD; while the low-VDS-curve from DD can be fitted to SBTE, DD underestimates drain current at high VDS by a factor of two, since it does not reproduce the velocity-overshoot-effect.

uses the built-in models for mobility and correction potential. Thus, no interface conditions for fluxes/currents need to be applied at the SBTE domain boundary. The SBTE correction potential is continuously extended outside the SBTE domain using an $\exp(-d/\lambda)$-law, where d is the distance from the SBTE boundary and λ is the characteristic length of the DG equation. This ensures a continuous carrier concentration and current density across the device.

In Fig. 10, we show transfer characteristics for the device based on SBTE and DD/DG for comparison. It should be noted that V_G should not approach flat-band voltage, at which point confinement is lost and the OFF-current starts flowing deep in the bulk and thus outside of the SBTE domain.

IV. Conclusion

In this work, we have presented a major update on the methodology behind the NDS. We have provided detailed explanations of the key models and computational methods that are part of the methodology and discussed the practical aspects of device simulation with SBTE. We have demonstrated the breadth of technologies and applications that it can handle ranging from path finding in advanced nodes to device optimization in established nodes. Its versatility allows the NDS methodology to easily fit into TCAD work-flows where it can enhance or replace DD-based simulators and serve as a basis for TCAD and compact model building: The solution of the SBTE provides data such as ballistic ratios and velocity profiles critical for the calibration of advanced DD-based models such as the ballistic mobility and kinetic velocity models [25].

Acknowledgment

The authors would like to thank Dr. Edward Chen for many fruitful discussions and the continued valuable feedback.

References

[1] Global TCAD Solutions. *Nano Device Simulator*. Accessed: May 6, 2021. [Online]. Available: http://www.globaltcad.com/nds

[2] Z. Stanojevic *et al.*, "Physical modeling—A new paradigm in device simulation," in *IEDM Tech. Dig.*, Dec. 2015, pp. 5.1.1–5.1.4.

[3] M. Karner *et al.*, "Vertically stacked nanowire MOSFETs for sub-10 nm nodes: Advanced topography, device, variability, and reliability simulations," in *IEDM Tech. Dig.*, Dec. 2016, pp. 30.7.1–30.7.4.

[4] S.-K. Su, J. Cai, E. Chen, L.-J. Li, and H.-S. Philip Wong, "Impact of Schottky Barrier on the performance of two-dimensional material transistors," in *Proc. Int. Conf. Simul. Semiconductor Processes Devices (SISPAD)*, Sep. 2020, pp. 285–287.

[5] S. Das, S. Dey, E. Mohapatra, J. Jena, T. P. Dash, and C. K. Maiti, "Role of stress/strain mapping and random dopant fluctuation in advanced CMOS process technology nodes," *Int. J. Nano Biomater.*, vol. 9, nos. 1–2, p. 18, 2020.

[6] Q. Huo *et al.*, "Physics-based device-circuit cooptimization scheme for 7-nm technology node SRAM design and beyond," *IEEE Trans. Electron Devices*, vol. 67, no. 3, pp. 907–914, Mar. 2020.

[7] Y. Li *et al.*, "1.5-nm node surrounding gate transistor (SGT)-SRAM cell with staggered pillar and self-aligned process for gate, bottom contact, and pillar," in *Proc. Int. Memory Workshop (IMW)*, 2021, pp. 99–102.

[8] M. Vandemaele *et al.*, "Full (V_g, V_d) bias space modeling of hot-carrier degradation in nanowire FETs," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Mar. 2019, pp. 1–7.

[9] Z. Stanojevic *et al.*, "Scaling FDSOI technology down to 7 nm—A physical modeling study based on 3D phase-space subband Boltzmann transport," in *Proc. Joint Int. EUROSOI Workshop Int. Conf. Ultimate Integr. Silicon (EUROSOI-ULIS)*, Mar. 2018, pp. 1–4.

[10] M. Noei, T. Linn, and C. Jungemann, "A numerical approach to quasi-ballistic transport and plasma oscillations in junctionless nanowire transistors," *J. Comput. Electron.*, vol. 19, no. 3, pp. 975–986, Sep. 2020, doi: 10.1007/s10825-020-01488-4.

[11] 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," *IEEE Trans. Electron Devices*, vol. 55, no. 11, pp. 2886–2897, Nov. 2008.

[12] M. Lenzi *et al.*, "Investigation of the transport properties of silicon nanowires using deterministic and Monte Carlo approaches to the solution of the Boltzmann transport equation," *IEEE Trans. Electron Devices*, vol. 55, no. 8, pp. 2086–2096, Aug. 2008.

[13] E. Gnani, A. Gnudi, S. Reggiani, M. Luisier, and G. Baccarani, "Band effects on the transport characteristics of ultrascaled SNW-FETs," *IEEE Trans. Nanotechnol.*, vol. 7, no. 6, pp. 700–709, Nov. 2008.

[14] Z. Stanojevic, 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 *Proc. Int. Conf. Simul. Semiconductor Processes Devices (SISPAD)*, Sep. 2020, pp. 63–66.

[15] Z. Stanojević *et al.*, "Consistent low-field mobility modeling for advanced MOS devices," *Solid-State Electron.*, vol. 112, pp. 37–45, Oct. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038110115000532

[16] Z. Stanojević. *Physical Mobility Modeling for TCAD Device Simulation*. Accessed: May 6, 2021. [Online]. Available: https://resolver.obvsg.at/urn:nbn:at:at-ubtuw:1-6530

[17] C. Jacoboni and L. Reggiani, "The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials," *Rev. Modern Phys.*, vol. 55, no. 3, pp. 645–705, Jul. 1983. [Online]. Available: http://link.aps.org/doi/10.1103/RevModPhys.55.645

[18] S. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the universality of inversion layer mobility in Si MOSFET's: Part II-effects of surface orientation," *IEEE Trans. Electron Devices*, vol. 41, no. 12, pp. 2363–2368, Dec. 1994.

[19] S. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the universality of inversion layer mobility in Si MOSFET's: Part I-effects of substrate impurity concentration," *IEEE Trans. Electron Devices*, vol. 41, no. 12, pp. 2357–2362,Dec. 1994.

[20] H. Si, "TetGen, a delaunay-based quality tetrahedral mesh generator," *ACM Trans. Math. Softw.*, vol. 41, no. 2, pp. 1–36, Feb. 2015, doi: 10.1145/2629697.

[21] Global TCAD Solutions. *Vienna Schrödinger-Poisson*. Accessed: May 6, 2021. [Online]. Available: http://www.globaltcad.com/vsp

[22] T. Chiarella *et al.*, "Towards high performance sub-10 nm finW bulk FinFET technology," in *Proc. 46th Eur. Solid-State Device Res. Conf. (ESSDERC)*, Sep. 2016, pp. 131–134.

[23] T. Karatsori *et al.*, "Statistical characterization and modeling of drain current local and global variability in 14 nm bulk FinFETs," in *Proc. Int. Conf. Microelectron. Test Struct. (ICMTS)*, Mar. 2017, pp. 1–5.

[24] E. Capogreco *et al.*, "First demonstration of vertically stacked gate-all-around highly strained germanium nanowire pFETs," *IEEE Trans. Electron Devices*, vol. 65, no. 11, pp. 5145–5150, Nov. 2018.

[25] O. Penzin, L. Smith, A. Erlebach, M. Choi, and K.-H. Lee, "Kinetic velocity model to account for ballistic effects in the drift-diffusion transport approach," *IEEE Trans. Electron Devices*, vol. 64, no. 11, pp. 4599–4606, Nov. 2017.