Simulation of Semiconductor Processes and Devices 2016
Edited by E. Bär, J. Lorenz, and P. Pichler

On The Efficient Methods To Solve Multi-Subband BTE in 1D Gas Systems: Decoupling Approximations Versus The Accurate Approach

Anh-Tuan Pham, Zhengping Jiang, Seonghoon Jin, Jing Wang, Woosung Choi
Samsung Semiconductor, Inc., San Jose, CA, USA
Email: at.pham@samsung.com

Mohammad Ali Pourghaderi, Jongchol Kim, Keun-Ho Lee
Semiconductor R&D Center, Samsung Electronics, Hwasung-si, Gyeonggi-do, Korea

***Abstract*—For the deterministic MSBTE solver in 1D gas systems, efficient methods including the accurate approach, the decoupling of even-odd components of the distribution function (f_e–f_o decouple), and the decoupling of subband index and subband energy (RTA) have been implemented. For 3D MOSFET simulations, compared to the accurate approach, the f_e–f_o decouple approach helps to retain the accuracy of drain current calculation, and, at the same time, significantly reduce the turn-around-time as well as the memory usage. The hybrid scheme combining the decouple and the RTA approach is a good option to improve further the efficiency, while high accuracy of the drain current calculation is remained.**

I. Introduction

For nano-scaled device simulations, the deterministic multi-subband BTE (MSBTE) approach, which involves the self-consistent solution of the MSBTE, the Schrödinger equation (SE), and the Poisson equation (PE), has been used in the literatures [1]–[3]. The MSBTE approach has become even more important for simulations of MOSFET at sub 20 nm technology nodes, due to flexibilities for incorporating essential effects, which influence the performance of such a small device, such as the 2D quantization, the scatterings, the screening, the non-parabolic (NMOS) or warped (PMOS) band structure, and even the direct source-drain tunneling [4]. In this paper, the decoupling approximations in order to improve the efficiency of the MSBTE approach are focused. For a 1DEG or 1DHG system for NMOSFETs or PMOSFETs, respectively, we implement for the MSBTE solver (i) the accurate approach, and the approximated approaches including (ii) the decoupling of even-odd components of the distribution function (f_e–f_o decouple), and (iii) the decoupling of subband index and subband energy, i.e. the relaxation time approximation (RTA). We also implement (iv) the hybrid scheme, where the accurate (i) or decoupling (ii) approach is applied for the main channel region and the RTA approach (iii) is applied for the remaining regions of the device. We examine the efficiency and the accuracy of the 4 methods. In section II, the method is described. Results are shown in section III. Finally, conclusion are drawn in section IV.

978-1-5090-0818-6/16/$31.00

II. Method

The H-transformation is used for the discretization of the MSBTE [2]. Using the H-transformation the free-streaming operator of the MSBTE in 1D $\vec{k}$-space can be derived to take the following forms:

$$\frac{\partial}{\partial x}\left\{\{v_x Z\}^\nu(x,H) f_o^\nu(x,H)\right\}, \qquad \text{for } f_e^\nu(x,H) \tag{1}$$

$$\frac{\partial}{\partial x}\left\{\{v_x Z\}^\nu(x,H) f_e^\nu(x,H)\right\}, \qquad \text{for } f_o^\nu(x,H) \tag{2}$$

where f_e and f_o are the even and odd components of the distribution function f ($f_e(k) = (f(k) + f(-k))/2$ and $f_o(k) = (f(k) - f(-k))/2$). In (1) and (2), x is the transport direction, ν is the subband index, v is the group velocity, Z the generalized DOS for a single spin direction, H is the Hamilton (i.e. the total energy). Note that there is no same parity coupling in the free-streaming operator because f_e only couples to f_o in (2), and f_o only couples to f_e in (1).

However, the distribution function parity coupling within the scattering integral is much more complicated depending on the type of scattering mechanisms whether it is anisotropic or isotropic. In general, the following coupling is possible in the anisotropic scattering integral: the inverse parity coupling ($f_o^{\nu'}(x,H') \leftrightarrow f_e^\nu(x,H)$, $f_e^{\nu'}(x,H') \leftrightarrow f_o^\nu(x,H)$), and the same parity coupling ($f_e^{\nu'}(x,H') \leftrightarrow f_e^\nu(x,H)$, $f_o^{\nu'}(x,H') \leftrightarrow f_o^\nu(x,H)$). Here, unprimed and primed notations are associated with the initial and final states, respectively.

For the special case of isotropic scattering the inverse parity coupling is canceled out, and only the same parity coupling remains. However, there is still coupling between different energies H', H where $H' \neq H$ due to the in-elasticity. For the f_o equation, if the value of f_e from the previous iteration is used for the in-elastic scattering term, then the scattering integral can be expressed in terms of a relaxation time $\tau^\nu(x,H)$.

$$\hat{S}_{\text{iso.}}(f_o) = -\frac{Z^\nu(x,H) f_o^\nu(x,H)}{\tau^\nu(x,H)} \tag{3}$$

Putting the free-streaming (2) and scattering integral (3) together, we obtain a closed form for f_o as follows:

$$f_o^\nu(x,H) = -\frac{\tau^\nu(x,H)}{Z^\nu(x,H)}\frac{\partial}{\partial x}\left\{\{v_x Z\}^\nu(x,H) f_e^\nu(x,H)\right\} \quad (4)$$

Substituting (4) into (1), and putting the resultant free streaming operator together with the scattering integral, we obtain the final equation for f_e:

$$\frac{\partial}{\partial x}\left\{\{v_x Z\}^\nu \frac{\tau^\nu}{Z^\nu}\frac{\partial}{\partial x}\left\{\{v_x Z\}^\nu f_e^\nu\right\}\right\} = \hat{S}_{\text{iso.}}(f_e) \quad (5)$$

It can be seen that the free streaming operator in (5) now contains only f_e because f_o is decoupled. Moreover, within the isotropic scattering integral $\hat{S}_{\text{iso.}}(f_e)$ in (5), the f_o is fully decoupled from f_e due to the isotropicity approximation. Note that, there is still energy coupling in (5) because the elasticity assumption is not used within $\hat{S}_{\text{iso.}}(f_e)$. Therefore, we only need to solve (5) to obtain f_e, and use (4) to determine f_o. This approach is called f_e–f_o decouple in order to distinguish with the accurate approach (f_e–f_o couple) where the full scattering without any approximation and simplification for the scattering mechanisms is considered. The advantage of the f_e–f_o decouple approach is that the number of unknowns is reduced by a factor of 2 compared to the accurate approach. Therefore, the decouple approach consumes less CPU time and less memory. If the $f_e^{\nu'}(x,H')$-term within the $S_{\text{iso.}}(f_e^\nu(x,H))$ is calculated with $f_e^{\nu'}(x,H')$ from the previous iteration and the equilibrium f_{eq} is introduced, then $S_{\text{iso.}}(f_e^\nu(x,H))$ takes the same form as (3), where $f_o \leftarrow f_e - f_{\text{eq}}$. Consequently, the energy coupling as well as subband coupling vanishes in (5). This results in the RTA MSBTE. For the RTA, an additional equation for subband quasi-Fermi energy associated with f_{eq} needs to be solved in order to conserve the number of particle within the subband [5].

In order to retain the accuracy, and, at the same time, improve the efficiency, a hybrid scheme is considered. Within the hybrid scheme, the accurate or the $f_e - f_o$ decouple approach is applied for the main channel region and the RTA approach is applied for the remaining regions of the device.

In order to correct the effects of anisotropic scattering, the scattering rate is multiplied with a normalized correction factor (e.g. [6]). Scatterings due to phonons, SR, and ionized impurity (Coulomb) are included, where the anisotropicity of SR and Coulomb scatterings is accounted for. Rigorous Lindhard screening treatment based on tensorial dielectric function is used. The MSBTE is solved self-consistently with the EMA (for NMOSFET) or the 6×6 $\vec{k}\cdot\vec{p}$ (for PMOSFET) SE, and the PE using the Gummel liked iteration scheme.

III. Results

For comparison, rectangular 5×7 nm GAA Si NMOSFET and PMOSFET is simulated. Surface orientation of the 5 nm wall is (110) direction and channel orientation is <110> direction. Gate length is $L_g = 13$ nm. Source, drain extension is $L_s = L_d = 10$ nm. Effective oxide thickness is EOT = 0.8 nm. 2 GPa tensile (NMOSFET) or compressive (PMOSFET) uniaxial stress is applied along the transport direction.

TABLE I
Normalized TAT and memory usage for |VD| = 0.7 V

		couple	decouple	RTA
NMOSFET	TAT [au]	7.67	2.12	1.0
	RAM [au]	3.4	2.8	1.0
PMOSFET	TAT [au]	0.88	0.8	1.0
	RAM [au]	1.69	1.32	1.0

Fig. 1 shows ID-VG curves for linear and high-field transport regimes. The RTA approach causes up to 20% error for ID compared to the accurate (couple) method, while the error is only 2% for the decouple method case.

Normalized turn-around-time (TAT) and peak memory usage is shown in Tab. I for |VD| = 0.7 V. For NMOSFET, the decouple approach helps to reduce the TAT significantly by about 3.6 X compared to the accurate approach, and the TAT enhancement factor for the RTA approach vs the accurate approach is about 7.7 X. The peak memory usage for the decouple and RTA approach is reduced by a factor of 1.2 X and 3.4 X, respectively, compared to the couple approach case.

For PMOSFET, the computational efficiency gain is not much for the decoupling approximations due to the fact that the 6×6 $\vec{k}\cdot\vec{p}$ SE eigen solver consumes much CPU time to compute the eigen states for many k-point in the 1D k-space. Additionally, the number of unknowns of MSBTE for PMOSFET is smaller than the one for NMOSFET case because we need to solve f_e, f_o for 3 degenerate ladders of X-valleys of the conduction band, while for holes we need to solve f_e, f_o for only Γ-valley. Due to this reduction in the number of f_e, f_o unknowns, for PMOSFET, the reduction of peak memory usage for decoupling approximation approaches is also smaller than the one for NMOSFET case, as shown in Tab. I.

Electron inversion charge (Ninv) and average electron drift velocity (v_{drift}) is shown in Fig. 2 for NMOSFET for VD = 50 mV (top) and VD = 0.7 V (bottom) @ VG = 0.8 V. For the linear transport regime, the decoupling approximations capture well the details of transport like Ninv and v_{drift} calculated based on the accurate approach. For the high-field transport regime, only the decouple approach can capture the Ninv and v_{drift} from the accurate approach, while the RTA approach strongly overestimates v_{drift}.

Phase space diagram of $f_o(x,\pm k)$ for first subband of the most occupation ladder is shown in Fig. 3 for NMOSFET @ VD = 0.7 V, VG = 0.8 V. The decouple approach reproduces well the details of $f_o(x,\pm k)$ calculated based on the accurate approach, while the RTA approach strongly overestimates the accurate result of $|f_o(x,\pm k)|$.

For the hybrid scheme, the decouple approach is applied for the main channel region and the RTA for the remaining regions. The decouple approach region, which covers the channel region, is ranging from $-(L_g/2 + L_{\text{margin}})$ to $(L_g/2 + L_{\text{margin}})$, where L_{margin} is the marginal length. The

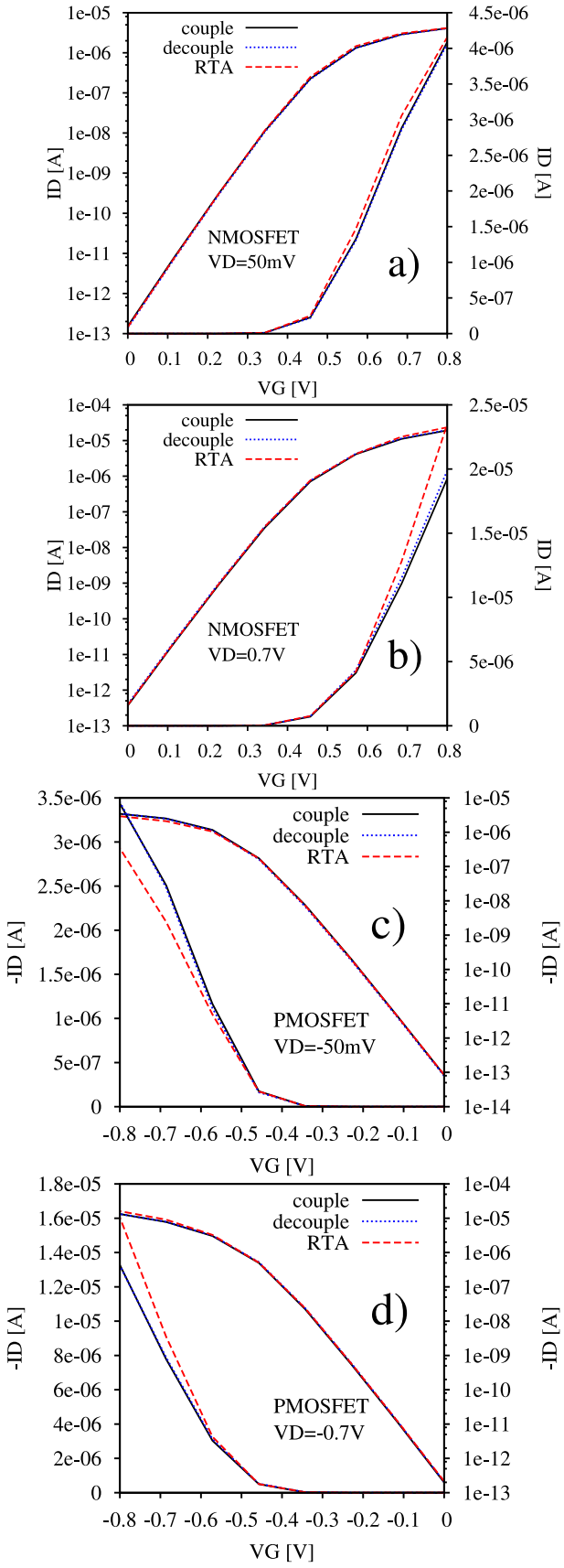

Fig. 1. ID-VG curves for NMOSFET (a, b) and PMOSFET (c, d) for |VD| = 50 mV (a, c) and |VD| = 0.7 V (b, d).

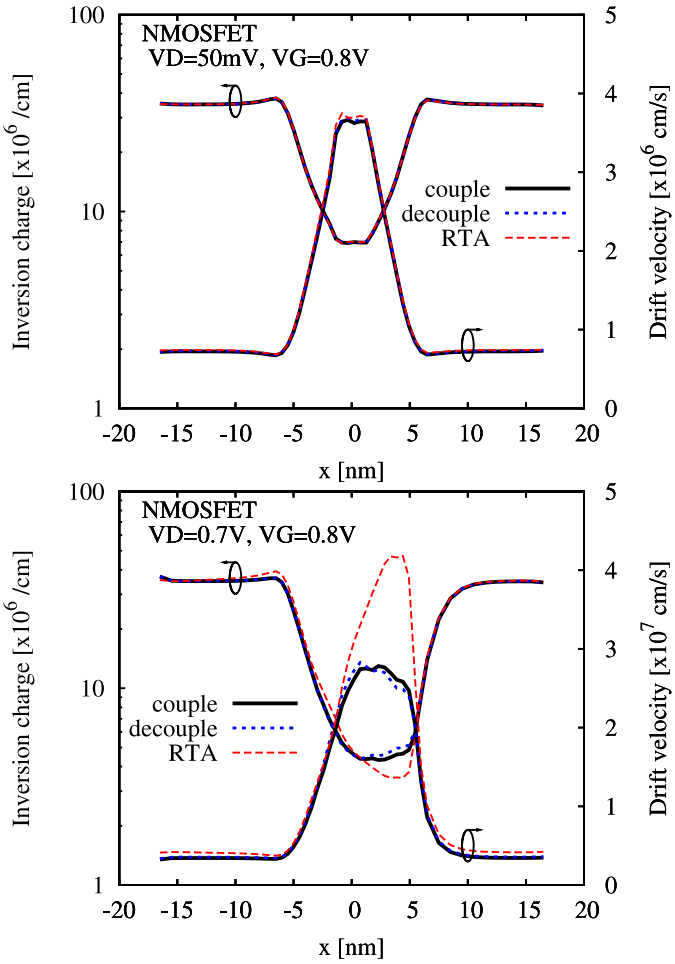

Fig. 2. Electron inversion charge and average electron drift velocity for NMOSFET. VD = 50 mV (top) and 0.7 V (bottom).

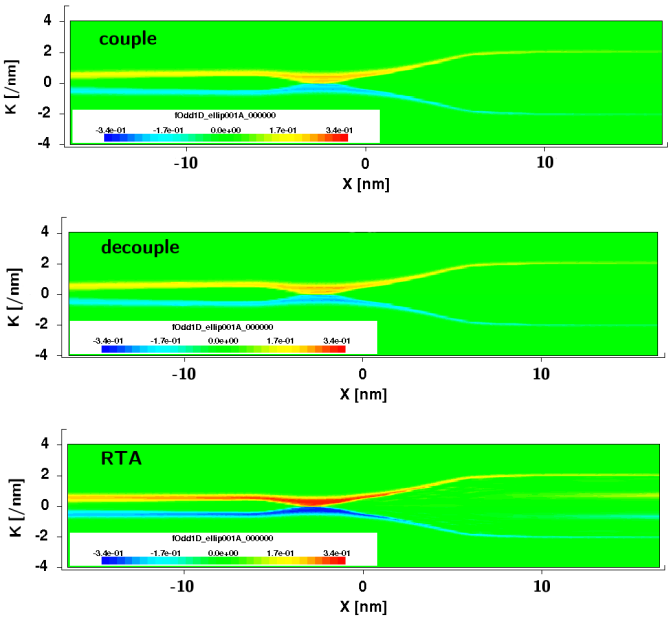

Fig. 3. $f_o(x, \pm k)$ for NMOSFET. @VD = 0.7 V, VG = 0.8 V.

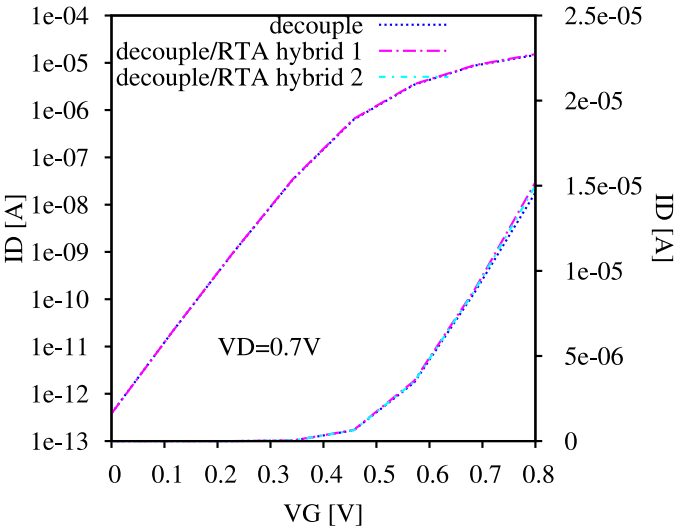

Fig. 4. ID-VG curves based on decouple/RTA hybrid 1 ($L_{\text{margin}} = L_s/4$) and 2 ($L_{\text{margin}} = L_s/2$) scheme vs decouple scheme ($L_{\text{margin}} = L_s$) for NMOSFET for VD = 0.7 V.

TABLE II
NORMALIZED TAT AND MEMORY USAGE OF HYBRID SCHEME FOR VD = 0.7 V.

	decouple/RTA hybrid 2 ($L_{\text{margin}} = L_s/2$)	decouple/RTA hybrid 1 ($L_{\text{margin}} = L_s/4$)	decouple ($L_{\text{margin}} = L_s$)
TAT [au]	0.8	0.9	1.0
RAM [au]	0.6	0.7	1.0

source, drain extension is now extended to $L_s = L_d = 20$ nm. We compare the 2 cases: $L_{\text{margin}} = L_s/4 = 5$ nm and $L_{\text{margin}} = L_s/2 = 10$ nm with the case where the decouple approach is applied for the entire device ($L_{\text{margin}} = L_s$).

Fig. 4 shows ID-VG curves for NMOSFET at high VD = 0.7 V for 3 cases of L_{margin}. It can be seen that the hybrid scheme ($L_{\text{margin}} < L_s$) causes a small error of 3-4% compared to the case where the decouple approach is applied for the entire device ($L_{\text{margin}} = L_s$).

As shown in table II, the hybrid scheme helps to reduce the TAT up to 20% and it also helps to reduce the memory usage up to 40% compared to the decouple approach for the entire device.

Ninv and v_{drift} is shown in Fig. 5 for VD = 0.7 V @ VG = 0.8 V. For such the high-field transport regime, the hybrid scheme can still capture the Ninv and v_{drift} from the decouple approach in the source, drain, and even in the channel near the source. However, the difference is larger, once the electrons are accelerated in the middle of channel. The difference is largest, when the electrons get hottest near the drain/channel interface. The main reason for this is that the RTA approximation in the drain side underestimates the second order effect on the distribution function, while the electrons are still not fully thermalized.

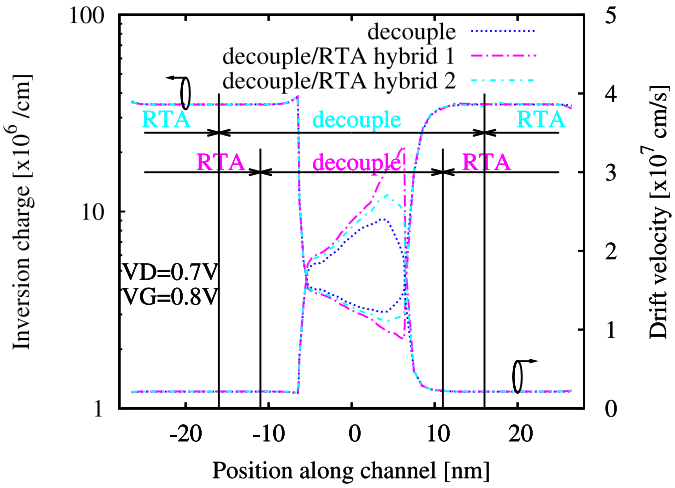

Fig. 5. Electron inversion charge and average electron drift velocity based on decouple/RTA hybrid 1 ($L_{\text{margin}} = L_s/4$) and 2 ($L_{\text{margin}} = L_s/2$) scheme vs decouple scheme ($L_{\text{margin}} = L_s$) for NMOSFET for VD = 0.7 V, VG = 0.8 V.

IV. CONCLUSION

For 3D MOSFET device simulations, the accurate approach together with decoupling approximations including the decoupling of even-odd components of the distribution function (f_e–f_o decouple), and the decoupling of subband index and subband energy (RTA) has been implemented for the deterministic MSBTE solver in 1D gas systems. Compared to the accurate approach, the f_e–f_o decouple approach helps to retain the accuracy of drain current calculation, and, at the same time, significantly reduce the turn-around-time as well as the memory usage. The hybrid scheme combining the decouple and the RTA approach is a good option to improve further the efficiency, while high accuracy of the drain current calculation is remained.

ACKNOWLEDGMENT

The authors express appreciation to Prof. S. M. Hong and Prof. C. Jungemann for many helpful discussions.

REFERENCES

[1] S. Jin, A. T. Pham, W. Choi, Y. Nishizawa, Y. Kim, K. H. Lee, Y. Park, and E. S. Jung, "Performance evaluation of InGaAs, Si, and Ge nFinFETs based on coupled 3D drift-diffusion/multisubband Boltzmann transport equations solver," *IEDM Tech. Dig.*, 2014.

[2] S. M. Hong, A. T. Pham, and C. Jungemann, *Deterministic solvers for the Boltzmann transport equation*. SpringerWienNewYork: Springer, 2011.

[3] Z. Stanojevic, O. Baumgartner, F. Mitterbauer, H. Demel, C. Kernstock, M. Karner, V. Eyert, A. France-Lanord, P. Saxe, C. Freeman, and E. Wimmer, "Physical modeling - a new paradigm in device simulation," *IEDM Tech. Dig.*, 2015.

[4] P. Palestri, L. Lucci, S. D. Tos, D. Esseni, and L. Selmi, "An improved empirical approach to introduce quantization effects in the transport direction in multi-subband monte carlo simulations," *Semicond. Sci. Technol.*, vol. 25, pp. 055011–055020, 2010.

[5] S. Jin, T. Tang, and M. V. Fischetti, "Simulation of silicon nanowire transistors using Boltzmann Transport Equation under relaxation time approximation," *IEEE Trans. Electron Devices*, vol. 55, no. 3, pp. 727–736, 2008.

[6] Z. Stanojevic, O. Baumgartner, M. Karner, L. Filipovic, C. Kernstock, and H. Kosina, "On the validity of momentum relaxation time in low-dimensional carrier gases," in *Proc. SISPAD*, pp. 181–184, 2014.