TCAD Models of the Ballistic Mobility in the Source-to-Drain Tunneling Regime

A. Schenk, P. Aguirre

Integrated Systems Laboratory ETH Zürich, Gloriastrasse 35, 8092 Zürich, Switzerland

Abstract

TCAD models of the ballistic mobility are developed where the mean ballistic velocity is not a fitting parameter, but a function of either the quasi-Fermi potential or the density. In the first case, a local version can be derived which is more robust when used together with a model for source-to-drain tunneling. The second form conserves the thermionic ballistic current and better matches the on-currents of short-channel FETs obtained from a quantum-transport solver, in particular at low source-drain bias. It requires the iterative extraction of the top-of-the-barrier density. This is the only non-local remnant of the hydrodynamic term in the balance equation for the mean velocity which is discarded in all commercial 2D/3D device simulators. The ballistic mobility, used with the Matthiessen rule, substitutes for this term and prevents that the drift-diffusion current diverges in the limit of zero gate length. The numerical integration of the models with the TCAD simulator S-Device is set out, and simulated transfer characteristics of $In_{0.53}Ga_{0.47}As$ double-gate ultra-thin-body FETs with gate lengths ranging from 7 nm to 40 nm are compared with the corresponding quantum-transport results. It is shown that under conditions of dominant source-to-drain tunneling, the concepts of local quasi-Fermi potential and mean ballistic velocity break down. Suggestions for non-local modifications of both the mobility and tunneling models are given that would allow to use the same setup for all gate voltages from deep sub-threshold to deep inversion.

1. Introduction

As the length of transistor channels is scaled down into the range of a few nanometers, quantum [1, 2] and ballistic effects [3] start to play a major role. Quantum transport (QT) solvers (e.g. [4])

Email address: `schenk@iis.ee.ethz.ch` (A. Schenk) can accurately simulate such devices, but they are computationally expensive and still not mature for industrial environments. In drift-diffusion (DD) TCAD one tries to mimic the above effects with models that depend on the continuous solution variables of the van Roosbroeck equation system [5]. In [6] the quantum drift-diffusion (QDD) tool Sentaurus-Device from Synopsys [7] was used to simulate source-to-drain tunneling (STDT) in $In_{0.53}Ga_{0.47}As$ double-gate ultra-thin-body (DG UTB) FETs with gate lengths L_G ranging from 10 nm to 25 nm (see Fig. 1). Their $I_D V_{GS}$-characteristics exhibit a pronounced current overshoot after the onset of inversion due to the exclusive usage of a diffusive mobility (μ_d). Since electron-phonon scattering is weak in these devices, the so-called ballistic QT solution can be taken as reference. The aim of this work is to develop TCAD models of the "ballistic mobility" capable to limit the current to the correct QT values and to test their applicability together with source-to-drain tunneling (STDT). As distinct from various models in the literature [8, 9, 10, 11, 12], the ballistic velocity is modeled explicitly either as a function of the quasi-Fermi potential (QFP) or the density (n). Such choices are "TCAD-friendly" and hence preferred to other dependencies e.g. on the electric field [13]. Since the mean ballistic velocity is the key quantity of the models, the physical consequences of the different expressions are analyzed and the resulting transfer characteristics are compared with each other. As shown, the origin of the "ballistic mobility" is the omitted quadratic term in the conservation law of the mean velocity. A factor with the dimension of a mobility can be singled out in this term, hence the quotation marks in "ballistic mobility" will be skipped as usually done in the community.

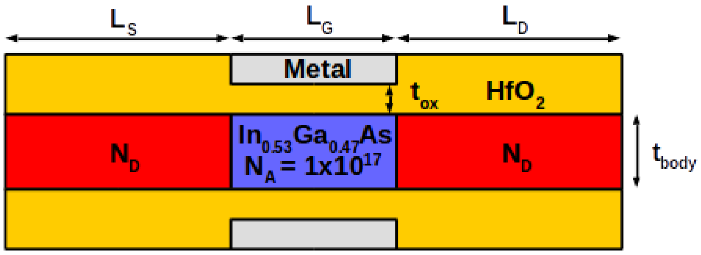

Figure 1: Schematic of the $In_{0.53}Ga_{0.47}As$ double-gate UTB FET used as test device.

The paper starts with a rigorous classification of the different transport regimes in Section 2. In particular, the difference between “kinetic” and “ballistic” is worked out since it is essential to understand the steps towards a *local* model of the ballistic mobility. Then, the kinetic limit is used in Section 3 to derive the mean ballistic velocity as function of the QFP in local form. The necessary assumptions and simplifications are pointed out and the induced error is discussed. This also provides a natural explanation for the observed weakness of such a model in the linear regime [13]. In Section 4, a simple model of the mean ballistic velocity as function of density is proposed which conserves the current. In this model, the position of the top of the source-to-drain potential barrier must be found numerically. With this non-locality the injection velocity becomes independent of source resistance and bias. Details of the numerical implementation are provided in Section 5. The interplay between the ballistic mobility models and a model for STDT with local QFPs is studied in Section 6. Since in the regime of dominant STDT the TCAD variables QFP and density are determined by the tunneling carriers, their usage in the ballistic mobility leads to an artificial suppression of the tunneling current. To circumvent this problem, a modification of the STDT model is suggested. Finally, Section 7 discusses the shortcomings of the models implicated by the locality, the problem of an inhomogeneous band structure, and the difficulty to compare QDD with QT characteristics due to the different density-of-states (DOS) models.

2. Transport Regimes and Ballistic Mobility

The first moment of the Boltzmann transport equation in the relaxation time approximation, reads in the stationary, iso-thermal, and non-degenerate case, as [14]

$$\left(\mu_{\mathrm{d}}^{-1} + \frac{m_{\mathrm{e}}}{q}\vec{v}\cdot\nabla\right)\vec{v} = \nabla\psi_{\mathrm{n}}\,, \tag{1}$$

where μ_{d} is the diffusive mobility, $\vec{v}$ the mean velocity, m_{e} the effective mass, and q the elementary charge. The function $\psi_{\mathrm{n}} = \Phi - V_{\mathrm{T}} \ln(n/n_{\mathrm{i}})$ on the right-hand side contains the electrostatic potential Φ, the density n, the intrinsic density n_{i}, and the thermal voltage $V_{\mathrm{T}} = k_{\mathrm{B}}T/q$. It has the meaning of the QFP, provided the concept of local equilibrium is (at least approximately) applicable. The second term on the left-hand side, typical for hydrodynamic equations, is generally skipped in

2D/3D device simulators to ensure convergence. Simplifying Eq. (1) to the 1D case (channel direction x) one obtains

$$\frac{v}{\mu_\mathrm{d}} + \frac{m_\mathrm{e}}{q} v\, v' = \psi_\mathrm{n}' \,. \tag{2}$$

The second term becomes larger than the first one if $\mu_\mathrm{d} > q/(m_\mathrm{e}|v'|)$. For an estimate one can replace v' by $v_\mathrm{th}/L_\mathrm{G}$ with the root mean square (r.m.s.) 1D thermal velocity $v_\mathrm{th} = \sqrt{k_\mathrm{B}T/m_\mathrm{e}}$ which gives approximately $\mu_\mathrm{d} > L_\mathrm{G}\sqrt{q/(V_\mathrm{T}m_\mathrm{e})} \equiv \mu_\mathrm{lim}$. For example, $L_\mathrm{G} = 15\,\mathrm{nm}$ and $m_\mathrm{e} = 0.0516\,m_0$ yield $\mu_\mathrm{lim} = 1719\,\mathrm{cm^2/Vs}$.

The proportionality to v in the second term of Eq. (2) suggests to introduce a quantity called *ballistic mobility* in the form

$$\mu_\mathrm{b} = \frac{q}{m_\mathrm{e}} \frac{1}{v_\mathrm{b}'} \tag{3}$$

which requires to define a local model for the mean ballistic velocity $v_\mathrm{b}(x)$.

The *ballistic* regime is defined by μ_d " $\gg$ " μ_lim. The quotation marks mean that μ_d still has to be small enough such that a local, non-constant QFP $\psi_\mathrm{n}(x)$, i.e. the concept of local thermodynamic equilibrium, can be assumed and used. In the simulation samples sketched in Fig. 1, the measured bulk mobility of $\mathrm{In_{0.53}Ga_{0.47}As}$ ($\sim 10^4\,\mathrm{cm^2/Vs}$) could be seen as such a limit.

There is no analytical solution of the differential equation (2) even for constant mobility μ_d. Setting $\mu_\mathrm{d} = \infty$ makes Eq. (2) to an Euler equation and defines the *kinetic* transport regime (see Table 2) where in analogy to Eq. (3) a *kinetic mobility* could be introduced in the form

$$\mu_\mathrm{k} = \frac{q}{m_\mathrm{e}} \frac{1}{v_\mathrm{k}'} \tag{4}$$

with a mean kinetic velocity v_k. This expression for μ_k can also be motivated from Heisenberg's uncertainty principle (a method often used in the past for the development of TCAD mobility models, e.g. [17]). In the limit $L_\mathrm{G} \to 0$ one can write $\Delta t\,\Delta E \approx \tau_\mathrm{k}\Delta E_\mathrm{k} \approx \hbar/2 \approx \Delta x\, m_\mathrm{e} v_\mathrm{k} \approx \Delta x\,\Delta p$, where τ_k is a 'kinetic relaxation time' related to the kinetic mobility by $\mu_\mathrm{k} = q\,\tau_\mathrm{k}/m_\mathrm{e}$. By division it follows that $\tau_\mathrm{k} = (m_\mathrm{e}v_\mathrm{k}\Delta x/\Delta k)(\Delta E_\mathrm{k}/\Delta k)^{-1}$, and with the parabolic band approximation $\Delta E_\mathrm{k}/\Delta v_\mathrm{k} = m_\mathrm{e}v_\mathrm{k}$ one obtaines Eq. (4).

Since $\mu_\mathrm{d} \gg \mu_\mathrm{b}$ holds in the ballistic transport regime, the left-hand side of Eq. (2) can be approximated by $v_\mathrm{b}/\mu_\mathrm{b}$ (neglecting the first term) which yields another definition of the ballistic mobility:

$$\mu_{\mathrm{b}}(x) = \frac{v_{\mathrm{b}}(x)}{\psi_{\mathrm{n}}'(x)}. \tag{5}$$

The nomenclature for the different transport regimes in relation to QFP and mobilities is summarized in Table 2. In the emission regime, the electron is decoupled from any thermodynamic bath, i.e. short-range electron-electron interaction is not efficient to ensure particle and energy exchange, hence a QFP is not defined. In textbooks, the course of the QFP in emission regions is often symbolized by crosses. Diodes (pn-junction and Schottky) have been treated in the emission theory with the result that the slope of the current-voltage characteristics is the same as in the diffusion theory, only the pre-factors differ (by a moderate factor).

In the kinetic transport regime, electrons keep the QFP of the contact from which they are injected until they hit the opposite contact. Short-range electron-electron interaction is efficient to ensure particle and energy exchange in the ensemble of moving electrons. The QFP is constant in the channel because of the absence of scatterings that can relax the total momentum of the system. At the channel-drain junction, due to the vast number of electrons in the drain reservoir and due to the very quick thermalization by short-range electron-electron interaction, the incoming electrons rapidly become members of the drain ensemble, i.e. they attain the Fermi level of the drain. Thus, the change of the QFP is step-like. In most of state-of-the-art NEGF QT simulators, short-range electron-electron interaction is not included in the Hamiltonian, and the position of the step remains undetermined.

Table 1: Nomenclature for different transport regimes in relation to QFP and mobilities.

emission	QFP in contacts, no QFP defined in channel
kinetic	QFP in contacts, constant QFP in channel, $\mu_{\mathrm{d}} = \infty$
ballistic	non-constant QFP $\psi_{\mathrm{n}}(x)$ in channel can be assumed and used, μ_{d} “ $\gg$ ” μ_{lim}
quasi-ballistic	intermediate regime, $\mu_{\mathrm{d}} \approx \mu_{\mathrm{b}}$
diffusive	dissipative regime, $\mu_{\mathrm{d}} \ll \mu_{\mathrm{b}}$

In the ballistic regime, a weak relaxation of the total momentum disturbs the kinetic motion (similar to the trajectory of a cannonball which is not parabolic but ballistic due to the air drag). In the community the name “ballistic” is commonly used instead of “kinetic/emission”. DD TCAD simulators are based on the concept of local thermodynamic equilibrium. The value of the diffusive mobility μ_{d} must not be much larger than the value of μ_{lim}. The latter corresponds to the case where the momentum relaxation time is approximately equal to the transit time. As shown by Frensley in his pioneering paper [3], the shape of the QFP along the channel is essentially given by an error function if the barrier is treated with the saddle-point method. This analytical solution for $\psi_{\mathrm{n}}(x)$ does not depend at all on the value of the diffusive mobility μ_{d}, only $\mu_{\mathrm{d}} = \infty$ is excluded. However, the existence of a continuous function $\psi_{\mathrm{n}}(x)$ warrants the limitation of the diffusive mobility to μ_{lim}. When $L_{\mathrm{G}} \to 0$ (kinetic case), the error function is “squeezed” to a step function.

A step-like change of the QFP can also occur in DD transport as the result of sharp density gradients, as visible in Fig. 2b). The figure shows a TCAD simulation of the QFP along the channel of the test transistor of Fig. 1. A gate length of 40 nm was used which is long enough to safely neglect STDT. The softened step function is symmetrical in the linear regime (50 mV) with an almost linear potential drop typical for an Ohmic resistor, but strongly asymmetrical in the saturation regime (0.63 V) with a sharp edge at the drain-side pn-junction ($x = x_{\mathrm{j}} = 80\,\mathrm{nm}$). To the left of x_{j}, the QFP belongs to the moving electrons coming from the source which have a very small density (that nevertheless determines the total value of n in this region), but to the right of x_{j} the QFP belongs to the thermalized high-density electrons of the drain. A heavy S/D doping of $N_{\mathrm{D}} = 5 \times 10^{19}\,\mathrm{cm}^{-3}$ was chosen to demonstrate the degeneracy effect. It shows up as a slight difference between $\psi_{\mathrm{n}}(x)$ and $\Phi(x) - V_{\mathrm{T}} \ln{(n(x)/n_{\mathrm{i}})}$ in the pn-junctions (see the insert of Fig. 2(a)). Whereas degeneracy is negligible for the shape of the QFP, it impacts the on-current of a FET due to a Fermi correction term $-\mu_{\mathrm{n}} n k_{\mathrm{B}} T \nabla \ln(\gamma_{\mathrm{n}})$ in the equation for the current density, where $\gamma_{\mathrm{n}} = n/n_{\mathrm{B}}$ is the ratio between the actual density and its Boltzmann form. This term tends to reduce the on-current (negative sign), thus it has the same effect as the ballistic mobility. For a better decoupling of both effects, the S/D doping is kept at $N_{\mathrm{D}} = 6 \times 10^{18}\,\mathrm{cm}^{-3}$ in Sections 3 and 4.

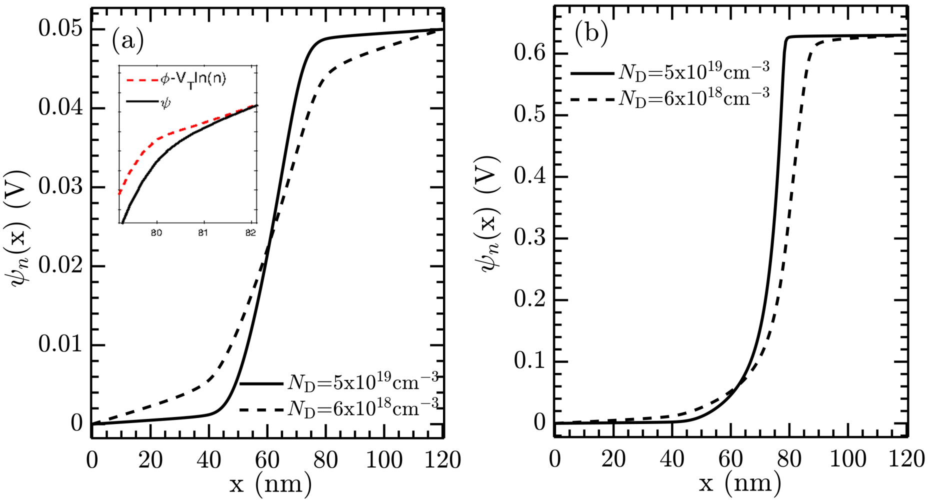

Figure 2: Quasi-Fermi potential $\psi_{\mathrm{n}}(x)$ along the channel of the transistor shown in Fig. 1 at (a) $V_{\mathrm{DS}} = 50\,\mathrm{mV}$, (b) $V_{\mathrm{DS}} = 0.63\,\mathrm{V}$ obtained from TCAD simulations with constant mobility. Parameters: $L_{\mathrm{G}} = 40\,\mathrm{nm}$, $V_{\mathrm{GS}} = 0.31\,\mathrm{V}$, $\mu_{\mathrm{d}} = 10^4\,\mathrm{cm^2/Vs}$. The small difference between $\psi_{\mathrm{n}}(x)$ and $\Phi(x) - V_{\mathrm{T}} \ln n(x)$ in the case $N_{\mathrm{D}} = 5 \times 10^{19}\,\mathrm{cm^{-3}}$ is shown in the insert of (a).

3. Ballistic Velocity as Function of Quasi-Fermi Potential

Two expressions for the ballistic mobility were suggested in Section 2: Eq. (3) and Eq. (5). Equating them leads to the balance equation

$$q\psi_{\mathrm{n}}'(x) = \frac{m_{\mathrm{e}}}{2}\left(v_{\mathrm{b}}^2\right)' \,, \tag{6}$$

stating that the loss of electro-chemical energy of the ballistic electrons is compensated by the gain of their kinetic energy. Integrating this equation from a starting point x_{S} near the source-channel junction to a point x in the channel results in a solution for the mean ballistic velocity:

$$v_{\mathrm{b}}(x) = \sqrt{v_{\mathrm{b}}^2(x_{\mathrm{S}}) + \frac{2q}{m_{\mathrm{e}}}\left[\psi_{\mathrm{n}}(x) - \psi_{\mathrm{n}}(x_{\mathrm{S}})\right]} \qquad (x > x_{\mathrm{S}})\,. \tag{7}$$

This solution contains two non-localities: the QFP and the velocity at the starting point x_{S}. In order to obtain a *local* model, one has to (i) define the starting point as the boundary of ballistic motion, (ii) to find a reasonable expression for $v_{\mathrm{b}}(x_{\mathrm{S}})$, and (iii) to replace $\psi_{\mathrm{n}}(x_{\mathrm{S}})$ by a reasonable value. For the choice of x_{S} one can proceed as in the compact modeling of the ballistic transistor and set $x_{\mathrm{S}} = x_{\mathrm{TOB}}$, where x_{TOB} is the so-called virtual source, i.e. the top-of-the-barrier (TOB) point. The only way to remove the non-locality $v_\mathrm{b}(x_\mathrm{TOB})$ is to consider the kinetic limit ($L_\mathrm{G} \to 0$), where only thermionic electrons with energies higher than the TOB contribute and are injected with the mean thermal velocity $v_\mathrm{th,i}$. Then,

$$v_\mathrm{k}(x_\mathrm{TOB}) = v_\mathrm{th,i} \frac{1 - e^{-V_\mathrm{DS}/V_\mathrm{T}}}{1 + e^{-V_\mathrm{DS}/V_\mathrm{T}}} = v_\mathrm{th,i} \tanh\left(\frac{V_\mathrm{DS}}{2V_\mathrm{T}}\right) \equiv \langle v_\mathrm{inj} \rangle \,. \tag{8}$$

Note that $v_\mathrm{k}(x_\mathrm{TOB})$ vanishes at thermodynamic equilibrium ($V_\mathrm{DS} = 0\,V$, $\psi_\mathrm{n}(x) = \mathrm{const}$). The mean thermal velocity $v_\mathrm{th,i}$ at the injection point could be set to the r.m.s. value in one dimension v_th or to the mean of the value in one dimension $v_\mathrm{th}\sqrt{2/\pi} \approx 0.8 v_\mathrm{th}$. The latter option is used in the following simulations. The actual value needed for 2D/3D device simulation might be slightly different. This will be further discussed in Section 7.

Finally, the value of $\psi_\mathrm{n}(x_\mathrm{TOB})$ is approximated by zero (grounded source contact). This removes the second non-locality, however it neglects the voltage drop up to the virtual source caused by the finite series resistance between contact and x_TOB. Inserting the above simplifications ($v_\mathrm{b}(x_\mathrm{S}) \to \langle v_\mathrm{inj} \rangle$, $\psi_\mathrm{n}(x_\mathrm{S} = x_\mathrm{TOB}) \to 0$) into Eq. (7) one obtains

$$v_\mathrm{b}(x) = \sqrt{\langle v_\mathrm{inj} \rangle^2 + 2 v_\mathrm{th}^2 \psi_\mathrm{n}(x)/V_\mathrm{T}} \,, \tag{9}$$

which is a local model up to the V_DS-dependence of the mean injection velocity $\langle v_\mathrm{inj} \rangle$. The same form of the ballistic velocity was suggested in Ref. [13], with a fitting parameter instead of the first term under the square root, and called Kinetic Velocity Model (KVM). The non-local factor $\tanh(V_\mathrm{DS}/2V_\mathrm{T})$ becomes 0.75 at 300 K and $V_\mathrm{DS} = 50\,\mathrm{mV}$, the standard voltage for the linear regime. It quickly approaches unity, when the source-drain bias exceeds a few V_T.

The mean ballistic velocity Eq. (9) was implemented in the Physical Model Interface (PMI) of S-Device (for details see Section 5). In Fig. 3 it is extracted from the self-consistent TCAD simulations and compared with the mean kinetic velocity from the effective-mass QT solver QTx [18]. The extraction of velocity profiles from QTx is delicate. To compare with the ballistic velocity of thermionic electrons (9), the QTx kinetic velocity is computed by

$$v_\mathrm{k} = \frac{I_\mathrm{T}}{q\, n_\mathrm{1D-T}} \,, \tag{10}$$

where I_T is the thermionic current and $n_\mathrm{1D-T}$ is the 1D density of thermionic electrons, obtained by integration of the thermally weighted LDOS over energies above the TOB. The thermionic (kinetic) electrons are assumed to possess the QFP of the source up to the drain contact. For the high source-drain bias, the velocity profiles agree well in the channel region (from 40 nm to 80 nm). The oscillatory behavior in the source found with QTx is caused by quantum reflection. In the drain, the velocity extracted from QTx becomes much larger than the one computed by Eq. (9) which is attributed to the 2D DOS model of QTx.

Note, that the total increase of the mean ballistic velocity is approximately given by $\sqrt{2\,V_{\mathrm{DS}}/V_{\mathrm{T}}}$. At x_{TOB} (indicated by an arrow in Fig. 3), the mean ballistic velocity has already increased from its initial value $\langle v_{\mathrm{inj}}\rangle = 1.78 \times 10^7$ cm/s at $V_{\mathrm{DS}} = 50$ mV and $\langle v_{\mathrm{inj}}\rangle = 2.37 \times 10^7$ cm/s at $V_{\mathrm{DS}} = 0.63$ V. In the linear regime the increase is $\approx 85\%$ and in the saturation regime $\approx 60\%$. This increase is due to the non-zero value $\psi_{\mathrm{n}}(x_{\mathrm{TOB}})$ (see Fig. 2), i.e. the price for the negligence of the last term in Eq. (7). It not only depends on source doping and source extension, but also on the gate bias V_{GS}. The latter dependence is uncritical because $\psi_{\mathrm{n}}(x_{\mathrm{TOB}})$ changes by less than $V_{\mathrm{DS}}/10$ over the whole gate voltage range. The three assumptions that lead to a local model will be further discussed in Section 7. In Section 4 another model of the ballistic velocity will be proposed which requires the numerical determination of x_{TOB}. Introducing this non-locality in the above model would allow to keep the term $\psi_{\mathrm{n}}(x_{\mathrm{TOB}})$ with the result that $v_{\mathrm{b}}(x_{\mathrm{TOB}}) = \langle v_{\mathrm{inj}}\rangle$.

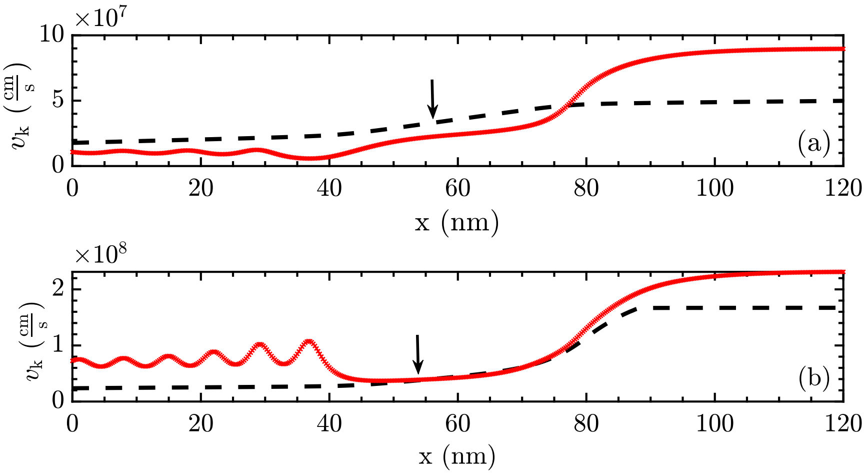

Figure 3: Comparison of the mean ballistic velocity Eq. (9) (black dashed, extracted from the self-consistent TCAD simulations) with the mean kinetic velocity extracted from QTx (red symbols). (a) $V_{\mathrm{DS}} = 50$ mV, (b) $V_{\mathrm{DS}} = 0.63$ V. Parameters: $N_{\mathrm{D}} = 6 \times 10^{18}$ cm^{-3}, $L_{\mathrm{G}} = 40$ nm, $V_{\mathrm{GS}} = 0.31$ V. The arrow points to x_{TOB}, the position of the TOB.

Frensley [3] suggested to make use of $\nabla \cdot \vec{j}_{\mathrm{n}} = 0$ in the Euler equation and to replace $(v^2)'$ by $-2v^2(\ln n)'$. Based on this he derived the kinetic (actually: ballistic) current density of a model problem with the essential feature that the current depends on the QFP at x_{TOB}, in contrast to the emission theory, where no QFP exists in the barrier region. Applying the same replacement to the more general equation (2) it can be cast into the form [13]

$$\frac{v}{v_{\mathrm{d}}} + \frac{v^2}{v_{\mathrm{b}}^2} = 1 \tag{11}$$

with

$$v_{\mathrm{b}}^2 = -\frac{q\psi_{\mathrm{n}}'}{m_{\mathrm{e}}(\ln n)'} = v_{\mathrm{th}}^2\left(1 + \frac{\Phi'}{\psi_{\mathrm{n}}' - \Phi'}\right) = v_{\mathrm{th}}^2\frac{\psi_{\mathrm{n}}'}{\psi_{\mathrm{n}}' - \Phi'}\ . \tag{12}$$

This model for v_{b} could be used for $x \geq x_{\mathrm{TOB}}$ because $\psi_{\mathrm{n}}'(x_{\mathrm{TOB}}) \neq 0$. However, it depends on the electric field, and gives 0/0 in the heavily doped drain region.

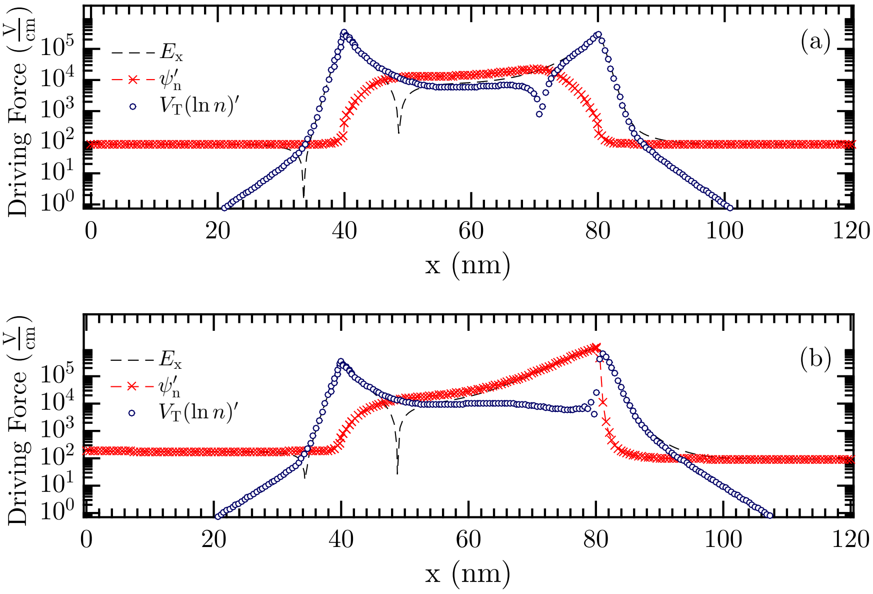

Figure 4: Comparison of driving forces along the channel obtained from TCAD simulations with a constant mobility of $\mu_{\mathrm{d}} = 10^4\,\mathrm{cm^2/Vs}$. (a) $V_{\mathrm{DS}} = 50\,\mathrm{mV}$, (b) $V_{\mathrm{DS}} = 0.63\,\mathrm{V}$. Parameters: $L_{\mathrm{G}} = 40\,\mathrm{nm}$, $V_{\mathrm{GS}} = 0.31\,\mathrm{V}$.

Fig. 4 compares the driving forces ψ_{n}', E_{x}, and $V_{\mathrm{T}}(\ln n)'$. As can be seen, in the source-side part of the channel ψ_{n}' is dominated by the density gradient whereas in the drain-side part it is dominated by the electric field. Therefore, ψ_{n}' cannot be replaced by one of the two. This becomes clear if one plots the local inverse mobility $\mu_{\mathrm{b}}^{-1}(x)$ and the local resistivity $\rho(x) = (q\,\mu_{\mathrm{b}}(x)\,n(x))^{-1}$ using the ballistic velocity (9). Fig. 5 shows that the whole channel contributes to both the inverse ballistic mobility and the ballistic resistivity. Furthermore, a sharp drop is observed towards the gate edges so that TCAD artifacts from model simplifications can be tolerated here. Fig. 6 presents the simulated $I_{\mathrm{D}}V_{\mathrm{GS}}$-curves for $L_{\mathrm{G}} = 15\,\mathrm{nm}$, computed with $\mu_{\mathrm{b}}(x)$ (Eq. (5) with Eq. (9)) and $\mu_{\mathrm{d}} = 10^4\,\mathrm{cm}^2/\mathrm{Vs}$, respectively, for $V_{\mathrm{DS}} = 50\,\mathrm{mV}$ (a) and for $V_{\mathrm{DS}} = V_{\mathrm{D,sat}} = 0.63\,\mathrm{V}$ (b). The QTx reference characteristics are plotted for comparison. The shift of the threshold voltage V_{th} due to geometrical confinement [16] found with QTx was modeled by a work function fit in S-Device. All transfer curves in this paper are plotted as function of gate overdrive $V_{\mathrm{GS}} - V_{\mathrm{th}}$, where V_{th} is defined as the gate voltage at which the current becomes $100\,\mathrm{nA}/\mu\mathrm{m}$. Geometry parameters [19], effective masses, and applied $V_{\mathrm{DS,sat}}$ are summarized in Table 2. The QDD current after the onset of inversion is still overestimated compared to QTx. The relative deviation is quite strong in the linear regime, but much smaller in the saturation regime. This behavior can be attributed to the shape of the QFP $\psi_{\mathrm{n}}(x)$ (see Fig. 2). At high drain bias, it is much closer to a step function, i.e. to the kinetic case where the electrons keep their QFP from the contact. Note, that the latter had been one of the initial assumptions to construct the local model (9). In contrast, the channel behaves like

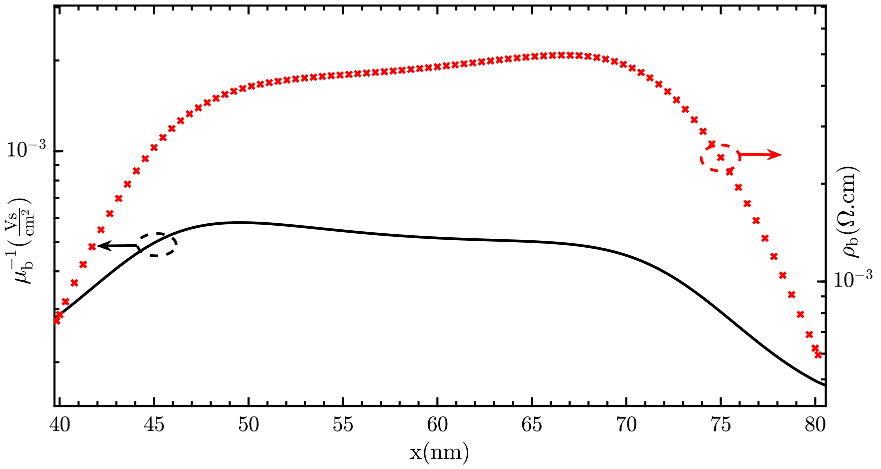

Figure 5: Inverse ballistic mobility (using Eq. (5) with the velocity model Eq. (9) in the TCAD simulations) and the corresponding resistivity along the channel. Parameters: $V_{\mathrm{DS}} = 50\,\mathrm{mV}$, $N_{\mathrm{D}} = 6 \times 10^{18}\,\mathrm{cm}^{-3}$, $L_{\mathrm{G}} = 40\,\mathrm{nm}$, $V_{\mathrm{GS}} = 0.31\,\mathrm{V}$.

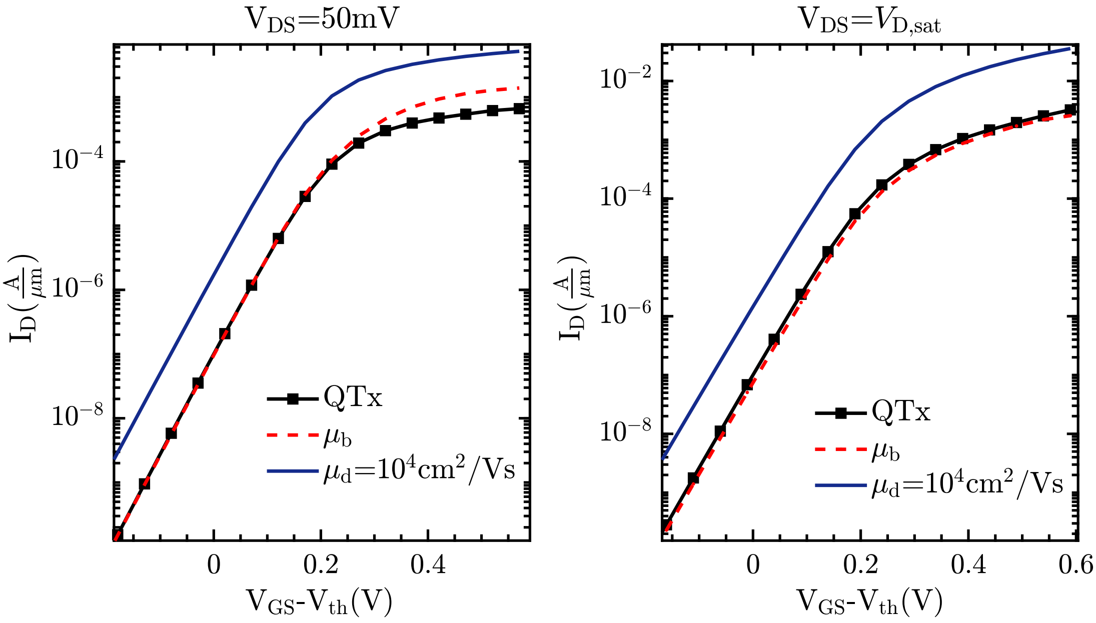

Figure 6: $I_\mathrm{D}V_\mathrm{GS}$-characteristics of an In$_{0.53}$Ga$_{0.47}$As DG UTB FET (L_G=15 nm) computed with μ_b from Eq. (5) with the velocity model Eq. (9) (red-dashed curves) and with constant mobility μ_d = 10^4 cm^2/Vs (solid blue curves) for (a) V_DS = 50 mV and (b) V_DS = $V_\mathrm{D,sat}$ = 0.63 V. Parameters: N_D = 6×10^{18} cm^{-3}, $v_\mathrm{th,i}$ = 2.37×10^7 cm/s. The self-consistent TCAD simulations are compared with the QTx reference characteristics (black lines with squares). STDT was turned-off in all simulations.

a low-Ohmic resistor with an almost linear $\psi_\mathrm{n}(x)$ in the linear regime, a situation furthest from the kinetic case. Second, due to the large series resistance, the injection velocity at the virtual source is overestimated which contributes to the overestimation of the on-current at low V_GS. This calls for improvements in modeling the ballistic velocity. Penzin et al. [13] added an empirical function of ψ_n in their KVM model to achieve the necessary drop of the on-current in the linear regime. In the next section, another option for the ballistic velocity will be considered.

4. Ballistic Velocity as Function of Density

An obvious feature of the ballistic velocity (9) is, that the ballistic current density j_b is not conserved when used in the mobility μ_b. Evaluating the continuity equation with a density that depends on a *local* QFP $\psi_\mathrm{n}(x)$ results in a position-dependent $j_\mathrm{b}(x)$. For $j_\mathrm{b}(x_\mathrm{TOB}) = j_\mathrm{b}(x_\mathrm{D})$ to hold, the densities of thermionic electrons would have to fulfill $n_\mathrm{TOB}(x_\mathrm{TOB})/n_\mathrm{TOB}(x_\mathrm{D}) = v_\mathrm{b}(x_\mathrm{D})/v_\mathrm{b}(x_\mathrm{TOB}) \approx$

Table 2: Summary of geometry parameters, effective masses, and $V_{\mathrm{DS,sat}}$ of the simulated devices.

L_{G}	t_{body}	t_{ox}	m_{e}/m_0	$V_{\mathrm{DS,sat}}$
7 nm	2.8 nm	2.6 nm	0.080	0.56 V
11.5 nm	4.6 nm	3.2 nm	0.0678	0.61 V
15 nm	7 nm	3.7 nm	0.0516	0.63 V

$\sqrt{2\,V_{\mathrm{DS}}/V_{\mathrm{T}}}$ which is not the case. Here, x_{D} denotes a point near the channel-drain junction. Densities, whether TOB or total, depend exponentially on V_{DS}. A ballistic velocity that conserves the current is

$$v_{\mathrm{b}}(x) = \langle v_{\mathrm{inj}}\rangle \,\frac{n_{\mathrm{TOB}}(x_{\mathrm{TOB}})}{n_{\mathrm{TOB}}(x)} = \frac{j_{\mathrm{k}}}{q n_{\mathrm{TOB}}(x)}\,, \tag{13}$$

where the mean injection velocity is divided by the normalized density of thermionic electrons in channel direction. Because of current conservation,

$$j_{\mathrm{b}}(x_{\mathrm{TOB}}) = q\,v_{\mathrm{b}}(x_{\mathrm{TOB}})\,n_{\mathrm{TOB}}(x_{\mathrm{TOB}}) = q\,\langle v_{\mathrm{inj}}\rangle\,n_{\mathrm{TOB}}(x_{\mathrm{TOB}}) = q\,\langle v_{\mathrm{inj}}\rangle\,n(x_{\mathrm{TOB}}).$$

The last identity is due to the fact that at x_{TOB} the TOB density is exactly equal to the total density. The replacement $v_{\mathrm{b}}(x_{\mathrm{TOB}}) \to \langle v_{\mathrm{inj}}\rangle$ is the same local approximation as in Section 3. Inside the channel region, using parabolic bands, a relation between $n_{\mathrm{TOB}}(x)$ and the total density $n(x)$ can be found:

$$n_{\mathrm{TOB}}(x) = n(x)\,\frac{2}{\sqrt{\pi}}\,\Gamma\!\left(\frac{3}{2}\,, \frac{\Phi(x) - \Phi(x_{\mathrm{TOB}})}{V_{\mathrm{T}}}\right)\,. \tag{14}$$

Here, $\Gamma(v, b)$ denotes the incomplete Gamma function. The derivation of Eq. (14) is given in Appendix A. Fig. 7 compares $n_{\mathrm{TOB}}(x)$ and $n(x)$ in the channel region. As one can see, both practically coincide in a large part of the channel. Deviations only occur where the local resistivity quickly drops to zero (compare Fig. 5). Hence, the model (13) can be simplified in the following way: the density $n_{\mathrm{TOB}}(x)$ in the denominator can be replaced by the actual density $n(x)$ which avoids to evaluate Eq. (14) and results in

$$v_{\mathrm{b}}(x) = \langle v_{\mathrm{inj}}\rangle\,\frac{n(x_{\mathrm{TOB}})}{n(x)}\,. \tag{15}$$

Fig. (8) compares the ballistic velocities (9) and (15). Where needed, i.e. in the channel region, the n-dependent model Eq. (15) yields a slower increase of the mean velocity towards the drain which

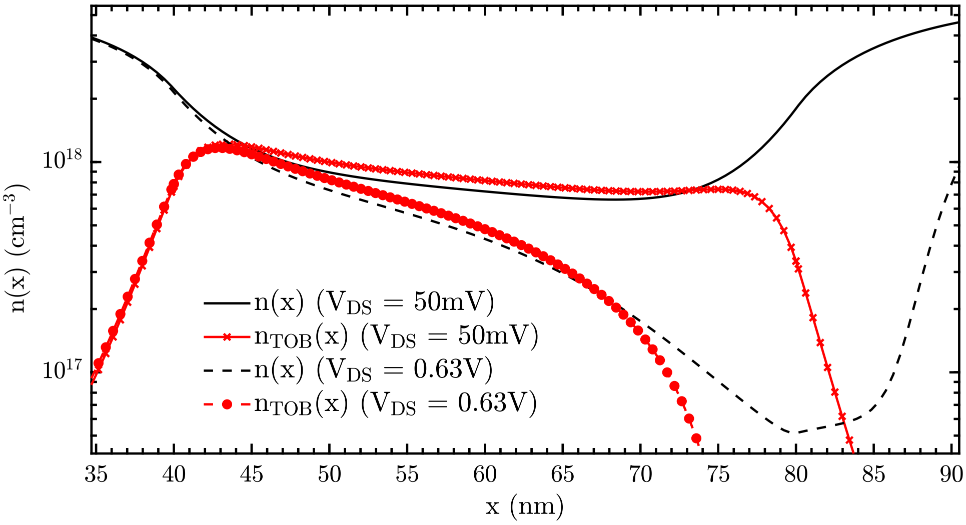

Figure 7: Comparison between $n_{\mathrm{TOB}}(x)$ and $n(x)$ along the channel in the middle of the body. Parameters: $N_{\mathrm{D}} = 6 \times 10^{18}\,\mathrm{cm}^{-3}$, $L_{\mathrm{G}} = 40\,\mathrm{nm}$, $V_{\mathrm{GS}} = 0.31\,\mathrm{V}$. The $n(x)$-profiles are direct extractions from the self-consistent TCAD simulation, whereas the $n_{\mathrm{TOB}}(x)$-profiles are a post-processing evaluation of Eq. (14) with all densities imported from TCAD.

results in a lower ballistic mobility. Consequently, the on-current in the linear regime is reduced. This is demonstrated by the comparison of the transfer characteristics of an $\mathrm{In}_{0.53}\mathrm{Ga}_{0.47}\mathrm{As}$ DG UTB FETs with gate lengths of 40 nm in Fig. 9, where STDT is absent. The agreement with the QTx-curves is reasonable considering the fact that the DOS models are different. In contrast to the ψ_{n}-dependent model in the previous section, the n-dependent model (15) makes explicit use of the injection point with the obvious advantage that $v_{\mathrm{b}}(x_{\mathrm{TOB}}) = \langle v_{\mathrm{inj}} \rangle$. The non-locality $n(x_{\mathrm{TOB}})$ cannot be removed, since the point x_{TOB} has to be under full gate control. Though replacing $n(x_{\mathrm{TOB}})$ by the source doping N_{D} would make the model local (analogous to setting $\psi_{\mathrm{n}}(x_{\mathrm{TOB}}) = 0$ in the previous section), this would lead to bias-dependent velocities with extreme values. The replacement of the TOB density by the density overestimates the TOB density in the region between the source-channel junction and x_{TOB}. This artificially drops the velocity and underestimates the ballistic mobility. Thus, a possible refinement of the model could be to replace v_{b} by $\langle v_{\mathrm{inj}} \rangle$ for all $x < x_{\mathrm{TOB}}$, leading to $v_{\mathrm{b}}(x) = \langle v_{\mathrm{inj}} \rangle [\Theta(x_{\mathrm{TOB}} - x) + \Theta(x - x_{\mathrm{TOB}}) n(x_{\mathrm{TOB}})/n(x)]$.

Devices with shorter gate were at first simulated without STDT. For this, the tunneling part of

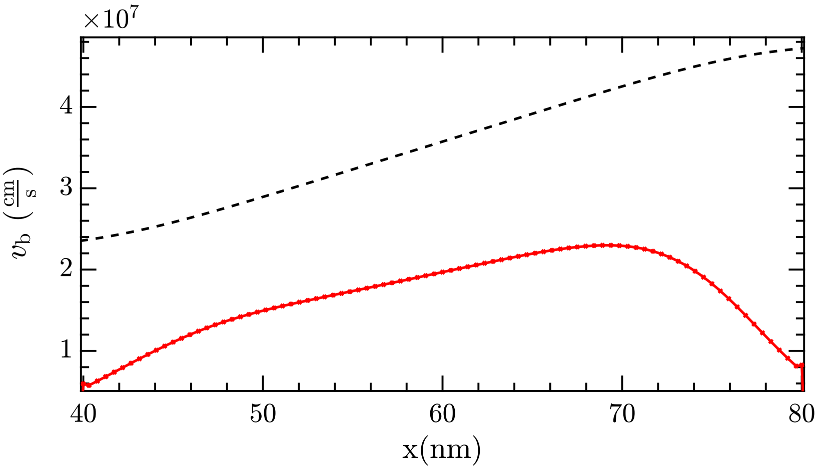

Figure 8: Profiles of the ballistic velocities along the channel with model (9) (black dashed) and with model (15) (red symbols) extracted from self-consistent TCAD simulations. Parameters: $V_{\mathrm{DS}} = 50\,\mathrm{mV}$, $N_{\mathrm{D}} = 6 \times 10^{18}\,\mathrm{cm}^{-3}$, $L_{\mathrm{G}} = 40\,\mathrm{nm}$, $V_{\mathrm{GS}} = 0.31\,\mathrm{V}$, $v_{\mathrm{th,i}} = 2.37 \times 10^7\,\mathrm{cm/s}$.

the spectral current in QTx was filtered out, and no model for STDT was used in the simulations with S-Device. Fig. 10 presents transfer characteristics for DG FETs with $L_{\mathrm{G}} = 11.5\,\mathrm{nm}$ and $L_{\mathrm{G}} = 7\,\mathrm{nm}$ with the mean ballistic velocity Eq. (15). Note, that there is no explicit fitting except the electrostatics in the sub-threshold range at $V_{\mathrm{DS}} = 0\,\mathrm{V}$ (no influence of the mobility model). The agreement with QTx is very good up to an overdrive voltage of 0.2 V. In deep inversion, the deviation increases with decreasing gate length and becomes more significant in the saturation regime. This behavior is attributed to the different DOS models which will be discussed in more detail in Section 7. The shorter the gate, the thinner the body (see Table 2) and the stronger the impact of the true 2D DOS of QTx. The dashed curves in Fig. 11 represent the corresponding profiles $\psi_{\mathrm{n}}(x)$ and $\mu_{\mathrm{b}}^{-1}(x)$ along a cut line in the middle of the body. The solid curves in this figure are obtained with the ψ_{n}-dependent model of the ballistic velocity.

5. Implementation Details

All above considerations apply to the ballistic transport regime. Of practical interest is the quasi-ballistic regime, defined by $\mu_{\mathrm{d}} \approx \mu_{\mathrm{b}}$ (see Table 2). Eq. (2) yields $v = v_{\mathrm{qb}} = \mu_{\mathrm{n}}\,\psi_{\mathrm{n}}'$, and μ_{n} is given by the Matthiessen rule in its common form $\mu_{\mathrm{n}}^{-1} = \mu_{\mathrm{d}}^{-1} + \mu_{\mathrm{b}}^{-1}$. Weighting factors for diffusive and ballistic sub-populations [10, 15] are not considered here. For μ_{b} it would mean an additional

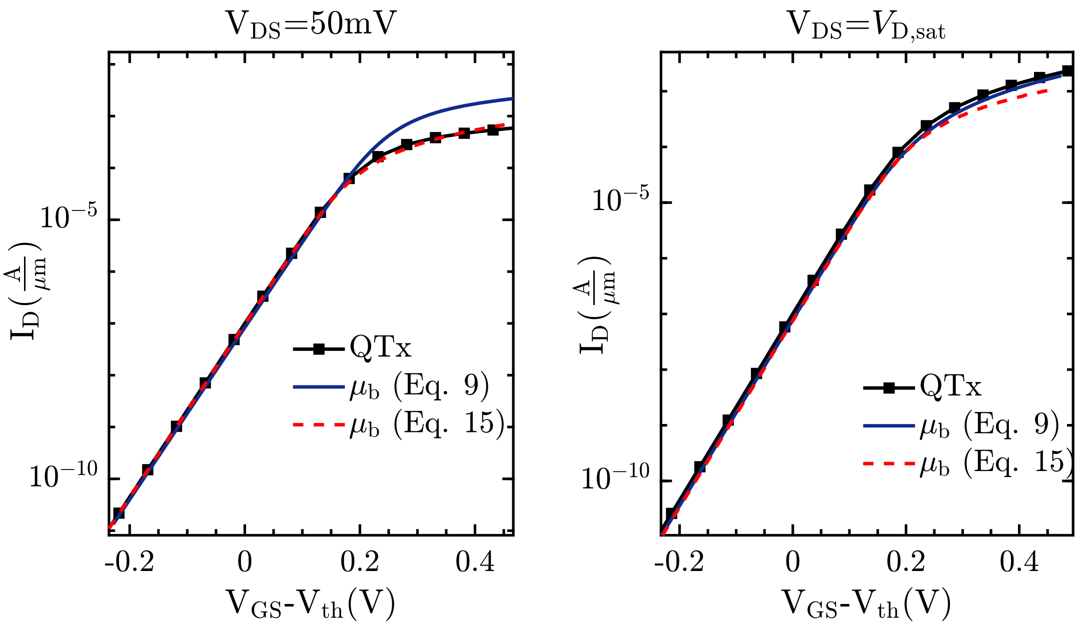

Figure 9: $I_D V_{GS}$-characteristics of an $\mathrm{In_{0.53}Ga_{0.47}As}$ DG UTB FET with $L_G = 40\,\mathrm{nm}$ at (a) $V_{DS} = 50\,\mathrm{mV}$ and (b) $V_{DS} = 0.63\,\mathrm{V}$ computed with μ_b (Eq. (5)). Blue solid curve: v_b-model as function of ψ_n (Eq. (9)), red dashed curve: v_b-model as function of n (Eq. (15)). Parameters: $N_D = 6 \times 10^{18}\,\mathrm{cm^{-3}}$, $v_{th,i} = 2.37 \times 10^7$ cm/s. STDT was suppressed by the sufficiently long gate. Other parameters are given in Table 2.

fitting factor, and the PMI of S-Device internally combines mobility models with the common Matthiessen rule.

The models (4) with (9) and with (15) were implemented in the PMI “HighFieldMobility” which combines the ballistic mobility model with the user-defined diffusive mobility μ_d by the Matthiessen rule. Generalization of Eq. (4) to the 2D/3D case is done by $\mu_b = v_b/(|\nabla\psi_n| + \epsilon)$, where ϵ is an appropriate cut-off. The computation of an average 1D density $n(x)$ is prohibitive, hence $n(\vec{r})$ is used. Density minima perpendicular to the channel yield high local values of μ_b and do not contribute. The obviously better approach would be to compute n_{TOB} as average over a line/slice perpendicular to the transport direction. However, this would require the availability of a tensor-product grid in the PMI of S-Device which is not the case. To extract $n(\vec{r}_{TOB})$, a search algorithm was implemented to find the point where the conduction band (CB) energy is maximum. In each step of the Newton iteration, a variable $E_{c,ref}$ is first set to a large negative value. Then, the local value of the CB energy is compared with $E_{c,ref}$ on every vertex $\vec{r}_n$ of the search path.

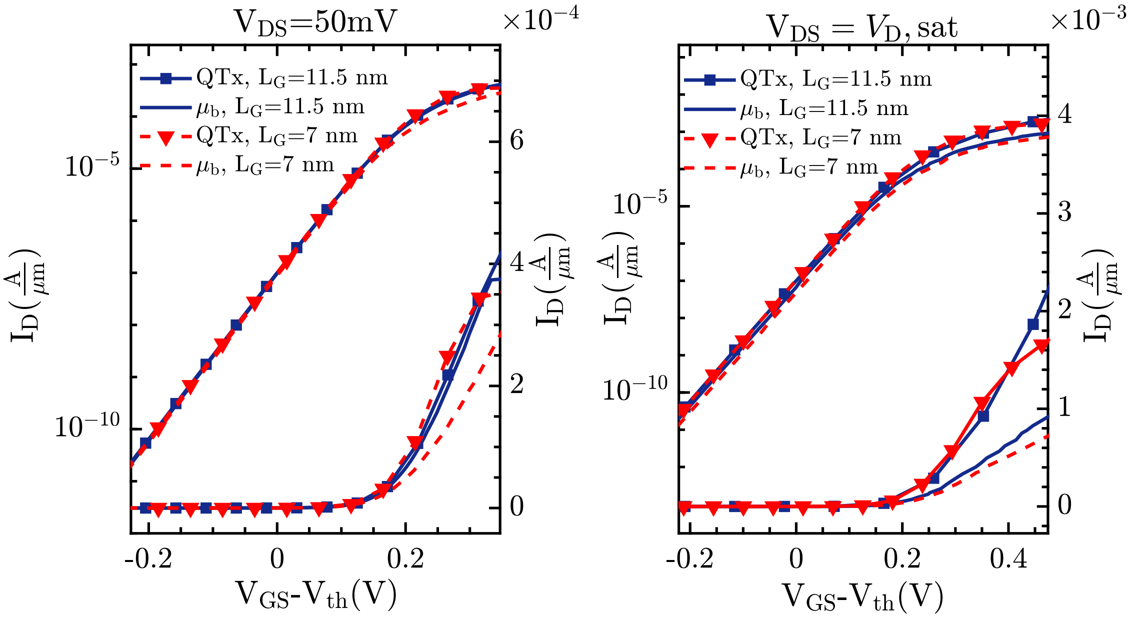

Figure 10: $I_D V_{GS}$-characteristics of $In_{0.53}Ga_{0.47}As$ DG UTB FETs with L_G = 11.5 nm and L_G = 7 nm at V_{DS} = 50 mV (left) and $V_{DS} = V_{D,sat}$ (right) computed with μ_b (Eq. (5)) and $v_b(n)$ (Eq. (15)). Parameters: $N_D = 6 \times 10^{18}\,\mathrm{cm}^{-3}$, $v_{th,i} = 2.37 \times 10^7$ cm/s. STDT was turned off. Other parameters are given in Table 2.

Only if $E_c(\vec{r}_n) > E_{c,ref}$, $E_{c,ref}$ is updated and set to the value of $E_c(\vec{r}_n)$. The vertex $\vec{r}_{TOB}$ is found for $E_c(\vec{r}_{TOB}) = E_{c,ref}$. The electron density is extracted at this vertex and used for the mobility model in the next Newton step. A sufficient mesh refinement, in particular at the source-channel junction, is required to ensure convergence.

6. Model Behavior in the Case of Strong Source-to-Drain Tunneling

Below 20 nm gate length, STDT is not negligible and notably increases the leakage current in the sub-threshold regime [21, 22]. In the FET with L_G = 7 nm and $N_D = 5 \times 10^{19}\,\mathrm{cm}^{-3}$, at V_{DS} = 50 mV and V_{GS} = 0 V, over 95% of the spectral current is tunnel current. STDT can be simulated in S-Device with the Nonlocal Tunneling (NLT) model [7] using the effective transport mass extracted from QTx for the tunneling mass. The used values that change with body thickness are given in Table. 2. Details of the NLT model can be found in Appendix B. The choice of the model for the ballistic velocity has a strong effect on the sub-threshold current, which can be explained as follows. When using the NLT model in S-Device, electrons "recombine" at the

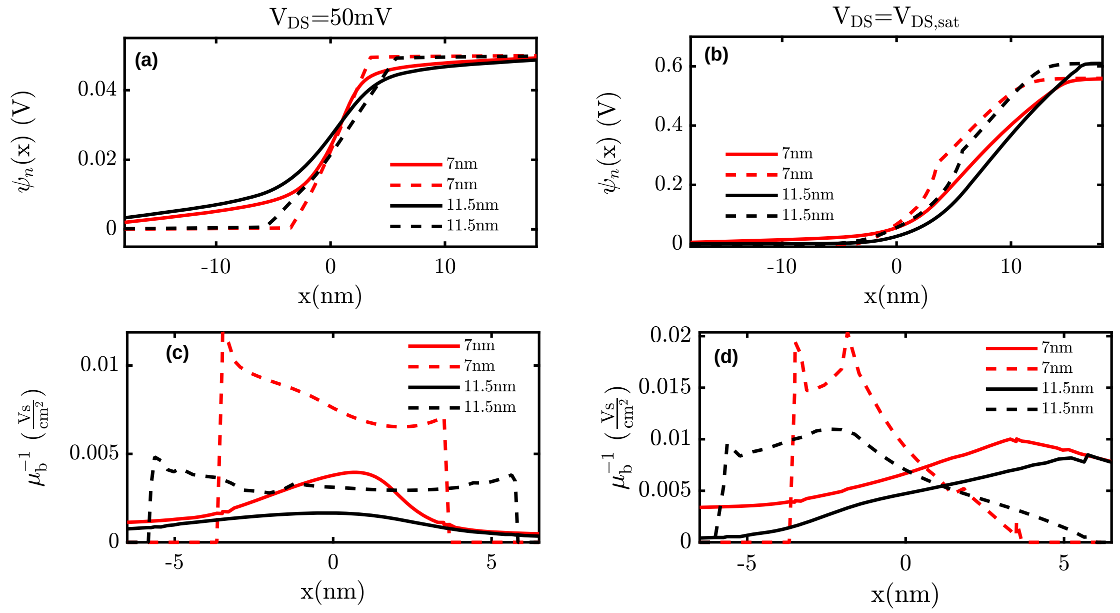

Figure 11: Profiles of $\psi_n(x)$ and $\mu_b^{-1}(x)$ along a cut line in the middle of the body of the $In_{0.53}Ga_{0.47}As$ DG UTB FETs with $L_G = 11.5\,nm$ (black) and $L_G = 7\,nm$ (red) at $V_{DS} = 50\,mV$ (left) and $V_{DS} = V_{D,sat}$ (right). Solid curves: $v_b(\psi_n)$ (Eq. (9)), dashed curves: $v_b(n)$ (Eq. (15)). Parameters: $V_{GS} = 0.45\,V$ ($L_G = 7\,nm$), $V_{GS} = 0.35\,V$ ($L_G = 11.5\,nm$, $V_{DS} = 50\,mV$), $V_{GS} = 0.27\,V$ ($L_G = 11.5\,nm$, $V_{DS} = V_{D,sat}$), $N_D = 6 \times 10^{18}\,cm^{-3}$, $v_{th,i} = 2.37 \times 10^7\,cm/s$. Other parameters are given in Table 2.

beginning of a tunnel path, whereas they are “generated” at its end. The generation/recombination rates, shown in Fig. (12), are computed with the local QFPs at the classical turning points. These G/R rates are the source/sink of an additional DD current which adds to the thermionic current and determines the shape of the QFP in the self-consistent solution. As a result, the DD current is driven by the gradient of this QFP, and the current level is the integral over the G/R rates with the corresponding QFP values at the classical turning points. The profile of the current density in Fig. (12) reveals a strong compensation of the oppositely flowing partial currents of recombining and generated electrons in the channel region which leads to a sharp drop of the total current under the gate. The small thermionic current (shown by the black curve in Fig. (12)) is not perfectly constant due to a x-dependent y-component that still exists even in the narrow slab of only 2.8 nm width. As the tunnel-generated density is much smaller than the channel doping in the simulated transistors, it has no electrostatic effect, i.e. the tunnel barrier remains unchanged. Hence, the

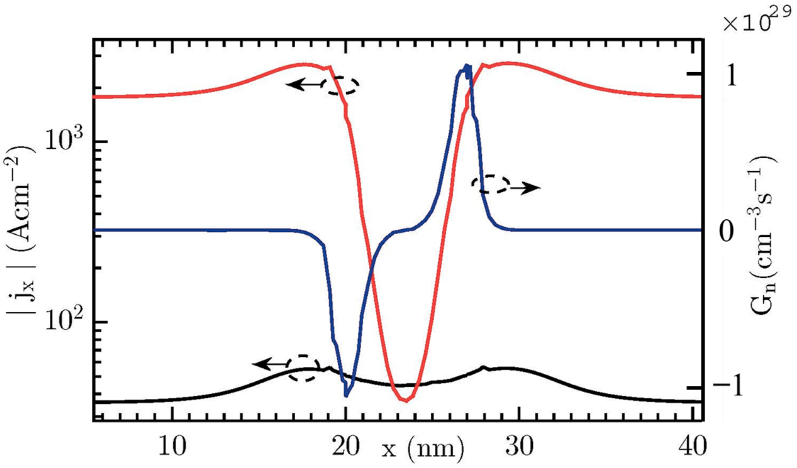

Figure 12: Profiles of the STDT rate (blue curve) and the current densities with STDT (red curve) and without STDT (black curve) along a cut in the middle of the channel of the $In_{0.53}Ga_{0.47}As$ DG UTB FET with $L_G = 7\,\mathrm{nm}$. Parameters: $N_D = 5 \times 10^{19}\,\mathrm{cm}^{-3}$, $V_{DS} = V_{GS} = 50\,\mathrm{mV}$. The v_b-model as function of ψ_n (Eq. (9)) was used.

strong dependence of the off-current on the mobility model originates from the changed values of $\psi_n(x)$ at the classical turning points.

If the ballistic mobility, modeled by Eq. (5), becomes dominant, it strongly changes the shape of the QFP $\psi_n(x)$ ($\sim \ln n(x)$) in the sub-threshold region. This is a consequence of current conservation enforced by the continuity equation and the small electron concentration in the channel. Only a constant mobility has no effect (see Fig. 2). The high sensitivity of $\psi_n(x)$ to the ballistic mobility is demonstrated in Fig. 13 for $L_G = 11.5\,\mathrm{nm}$. If STDT is absent (Fig. 13 a)), the edge in the error-function-like profile becomes steeper with the ballistic mobilities - more pronounced when the n-dependent model of the ballistic velocity is used. The position of the edge is not changed and remains at the center of the device. If the strong STDT is now turned on (Fig. 13 b)), the slope further increases and the edge is shifted towards the drain. The electron quasi-Fermi energies at the drain side increase which reduces the STDT rate and the off-current. In the case of the n-dependent model of the ballistic velocity, the QFP shape degenerates into a step function located at the channel-drain junction. The constant QFP under the gate suppresses the STDT rate and, therefore, the sub-threshold current as shown in Fig. 14. This behavior can also be understood analytically inserting the models of the ballistic mobility into the continuity equation and using a simplistic x-dependence of the STDT rate.

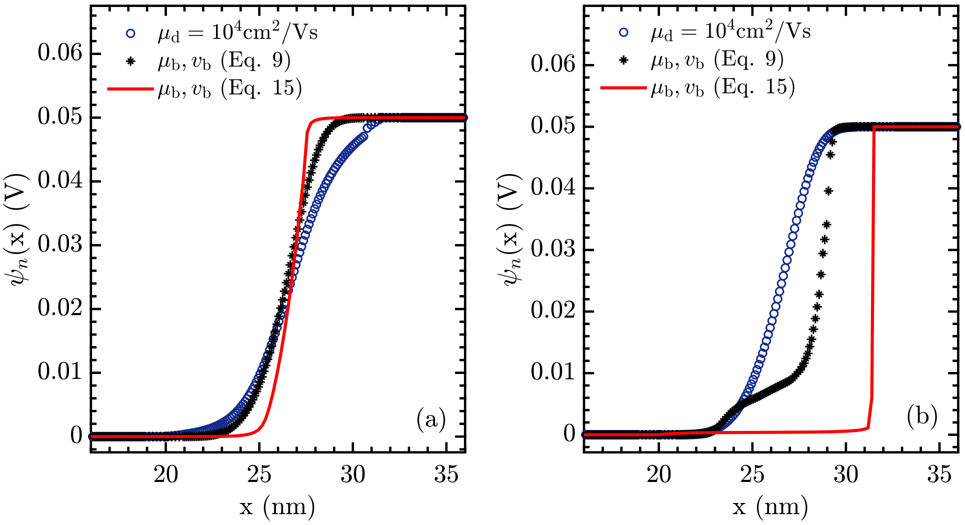

Figure 13: Profiles of the quasi-Fermi potential along a cut in the middle of the channel of the $\mathrm{In}_{0.53}\mathrm{Ga}_{0.47}\mathrm{As}$ DG UTB FET with L_G = 11.5 nm for different mobility models. a) STDT turned off, b) with STDT. Parameters: N_D = $5 \times 10^{19}\,\mathrm{cm}^{-3}$, V_DS = 0.05 V, V_GS = -0.23 V.

7. Discussion

Introducing a ballistic mobility can be seen as the re-introduction of the hydrodynamic term in the balance equation for the mean velocity. In FETs with ultra-short channels this term is crucial to prevent the divergence of the DD current when $L_\mathrm{G} \to 0$ and no series resistance is present. The expression of the ballistic mobility contains the mean ballistic velocity. It's determination would necessitate the iterative solution of the whole equation system in the ballistic regime. In order to avoid this, two explicit models of the mean ballistic velocity as function of solution variables were suggested, $v_\mathrm{b}(\psi_\mathrm{n})$ and $v_\mathrm{b}(n)$. Utilizing the kinetic limit, locality can be approximately achieved in the case of $v_\mathrm{b}(\psi_\mathrm{n})$. In the $v_\mathrm{b}(n)$-model, the TOB density $n(x_\mathrm{TOB})$ has to be extracted which requires to find the virtual source x_TOB numerically. This is the only non-local remnant of the hydrodynamic term. The models have no free parameters, but the mean injection velocity is not unique and can serve as TCAD parameter with the mean thermal velocity as default.

The models were implemented in the Physical Model Interface of S-Device, and transfer characteristics of InGaAs DG FETs with L_G ranging from 7 nm to 40 nm were simulated. InGaAs was chosen because of the small transport and tunnel mass that both highlight the ballisticity effect and

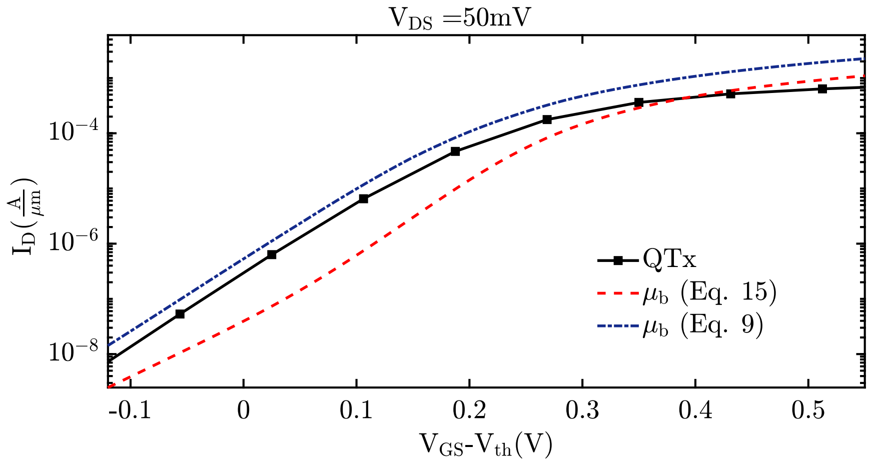

Figure 14: $I_{\mathrm{D}}V_{\mathrm{GS}}$-characteristics of an In$_{0.53}$Ga$_{0.47}$As DG UTB FET with $L_{\mathrm{G}} = 11.5$ nm for different mobility models. STDT is included by the NLT model. Parameters: $N_{\mathrm{D}} = 5 \times 10^{19}$ cm^{-3}, $V_{\mathrm{DS}} = 50$ mV.

the interplay with STDT. Benchmarking against real devices would require a fully characterized (electrical and physical) ultra-short-channel FET which is not available. Instead, the simulations were benchmarked against quantum-transport results obtained with QTx, but also here a precise one-to-one comparison is impossible. This is mainly due to the different DOS models. The DOS in S-Device is the common 3D DOS of a 3DEG populating one valley. (In all released versions of the commercial simulator there are no low-dimensional DOS models.) QTx features the correct 2D DOS of a 2DEG, and the sub-band dispersions are self-consistent Poisson-Schrödinger solutions. For any comparison between S-Device and QTx the first step is always to match the electrostatics. This is done at zero current to avoid any influence of the mobility model. Furthermore, a longer gate is used in order to suppress the effect of charge penetration into the S-D potential barrier. However, it is not possible to match the electrostatics over the entire V_{GS}-range. In deep inversion, not only the relative contribution of different sub-bands has changed, but also their dispersion. In other words, the DOS in QTx self-consistently depends on the gate voltage, but is a constant in S-Device.

The developed models of the ballistic velocity are oversimplified in the following sense. First, the starting point of ballistic motion is not well defined. The virtual source has been chosen, but a better option could be a point between the source-channel junction and x_{TOB}. The practical advantage of x_{TOB} is that it can be easily found numerically because of $\nabla\Phi(x = x_{\mathrm{TOB}}) = 0$. In 2D/3D devices, under flat-band conditions, the point can degenerate to a contour line/surface. Second, the injection velocity at the virtual source is, in general, not a mean thermal velocity. Even the mean *kinetic* velocity has already increased here from the latter due to the steep density gradient (continuity equation holds in any transport regime). The mean ballistic velocity differs from the mean kinetic velocity as result of the weak scattering related to the finite value of μ_{d}. The corresponding small voltage drop adds to the possibly much greater voltage drop due to the source resistance. Third, a local model requires to replace the QFP at x_{TOB} by zero. This artificially increases the mean ballistic velocity at the virtual source. Fourth, the locality in the range $x < x_{\mathrm{TOB}}$ has the price that the ballistic velocity is overestimated in the ψ_{n}-dependent model, but underestimated in the n-dependent model.

The models of the ballistic velocity were derived for a homogeneous band structure (BS). However, alloy- or strain-induced gradients can be present in the channel of modern FETs. In the TCAD simulator S-Device those gradients are treated by the driving force they induce in the current equation. A similar term, the position-dependent Fermi statistics correction, was included in the above simulations of the IV-characteristics. All those terms self-consistently change n, ψ_{n}, ψ_{n}', and x_{TOB}. As the models directly depend on them, they capture the influence of an inhomogeneous BS to some extent. However, BS gradients, or even band edge discontinuities, could change the injection velocity at x_{TOB} significantly. A deeper investigation of BS inhomogeneities is beyond the scope of the paper.

Under conditions where STDT dominates the current (high S/D doping, ultra-short channels, deep sub-threshold regime) the ballistic mobility models cannot be used. The sharp maximum of the STDT rate leads to a high local electron density exceeding the density of thermionic electrons. This high local density has a catastrophic impact on the ballistic mobility which depends on ψ_{n}' in the denominator and either on the inverse density or on $\sqrt{\psi_{\mathrm{n}}}$ in the numerator. In turn, the deformed QFPs shrink or even extinguish the tunnel current. This is an artifact, because the true tunnel current is a spectral current that flows *below* the TOB. The generated density which mimics this current *has nothing to do* with the thermionic density above the TOB in the model of the ballistic velocity. One faces a fundamental TCAD problem here. The off-current becomes corrupted by the mobility model which is needed for the on-current, where, on the other hand, the influence of STDT fades away. Thus, one cannot simulate the entire transfer characteristics with the same model set. Obviously, if in the implemented NLT model of S-Device just the Fermi levels of the S/D contact regions would be used (emission approach, $\psi_{\mathrm{n}}(x) \to \psi_{\mathrm{n}}(x_{\mathrm{S}})\Theta(x_{\mathrm{TOB}} - x) + \psi_{\mathrm{n}}(x_{\mathrm{D}})\Theta(x - x_{\mathrm{TOB}})$), the mobility would not change the STDT rate *at all*. This shows that not only the concept of a local QFP breaks down in the STDT regime but also the concept of a mean ballistic velocity. In other words, the situation in the deep sub-threshold region of an ultra-short FET becomes comparable to a MIM (metal-insulator-metal) structure where a $\psi_{\mathrm{n}}(x)$ does not exist in the insulator.

Appendix A

This appendix presents the derivation of Eq. (14) which relates $n_{\mathrm{TOB}}(x)$ to $n(x)$ in the channel region. Assuming Maxwell-Boltzmann statistics and a parabolic conduction band (CB), the density is given by

$$n(x) = N_{\mathrm{c}} \frac{2}{\sqrt{\pi}} \int_{\frac{E_{\mathrm{c}}(x)}{k_{\mathrm{B}}T}}^{\infty} d\epsilon \sqrt{\epsilon - \frac{E_{\mathrm{c}}(x)}{k_{\mathrm{B}}T}} \, \exp\left(-\frac{\epsilon - E_{\mathrm{F,n}}(x)}{k_{\mathrm{B}}T}\right) = N_{\mathrm{c}} \exp\left(-\frac{\psi_{\mathrm{n}}(x) + E_{\mathrm{c}}(x)/q}{V_{\mathrm{T}}}\right), \tag{A1}$$

where $E_{\mathrm{c}}(x)$ is the energy of the CB edge, N_{c} the effective CB DOS, and $E_{\mathrm{F,n}}(x)$ the quasi-Fermi energy $E_{\mathrm{F,n}}(x) = -q\psi_{\mathrm{n}}(x)$. The TOB density is computed with the same integrand, but the lower integration limit has to be replaced by the energy of the TOB $E_{\mathrm{TOB}} = E_{\mathrm{c}}(x_{\mathrm{TOB}})$:

$$\begin{aligned} n_{\mathrm{TOB}}(x) &= N_{\mathrm{c}} \frac{2}{\sqrt{\pi}} \int_{\frac{E_{\mathrm{TOB}}}{k_{\mathrm{B}}T}}^{\infty} d\epsilon \sqrt{\epsilon - \frac{E_{\mathrm{c}}(x)}{k_{\mathrm{B}}T}} \, \exp\left(-\frac{\epsilon - E_{\mathrm{F,n}}(x)}{k_{\mathrm{B}}T}\right) \\ &= N_{\mathrm{c}} \frac{2}{\sqrt{\pi}} \int_{0}^{\infty} d\epsilon \sqrt{\epsilon + \frac{E_{\mathrm{TOB}} - E_{\mathrm{c}}(x)}{k_{\mathrm{B}}T}} \, e^{-\epsilon} \exp\left(-\frac{\psi_{\mathrm{n}}(x) + E_{\mathrm{TOB}}/q}{V_{\mathrm{T}}}\right). \end{aligned} \tag{A2}$$

The last integral is proportional to the incomplete Gamma function [20]

$$\Gamma\left(\frac{3}{2}, b\right) = e^{-b} \int_{0}^{\infty} d\epsilon \sqrt{\epsilon + b} \, e^{-\epsilon} . \tag{A3}$$

Expressing energies by the electrostatic potential, i.e. $E_{\mathrm{TOB}} - E_{\mathrm{c}}(x) = q\Phi(x) - q\Phi(x_{\mathrm{TOB}})$, the TOB density can be written with (A3) as

$$\begin{aligned} n_{\mathrm{TOB}}(x) &= \exp\left(\frac{E_{\mathrm{TOB}} - E_{\mathrm{c}}(x)}{qV_{\mathrm{T}}}\right) N_{\mathrm{c}} \frac{2}{\sqrt{\pi}} \Gamma\left(\frac{3}{2}, \frac{E_{\mathrm{TOB}} - E_{\mathrm{c}}(x)}{qV_{\mathrm{T}}}\right) \exp\left(-\frac{q\psi_{\mathrm{n}}(x) + E_{\mathrm{TOB}}}{qV_{\mathrm{T}}}\right) \\ &= N_{\mathrm{c}} \frac{2}{\sqrt{\pi}} \Gamma\left(\frac{3}{2}, \frac{\Phi(x) - \Phi(x_{\mathrm{TOB}})}{V_{\mathrm{T}}}\right) \exp\left(-\frac{\psi_{\mathrm{n}}(x) - E_{\mathrm{c}}(x)/q}{V_{\mathrm{T}}}\right) \\ &= \frac{2}{\sqrt{\pi}} \Gamma\left(\frac{3}{2}, \frac{\Phi(x) - \Phi(x_{\mathrm{TOB}})}{V_{\mathrm{T}}}\right) n(x) \end{aligned} \tag{A4}$$

which is Eq. (14). The assumption of Maxwell-Boltzmann statistics holds in a large part of the channel, but breaks down where $E_{\mathrm{c}}(x) - E_{\mathrm{F,n}}(x) < k_{\mathrm{B}}T$, i.e. near the pn-junctions (see Fig. 7). Since $\Gamma\left(\frac{3}{2}, 0\right) = \sqrt{\pi}/2$ it follows that $n(x_{\mathrm{TOB}}) = n_{\mathrm{TOB}}(x_{\mathrm{TOB}})$ as required.

Appendix B

The Nonlocal Tunneling (NLT) model in S-Device [7] is an adaptation of the Schottky barrier model originally proposed in Ref. [23] to a general barrier that carriers encounter in a device. The actual spectral tunnel current through the source-to-drain potential barrier up to E_{TOB} is mimicked by a generation-recombination current with the net rate

$$\begin{aligned} R_{\mathrm{n}}(u, l, \epsilon) - G_{\mathrm{n}}(u, l, \epsilon) = \frac{4\pi q^2 m_0 k_{\mathrm{B}} T}{h^3} \delta\left[\epsilon - E_{\mathrm{c}}(u)\right] \delta\left[\epsilon - E_{\mathrm{c}}(l)\right] |F(u)||F(l)| \Theta\left[-F(u)\right] \Theta\left[F(l)\right] \times \\ \times\ \mathcal{T}_{\mathrm{n}}(u, l, \epsilon) \left\{ \ln\left(1 + \exp\left[\frac{E_{\mathrm{F,n}}(u) - \epsilon}{k_{\mathrm{B}}T}\right]\right) - \ln\left(1 + \exp\left[\frac{E_{\mathrm{F,n}}(l) - \epsilon}{k_{\mathrm{B}}T}\right]\right) \right\}, \end{aligned} \tag{B1}$$

where u and l are positions on a tunnel path with energy ϵ, $F(x) = d\Phi(x)/dx$ is the local field strength, and

$$\mathcal{T}_{\mathrm{n}}(u, l, \epsilon) = \exp\left(-\frac{2}{\hbar} \int_l^u dx\, \sqrt{2m_{\mathrm{e}}[E_{\mathrm{c}}(x) - \epsilon]}\, \Theta[E_{\mathrm{c}}(x) - \epsilon]\right) \tag{B2}$$

the WKB transmission probability. The contribution to the STDT current density by electrons that tunnel from the CB edge at points ahead u, to the CB edge at the point u is obtained by a double integral over the net rate:

$$\frac{dj_{\mathrm{tun,n}}}{du}(u) = -q \int_{-\infty}^{u} dl \int_{-\infty}^{\infty} d\epsilon \left[R_{\mathrm{n}}(u, l, \epsilon) - G_{\mathrm{n}}(u, l, \epsilon)\right] . \tag{B3}$$

Acknowledgment

The authors would like to thank A. Erlebach (Synopsys Switzerland LCC) and O. Penzin (Synopsys, Inc.) for valuable discussions and technical advice during the implementation of the PMI model. The QTx simulations were supported by Prof. Mathieu Luisier and Martin Rau. The research leading to the results of this paper has received funding from the European Commissions Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No. 619326 (III-V-MOS Project).

References

[1] A. Wettstein, A. Schenk, and W. Fichtner, "Quantum Device-Simulation with the Density-Gradient Model on Unstructured Grids", IEEE Trans. Electron Devices, vol. 48 (2), pp. 279-84, 2001.

[2] M. Rau, E. Caruso, D. Lizzit, P. Palestri, D. Esseni, A. Schenk, L. Selmi, and M. Luisier, "Performance Projection of III-V Ultra-Thin-Body, FinFET, and Nanowire MOSFETs for two Next- Generation Technology Nodes", IEDM Tech. Digest pp. 758 - 761 (2016).

[3] W. R. Frensley, IEEE Trans. Electron Devices, vol. 30 (12), pp. 1619-1623, 1983.

[4] M. Luisier, A. Schenk, and W. Fichtner, "Atomistic simulation of nanowires in the $sp^3d^5s^*$ tight-binding formalism: From boundary conditions to strain calculations", Phys. Rev. B 74, 205323, 2006.

[5] W. van Roosbroeck, "Theory of Flow of Electrons and Holes in Germanium and Other Semiconductors", Bell System Techn. J. 29, 560 - 607, 1950.

[6] P. Aguirre, H. Carrillo-Nuñez, A. Ziegler, M. Luisier, and A. Schenk, "Drift-Diffusion Quantum Corrections for $In_{0.53}Ga_{0.47}As$ Double Gate Ultra-Thin-Body FETs", Proc. 22th Int. Conf. on Simulation of Semiconductor Processes and Devices (SISPAD), Nuremberg, Germany, Sep. 6 - 8, 2016, pp. 53 - 26.

[7] Sentaurus-Device User Guide, V-2016.03, Synopsys Inc., Mountain View, CA, 2016.

[8] M. S. Shur, "Low ballistic mobility in submicron HEMTs", IEEE Electron Device Letters, vol. 23 (9), pp. 511-513, 2002.

[9] M. Lundstrom and X. Sun, "Some Useful Relations for Analyzing Nanoscale MOSFETs Operating in the Linear Region", arXiv:1603.03132, 2016.

[10] R. Kotlyar, R. Rios, C. E. Weber, T. D. Linton, M. Armstrong, and K. Kuhn, "Distributive Quasi-Ballistic Drift Diffusion Model Including Effects of Stress and High Driving Field", IEEE Trans. Electron Devices, vol. 62 (3), pp. 743-750, 2015.

[11] S. Martinie, G. Le Carval, D. Munteanu, S. Soliveres, Jean-Luc Autran, "Impact of Ballistic and Quasi-Ballistic

Transport on Performances of Double-Gate MOSFET-Based Circuits", IEEE Trans. Electron Devices, vol. 55 (9), pp. 2443 - 2453, 2008.

[12] E. Gnani, A. Gnudi, S. Reggiani, and G. Baccarani, "Effective Mobility in Nanowire FETs Under Quasi-Ballistic Conditions", IEEE Trans. Electron Devices, vol. 57 (1), pp. 336 - 344, 2010.

[13] 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. on Electron Devices* vol. 64 (11), 4599 - 4606, 2017.

[14] K. Bløtekjær, "Transport Equations for Electrons in Two-Valley Semiconductors", IEEE Trans. Electron Devices, vol. 17 (1), pp. 38-47, 1970.

[15] A. Erlebach, K. H. Lee, and F. M. Bufler, "Empirical Ballistic Mobility Model for Drift-Diffusion Simulation", Proc. ESSDERC, pp. 420 - 423, 2016.

[16] A. Schenk and A. Wettstein, "Simulation of DGSOI MOSFETs with a Schrödinger-Poisson Based Mobility Model", in Proc. 7th Int. Conf. on Simulation of Semiconductor Processes and Devices (SISPAD), pp. 21-24, Kobe (Japan), Sept. 4-6, 2002.

[17] S. A. Schwarz and S. E. Russek, "Semi-Empirical Equations for Electron Velocity in Silicon: Part II - MOS Inversion Layer", IEEE Trans. Electron Devices, vol. 30 (12), pp. 1634 - 1639, 1983.

[18] A. Ziegler; M. Frey; L. Smith; M. Luisier, "A Nonparabolic Bandstructure Model for Computationally Efficient Quantum Transport Simulations", IEEE Trans. Electron Devices, vol. 63 (5), pp. 2050 - 2056, 2016.

[19] Extended Intel Technology Road Map, private communication with IMEC, see also www.iii-v-mos-project.eu.

[20] M. Abramowitz and I. A. Stegun, "Handbook of Mathematical Functions", Dover Publications Inc., New York, p. 260.

[21] C. Convertino, C. Zota, S. Sant, F. Eltes, M. Sousa, D. Caimi, A. Schenk and L. Czornomaz, "nGaAs-on-Insulator FinFETs with Reduced Off-Current and Record Performance", IEDM Tech. Digest, 899-902, 2018.

[22] C. Medina-Bailon, J. L. Padilla, T. Sadi, C. Sampedro, A. Godoy, L. Donetti, V. P. Georgiev, F. Gámiz, and A. Asenov, "Multisubband Ensemble Monte Carlo Analysis of Tunneling Leakage Mechanisms in Ultrascaled FDSOI, DGSOI, and FinFET Devices", IEEE Trans. Electron Devices, vol. 66 (3), pp. 1145 - 1152 , 2019.

[23] M. Ieong, P. M. Solomon, S. E. Laux, H. P. Wong, and D. Chidambarrao, "Comparison of Raised and Schottky Source/Drain MOSFETs Using a Novel Tunneling Contact Model", IEDM Tech. Digest, 733-735, 1998.