An Energy Transport Gate Current Model Based on a Non-Maxwellian Energy Distribution

A. Gehring, T. Grasser, H. Kosina, and S. Selberherr

Institute for Microelectronics, TU Vienna, Gusshausstr. 27–29, A-1040 Vienna, Austria

ABSTRACT

We report on a new formulation of hot electron tunneling through dielectrics. It is based on an expression of the electron energy distribution function which accounts for the non-Maxwellian shape and is determined by the first three even moments n, T_n, and β_n. We present a simplified model applicable within the framework of the energy-transport model which only provides two even moments, n and T_n. Simulation results of long channel EEPROM devices and short channel MOSFETs show excellent agreement with Monte Carlo results and measurements.

Keywords: tunneling, moment equations, BOLTZMANN equation, distribution function model.

1 INTRODUCTION

In contemporary sub-quartermicron CMOS devices the prediction of gate oxide tunneling is crucial to enable further device scaling. Tunneling models usually assume a Fermi-Dirac or Maxwellian shape of the electron energy distribution function (EED). This assumption is clearly violated in the channel of miniaturized MOS devices, where the deviation from the Maxwellian shape may reach several orders of magnitude due to a pronounced high energy tail with carrier temperatures exceeding 1000 K. A correct modeling of the shape of the distribution function is thus necessary to reliably predict gate currents due to hot electron injection.

2 TUNNELING MODEL

The most prominent and almost exclusively used expression to describe oxide tunneling is based on Tsu-Esaki's equation [1], where the gate current density J_g is computed as

$$J_\mathrm{g} = \frac{m_\mathrm{eff}\mathrm{q}}{2\pi^2\hbar^3} \cdot \int\limits_0^\infty \left[f_1(\mathcal{E}) - f_2(\mathcal{E} + \Delta\mathcal{E}_\mathrm{C})\right] \mathrm{d}\mathcal{E}_\mathrm{l} \int\limits_0^\infty TC(\mathcal{E}_\mathrm{t})\,\mathrm{d}\mathcal{E}_\mathrm{t}. \tag{1}$$

In this expression the total energy $\mathcal{E}$ is the sum of a longitudinal component $\mathcal{E}_\mathrm{l}$ and a transversal component $\mathcal{E}_\mathrm{t}$. The transversal component is perpendicular to the semiconductor-oxide interface. The electron energy distribution functions in the gate and substrate are denoted by f_1 and f_2, the difference in the conduction band edges $\Delta\mathcal{E}_\mathrm{C}$ is calculated as $\mathcal{E}_\mathrm{C1} - \mathcal{E}_\mathrm{C2}$, and the quantum-mechanical transmission coefficient is denoted by TC. It is assumed that the transmission coefficient only depends on the transversal energy component. The dependence on the Fermi energies has been lumped into an energy-independent prefactor of the respective distribution functions. (1) can be written as an integral over a transmission coefficient $TC(\mathcal{E}_\mathrm{t})$ and a supply function $N(\mathcal{E}_\mathrm{t})$

$$J_\mathrm{g} = \frac{4\pi m_\mathrm{eff}\mathrm{q}}{\mathrm{h}^3} \int\limits_0^\infty TC(\mathcal{E}_\mathrm{t}) N(\mathcal{E}_\mathrm{t})\,\mathrm{d}\mathcal{E}_\mathrm{t}. \tag{2}$$

The Wentzel-Kramers-Brillouin (WKB) approximation yields a simple expression for the transmission coefficient of trapezoidal and triangular barriers

$$TC(\mathcal{E}_\mathrm{t}) = \exp\left\{-4\frac{\sqrt{2m_\mathrm{ox}}}{3\hbar\mathrm{q}F_\mathrm{ox}} \cdot \phi(\mathcal{E}_\mathrm{t})\right\} \tag{3}$$

where F_ox is the electric field and m_ox the electron mass in the oxide. The function $\phi(\mathcal{E}_\mathrm{t})$ is defined as

$$\phi(\mathcal{E}_\mathrm{t}) = \begin{cases} (\Phi - \mathcal{E}_\mathrm{t})^{3/2} & \Phi_0 < \mathcal{E}_\mathrm{t} < \Phi \\ (\Phi - \mathcal{E}_\mathrm{t})^{3/2} - (\Phi_0 - \mathcal{E}_\mathrm{t})^{3/2} & \mathcal{E}_\mathrm{t} < \Phi_0 \end{cases} \tag{4}$$

where Φ and Φ_0 are the SiO_2 barrier heights at the boundaries of the gate oxide. The value of Φ_0 is calculated from $\Phi_0 = \Phi - \mathrm{q} \cdot F_\mathrm{ox} \cdot t_\mathrm{ox}$ where q is the electron charge and t_ox denotes the gate oxide thickness. If a Fermi-Dirac distribution is used for f_1 and f_2, the supply function $N(\mathcal{E}_\mathrm{t})$ evaluates to

$$N(\mathcal{E}_\mathrm{t}) = \mathrm{k_B}T \ln\left[\frac{1 + \exp\left((\mathcal{E}_\mathrm{f1} - \mathcal{E}_\mathrm{t})/\mathrm{k_B}T\right)}{1 + \exp\left((\mathcal{E}_\mathrm{f2} - \mathcal{E}_\mathrm{t})/\mathrm{k_B}T\right)}\right] \tag{5}$$

where $\mathcal{E}_\mathrm{f1}$ and $\mathcal{E}_\mathrm{f2}$ denote the Fermi energies at the semiconductor-oxide interfaces. This expression is frequently used in literature and implemented in all major device simulators. However, the assumption of a Fermi-Dirac distribution is not valid in the channel of turned-on MOSFET devices. It is therefore necessary to use advanced models to describe the distribution function.

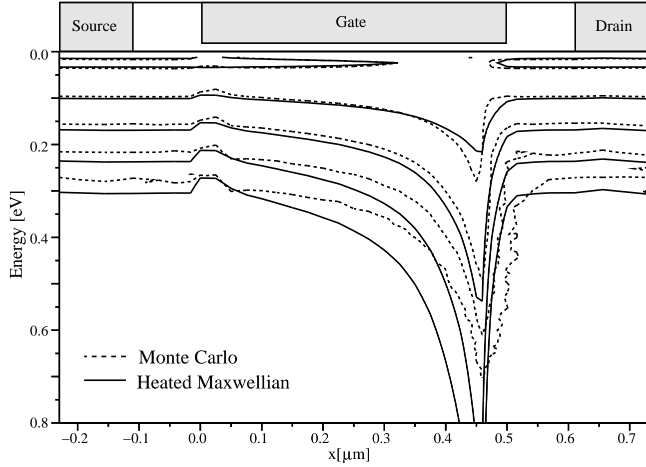

Figure 1: Contour lines of the heated Maxwellian EED compared to Monte Carlo results.

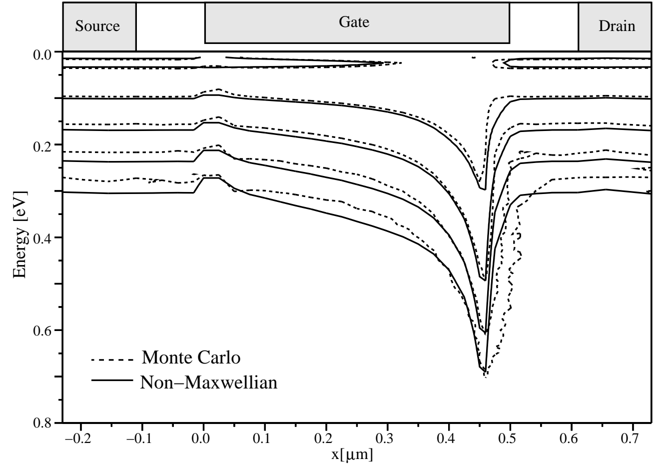

Figure 2: Contour lines of the non-Maxwellian EED compared to Monte Carlo results.

3 DISTRIBUTION FUNCTION

Distribution function modeling of hot carriers in the channel region of a MOSFET has been studied intensively [2], [3]. The topic is crucial because the assumption of a cold Maxwellian distribution function

$$f(\mathcal{E}) = A \cdot \exp\left(-\frac{\mathcal{E}}{\mathrm{k_B} \cdot T_\mathrm{L}}\right), \tag{6}$$

with T_L being the lattice temperature and A a normalization constant depending on the Fermi energy, underestimates the high-energy tail of the electron energy distribution near the drain region. The straightforward approach is to use a heated Maxwellian distribution

$$f(\mathcal{E}) = A \cdot \exp\left(-\frac{\mathcal{E}}{\mathrm{k_B} \cdot T_n}\right) \tag{7}$$

where the lattice temperature T_L is simply replaced by the electron temperature T_n calculated from a suitable transport model. We applied a Monte Carlo simulator employing analytical non-parabolic bands to check the validity of the heated Maxwellian approximation. Fig. 1 shows the contour lines of the heated Maxwellian EED in comparison to Monte Carlo results for a MOSFET device with a gate length of $L_\mathrm{g} = 500\,\mathrm{nm}$ at $V_\mathrm{DS} = V_\mathrm{GS} = 1\,V$. Neighboring lines differ by a factor of 10. It can be clearly seen that the heated Maxwellian distribution (full lines) yields only poor agreement with the Monte Carlo results (dashed lines). Particularly the high-energy tail near the drain side of the channel is heavily overestimated by the heated Maxwellian model.

A generalized expression for the EED has been proposed to account for the high-energy tail [4]:

$$f(\mathcal{E}) = A \exp\left[-\left(\frac{\mathcal{E}}{a}\right)^b\right]. \tag{8}$$

It was found that the values of a and b can be mapped to the solution variables T_n and β_n of a six moments transport model [5]. These solution variables are defined by the first and second order moments $\langle\mathcal{E}\rangle$ and $\langle\mathcal{E}^2\rangle$ of the distribution function:

$$\frac{3\mathrm{k_B}T_n}{2} = \langle\mathcal{E}\rangle \tag{9}$$

$$\frac{5\beta_n}{3} = \frac{\langle\mathcal{E}^2\rangle}{\langle\mathcal{E}\rangle^2}. \tag{10}$$

(8) has been shown to accurately reproduce Monte Carlo results of turned-on MOSFETs. It was successfully applied to the calculation of impact ionization coefficients [4] and gate currents [6].

This model would require the solution of a six moments transport model which may not be feasible for everyday TCAD applications. We therefore approximate the kurtosis β_n by an expression obtained from the local balance equation in homogeneously doped bulk silicon where a fixed relationship between β_n, T_n and the lattice temperature T_L exists:

$$\beta_\mathrm{Bulk}(T_n) = \frac{T_\mathrm{L}^2}{T_n^2} + 2\frac{\tau_\beta}{\tau_\epsilon}\frac{\mu_S}{\mu_n}(1 - \frac{T_\mathrm{L}}{T_n}) \tag{11}$$

In this expression τ_ϵ, τ_β, μ_n, and μ_S are the energy relaxation time, the kurtosis relaxation time, the electron mobility, and the energy flux mobility, respectively. We used a fit to Monte Carlo data for homogeneously doped bulk silicon for $\tau_\beta\mu_S/\tau_\epsilon\mu_n$ [3]. As the influence of the band structure was found to be negligible for the total tunneling current we restrict ourselves to the parabolic case where we find the moments of (8) [4]

$$T_n = \frac{2}{3}\frac{\Gamma(5/2b)}{\Gamma(3/2b)}\frac{a}{\mathrm{k_B}} \tag{12}$$

$$\beta_n = \frac{3}{5}\frac{\Gamma(3/2b)\,\Gamma(7/2b)}{\Gamma(5/2b)^2}. \tag{13}$$

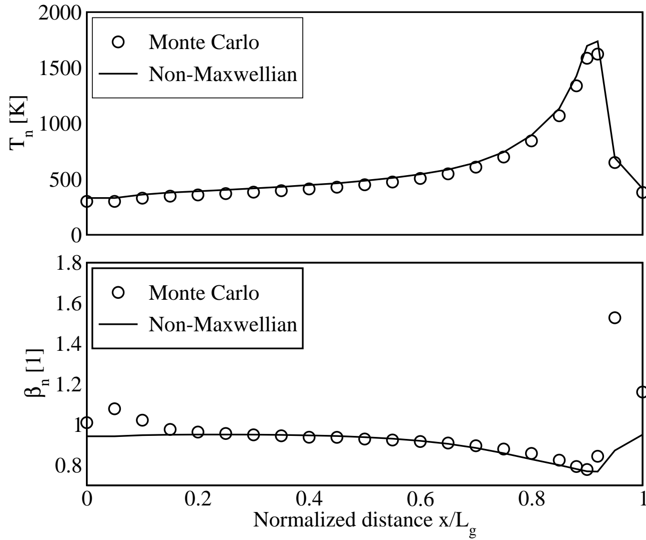

Figure 3: The values of β_n and T_n along the channel of a 0.5 µm gate length MOSFET.

(12) can be easily inverted to find $a(T_n)$ whereas the inversion of (13) to find $b(T_n)$ at $\beta_n(b) = \beta_{\mathrm{Bulk}}(T_n)$ cannot be given in a closed form. We therefore used the fit expression

$$b(T_n) = 1 + b_0 z^{b_1} + b_2 z^{b_3} \tag{14}$$

with $z = 1 - T_{\mathrm{L}}/T_n$ and the parameters $b_0 = 38.82$, $b_1 = 101.11$, $b_2 = 3.40$, and $b_3 = 12.93$. Using $a(T_n)$ and $b(T_n)$ the Monte Carlo EED can be approximated without knowledge of β_n as shown in Fig. 2. Thus, the self-consistent approach can be used within the energy-transport model which provides the carrier concentration and temperature.

With this generalized expression for the distribution function and the assumption of a Maxwellian EED in the poly gate, the supply function $N(\mathcal{E}_{\mathrm{t}})$ becomes

$$N(\mathcal{E}_{\mathrm{t}}) = A_1 \frac{a}{b} \Gamma_{\mathrm{i}} \left[\frac{1}{b}, \left(\frac{\mathcal{E}_{\mathrm{t}}}{a} \right)^b \right] - A_2 \mathrm{k_B} T_{\mathrm{L}} \exp \left[- \frac{\mathcal{E}_{\mathrm{t}} + \Delta\mathcal{E}_{\mathrm{C}}}{\mathrm{k_B} T_{\mathrm{L}}} \right] \tag{15}$$

where $\Gamma_{\mathrm{i}}(\alpha, \beta)$ denotes the incomplete gamma function. This expression represents a generalized supply function which, for $a = \mathrm{k_B} T_{\mathrm{L}}$ and $b = 1$, describes cold carrier tunneling as well.

Due to the simple shape of the distribution function the model does not account for the emerging population of cold carriers near the drain end of the channel which leads to a significant inaccuracy in the shape of the EED

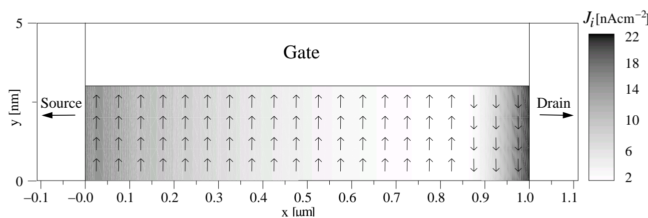

Figure 4: Distribution of the gate current density.

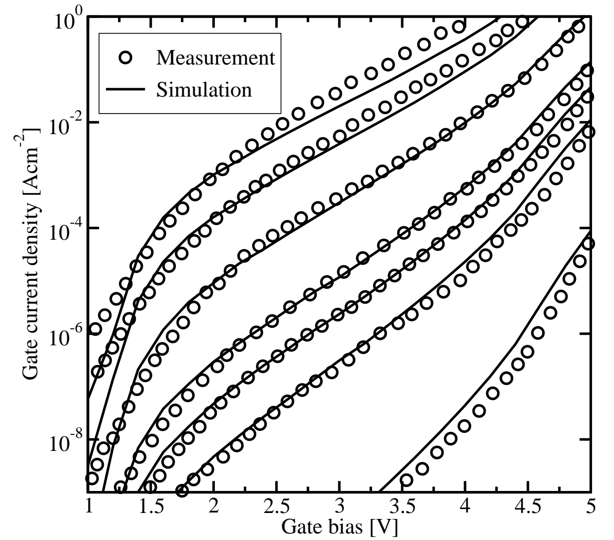

Figure 5: Gate current density of pMOS devices compared to measurements.

[3]. Looking at the values of the electron temperature T_n and kurtosis β_n shown in Fig. 3 it can be seen, however, that this error is confined to a very small part of the gate length, thus being negligible for the total gate current.

4 RESULTS

The outlined model with the generalized expression for the supply function has been implemented in the general-purpose device simulator MINIMOS-NT and compared to Monte Carlo results and measurements. The tunneling current is calculated between two specified boundaries of insulator or semiconductor segments with N interface nodes. The total current is found by

$$I = w \cdot \int_0^{L_g} J(x)\,\mathrm{d}x \approx w \cdot \sum_{i=1..N} J_i \Delta x_i \tag{16}$$

where w is the gate width, J_i the local tunneling current density calculated from (2), and Δx_i the interface length associated with the node i. The local tunneling current density J_i is added self-consistently to the continuity equation of the neighboring segment.

Fig. 4 shows the local electron tunneling current density in the gate oxide of a 1 µm MOSFET with a gate bias of 1.5 V and a drain bias of 3 V. Arrows indicate the current direction. Near the drain electrons tunnel from the gate contact to the semiconductor, while in the middle of the channel and near the source, electrons tunnel in the opposite direction. This is the reason why for turned-on nMOS devices, the gate current density shows its minimum at positive gate bias.

A comparison of our model with measurements is shown in Fig. 5 for pMOS devices with gate and source grounded [7]. The gate oxide thicknesses from top to bottom are 2.29 nm, 2.45 nm, 2.73 nm, 3.04 nm, 3.22 nm, 3.44 nm, and 4.18 nm, respectively. A substrate doping

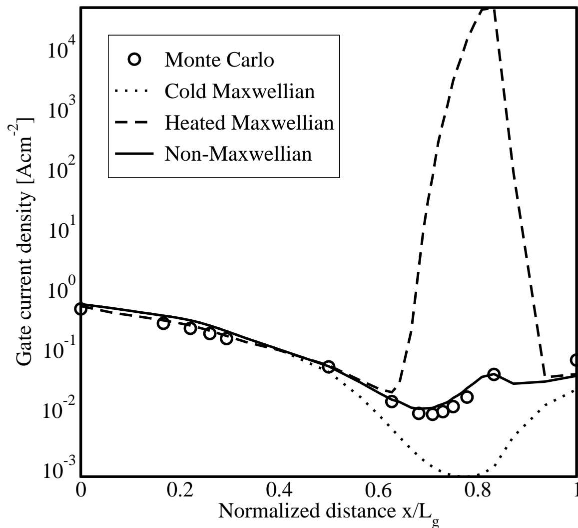

Figure 6: Gate current density along the channel of a 90 nm MOSFET.

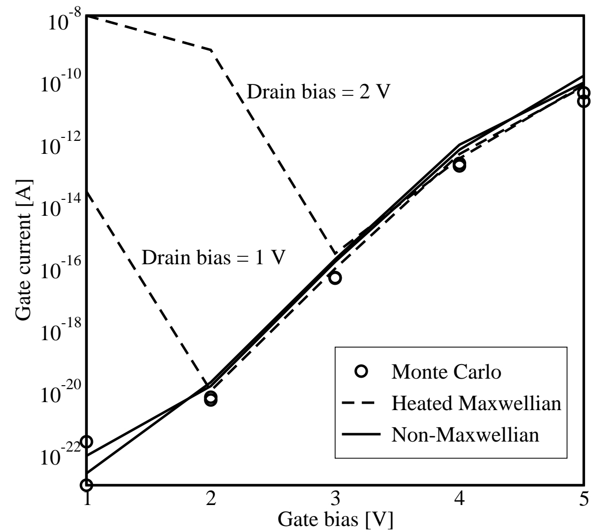

Figure 7: Gate current of a 500 nm MOSFET with 4 nm oxide thickness.

of $10^{17}\,\mathrm{cm}^{-3}$ and an electron effective mass of $m_{\mathrm{ox}} = 0.32\,\mathrm{m}_0$ was used. The measured values can be reproduced over a wide range of oxide thicknesses with a single set of physical parameters.

Fig. 6 shows the gate current density of a 90 nm nMOS device with 2 nm oxide thickness at $V_{GS} = V_{DS} = 1$ V. The values of a and b are calculated from (12) and (14), respectively, and the resulting gate current density along the channel is compared to Monte Carlo results. The model yields excellent agreement, while the heated Maxwellian approximation substantially overestimates the gate current density especially near the drain region. Instead of the heated Maxwellian EED it may even be better to use a cold Maxwellian EED in that regime. The effect of this overestimation can be seen in Fig. 7 where the model was applied to the simulation of a 0.5 µm EEPROM device with an oxide thickness of 4 nm. For low gate voltages the electron temperature is high and the heated Maxwellian approximation massively overestimates the total gate current. As the gate voltage increases, the peak electric field in the channel is reduced. Hence, the high energy tail is suppressed and the models deliver similar results.

5 CONCLUSIONS

We presented a tunneling model which is based on Tsu-Esaki's expression but accounts for a generalized non-Maxwellian electron energy distribution function in the channel region. The model is applied to the simulation of short and long channel MOSFET devices. Good agreement with Monte Carlo simulations and measurements is achieved. In contrast to the heated Maxwellian model which heavily overestimates the gate current density of turned-on MOSFETs, our model delivers accurate results for both hot and cold electron tunneling. The only adjustable parameter is the electron mass in the oxide. The model is thus well suited for the implementation in device simulators solving the energy-transport model.

REFERENCES

[1] R. Tsu and L. Esaki, "Tunneling in a Finite Superlattice," *Appl.Phys.Lett.*, vol. 22, no. 11, pp. 562–564, 1973.

[2] D. Cassi and B. Riccó, "An Analytical Model of the Energy Distribution of Hot Electrons," *IEEE Trans.Electron Devices*, vol. 37, no. 6, pp. 1514–1521, 1990.

[3] T. Grasser, H. Kosina, C. Heitzinger, and S. Selberherr, "Characterization of the Hot Electron Distribution Function Using Six Moments," *J.Appl.Phys.*, vol. 91, no. 6, pp. 3869–3879, 2002.

[4] T. Grasser, H. Kosina, and S. Selberherr, "Influence of the Distribution Function Shape and the Band Structure on Impact Ionization Modeling," *J.Appl.Phys.*, vol. 90, no. 12, pp. 6165–6171, 2001.

[5] T. Grasser, H. Kosina, M. Gritsch, and S. Selberherr, "Using Six Moments of Boltzmann's Transport Equation for Device Simulation," *J.Appl.Phys.*, vol. 90, no. 5, pp. 2389–2396, 2001.

[6] A. Gehring, T. Grasser, H. Kosina, and S. Selberherr, "Simulation of Hot-Electron Oxide Tunneling Current Based on a Non-Maxwellian Electron Energy Distribution Function," *J.Appl.Phys.*, vol. 92, no. 10, pp. 6019–6027, 2002.

[7] S. H. Lo, D. A. Buchanan, and Y. Taur, "Modeling and Characterization of Quantization, Polysilicon Depletion and Direct Tunneling Effects in MOSFETs with Ultrathin Oxides," *IBM J.Res.Dev.*, vol. 43, no. 3, pp. 327–337, 1999.