Consistent Comparison of Tunneling Models for Device Simulation

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

Institute for Microelectronics, TU Vienna, Gusshausstr. 27–29, A-1040 Vienna, Austria
Phone +43-1-58801-36016, FAX +43-1-58801-36099, e-mail: gehring@iue.tuwien.ac.at

Abstract

We present a survey of tunneling models describing carrier transport through insulating layers for semiconductor device simulation. Based on Tsu-Esaki's equation we separately discuss models for the energy distribution function and the transmission coefficient. We use a generalized non-Maxwellian distribution function to account for hot carrier tunneling. We show how to correctly calculate the transmission coefficient of energy barriers when the transfer-matrix methods fail and compare the models to commonly used compact models.

1 Introduction

For the proper prediction of device performance in state-of-the-art sub-quartermicron devices the simulation of quantum-mechanical tunneling effects is of increasing importance. The application area of such models ranges from the evaluation of gate stacks for advanced high-k gate insulator materials to source-drain tunneling which will become a matter of concern with further device size reduction.

However, tunneling model implementations in state-of-the-art device simulators rely almost exclusively on Tsu-Esaki's equation [1] using a Fermi-Dirac or Maxwellian distribution function and transmission coefficients derived either from the WKB approximation or Gundlach's formula [2]. In miniaturized devices these assumptions are violated in two important aspects.

First, the electron energy distribution function (EED) can in general not be described by a Fermi-Dirac or Maxwellian distribution. Higher order moments are necessary to more accurately characterize the distribution of hot carriers [3]. We propose a simplification to a recently proposed model where the EED is characterized by the carrier temperature and concentration, making it applicable for device simulators which solve the energy-transport equations.

The second weakness lies in the estimation of the transmission coefficient by the WKB or Gundlach method. Energy barriers which are not of triangular or trapezoidal shape are not treated correctly by these models. To accurately describe tunneling in such cases, Schrödinger's equation must be solved. This can be achieved using the transfer-matrix method [1].

It was, however, shown that the transfer-matrix method is numerically stable only for layer thicknesses up to a few nanometers. We therefore propose to use the quantum transmitting boundary method for this purpose [4]. We show that this method allows a stable and reliable evaluation of transmission coefficients.

2 Theory of Tunneling

The most prominent and almost exclusively used expression has been developed by Duke [5] and used by Tsu and Esaki to describe tunneling through a one-dimensional superlattice [1]. It is commonly known as Tsu-Esaki expression. The current density is given as

$$J_{\mathrm{g}} = \frac{4\pi m_{\mathrm{eff}} \mathrm{q}}{h^3} \int_0^\infty TC(\mathcal{E}_{\mathrm{t}}) N(\mathcal{E}_{\mathrm{t}}) \, \mathrm{d}\mathcal{E}_{\mathrm{t}} \tag{1}$$

with a transmission coefficient $TC(\mathcal{E}_{\mathrm{t}})$ and a supply function $N(\mathcal{E}_{\mathrm{t}})$ which is defined as

$$N(\mathcal{E}_{\mathrm{t}}) = \int_0^\infty \left[f_1(\mathcal{E}_{\mathrm{t}} + \mathcal{E}_{\mathrm{l}}) - f_2(\mathcal{E}_{\mathrm{t}} + \mathcal{E}_{\mathrm{l}} + \Delta\mathcal{E}_{\mathrm{C}}) \right] \mathrm{d}\mathcal{E}_{\mathrm{l}}. \tag{2}$$

In these expressions 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 latter is perpendicular to the substrate-oxide interface. The electron energy distribution functions in the gate and substrate are denoted by f_1 and f_2, respectively. The difference in the conduction band edges $\Delta\mathcal{E}_{\mathrm{C}}$ is calculated as $\mathcal{E}_{\mathrm{C1}} - \mathcal{E}_{\mathrm{C2}}$, and TC is the transmission coefficient. It is assumed that the transmission coefficient only depends on the transversal energy component and can therefore be treated independently of the supply function. For a Fermi-Dirac EED the supply function evaluates to

$$N(\mathcal{E}_{\mathrm{t}}) = \mathrm{k_B} T \ln \left[\frac{1 + \exp\left(\frac{\mathcal{E}_{\mathrm{f}} - \mathcal{E}_{\mathrm{t}}}{\mathrm{k_B} T} \right)}{1 + \exp\left(\frac{\mathcal{E}'_{\mathrm{f}} - \mathcal{E}_{\mathrm{t}})}{\mathrm{k_B} T} \right)} \right] \tag{3}$$

where $\mathcal{E}_{\mathrm{f}}$ and $\mathcal{E}'_{\mathrm{f}}$ denote the Fermi energies at the semiconductor-oxide interfaces. This expression is frequently used in the literature and implemented in device simulators. However, the assumption of a Fermi-Dirac distribution is not valid in the channel of a turned-on submicron MOSFET. Advanced models for the distribution function are necessary.

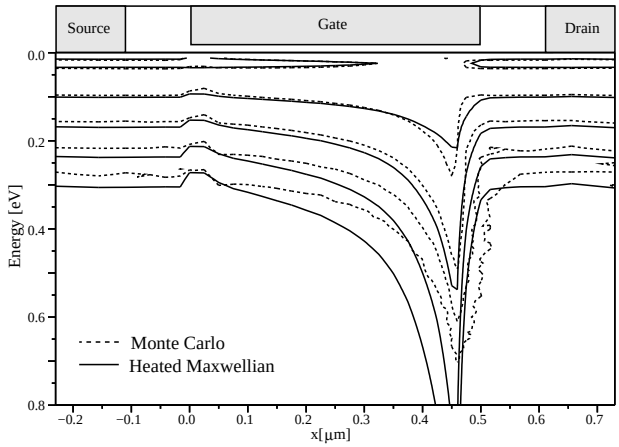

Figure 1: Contour lines of the heated Maxwellian EED compared to Monte Carlo results.

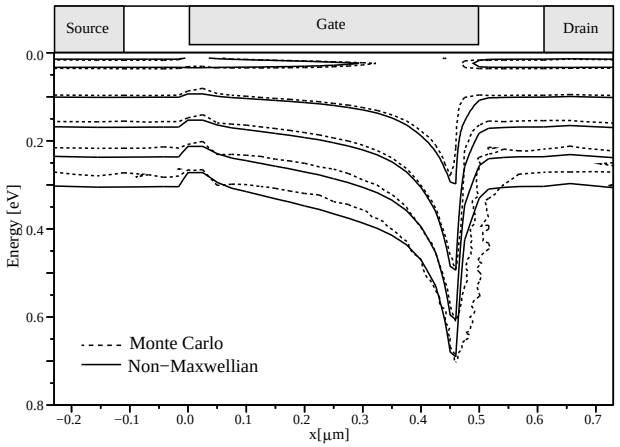

Figure 2: Contour lines of the non-Maxwellian EED compared to Monte Carlo results.

2.1 Distribution Function Modeling

Models for the EED of hot carriers in the channel region of a MOSFET have been studied by numerous authors [6, 7]. The topic is crucial because the assumption of a cold Maxwellian distribution function

$$f(\mathcal{E}) = A \cdot \exp\left(-\frac{\mathcal{E}}{\mathrm{k_B} \cdot T_\mathrm{L}}\right), \tag{4}$$

where T_L denotes the lattice temperature and A a normalization constant, underestimates the high-energy tail of the EED near the drain region. The straightforward approach is to use a heated Maxwellian distribution function where the lattice temperature T_L is simply replaced by the electron temperature T_n. 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 with a gate length of $L_\mathrm{g} = 500\,\mathrm{nm}$ at $V_\mathrm{DS} = V_\mathrm{GS} = 1\,V$. 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 by Grasser *et al.* [8]:

$$f(\mathcal{E}) = A \exp\left[-\left(\frac{\mathcal{E}}{a}\right)^b\right]. \tag{5}$$

The values of a and b are mapped to the solution variables T_n and β_n of a six moments transport model [8]. The electron concentration n, electron temperature T_n, and kurtosis β_n are derived from the first three even moments of the distribution function (5). This expression accurately reproduces Monte Carlo results in the source and the middle region of the channel of a turned-on MOSFET. The model, however, requires the solution of a six moments transport model, which is computationally expensive. We therefore approximate the kurtosis β_n by an expression obtained for a bulk semiconductor where a fixed relationship between β_n, T_n and the lattice temperature T_L exists:

$$\beta_\mathrm{Bulk}(T_n) = \frac{T_\mathrm{L}^2}{T_n^2} + 2\frac{\tau_\beta}{\tau_\epsilon}\frac{\mu_S}{\mu_n}\left(1 - \frac{T_\mathrm{L}}{T_n}\right) \tag{6}$$

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]. Estimating the kurtosis from (6), the EED (5) can be used within the energy-transport model. Restricting ourselves to the case of a parabolic band structure, we find [8]

$$T_n = \frac{2}{3}\frac{\Gamma(5/2b)}{\Gamma(3/2b)}\frac{a}{\mathrm{k_B}} \tag{7}$$

$$\beta_n = \frac{3}{5}\frac{\Gamma(3/2b)\,\Gamma(7/2b)}{\Gamma(5/2b)^2}. \tag{8}$$

While (7) can easily be inverted to obtain $a(T_n)$, the inversion of (8) to find $b(T_n)$ at $\beta_n(b) = \beta_\mathrm{Bulk}(T_n)$ cannot be given in a closed form. We therefore use the fit expression $b(T_n) = 1 + b_0 z^{b_1} + b_2 z^{b_3}$ 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. With the generalized distribution (5) in the channel and a Maxwellian EED in the poly gate, the supply function (2) finally becomes

$$N(\mathcal{E}_\mathrm{t}) = A_1\frac{a}{b}\,\Gamma_\mathrm{i}\left[\frac{1}{b}, \left(\frac{\mathcal{E}_\mathrm{t}}{a}\right)^b\right] - N_2 \exp\left[-\frac{\mathcal{E}_\mathrm{t} + \Delta\mathcal{E}_\mathrm{C}}{\mathrm{k_B}T_\mathrm{L}}\right] \tag{9}$$

where $N_2 = A_2 \mathrm{k_B} T_\mathrm{L}$ and $\Gamma_\mathrm{i}(\alpha, \beta)$ denotes the incomplete gamma function. This expression has been implemented in the device simulator MINIMOS-NT.

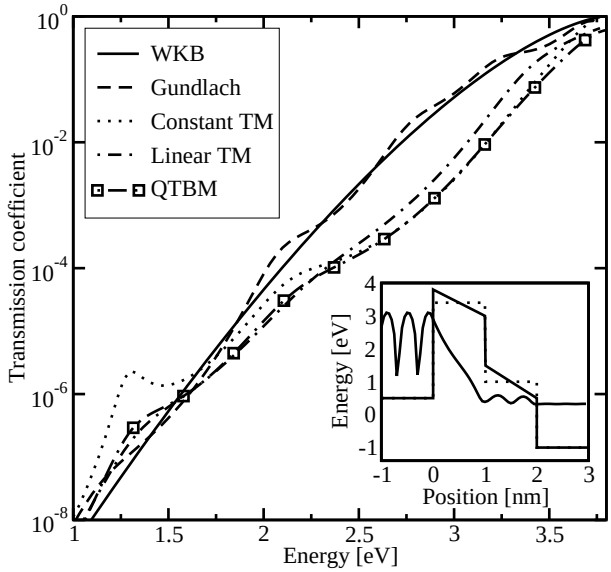

Figure 3: Transmission coefficient TC as a function of energy for a stacked oxide.

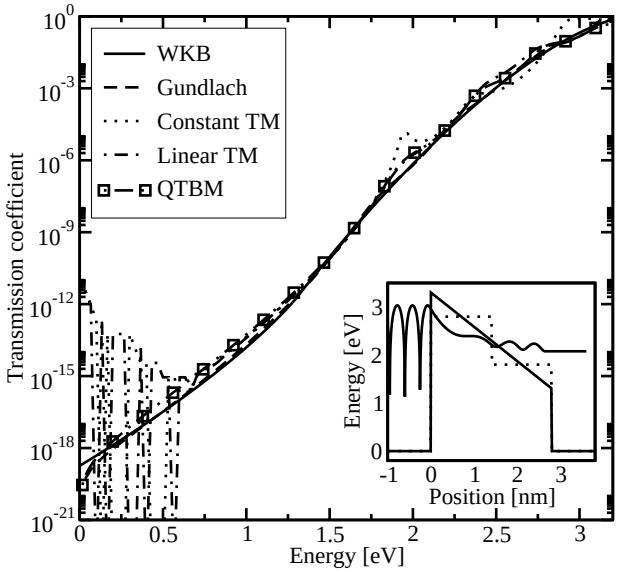

Figure 4: Transmission coefficient TC as a function of energy for a 3 nm thick layer of SiO_2.

2.2 Transmission Coefficient Modeling

Apart from the distribution function the quantum-mechanical transmission coefficient is the second building block of any tunneling model. It is based on the probability flux

$$j = \frac{\hbar}{2im} \cdot (\Psi^* \cdot \nabla\Psi - \nabla\Psi^* \cdot \Psi) \qquad (10)$$

where Ψ is the wave function, m the carrier mass, and $i = \sqrt{-1}$. The transmission coefficient is defined as the ratio of the fluxes j_2 and j_1 due to an incident wave in Region 1 and a transmitted wave in Region 2. The respective wave functions can be found by solving the stationary one-dimensional Schrödinger equation in the barrier region.

This can be achieved using different numerical methods, such as the commonly applied Wentzel-Kramers-Brillouin (WKB) approximation or Gundlach's method [2] which is accurate for triangular and trapezoidal barriers.

A more general approach is the transfer-matrix method [1] the basic principle of which is the approximation of an arbitrary-shaped energy barrier by a series of barriers with constant or linear potential. Since the wave function for such barriers can easily be calculated, the transfer matrix, which relates the incoming to the outgoing wave amplitudes, can be derived by a number of subsequent matrix computations. From the transfer matrix, the transmission coefficient can be calculated. However, the main shortcoming of the method is that it becomes numerically instable for thick barriers which is due to the multiplication of exponentially growing and decaying states, leading to rounding errors which eventually exceed the amplitude of the wave function itself [9].

An alternative method to compute the transmission coefficient is based on the quantum transmitting boundary method [4]. The method uses a finite-difference approximation of Schrödinger's equation with open boundary conditions. This results in a complex-valued linear equation system for the unknown values of the wave amplitudes. The method is easy to implement, fast, and more robust than the transfer-matrix method.

The different numerical methods are compared in Fig. 3 for a 2 nm gate stack comprised of SiO_2 and an adjacent material with a lower barrier height. The WKB and Gundlach methods, which approximate the barrier with a linear potential, overestimate the transmission coefficient as compared to the transfer-matrix based methods. It can also be seen that the two-step constant-potential approximation (dotted line) shows a resonance at 1.3 eV. The linear potential approximation shows no resonance and is more accurate. The quantum transmitting boundary method (QTBM) was applied with a resolution of 500 points in the x-direction and represents the most accurate solution. The inset in Fig. 3 shows the shape of the barrier and the squared wave function for an energy of 2.8 eV on a logarithmic scale.

While the transfer-matrix based methods yield reasonable results for a thin gate stack, they cannot be used for thick barriers as shown in Fig. 4 for a 3 nm SiO_2 gate oxide. For energies lower than ≈ 0.6 eV the transfer-matrix methods become unstable and give erroneous results. The transmitting boundary method, however, delivers the same results as the Gundlach method which provides an accurate analytical solution in this case. It therefore promises to be a reliable method for the estimation of the transmission coefficient of high-k gate stacks.

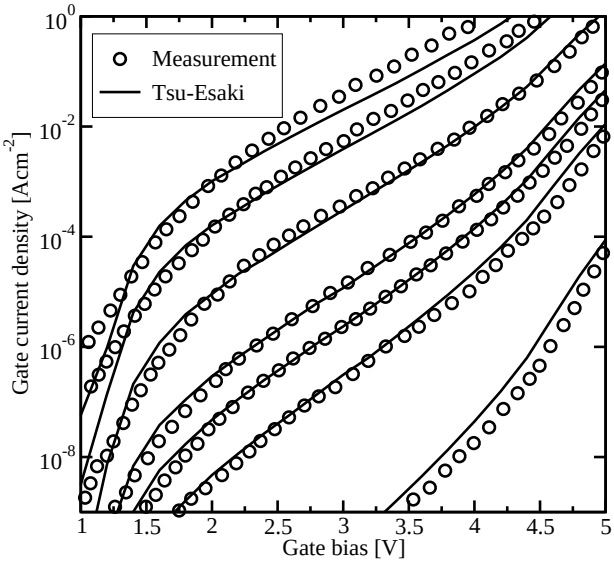

Figure 5: Comparison of Tsu-Esaki's model with measurements of a pMOS device [10].

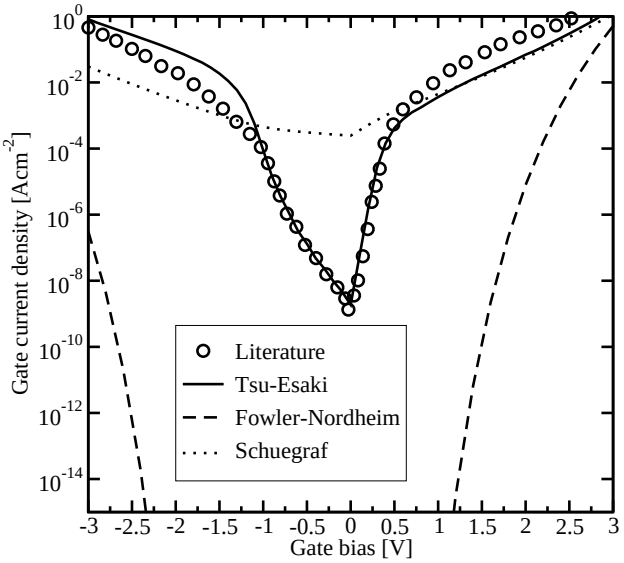

Figure 6: Comparison of compact tunneling models for a nMOS device [13].

3 Simulation Results

The Tsu-Esaki expression with a WKB transmission coefficient is in good agreement with measurements as shown in Fig. 5 for pMOS devices with oxide thicknesses of 2.29 nm, 2.45 nm, 2.73 nm, 3.04 nm, 3.22 nm, 3.44 nm, and 4.18 nm, respectively [10]. The gate current density can be reproduced over a wide range of oxide thicknesses with a single set of physical parameters. The model seems to be well suited to describe cold-carrier tunneling through single-layer dielectrics.

For the use in practical device simulation it is desirable to use compact models which do not require large computational resources and may be used for a quick estimation of the oxide thickness from IV-data. The most commonly used model to describe tunneling is the Fowler-Nordheim formula [11]:

$$J = \frac{\mathrm{q}^3 m_{\mathrm{eff}}}{8\pi m_{\mathrm{ox}} h \Phi_{\mathrm{B}}} \cdot F_{\mathrm{ox}}^2 \cdot \exp\left(-\frac{4\sqrt{2m_{\mathrm{ox}}\Phi_{\mathrm{B}}^3}}{3\hbar \mathrm{q} F_{\mathrm{ox}}}\right) \qquad (11)$$

It can be derived from (1) by the assumption of zero temperature and a triangular barrier. Schuegraf and Hu [12] derived simple correction terms for this expression to make it applicable to the regime of trapezoidal barriers. The correction factors cause the minimum tunneling current to occur for minimum bias, and not for minimum field in the oxide, which is a major problem of the Fowler-Nordheim formula.

Fig. 6 shows the results of the compact models for the case of an nMOS device with 2 nm oxide thickness. The Schuegraf model fails to describe the tunneling current density at low bias, because it does not take the band structure into account. For high bias, however, it may be used to provide an estimation of the gate current. The Fowler-Nordheim formula totally fails for this application.

4 Conclusion

We presented a hierarchy of tunneling models for device simulation. Based on Tsu-Esaki's equation models for the supply function and the transmission coefficient are discussed. We described a method to correctly account for the electron energy distribution in the channel of a turned-on MOSFET. Furthermore, we outlined methods to estimate the transmission coefficient of an energy barrier, namely the WKB, Gundlach, constant and linear potential transfer-matrix, and transmitting boundary methods. The transmitting boundary method may be used to overcome inherent numerical problems of the transfer-matrix method and is the method of choice for the evaluation of tunneling through high-k gate stacks. A study of compact models showed that the commonly used Fowler-Nordheim formula fails for low bias, while the Schuegraf model shows reasonable agreement with measurements.

References

[1] R. Tsu and L. Esaki, Appl.Phys.Lett. **22**, 562 (1973).

[2] K. Gundlach, Solid-State Electron. **9**, 949 (1966).

[3] T. Grasser *et al.*, J.Appl.Phys. **91**, 3869 (2002).

[4] W. R. Frensley, Superlattices & Microstructures **11**, 347 (1992).

[5] C. B. Duke, *Tunneling in Solids* (Academic Press New York, 1969).

[6] D. Cassi and B. Riccó, IEEE Trans.Electron Devices **37**, 1514 (1990).

[7] K.-I. Sonoda *et al.*, J.Appl.Phys. **80**, 5444 (1996).

[8] T. Grasser *et al.*, J.Appl.Phys. **90**, 6165 (2001).

[9] T. Usuki *et al.*, Physical Review B **52**, 8244 (1995).

[10] S. H. Lo *et al.*, IBM J.Res.Dev. **43**, 327 (1999).

[11] M. Lenzlinger and E. Snow, J.Appl.Phys. **40**, 278 (1969).

[12] K. F. Schuegraf and C. Hu, IEEE Trans.Electron Devices **41**, 761 (1994).

[13] J. Cai and C.-T. Sah, J.Appl.Phys. **89**, 2272 (2001).