Simulation of hot-electron oxide tunneling current based on a non-Maxwellian electron energy distribution function

A. Gehring,[a] T. Grasser, H. Kosina, and S. Selberherr
Institute for Microelectronics, TU Vienna, Gusshausstrasse 27–29, A-1040 Vienna, Austria

(Received 27 November 2001; accepted 30 August 2002)

For the simulation of gate oxide tunneling currents in sub-quarter-micron devices, the correct modeling of the electron energy distribution function is crucial. Our approach is based on a recently presented transport model which accounts for six moments of the Boltzmann transport equation. A corresponding analytical model for the electron energy distribution function shows good agreement with Monte Carlo data. Using this model, we show that the gate current behavior of short-channel devices can be reproduced correctly. This is not the case for the heated Maxwellian approximation which leads to a massive overestimation of gate currents especially for devices with small gate lengths. We develop a formalism to distinguish between cases where the heated Maxwellian distribution delivers correct results and cases where it overestimates the tunneling current at low drain bias and find that for oxide thicknesses around 2 nm, the heated Maxwellian approximation is only valid for electron temperatures below about 1000 K. © *2002 American Institute of Physics.*

[DOI: 10.1063/1.1516617]

I. INTRODUCTION

The design of submicron semiconductor devices with gate oxide thicknesses around or below 2 nm depends increasingly on gate oxide tunneling currents which lead to additional energy consumption of logic devices and reduced retention time of memory devices. A proper approach to model the phenomenon of oxide tunneling is necessary and has to be implemented in device and circuit simulators. There are several approaches published which range from pure fit formulas[1] to more physics-based models based on the Fowler–Nordheim formula.[2] However, only the thermionic emission model[3] and the Bardeen model,[4] which are both based on the Tsu-Esaki equation,[5] are allowed to explicitly account for the electron energy distribution function (EED).

The approximation of the transmission coefficient ranges from the Wentzel–Kramer–Brillouin (WKB) approximation up to the calculation of wave functions in the oxide using analytical solutions via the Gundlach[6] or transfer-matrix approach.[7] With increasing sophistication of the models, the computational burden also increases and the need for a proper tradeoff between accuracy and simulation time has to be kept in mind to achieve models which are feasible for device simulators.

One of the commonly used assumptions in the derivation of tunneling models is that the electron energy distribution in the channel can be described by a heated Maxwellian distribution. This assumption is not valid for sub-quarter-micron MOSFET devices where it was shown that the distribution function cannot be described by its average energy or temperature alone.[8]

In this article we show that the assumption of a heated Maxwellian distribution leads to erroneous results if the electron temperature exceeds an oxide-thickness dependent threshold. A recently developed expression for the EED based on a transport model using six moments of the Boltzmann transport equation overcomes this problem and delivers correct results independently of the carrier temperature.

The article is organized as follows. In Sec. II we describe our hot carrier tunneling model and elaborate on the distribution function used. In Sec. III we present simulation results and a comparison to measurement data. Section IV is dedicated to the derivation of a temperature limit up to which the heated Maxwellian assumption can be used safely, and finally, some conclusions are given in Sec. V.

II. HOT CARRIER TUNNELING

Tsu and Esaki[5] presented an expression to describe the tunneling current density through a dielectric layer which reads

$$J_g = \frac{q}{4\pi^3\hbar} \cdot \int d^2k_t \int_0^\infty [f(\mathcal{E}) - f(\mathcal{E}')] \cdot T(\mathcal{E}) \frac{\partial \mathcal{E}}{\partial k_l} dk_l, \quad (1)$$

where $f(\mathcal{E})$ is the electron energy distribution function, $T(\mathcal{E})$ the tunneling probability, k_l and k_t the longitudinal and transversal wave numbers, and $\mathcal{E}' = \mathcal{E} - qV$. For the case of a Fermi–Dirac distribution, this expression can be integrated over the transverse plane, leading to

$$J_g = \frac{4\pi m_{ox} q k_B T}{h^3} \int_0^\infty T(\mathcal{E}) \ln\left(\frac{1+\exp(\mathcal{E}_f - \mathcal{E}/k_B T)}{1+\exp(\mathcal{E}_f' - \mathcal{E})/k_B T} \right) d\mathcal{E}, \quad (2)$$

where $\mathcal{E}_f$ and $\mathcal{E}_f'$ denote the Fermi energies next to the oxide layer. This expression is frequently used in literature (see Ref. 4, and references therein).

However, in the channel of a turned-on sub-quarter-micron MOSFET, the assumption of a Fermi–Dirac energy

[a]Electronic mail: gehring@iue.tuwien.ac.at

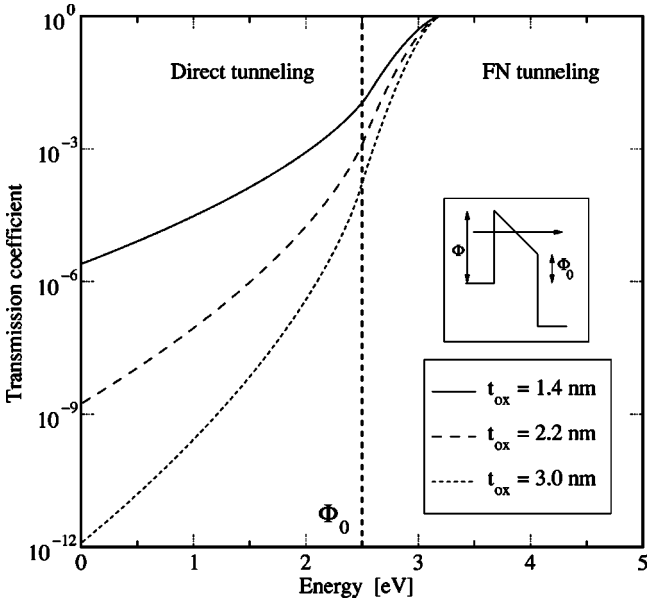

FIG. 1. Transmission coefficient for different oxide thicknesses as a function of energy for $\Phi_0 = 2.5$ and $\Phi = 3.2$ eV. The line shows the location of the transition between the Fowler-Nordheim and the direct tunneling regime.

distribution is not valid. Thus, we follow the approach presented by Fiegna *et al.* in Ref. 9 who assume $f(\mathcal{E}') \approx 0$ in Eq. (1). This implies that only electrons tunneling from the substrate to the gate are taken into account and the distribution of free states in the gate is neglected. The gate current density can then be written as

$$J_g = q \cdot \int_0^\infty f(\mathcal{E}) \cdot g(\mathcal{E}) \cdot v_\perp(\mathcal{E}) \cdot T(\mathcal{E})\, d\mathcal{E}, \tag{3}$$

where $g(\mathcal{E})$ is the density of states and $v_\perp(\mathcal{E})$ the electron velocity perpendicular to the interface. The integration is performed starting from the conduction band edge which serves as reference energy. This approach offers the possibility to explicitly take the non-Maxwellian shape of the electron energy distribution function into account. However, it must be kept in mind that for low oxide voltages, the assumption $f(\mathcal{E}') \approx 0$ might not be valid.

A. Transmission coefficient

A simple model for the tunneling probability can be derived using the WKB approximation[10] for trapezoidal and triangular barriers:

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

with F_{ox} being the electric field and m_{ox} the electron mass in the oxide. The function $\phi(\mathcal{E})$ is defined as

$$\phi(\mathcal{E}) = \begin{cases} (\Phi - \mathcal{E})^{3/2} & \text{for } \Phi_0 < \mathcal{E} < \Phi \\ (\Phi - \mathcal{E})^{3/2} - (\Phi_0 - \mathcal{E})^{3/2} & \text{for } \mathcal{E} < \Phi_0 \end{cases}. \tag{5}$$

Φ and Φ_0 are the upper and lower barrier heights, as shown in the inset in Fig. 1. The value of Φ_0 is calculated as

$$\Phi_0 = \Phi - q \cdot F_{ox} \cdot t_{ox}, \tag{6}$$

where q is the electron charge and t_{ox} denotes the gate oxide thickness. The transmission coefficient for different oxide thicknesses is depicted in Fig. 1 for silicon dioxide and several oxide thicknesses. The transition between direct tunneling and Fowler–Nordheim tunneling leads to the clearly visible change of the slope at $\mathcal{E} = \Phi_0$. The only fitting parameters of this model are the electron mass in the oxide and the barrier height Φ at the Si–SiO_2 interface.

B. Density of states

For the common assumption of a parabolic dispersion relation and with the conduction band edge as reference energy, the density of states is

$$g(\mathcal{E}) = g_0 \cdot \sqrt{\mathcal{E}}, \tag{7}$$

with

$$g_0 = 6 \frac{\sqrt{2} m_{eff}^{3/2}}{\pi^2 \hbar^3}, \tag{8}$$

where m_{eff} is the electron effective mass of the six lowest valleys of the silicon conduction band. As a first order correction to the parabolic band model, we use Kane's dispersion relation[11,12]

$$\frac{\hbar^2 k^2}{2 m_{eff}} = \mathcal{E} \cdot (1 + \alpha \cdot \mathcal{E}). \tag{9}$$

For this expression the density of states $g(\mathcal{E})$ evaluates to

$$g(\mathcal{E}) = g_0 \cdot \sqrt{\mathcal{E}} \cdot \sqrt{1 + \alpha \mathcal{E}} \cdot (1 + 2\alpha \mathcal{E}), \tag{10}$$

with the nonparabolicity factor α being 0.5 eV^{-1} for silicon.

C. Perpendicular velocity

The velocity perpendicular to the semiconductor-gate oxide interface is calculated as[9]

$$v_\perp(\mathcal{E}) = \frac{1}{4\hbar} \frac{\partial \mathcal{E}}{\partial k}. \tag{11}$$

The derivation of this expression is shown in Appendix A. This leads to the following expressions for a parabolic and Kane's dispersion relation:

$$\begin{aligned} &\text{Parabolic: } v_\perp(\mathcal{E}) = \sqrt{\frac{\mathcal{E}}{8 m_{eff}}} \\ &\text{Kane: } v_\perp(\mathcal{E}) = \sqrt{\frac{\mathcal{E}(1 + \alpha \mathcal{E})}{8 m_{eff} (1 + 2\alpha \mathcal{E})^2}}. \end{aligned} \tag{12}$$

Figure 2 shows a comparison of the density of states and the resulting normal velocity for the two dispersion relations. It can be seen that Kane's dispersion relation gives a higher density of states and a lower velocity than the parabolic dispersion relation. The total effect of the chosen dispersion relation on the gate current density will thus be small which explains why good results have been achieved using the parabolic dispersion relation.

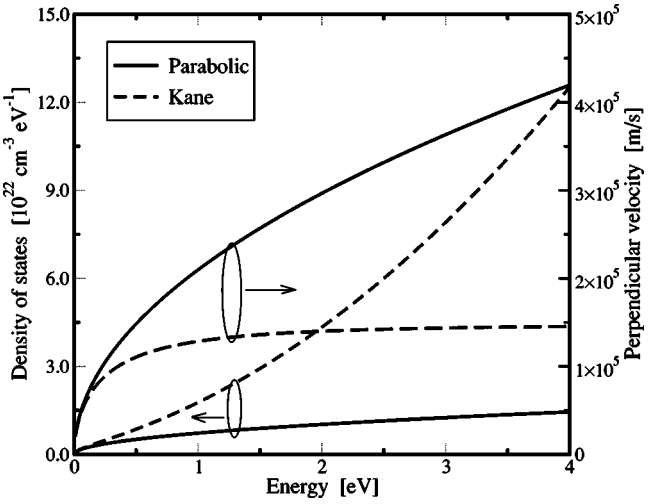

FIG. 2. Density of states $g(\mathcal{E})$ and the resulting normal velocity $v_{\perp}(\mathcal{E})$ as a function of energy for a parabolic and Kane's dispersion relation.

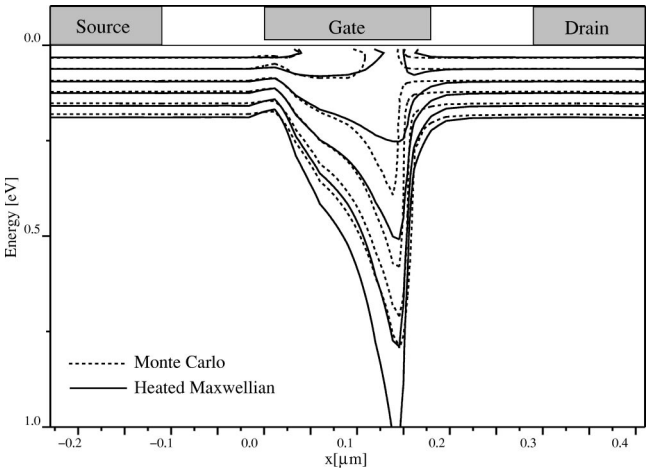

FIG. 3. Comparison of the heated Maxwellian distribution function (full lines) and the EED resulting from Monte Carlo simulations (dashed lines) for a MOSFET with a gate length of 180 nm. Neighboring lines differ by a factor of 10.

D. Distribution function

Various research deals with the problem of distribution function modeling for hot carriers in the channel region of a MOSFET.[13–15] The problem arises from the fact that the assumption of a cold Maxwellian distribution function

$$f(\mathcal{E}) = A\ \exp\left(-\frac{\mathcal{E}}{k_B \cdot T_L}\right), \tag{13}$$

with T_L being the lattice temperature and A a normalization constant accounting for 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 function

$$f(\mathcal{E}) = A\ \exp\left(-\frac{\mathcal{E}}{k_B \cdot T_n}\right), \tag{14}$$

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 nonparabolic bands to check the validity of the heated Maxwellian approximation. The effect of electron–electron interaction, which increases the population of the high energy tail,[13] was neglected in this study. Figure 3 shows the contour lines of the heated Maxwellian EED in comparison to Monte Carlo results for a MOSFET device with a gate length of $L_g = 180\,\text{nm}$ at $V_{DS} = V_{GS} = 1\,V$. It can be clearly seen that the heated Maxwellian distribution (full lines) yields only poor agreement with the Monte Carlo results (dashed lines). The heated Maxwellian distribution overestimates the high-energy tail in the channel. Furthermore, at the drain end of the channel hot electrons mix with cold electrons supplied from the drain region, which leads to an additional population of cold electrons which cannot be reproduced by this model.

Cassi *et al.*[14] presented the following expression for the electron energy distribution function:

$$f(\mathcal{E}) = A\ \exp\left(-\frac{\chi \mathcal{E}^3}{F_{ox}^{1.5}}\right), \tag{15}$$

with χ as fitting parameter and F_{ox} being the electric field in the oxide. This expression was also used by Fiegna *et al.*[9] to model the EEPROM writing process. However, they replaced the electric field by an effective field calculated from the average electron energy. This expression was questioned by Hasnat *et al.* in Ref. 3 where they presented another form for the distribution function:

$$f(\mathcal{E}) = A\ \exp\left(-\frac{\mathcal{E}^{\xi}}{\eta \cdot (k_B T_n)^n}\right). \tag{16}$$

They obtained values of $\xi = 1.3$, $\eta = 0.265$, and $n = 0.75$ by fitting simulation results to measurement data. However, these values fail to describe the shape of the distribution function along the channel.[8] A generalized expression for the EED has been proposed in Ref. 16:

$$f(\mathcal{E}) = A\ \exp\left[-\left(\frac{\mathcal{E}}{a}\right)^b\right]. \tag{17}$$

The values of a and b are mapped to the solution variables T_n and β_n of a six moments transport model[17] as described in Ref. 16. Equation (17) has been shown to accurately reproduce Monte Carlo results in the source and the middle region of the channel of a turned-on MOSFET. It was successfully applied to the calculation of impact ionization coefficients[16] and gate current densities for devices without LDD implants.[18] However, this model is still not able to reproduce the high energy tail of the distribution function near the drain side of the channel because it does not account for the population of cold carriers. A correct description of the high energy tail is crucial for the evaluation of hot-carrier injection at the drain-side used for programming and erasing of EEPROM or Flash devices, as indicated in Ref. 19.

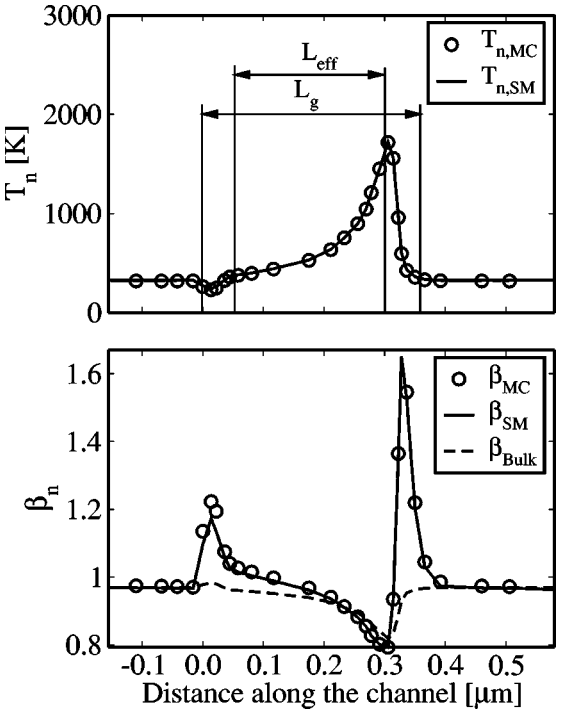

FIG. 4. Electron temperature T_n and the relative kurtosis β_n along the channel of a 350 nm MOSFET compared to the results of a Monte Carlo simulation. The dashed line shows the value of the relative kurtosis in the bulk which was used to indicate the emerging cold Maxwellian part of the EED near the drain side of the channel.

A distribution function accounting for the cold carrier population near the drain contact was proposed by Sonoda *et al.*,[15] and an improved model has been suggested by Grasser *et al.*:[8]

$$f(\mathcal{E}) = A\left\{ \exp\left[-\left(\frac{\mathcal{E}}{a}\right)^b \right] + c\ \exp\left[-\frac{\mathcal{E}}{k_B T_L} \right] \right\}, \tag{18}$$

where the pool of cold carriers in the drain region is correctly modeled by an additional cold Maxwellian subpopulation which leads to a reduced high-energy tail. The values of *a*, *b*, and *c* are again derived from the solution variables of a six moments transport model using the procedure described in Ref. 8. Figure 4 shows the resulting electron temperature along the channel for a 350 nm MOSFET and the relative kurtosis of the distribution function β_n compared to Monte Carlo results. The dotted lines show the value of the relative kurtosis in the bulk, which is used to locate the regions where the cold Maxwellian part of the distribution function emerges. Note that only a six moments transport model can provide information about the kurtosis of the distribution function. The electron concentration, electron temperature and kurtosis are derived from

$$n = \langle 1 \rangle, \tag{19}$$

$$\frac{3 k_B T_n}{2} = \langle \mathcal{E} \rangle, \tag{20}$$

$$\frac{5 \beta_n}{3} = \frac{\langle \mathcal{E}^2 \rangle}{\langle \mathcal{E} \rangle^2}, \tag{21}$$

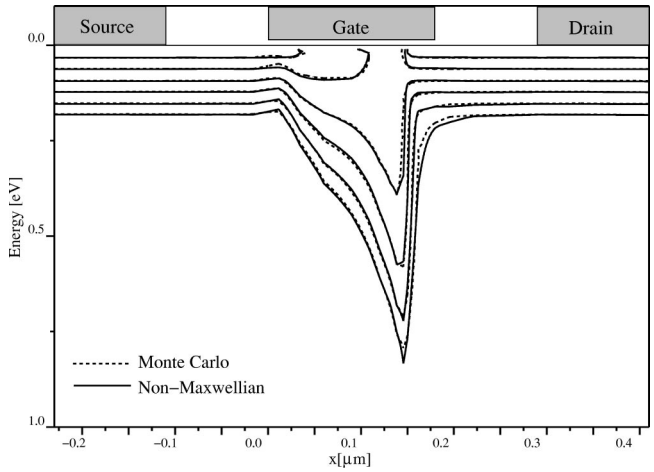

FIG. 5. Comparison of the non-Maxwellian distribution function (full lines) and the EED resulting from Monte Carlo simulations (dashed lines) for a MOSFET with a gate length of 180 nm.

where the moments of the distribution function are defined as

$$\langle \Phi \rangle = \int_0^\infty \Phi\, f(\mathcal{E})\, g(\mathcal{E})\, d\mathcal{E}. \tag{22}$$

The value of the normalization constant *A* can then be calculated from the carrier concentration *n* taken from the transport model. This assures consistency of the model, in contrast to the normalization which was used by Hasnat *et al.*:

$$A = \frac{1}{2\int_0^\infty f(\mathcal{E})\, d\mathcal{E}}. \tag{23}$$

This normalization is independent of the carrier concentration and inevitably leads to erroneous results for both the carrier concentration and the electron temperature. In our case it led to a massive overestimation of the distribution function at all points along the channel. Additionally, the population of cold carriers near the drain side of the channel cannot be reproduced using Hasnat's model.

In Fig. 5, expression (18) is compared to Monte Carlo results showing excellent agreement all along the channel. Figure 6 offers a closer look at the shape of the EED at three points in the channel of a 0.35 μm MOSFET device biased at $V_{DS} = V_{GS} = 1$ V. Near the source side of the channel, the cold Maxwellian, heated Maxwellian and non-Maxwellian distribution all deliver approximately the same result. In the middle of the channel, carriers have gained energy and the electron temperature is high. Thus, the cold Maxwellian underestimates the high energy tail, while the heated Maxwellian overestimates the amount of hot carriers. Near the drain side, the non-Maxwellian distribution exactly reproduces the emerging population of cold electrons, while neither the heated nor the cold Maxwellian can reproduce the Monte Carlo results.

III. RESULTS

For the evaluation of the tunneling model we apply our non-Maxwellian distribution function to the simulation of MOS transistors with varying gate lengths and oxide thick-

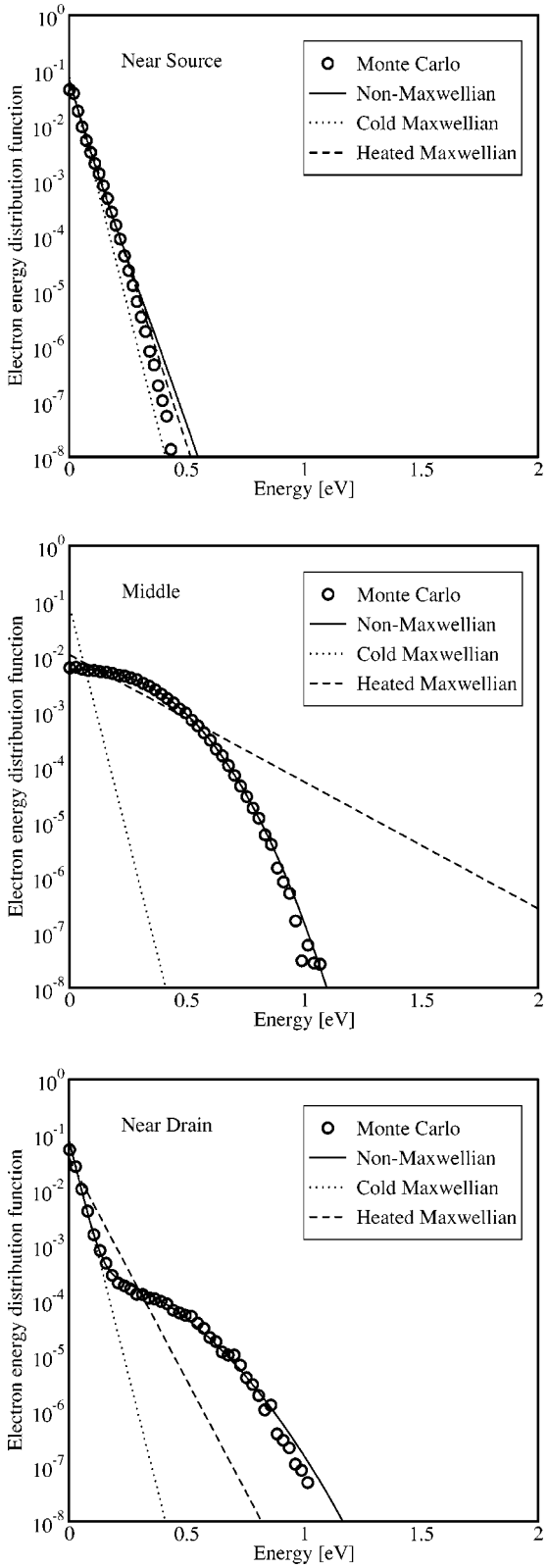

FIG. 6. Comparison of the non-Maxwellian, the heated Maxwellian, and the cold Maxwellian EED with Monte Carlo results along the channel of a MOSFET.

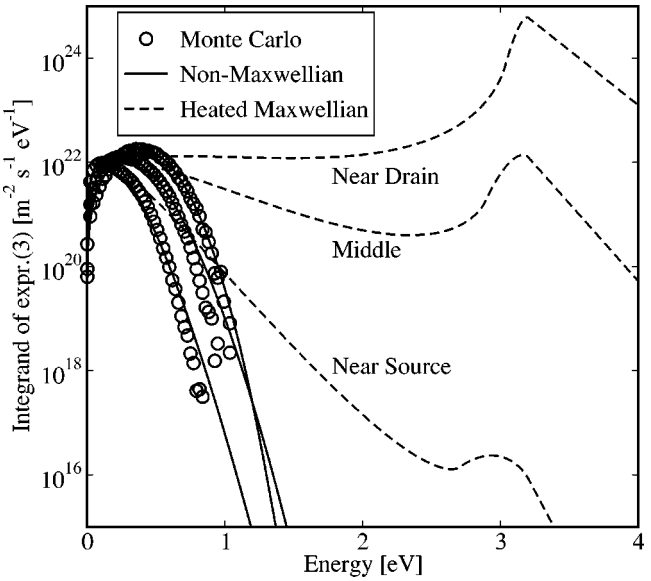

FIG. 7. Integrand of expression (3) for a Maxwellian and the non-Maxwellian distribution at different points in the channel. The electron temperatures are 976, 1585, and 2119 K near the source, in the middle, and near the drain side of the channel.

nesses. We simulated nMOS devices in on-state with $V_{GS}=1$ V and $V_{DS}=1$ V. Gate lengths of 350, 250, 180, and 100 nm with gate oxide thicknesses of 3.4, 3.0, 2.6, 2.2, 1.8, 1.4, and 1.0 nm have been assumed. Gaussian source and drain doping peaks of 10^{20} cm^{-3} with LDD extensions were used. In the following figures the results from Monte Carlo simulations will serve as reference.

Figure 7 shows the integrand of expression (3) as a function of the electron energy for the case of a heated Maxwellian distribution and the non-Maxwellian distribution function [expressions (14) and (18)]. The simulated device has a gate length of 100 nm and a gate oxide thickness of 3 nm. While at low energies the difference between the non-Maxwellian distribution function and the heated Maxwellian distribution seems to be negligible, the amount of overestimation of the incremental gate current density for the heated Maxwellian distribution reaches several orders of magnitude at 1 eV and peaks when the electron energy exceeds the barrier height. This spurious effect is clearly more pronounced for points at the drain end of the channel where the electron temperature is high.

The peak in the integrand for the heated Maxwellian approximation results in an increased gate current density near the drain side of the gate contact. Figure 8 shows the gate current density along the channel of a 180 nm gate length device for different gate oxide thicknesses simulated with the heated Maxwellian approximation. Near the drain side, the high electron temperature leads to a pronounced overestimation of the high-energy tail as seen in Fig. 7, which in turn leads to the peak in the gate current density. With lower oxide thickness the electric field in the oxide increases and leads to increased transmission coefficients and higher gate current densities. Hence, the spurious peak is more pronounced for thicker oxides.

In Fig. 9 the gate current density is depicted for the 100 and 180 nm device for an oxide thickness of 2.6 nm. While the non-Maxwellian distribution correctly reproduces the Monte Carlo results, the cold Maxwellian distribution leads to a sound underestimation reaching one order of magnitude,

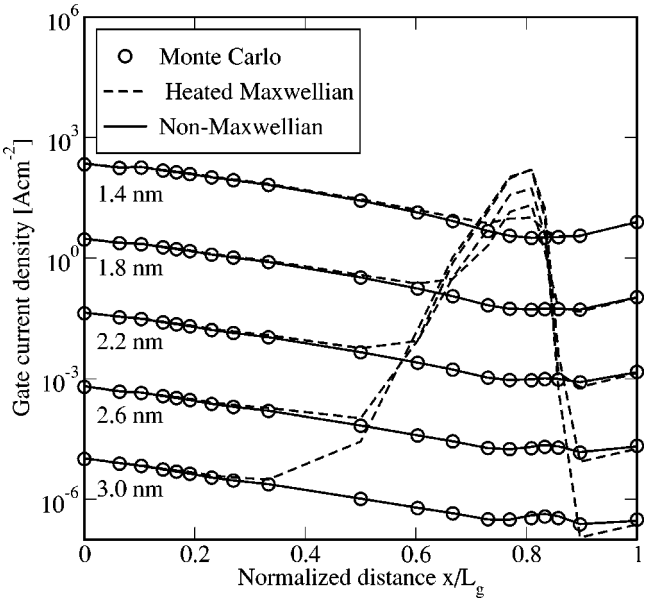

FIG. 8. Gate current density along the channel for a heated Maxwellian distribution, the non-Maxwellian distribution and Monte Carlo results for varying gate oxide thicknesses. The gate length L_g is 180 nm.

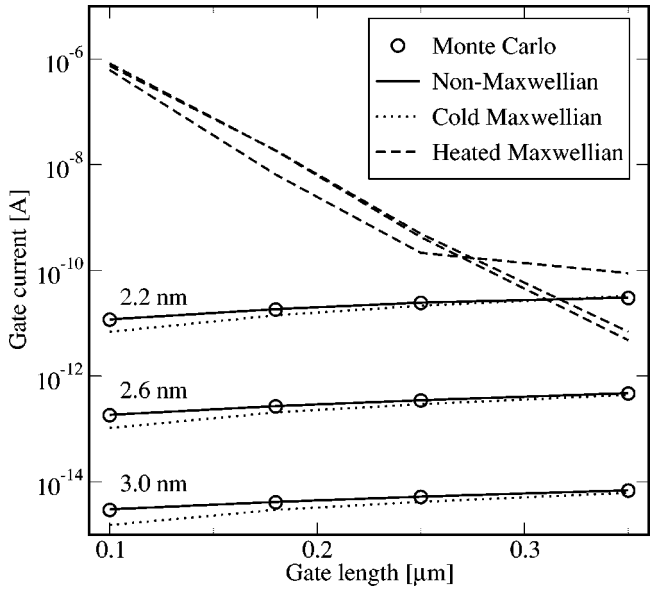

FIG. 10. Gate current for different gate oxide thicknesses as a function of the gate length L_g for a cold Maxwellian, heated Maxwellian, and the non-Maxwellian distribution function, compared to Monte Carlo results.

and the heated Maxwellian distribution predicts a much too high gate current density near the drain side of the channel. This effect is even more pronounced for smaller gate lengths. Note also that the current density predicted by Monte Carlo simulations shows only a small increase for reduced gate lengths.

The effect on the total gate current of the devices is shown in Fig. 10. In this figure, the gate current is given as a function of the gate length for different gate oxide thicknesses. It can be seen that the heated Maxwellian distribution delivers correct results only for large gate lengths and small oxide thicknesses, while it totally fails for smaller devices.

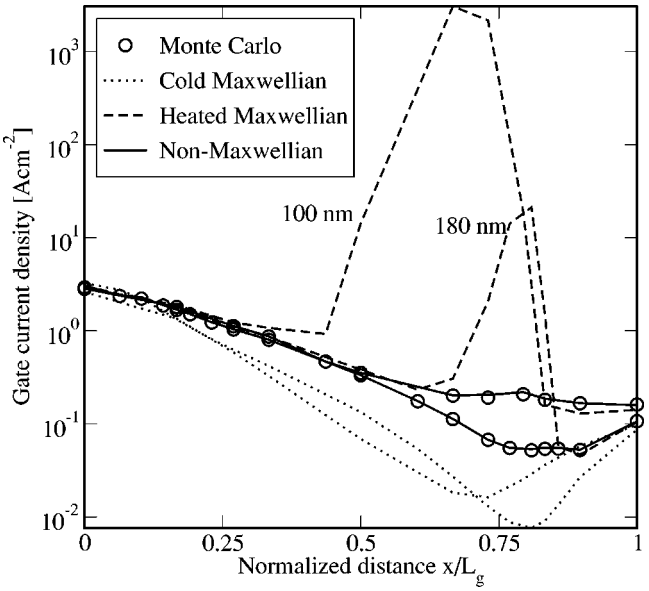

FIG. 9. Spatial distribution of the gate current density along the channel for different gate lengths L_g for a cold Maxwellian, a heated Maxwellian, and the non-Maxwellian distribution function, compared to Monte Carlo results.

Note that for a gate length of 100 nm, the gate current within the heated Maxwellian approximation depends very weakly on the oxide thickness which is at least questionable.

The use of a cold Maxwellian distribution, on the other hand, underestimates the gate current only slightly and seems to be the better choice if accurate modeling of the device physics is not that important or only a quick estimation is asked for. The non-Maxwellian model correctly reproduces the Monte Carlo results for all gate lengths and gate oxide thicknesses. It is thus well suited to analyze effects which are closely related to the shape of the distribution function.

A. Comparison with measurement data

The simulation results have been compared to data reported in recent publications. For a MOSFET with zero drain-source voltage, the cold Maxwellian, heated Maxwellian, and non-Maxwellian model deliver of course the same results which are shown in Fig. 11. The measurement values were taken from Ref. 20 (also published in Ref. 21). The electron mass in the oxide was used as fitting parameter and an excellent fit for all oxide thicknesses was achieved with $m_{ox}=0.65\,m_0$. Note that the result is independent of the gate length since no drain-source bias was applied.

For the case of hot carriers, however, the heated Maxwellian distribution fails to reproduce even the qualitative behavior observed in measurements. Figure 12 shows the gate current density as a function of the drain-source voltage for a 350 nm gate length, 1.8 nm oxide thickness MOSFET.[22] Perhaps due to differences in gate oxide thickness determination and the measurement setup, the gate oxide thickness had to be increased to give the same values for zero drain voltage as in Fig. 11. For increasing drain voltage, the electric field in the oxide is reduced which leads to a reduced transmission coefficient and lower gate current. The heated Maxwellian distribution overestimates the gate current den-

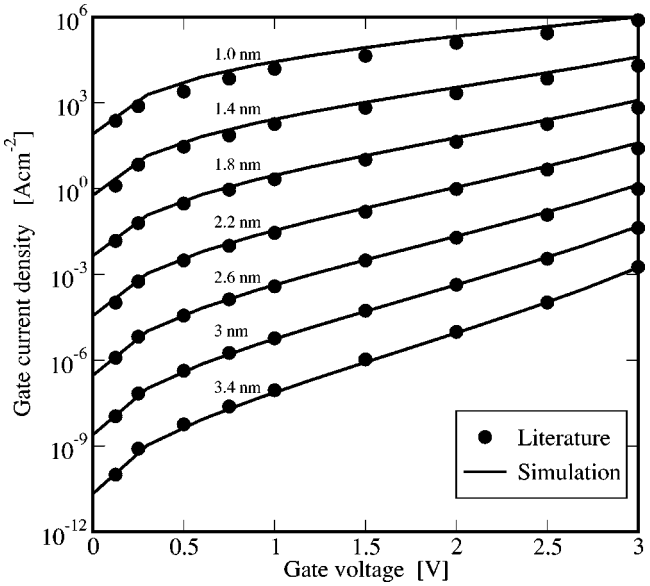

FIG. 11. Comparison of the gate current density for different gate oxide thicknesses with data reported in Ref. 21.

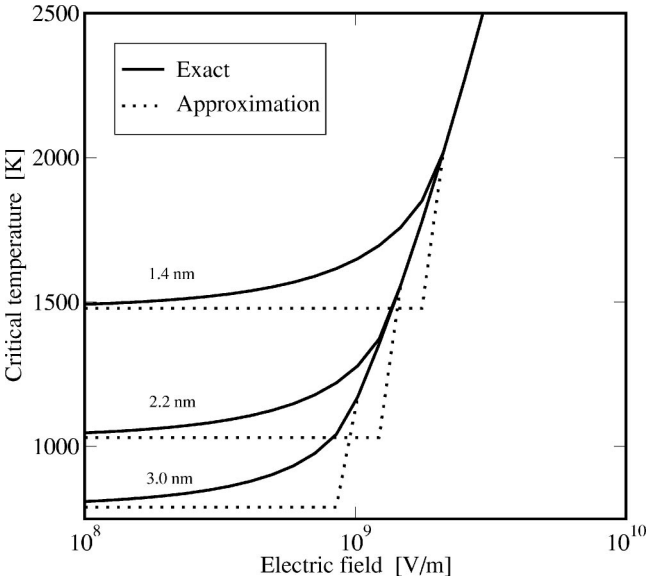

FIG. 13. Critical temperature T_{crit} as a function of the electric field in the oxide for different gate oxide thicknesses compared to the approximation Eq. (24).

sity especially for high bias. While this effect is not so strong for the 350 nm device, it is clearly visible for a 180 nm device and, due to the increase of the electric field, it will even be higher for increased drain voltages. The non-Maxwellian model, however, correctly reproduces the measurements and shows reasonable results for the 180 nm device.

IV. HEATED MAXWELLIAN DISTRIBUTION TEMPERATURE LIMIT

It was shown that the heated Maxwellian approximation delivers incorrect results if it is used for the gate current estimation of submicron devices at low drain bias because it overestimates the high-energy tail. To quantify this error we look at the ratio of the two maxima of the integrand of expression (3) in Fig. 7. It is clear that the second maximum which is located at $\mathcal{E}_2=\Phi$, is spurious and only appears for the heated Maxwellian distribution due to the overestimation of the high-energy tail. We assume the first maximum, which does not always appear at the same energy level, at $\mathcal{E}_1=\Phi/r$ and evaluate the ratio R of the integrands at the two maxima. We then introduce the critical temperature as the electron temperature where this ratio exceeds a certain value. A short calculation (see Appendix B) yields the following expression for the critical temperature:

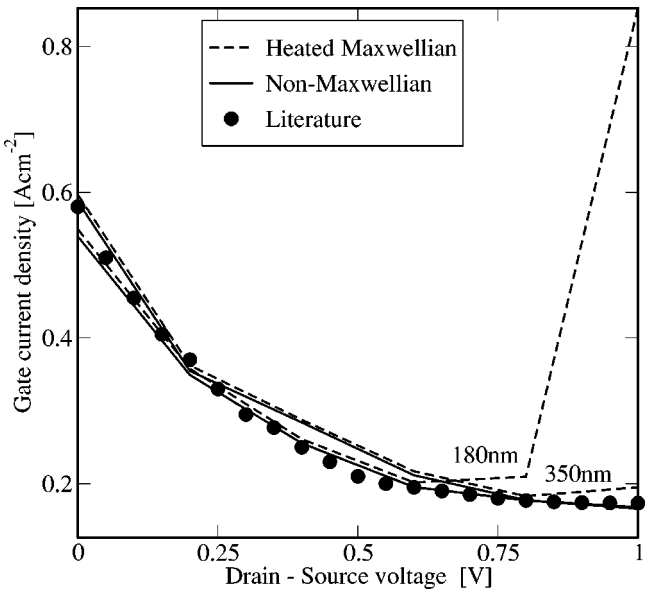

FIG. 12. Gate current density as a function of the drain bias compared to measurement data reported in Ref. 22. The gate oxide thickness is 1.8 nm.

$$T_{\mathrm{crit}} \approx \begin{cases} T_0 & \text{for} \quad F_{\mathrm{ox}} < \dfrac{2(\Phi-\mathcal{E}_1)}{3qt_{\mathrm{ox}}} \\ \dfrac{\overline{T}_0}{1+\dfrac{F_{\mathrm{c}}}{F_{\mathrm{ox}}}} & \text{for} \quad F_{\mathrm{ox}} > \dfrac{2(\Phi-\mathcal{E}_1)}{3qt_{\mathrm{ox}}}, \end{cases} \tag{24}$$

with the values of T_0, $\overline{T}_0$, and F_{c} being

$$T_0 = \frac{1}{k_{\mathrm{B}}} \frac{\Phi-\mathcal{E}_1}{\ln R - \ln C + \dfrac{2\sqrt{2m_{\mathrm{ox}}}}{\hbar}(\Phi-\mathcal{E}_1)^{1/2}t_{\mathrm{ox}}}, \tag{25}$$

$$\overline{T}_0 = T_0 \cdot \left(1 + \frac{2\sqrt{2m_{\mathrm{ox}}(\Phi-\mathcal{E}_1)}t_{\mathrm{ox}}}{\hbar(\ln R - \ln C)}\right), \tag{26}$$

$$F_{\mathrm{c}} = \frac{4\sqrt{2m_{\mathrm{ox}}}}{3\hbar q} \frac{(\Phi-\mathcal{E}_1)^{3/2}}{\ln R - \ln C}, \tag{27}$$

and C defined in the appendix. The critical temperature as a function of the oxide electric field is shown in Fig. 13 for values of $r=10$ and $R=10$. It can be seen that it increases for decreasing gate oxide thickness. Higher oxide fields (for example, by increasing the gate voltage) can increase the

critical temperature. However, for common device technologies, it is obvious that electron temperatures will exceed the critical temperature in most of the channel region. For devices in which the critical temperature is not exceeded, the heated Maxwellian assumption can safely be used without jeopardizing accuracy. For the non-Maxwellian distribution the definition of a critical temperature does not make any sense since the second maximum at $\mathcal{E}_2$ always stays well below the first maximum. This model will thus not lead to an overestimation of gate current even for high electron temperatures in the channel.

V. CONCLUSION

We presented a new model for the hot-electron gate tunneling current by taking the non-Maxwellian shape of the electron energy distribution function into account. We showed that the Maxwellian and heated Maxwellian assumptions for the distribution function deliver correct results for the case of cold carrier tunneling, but they fail to reproduce hot carrier tunneling where the heated Maxwellian assumption heavily overestimates the gate current density. We used a recently developed non-Maxwellian expression for the distribution function based on a six-moments transport model. Using the new expression we could accurately reproduce Monte Carlo results and measurement data of a turned-on MOSFET. We further introduced a critical electron temperature up to which the error due to the overestimation of the high-energy tail of the heated Maxwellian distribution at low drain bias is negligible. We derived a simple expression for the critical temperature and found a value of ~1000 K for the case of a 2.2 nm oxide thickness device. If the electron temperature in the channel exceeds this value, only the non-Maxwellian model is able to reproduce the device physics correctly.

APPENDIX A: DERIVATION OF THE PERPENDICULAR VELOCITY

Since we are only interested in the velocity component perpendicular to the interface $v_\perp$

$$v_\perp = \frac{1}{\hbar}\cdot\frac{\partial \mathcal{E}}{\partial k_\perp} = \frac{1}{\hbar}\cdot\frac{\partial k}{\partial k_\perp}\cdot\frac{\partial \mathcal{E}}{\partial k}, \tag{A1}$$

and the length of $\mathbf{k}$ is

$$k = \sqrt{k_\perp^2 + k_y^2 + k_z^2}, \tag{A2}$$

we can write

$$v_\perp = \frac{1}{\hbar}\cdot\frac{k_\perp}{k}\cdot\frac{\partial \mathcal{E}}{\partial k}. \tag{A3}$$

If only the upper half-space is taken into account for the $\mathbf{k}$ vector, we get the perpendicular component of the wave vector normalized to the sphere with $k=1$:

$$\langle k_\perp \rangle = \frac{1}{4\pi}\int_0^{2\pi} d\phi \int_0^{\pi/2} k\,\cos\,\vartheta\, d(\cos\,\vartheta) \tag{A4}$$

and finally arrive at

$$v_\perp = \frac{1}{4\hbar}\cdot\frac{\partial \mathcal{E}}{\partial k}. \tag{A5}$$

APPENDIX B: DERIVATION OF THE CRITICAL TEMPERATURE

We define two energy levels $\mathcal{E}_1$ and $\mathcal{E}_2$ with $\mathcal{E}_2$ being 3.2 eV (equal to the barrier height) and $\mathcal{E}_1 = \mathcal{E}_2/r$. The integrand of expression (3) is

$$I(\mathcal{E}) = f(\mathcal{E})\cdot g(\mathcal{E})\cdot T(\mathcal{E})\cdot v_\perp(\mathcal{E}). \tag{B1}$$

We define the ratio of the two integrand maxima as

$$R = \frac{I(\mathcal{E}_1)}{I(\mathcal{E}_2)}. \tag{B2}$$

For a heated Maxwellian distribution function and Kane's dispersion relation, the density of states, the distribution function and the perpendicular velocity at the energy levels $\mathcal{E}_1$ and $\mathcal{E}_2$ become

$$g(\mathcal{E}_i) = g_0\cdot\sqrt{\mathcal{E}_i}\cdot\sqrt{1+\alpha\mathcal{E}_i}\cdot(1+2\alpha\mathcal{E}_i), \tag{B3}$$

$$f(\mathcal{E}_i) = A\,\exp\left(-\frac{\mathcal{E}_i}{k_B T_n}\right), \tag{B4}$$

$$v_\perp(\mathcal{E}_i) = \sqrt{\frac{2\mathcal{E}_i(1+\alpha\mathcal{E}_i)}{16 m_{ox}(1+2\alpha\mathcal{E}_i)^2}}. \tag{B5}$$

The transmission coefficient at $\mathcal{E}=\mathcal{E}_1$ is

$$T(\mathcal{E}_1) = \begin{cases} \exp\left[-4\dfrac{\sqrt{2m_{ox}}}{3\hbar q F_{ox}}\cdot(\Phi-\mathcal{E}_1)^{3/2}\right] \\ \qquad \text{for} \quad \Phi_0 < \mathcal{E}_1 < \Phi \\ \exp\left\{-4\dfrac{\sqrt{2m_{ox}}}{3\hbar q F_{ox}}\cdot\left[(\Phi-\mathcal{E}_1)^{3/2}-(\Phi_0-\mathcal{E}_1)^{3/2}\right]\right\}, \\ \qquad \text{for} \quad \mathcal{E}_1 < \Phi_0 \end{cases} \tag{B6}$$

while $T(\mathcal{E}_2) = T(\Phi) = 1$. With the abbreviation

$$C = \frac{\mathcal{E}_1}{\mathcal{E}_2}\cdot\frac{1+\alpha\mathcal{E}_1}{1+\alpha\mathcal{E}_2} \tag{B7}$$

we get

$$R = C\cdot T(\mathcal{E}_1)\exp\left(\frac{\mathcal{E}_2-\mathcal{E}_1}{k_B T_{crit}}\right) \tag{B8}$$

and therefore

$$T_{crit} = \frac{1}{k_B}\cdot\frac{\mathcal{E}_2-\mathcal{E}_1}{\ln R - \ln C - \ln T(\mathcal{E}_1)}. \tag{B9}$$

Using a Taylor-series expansion of $T(\mathcal{E}_1)$ around $\mathcal{E}_1$ for the region $\mathcal{E}_1 < \Phi_0$

$$(\Phi-\mathcal{E}_1)^{3/2} - (\Phi_0-\mathcal{E}_1)^{3/2} \approx \frac{3}{2}(\Phi-\mathcal{E}_1)^{3/2}\cdot\frac{qF_{ox}t_{ox}}{\Phi-\mathcal{E}_1}, \tag{B10}$$

and with Φ_0 taken from expression (6), we get

$$T_{\mathrm{crit}}=\frac{\mathcal{E}_2-\mathcal{E}_1}{k_{\mathrm{B}}}\cdot\begin{cases}\left[\ln R-\ln C+\dfrac{4\sqrt{2m_{\mathrm{ox}}}}{3\hbar qF_{\mathrm{ox}}}\cdot(\Phi-\mathcal{E}_1)^{3/2}\right]^{-1} & \text{for } \Phi_0<\mathcal{E}_1<\Phi\\ \left[\ln R-\ln C+\dfrac{2\sqrt{2m_{\mathrm{ox}}}}{\hbar}\cdot(\Phi-\mathcal{E}_1)^{1/2}\cdot t_{\mathrm{ox}}\right]^{-1} & \text{for } \mathcal{E}_1<\Phi_0\end{cases} \tag{B11}$$

which can easily be rewritten to expression (24). The transition between the Fowler-Nordheim and the direct tunneling region can be formulated as a condition for the electric field in the oxide, namely

$$\begin{aligned}F_{\mathrm{ox}}&<\frac{2(\Phi-\mathcal{E}_1)}{3qt_{\mathrm{ox}}} \quad \text{for } \mathcal{E}_1<\Phi_0\\ F_{\mathrm{ox}}&>\frac{2(\Phi-\mathcal{E}_1)}{3qt_{\mathrm{ox}}} \quad \text{for } \Phi_0<\mathcal{E}_1<\Phi.\end{aligned} \tag{B12}$$

[1] S. Keeney *et al.*, IEEE Trans. Electron Devices **39**, 2750 (1992).
[2] K. F. Schuegraf and C. Hu, IEEE Trans. Electron Devices **41**, 761 (1994).
[3] K. Hasnat *et al.*, IEEE Trans. Electron Devices **44**, 129 (1997).
[4] J. Cai and C.-T. Sah, J. Appl. Phys. **89**, 2272 (2001).
[5] R. Tsu and L. Esaki, Appl. Phys. Lett. **22**, 562 (1973).
[6] K. Gundlach, Solid-State Electron. **9**, 949 (1966).
[7] A. Gehring, T. Grasser, and S. Selberherr, in *International Semiconductor Device Research Symposium* (Washington, DC, 2001), pp. 260–263.
[8] T. Grasser, H. Kosina, C. Heitzinger, and S. Selberherr, J. Appl. Phys. **91**, 3869 (2002).
[9] C. Fiegna *et al.*, IEEE Trans. Electron Devices **38**, 603 (1991).
[10] R. Shankar, *Principles of Quantum Mechanics* (Plenum, New York, 1994).
[11] E. O. Kane, J. Phys. Chem. Solids **1**, 249 (1957).
[12] W. Franz, in *Handbuch der Physik* (Springer, Berlin, 1956), Vol. XVII, p. 155.
[13] A. Abramo and C. Fiegna, J. Appl. Phys. **80**, 889 (1996).
[14] D. Cassi and B. Ricco, IEEE Trans. Electron Devices **37**, 1514 (1990).
[15] K.-I. Sonoda, M. Yamaji, K. Taniguchi, and C. Hamaguchi, J. Appl. Phys. **80**, 5444 (1996).
[16] T. Grasser, H. Kosina, and S. Selberherr, J. Appl. Phys. **90**, 6165 (2001).
[17] T. Grasser, H. Kosina, M. Gritsch, and S. Selberherr, J. Appl. Phys. **90**, 2389 (2001).
[18] A. Gehring, T. Grasser, and S. Selberherr, in *Modeling and Simulation of Microsystems*, San Juan, PR, 2002 (Computational Publications, Cambridge, MA, 2002), pp. 560–563.
[19] P. Cappelleti, *Flash Memories* (Kluwer Academic, Dordrecht, 2000).
[20] J. P. Shiely, Ph.D. dissertation, Duke University, 1999.
[21] H. Z. Massoud, J. P. Shiely, and A. Shanware, in *Materials Research Society Symposium Proceedings on Ultrathin SiO_2 and High-K Materials for ULSI Gate Dielectrics*, San Francisco, CA, 1999 (Materials Research Society, Warrendale, PA, 1999), pp. 227–239.
[22] S. Schwantes and W. Krautschneider, in *European Solid-State Device Research Conference*, Nuremberg, Germany, 2001 (Editions Frontieres, Paris, France, 2001), pp. 471–474.