*Journal of*
*Computational and Theoretical Nanoscience*
Vol. 2, 26–44, 2005

REVIEW

# Gate Current Modeling for MOSFETs

Andreas Gehring* and Siegfried Selberherr

*Institute for Microelectronics, TU Vienna, Gusshausstrasse 27–29, A-1040 Vienna, Austria*

We describe a set of models suitable for the two- and three-dimensional simulation of tunneling in logic and non-volatile MOS devices. The crucial modeling topics are comprehensively discussed. This comprises the modeling of the energy distribution function in the channel to account for hot-carrier tunneling, the calculation of the transmission coefficient of single and layered dielectrics, the influence of quasi-bound states in the inversion layer, the modeling of static and transient defect-assisted tunneling, and the modeling of dielectric degradation and breakdown. We propose a set of models to link the gate leakage to the creation of traps in the dielectric layer, the threshold voltage shift, and eventual dielectric breakdown. The simulation results are compared to commonly used compact models and measurements of logic and non-volatile memory devices.

**Keywords:** Semiconductor Device Simulation, MOS Tunneling, Energy Distribution Function, High-k Dielectrics, Transmitting-Boundary, Transfer-Matrix, Trap-Assisted Tunneling.

## CONTENTS



## 1. INTRODUCTION

For the prediction of the performance and for the optimization of MOS devices the accurate simulation of quantum-mechanical tunneling effects has always been of paramount interest.[1] The application area of such models ranges from the prediction of gate leakage in MOS transistors, the evaluation of gate stacks for advanced high-$k$ gate insulator materials, the optimization of programming and erasing times in non-volatile semiconductor memory cells up to the study of source-drain tunneling.

As shown in the silicon-dielectric-silicon structure sketched in Figure 1 a variety of tunneling processes can be identified.[2] Considering simply the shape of the energy barrier, Fowler-Nordheim (FN) tunneling and direct tunneling can be distinguished. However, a more rigorous classification distinguishes between ECB (electrons from the conduction band), EVB (electrons from the valence band), HVB (holes from the valence band), TAT (trap-assisted tunneling) processes, and QBS (quasi-bound state) tunneling processes. We denote direct tunneling all processes which are not defect-assisted. In the figure the electron (EED) and hole (HED) energy distribution functions are also indicated.

However, tunneling model implementations in state-of-the-art device simulators often rely on simplified models assuming Fermi-Dirac statistics and triangular energy

*Author to whom correspondence should be addressed.

 1546-198X/2005/2/026/019/$17.00+.25 doi:10.1166/jctn.2005.002

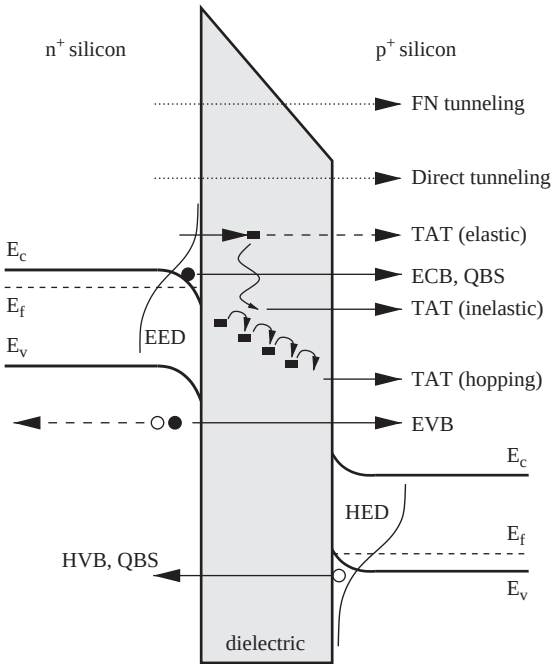


**Fig. 1.** Tunneling processes in a MOS structure. Direct tunneling processes (ECB, EVB, and HVB) are covered in Section 2, while Section 3 deals with TAT transitions. Bound and quasi-bound states are studied in Section 2.4.

barriers. In contemporary miniaturized devices these assumptions are violated in several 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. The second weakness lies in the estimation of the transmission coefficient. For this task the Wentzel-Kramers-Brillouin (WKB) or the Gundlach method is frequently used. These models, however, fail for energy barriers which are not of triangular or trapezoidal shape. To accurately describe tunneling in such cases, Schrödinger's equation must be solved. This is often achieved using the transfer-matrix method[3] or the quantum-transmitting boundary method.[4] Finally, a strong inaccuracy arises when tunneling current from the channel of inverted MOSFETs is calculated. In this case bound and quasi-bound states are formed, the latter giving rise to quasi-bound state tunneling. The use of the Tsu-Esaki formula which assumes a continuum of states, is questionable in this case.

The reduction of gate dielectric thicknesses makes the use of alternative gate dielectrics such as $ZrO_2$ necessary. These dielectrics often suffer from high defect densities,[5] which invalidates the application of tunneling models which assume coherent ballistic transport. Current transport by means of defect-assisted tunneling has been studied intensely.[5–7] In addition to the current, the gate dielectric reliability becomes a crucial issue not only for non-volatile memories but also for logic applications. In fact, the processes of leakage, trap creation, and dielectric breakdown are physically directly related. Thus, we recommend a set of models which directly link the simulation of direct and trap-assisted leakage current with the creation and occupation of traps and the occurrence of breakdown.

The paper is structured as follows. In Section 2 the theory of direct tunneling mechanisms with emphasis on modeling of the distribution function and the transmission coefficient is described. The calculation of tunneling in the presence of bound and quasi-bound states as encountered

**Andreas Gehring** was born in Mistelbach, Austria, in 1975. He studied communication engineering at the Technische Universität Wien where he received the Diplomingenieur and Ph.D. degrees in 2000 and 2003, respectively. He joined the Institute for Microelectronics in April 2000 and held visiting research positions at the Samsung Advanced Institute of Technology in Seoul, South Korea, in summer 2001, and at Cypress Semiconductor in San Jose, USA, in summer 2003. His scientific interests are focused on the modeling of quantum effects for semiconductor device simulation.

**Siegfried Selberherr** was born in Klosterneuburg, Austria, in 1955. He received the degree of Diplomingenieur in electrical engineering and the doctoral degree in technical sciences from the Technische Universität Wien in 1978 and 1981, respectively. Dr. Selberherr has been holding the venia docendi on Computer-Aided Design since 1984. Since 1988 he has been the head of the Institute for Microelectronics and since 1999 he has been dean of the Faculty of Electrical Engineering and Information Technology. His current topics are modeling and simulation of problems for microelectronics engineering.

REVIEW

in the inversion layer of a MOSFET is outlined. Typical results for MOS transistors are presented and compared with common compact models. We present an example non-volatile memory application utilizing a layered dielectric to allow independent tuning of on- and off-state currents. Section 3 presents a set of models which can be used to describe defect-assisted tunneling. We give a short overview of commonly used degradation models and show how to link the various tunneling models with the creation of defects, threshold voltage shift, and dielectric breakdown. A conclusion and model comparison section wraps up the main findings and gives directions for further research.

## 2. DIRECT TUNNELING

The most prominent and almost exclusively used expression to describe direct tunneling transitions has been developed by Duke[8] and used by Tsu and Esaki to describe tunneling through a one-dimensional superlattice.[3] It is commonly known as Tsu-Esaki expression. The current density reads

$$J = \frac{4\pi m_{3D} q}{h^3} \int_{\mathscr{E}_{min}}^{\mathscr{E}_{max}} TC(\mathscr{E}_x, m_{diel}) N(\mathscr{E}_x)\, d\mathscr{E}_x, \qquad (1)$$

with a transmission coefficient $TC(\mathscr{E}_x)$ and a supply function $N(\mathscr{E}_x)$ which is defined as

$$N(\mathscr{E}_x) = \int_0^\infty (f_1(\mathscr{E}) - f_2(\mathscr{E}))\, d\mathscr{E}_\rho. \qquad (2)$$

The total energy $\mathscr{E}$ is the sum of a transversal component parallel to the Si-$SiO_2$ interface $\mathscr{E}_\rho$ and a transversal component $\mathscr{E}_x$. The electron energy distribution functions in the gate and substrate are denoted by $f_1$ and $f_2$, respectively.

Two electron masses enter (1): the density-of-states mass in the plane parallel to the interface $m_{3D} = 2\,m_t + 4\sqrt{m_t m_1}$, which, for (100) silicon with $m_1 = 0.92\,m_0$ and $m_t = 0.19\,m_0$ equals $2.052\,m_0$, and the electron mass in the dielectric $m_{diel}$.[9]

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 distribution and the assumption of an isotropic distribution, the supply function evaluates to

$$N(\mathscr{E}_x) = k_B T \ln\left(\frac{1+\exp\left(-\frac{\mathscr{E}_x - \mathscr{E}_{F,1}}{k_B T}\right)}{1+\exp\left(-\frac{\mathscr{E}_x - \mathscr{E}_{F,2}}{k_B T}\right)}\right). \qquad (3)$$

where $\mathscr{E}_{F,1}$ and $\mathscr{E}_{F,2}$ denote the Fermi energies at the semiconductor-oxide interfaces. Note, however, that the assumption of an isotropic distribution may not be justified for short-channel devices.[10] Furthermore, the assumption of a Fermi-Dirac distribution is poor in the channel of a turned-on submicron MOSFET. Advanced models for the distribution function are necessary.

### 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, e.g.[11,12] The topic is of high importance, because the assumption of a 'cold' Maxwellian distribution function

$$f(\mathscr{E}) = A \cdot \exp\left(-\frac{\mathscr{E}}{k_B \cdot T_L}\right) \qquad (4)$$

underestimates the high-energy tail of the EED near the drain region.[13] The straightforward approach is to use a heated Maxwellian distribution function based on the electron temperature $T_n$. We applied a Monte Carlo simulator employing analytical non-parabolic bands to check the validity of this approximation. Figure 2 shows the contour lines of the heated Maxwellian EED in comparison to Monte Carlo results for a MOSFET with a gate length of $L_g = 180$ nm at $V_{DS} = V_{GS} = 1$ V. The electron temperature was calculated in a post-processing step as $T_n = 2\langle\mathscr{E}\rangle/3k_B$.

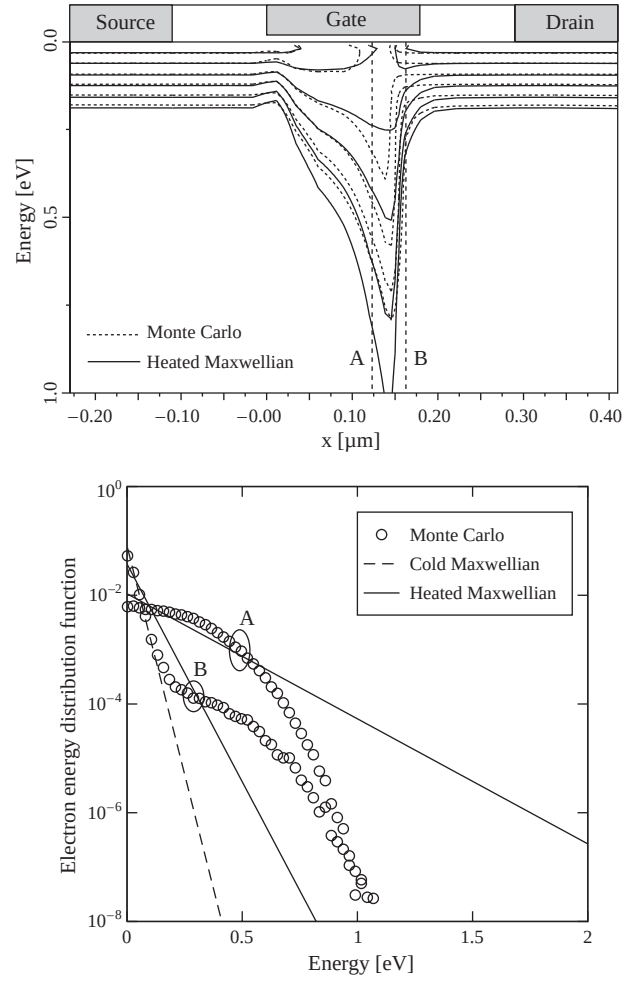


**Fig. 2.** Comparison of the heated Maxwellian distribution (full lines) with the results from a Monte Carlo simulation (dotted lines) in a turned-on 180 nm MOSFET. Neighboring lines differ by a factor of 10. The distributions at point A and B are compared with a cold Maxwellian in the lower figure.

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. Note that for increasing gate bias, namely $V_{GS} > V_{DS}$, the peak electric field in the channel is reduced, and the heated Maxwellian approximation delivers more reasonable results.[14]

A quite generalized approach for the EED has been proposed by Grasser et al.[15]

$$f(\mathscr{E}) = A \exp\left(-\left(\frac{\mathscr{E}}{\mathscr{E}_{\text{ref}}}\right)^b\right). \tag{5}$$

In this expression the values of $\mathscr{E}_{\text{ref}}$ and $b$ are mapped to the solution variables $T_n$ and $\beta_n$ of a six moments transport model.[16] The symbol $\beta_n$ denotes the normalized kurtosis of the distribution function ($\beta_n = 1$ for a Maxwellian distribution).

Expression (5) has been shown to appropriately reproduce Monte Carlo results in the source and the middle region of the channel of a turned-on MOSFET. However, this model is still not able to reproduce the high energy tail of the distribution function near the drain side of the channel. This is because it was shown that near the drain, the electron population consists of a mixture of hot electrons coming from the drain and a pool of cold carriers from the source.[13,17] Expression (5) does not explicitly account for this cold-carrier population. Therefore, when (5) is normalized to the actual carrier concentration, the high-energy tail is heavily overestimated.[18–20]

A distribution function accounting for this effect was proposed by Sonoda et al.,[12] and an improved model has been suggested by Grasser et al.:[13]

$$f(\mathscr{E}) = A\left(\exp\left(-\left(\frac{\mathscr{E}}{\mathscr{E}_{\text{ref}}}\right)^b\right) + c\exp\left(-\frac{\mathscr{E}}{k_B T_L}\right)\right). \tag{6}$$

Here the pool of cold carriers in the drain region is correctly modeled by an additional cold Maxwellian subpopulation. The values of $\mathscr{E}_{\text{ref}}$, $b$, and $c$ are again derived from the solution variables of a six moments transport model. Figure 3 shows again the results from Monte Carlo simulations in comparison to the analytical model. A good match between this non-Maxwellian distribution and the Monte Carlo results can be seen. The supply functions utilizing (5) and (6) are given in Appendix A. Note that the prefactor $A$ must be calculated from a normalization to the carrier concentration in the channel as shown in Appendix B. To check the impact of the distribution function, the integrand of the Tsu-Esaki formula, namely the expression $TC(\mathscr{E})N(\mathscr{E})$, has been evaluated as shown in Figure 4, and compared to post-processed Monte Carlo results. While at low energies the difference between the non-Maxwellian distribution function (6) and the heated Maxwellian distribution is negligible, the incremental gate current density is

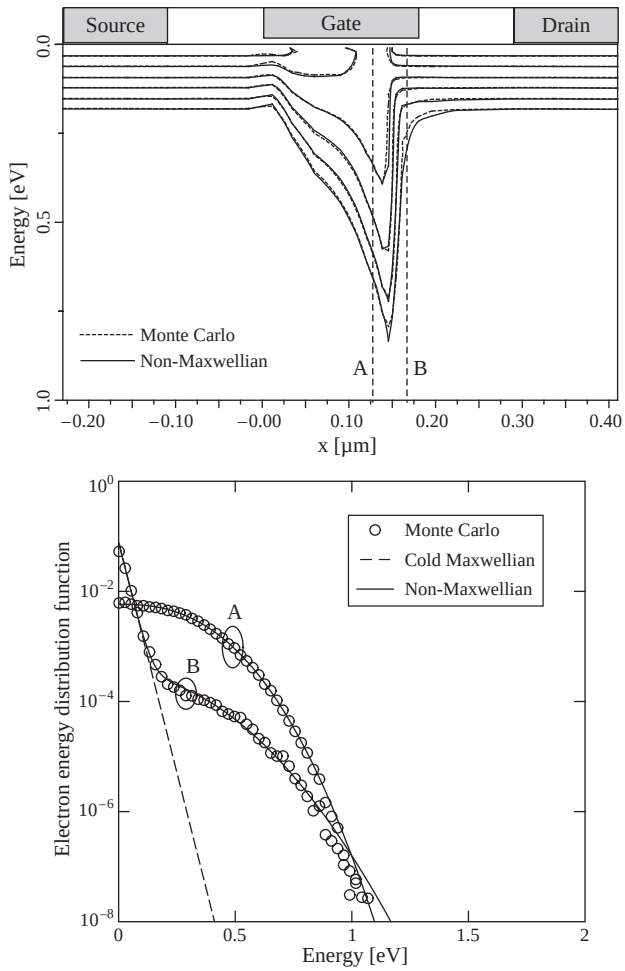


**Fig. 3.** Comparison of the non-Maxwellian distribution (full lines) with the results from a Monte Carlo simulation (dotted lines) in a turned-on 180 nm MOSFET. Neighboring lines differ by a factor of 10. The distributions at point A and B are compared with a cold Maxwellian in the lower figure.

heavily overestimated by the heated Maxwellian distribution 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 non-Maxwellian shape of the distribution function, indicated by the full line, reproduces the Monte Carlo results very well.

### 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

$$j = \frac{\hbar}{2\imath m} \cdot (\Psi^* \cdot \nabla\Psi - \nabla\Psi^* \cdot \Psi) \tag{7}$$

where $\Psi$ is the wave function, $m$ the carrier effective mass, and $\imath = \sqrt{-1}$. The transmission coefficient is defined as the ratio of the fluxes due to an incident and a reflected wave. These wave functions can

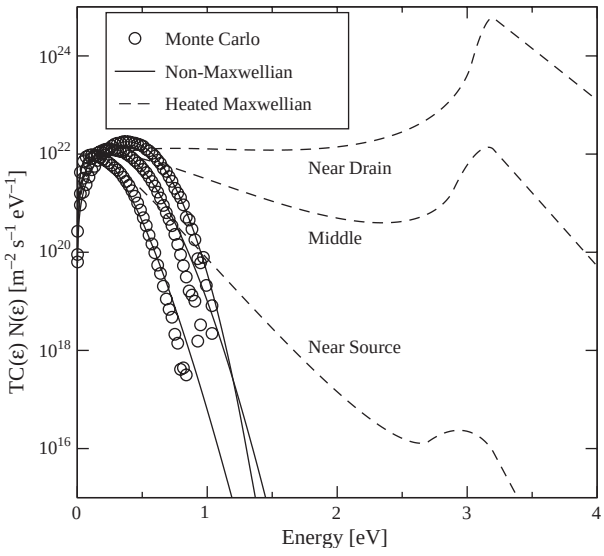


**Fig. 4.** Integrand of Tsu-Esaki's equation for a MOSFET with 100 nm gate length and 3 nm gate dielectric thickness at $V_{GS} = V_{DS} = 1$ V applying different models for the distribution function.

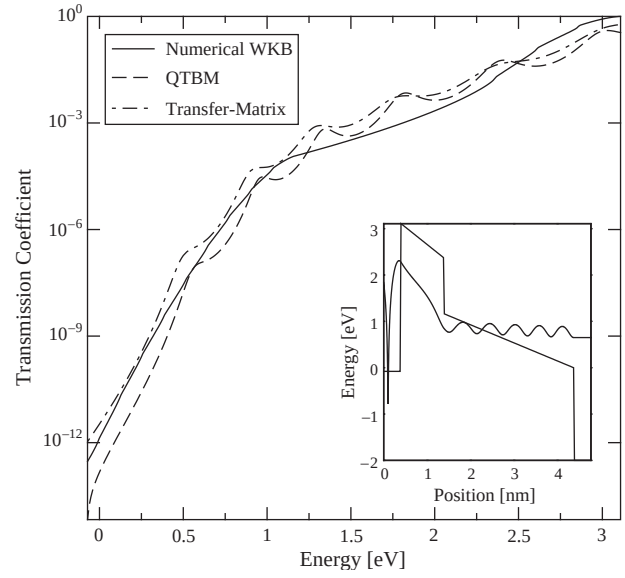


**Fig. 5.** The transmission coefficient using different methods for a dielectric consisting of two layers. The shape of the energy barrier and the wave function at 2.8 eV is shown in the inset.

be found by solving the stationary one-dimensional Schrödinger equation in the barrier region, which can be achieved using different numerical methods, such as the commonly applied Wentzel-Kramers-Brillouin (WKB) approximation[21] or Gundlach's method.[22] However, modern non-volatile memories often rely on nonlinear energy barriers to increase the device performance.[23] The WKB method does not account for wave function reflections in such structures, and the Gundlach method is accurate for triangular and trapezoidal barriers only.

A more general approach is the transfer-matrix method.[3] The basic principle of this method 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 can be derived by a number of subsequent matrix computations. From the transfer matrix, the transmission coefficient can be calculated (see Appendix C). However, several authors noted numerical problems applying this method for the computation of wave functions. These problems are due to the multiplication of matrices with exponentially growing and decaying states. For thick barriers, this leads to rounding errors which eventually exceed the amplitude of the wave function itself.[24–29]

An alternative method to compute the transmission coefficient is based on the quantum transmitting boundary method.[30–32] This 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. For one-dimensional calculations, as it is usually the case for gate dielectric tunneling, a fast recursive solution procedure has been proposed by Ravaioli[33] and is repeated in Appendix D.

Figure 5 shows the transmission coefficient for the different methods for a non-linear energy barrier. The inset shows the energy barrier and the values of $|\Psi|^2$ for an energy of 2.8 eV on a logarithmic scale. Note that at the left side of the barrier, the wave function consists of a superposition of incoming and reflected waves, which leads to the oscillating behavior of the absolute value. Right of the barrier, only a transmitted plain wave with constant $|\Psi|^2$ exists. The transfer-matrix and QTB methods deliver qualitatively similar results, while the WKB method does not resolve oscillations in the transmission coefficient.

### 2.3. Image Force Correction

When an electron approaches a dielectric layer, it induces a positive charge on the interface which acts like an image charge within the layer. This effect leads to a reduction of the barrier height for both electrons and holes:[34–36] The conduction band bends downward and the valence band bends upward, respectively. To account for this effect, the band edge energies must be modified

$$\begin{aligned} \mathscr{E}_c(x) &= \mathscr{E}_{c,0} - q\phi(x) + \mathscr{E}_{image}(x), \\ \mathscr{E}_v(x) &= \mathscr{E}_{v,0} - q\phi(x) + \mathscr{E}_{image}(x), \end{aligned} \tag{8}$$

where the image force correction in the dielectric with thickness $t_{diel}$ is calculated as[37]

$$\mathscr{E}_{image}(x) = -\frac{q^2}{16\pi k_{diel}} \sum_j^{\infty} (k_1 k_2)^j \left( \frac{k_1}{|x| + j t_{diel}} + \frac{k_2}{(j+1)t_{diel} - |x|} + \frac{2k_1 k_2}{(j+1)t_{diel}} \right), \tag{9}$$

where $x = 0$ is at the interface to the dielectric. The symbols $k_1$ and $k_2$ are calculated from the dielectric permittivities in the neighboring materials

$$k_1 = \frac{k_{\text{diel}} - k_{\text{si}}}{k_{\text{diel}} + k_{\text{si}}}, \quad k_2 = \frac{k_{\text{diel}} - k_{\text{metal}}}{k_{\text{diel}} + k_{\text{metal}}} = -1. \tag{10}$$

Here, $k_2$ accounts for the interface between the insulator and the metal and evaluates to $-1$. In the semiconductor the band edge energies are also altered

$$\mathscr{E}_{\text{image}}(x) = -\frac{\mathrm{q}^2}{16\pi k_{\text{si}}} \sum_{j}^{\infty} (k_1 k_2)^j \left( \frac{-k_1}{|x| + j t_{\text{diel}}} + \frac{k_2}{(j+1) t_{\text{diel}} + |x|} \right). \tag{11}$$

In practice it is sufficient to evaluate the sums in (9) and (11) up to $j = 11$.[38] Figure 6 shows the band edge energies in an MOS structure for a dielectric layer with a thickness of 2 nm and different dielectric permittivities for an applied bias of 2 V. A lower dielectric permittivity leads to a stronger band bending due to the image force and therefore strongly influences the transmission coefficient.

However, there is still some uncertainty if the image force has to be considered for tunneling calculations. While it is used in some works,[38–41] others neglect it or report only minor influence on the results.[42–46] For rigorous investigations, however, its necessary to include it in the simulations. This, however, raises the need for a high spatial resolution along the dielectric. Simple models like the analytical WKB formula or the Gundlach formula are not valid for this case. It may therefore be justified to account for the image force barrier lowering by correction factors.

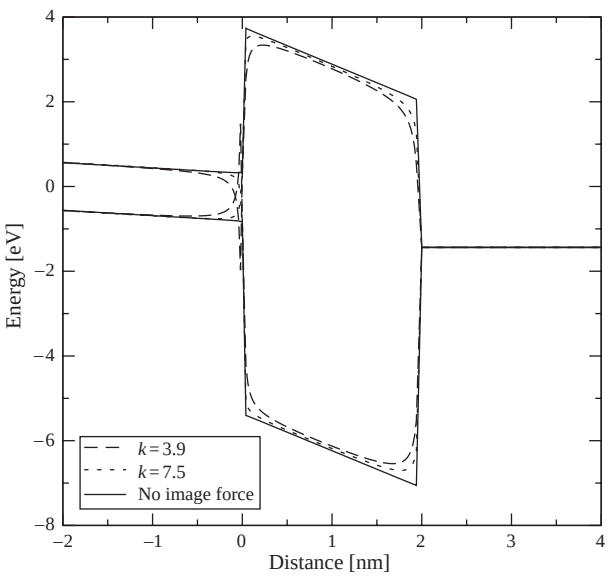


**Fig. 6.** Effect of the image force in an nMOS device with a dielectric thickness of 2 nm at a gate bias of 2 V.

### 2.4. Quasi-Bound State Tunneling

Up to now it has been assumed that all energetic states in the substrate contribute to the tunneling current. In the channel of small MOSFETs, however, the high electric field leads to a quantum-mechanical quantization of carriers.[47] If it is assumed that the wave function does not penetrate into the gate, discrete energy levels can be identified. However, taking a closer look at the conduction band edge of a MOSFET in inversion reveals that, depending on the boundary conditions, different types of quantized energy levels must be distinguished.[48] Bound states are formed at energies for which the wave function decays to zero at both sides of the dielectric layer. Quasi-bound states (QBS) have closed boundary conditions at one side and open boundary conditions on the other side of the dielectric. Only free states do not decay at any side. This can be seen in Figure 7 which shows the conduction band edge and two quasi-bound state wave functions. To account for tunneling current from both, free (3D) and quasi-bound (2D) states, the Tsu-Esaki equation must be replaced by

$$\begin{aligned} J = J_{\text{2D}} + J_{\text{3D}} = {} & \frac{\mathrm{k_B} T \mathrm{q}}{\pi \hbar^2} \sum_{i,\nu} \frac{g_\nu m_\parallel}{\tau_\nu(\mathscr{E}_{\nu,i}(m_{\mathrm{q}}))} \\ & \times \ln\left(1 + \exp\left(\frac{\mathscr{E}_{\mathrm{F}} - \mathscr{E}_{\nu,i}}{\mathrm{k_B} T}\right)\right) \\ & + \frac{4\pi \mathrm{q} m_{\text{3D}}}{h^3} \int_{\mathscr{E}_{\text{min},2}}^{\mathscr{E}_{\max}} TC(\mathscr{E}_x, m_{\text{diel}}) N(\mathscr{E}_x) \mathrm{d}\mathscr{E}_x, \end{aligned} \tag{12}$$

where the symbols $g_\nu$, $m_\parallel$, and $m_{\mathrm{q}}$ denote the valley degeneracy, parallel, and quantization masses ($g = 2$: $m_\parallel = m_{\mathrm{t}}$, $m_{\mathrm{q}} = m_{\mathrm{l}}$ and $g = 4$: $m_\parallel = \sqrt{m_{\mathrm{l}} m_{\mathrm{t}}}$, $m_{\mathrm{q}} = m_{\mathrm{t}}$), and $\tau_\nu(\mathscr{E}_{\nu,i})$ is

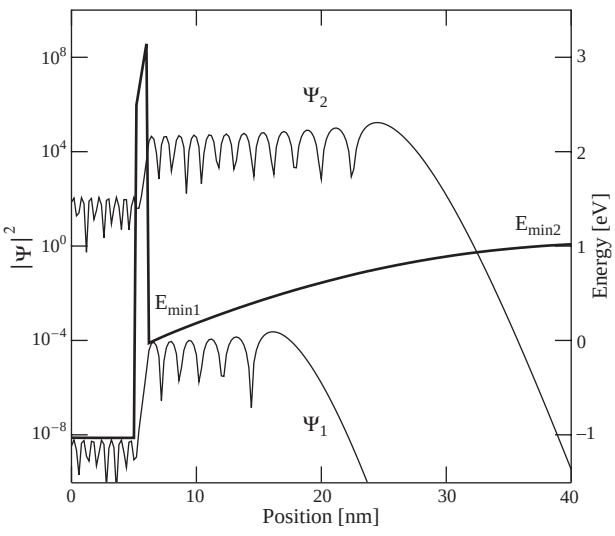


**Fig. 7.** The conduction band profile and two quasi-bound state wave functions. Quasi-bound state tunneling must be evaluated for $\mathscr{E}_{\text{min},1} < \mathscr{E} < \mathscr{E}_{\text{min},2}$, while the Tsu-Esaki expression must be used for $\mathscr{E} > \mathscr{E}_{\text{min},2}$.

the life time of the quasi-bound state $\mathscr{E}_{\nu,i}$. The life time can be interpreted as the time constant with which electrons in a quasi-bound state leak through the energy barrier. Several methods are, in principle, feasible for their calculation. They can be determined from the full-width half-maximum (FWHM) value of the phase of the reflection coefficient,[49] the FWHM value of the reflection coefficient itself,[50] or from the imaginary parts of the complex eigenvalues.[31] However, these methods are computationally demanding and therefore not suitable for implementation in general-purpose device simulators. Conventional device simulation packages even neglect the QBS tunneling component at all and use only the Tsu-Esaki formula (1) instead.[51–53] This formula, however, cannot reproduce the QBS tunneling component as shown in Figure 8, where the QBS current ($J_{2D}$) is compared to the continuum current ($J_{3D}$).

The dotted lines indicate the continuum current ($J_{3D}$) for $\mathscr{E}_{\min,2}$ as lower integration level (cf. Fig. 7), which is negligible for this case. The full lines show $J_{3D}$ using $\mathscr{E}_{\min,1}$ as lower integration level. Although the shape of the QBS component is reproduced, the absolute values differ significantly. It is thus necessary to account for QBS tunneling.

We propose to use (12) and calculate the life times from the quasi-classical approach

$$\tau_\nu(\mathscr{E}_{\nu,i}) = \int_0^x \frac{\sqrt{2m_\nu/(\mathscr{E}_{\nu,i} - \mathscr{E}_c(\xi))}}{TC(\mathscr{E}_{\nu,i})}\,\mathrm{d}\xi, \qquad (13)$$

with $\mathscr{E}_c(x) = \mathscr{E}_{\nu,i}$.[54] Furthermore, we keep the conventional shape of the Tsu-Esaki formula using $\mathscr{E}_{\min,2}$ as lower integration level. To further reduce the computation time, the eigenvalues of the triangular well approximation

$$\mathscr{E}_{\nu,i} = -z_i\left(\frac{\hbar^2}{2m_\nu}\right)^{1/3} E^{2/3} \qquad (14)$$

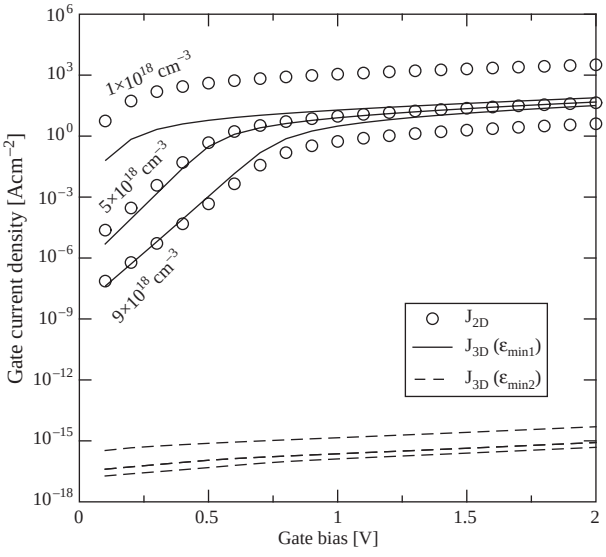


**Fig. 8.** Current density for different bulk doping and oxide thickness using only quasi-bound state tunneling ($J_{2D}$) and the Tsu-Esaki expression with $\mathscr{E}_{\min,1}$ ($J_{3D}(\mathscr{E}_{\min,1})$) or $\mathscr{E}_{\min,2}$ ($J_{3D}(\mathscr{E}_{\min,2})$) as lower integration level.

with $z_i$ being the zeros of the Airy function and $E$ the electric field can be used, instead of calculating the eigenvalues from the complex eigenvalue problem. Since the closed-boundary eigenvalues are higher than their open-boundary pendants, they must be corrected by an empirical fit factor in this case.[55]

## 2.5. Barrier Height and Tunneling Mass

The main parameters of the described tunneling models are the effective mass of the carrier in the dielectric layer and the barrier height of the dielectric material.

### *2.5.1. Barrier Height*

Table I shows the band gap energy and the dielectric permittivity of various dielectric materials considered as alternative dielectrics for MOS devices. Note the strong trade-off between the barrier height and the dielectric permittivity: Dielectrics with a high energy barrier have a low permittivity and *vice versa*. Hence, optimization becomes necessary to find the optimum material.

### *2.5.2. Tunneling Mass*

Table II shows a compilation of the effective electron ($m_{\mathrm{diel,e}}$) and hole ($m_{\mathrm{diel,h}}$) mass in $SiO_2$ layers given in the literature, which vary in the range of $0.3\,m_0$–$0.5\,m_0$ for electrons and $0.32\,m_0$–$0.77\,m_0$ for holes. Note that for the assumption of a Franz-type dispersion relation,[35] effective electron masses in the range of $0.41\,m_0$ to $0.61\,m_0$ have been found.[56–60] In the simulator MINIMOS-NT values of $m_{\mathrm{diel,e}} = 0.5\,m_0$ and $m_{\mathrm{diel,h}} = 0.8\,m_0$ have been applied, in accordance to the device simulator Dessis.[51]

Note, however, that the assumption of a constant electron mass in the dielectric is no more justified for ultra-thin $SiO_2$ layers. Here it was found both experimentally[61] and theoretically[62–64] by means of tight-binding simulations that the tunneling mass increases by almost 50% as the dielectric thickness is decreased down to 1 nm. They also present the fit formula $m'_{\mathrm{diel,e}}/m_{\mathrm{diel,e}} = c + (at_{\mathrm{diel}})^{-b}$ to

**Table I.** Dielectric permittivity $k/k_0$, band gap energy $\mathscr{E}_g$, conduction band offset $\Delta\mathscr{E}_c$, and valence band offset $\Delta\mathscr{E}_v$ of various dielectric materials. Values are taken from.[73, 116–121]

| | $k/k_0$ (1) | $\mathscr{E}_g$ (eV) | $\Delta\mathscr{E}_c$ (eV) | $\Delta\mathscr{E}_v$ (eV) |
|---|---|---|---|---|
| $SiO_2$ | 3.9 | 8.9–9.0 | 3.0–3.5 | 4.4–4.9 |
| $Si_3N_4$ | 7.0–7.9 | 5.0–5.3 | 2.0–2.4 | 1.5–2.0 |
| $Ta_2O_5$ | 23.0–26.0 | 4.4–4.5 | 0.3–1.50 | 1.9–3.0 |
| $TiO_2$ | 39.0–170.0 | 3.0–3.5 | 0.0–1.1 | 1.2–2.0 |
| $Al_2O_3$ | 8.0–10.0 | 8.7–9.0 | 2.7–2.8 | 4.8–5.1 |
| $ZrO_2$ | 12.0–25.0 | 5.0–7.8 | 1.4–2.5 | 2.2–5.3 |
| $HfO_2$ | 16.0–30.0 | 4.5–6.0 | 1.5 | 1.9–3.4 |
| $Y_2O_3$ | 4.4–18.0 | 5.5–6.0 | 1.3–2.3 | 2.2–3.6 |
| $ZrSiO_4$ | 3.8–12.6 | 4.5–6.0 | 0.7–1.5 | 2.7–3.4 |

**Table II.** Values of the effective electron ($m_{\mathrm{diel,e}}$) and hole ($m_{\mathrm{diel,h}}$) mass in $SiO_2$.[57]

| $t_{\mathrm{diel}}$ (nm) | $m_{\mathrm{diel,e}}/m_0$ (1) | $m_{\mathrm{diel,h}}/m_0$ (1) | Reference |
|---|---|---|---|
| 100 | 0.42 | | 65 |
| 100–12 | 0.5 | | 122 |
| 6–3 | 0.32 | | 123 |
| 3.5–1.5 | 0.5 | | 124 |
| 3.5–2.2 | 0.5 | | 60 |
| 6.5–1.56 | 0.5 | 0.42 | 125 |
| 5–2 | 0.437 | 0.437 | 69 |
| 3.6–1 | 0.4 | 0.32 | 2 |
| | 0.5 | 0.77 | 51 |

describe the thickness dependence of the tunneling mass, with $m'_{\mathrm{diel,e}}$ being the corrected value and parameter values of $c = 0.706$, $a = 0.708\,\mathrm{nm}^{-1}$, and $b = 1.004$ for parabolic effective-mass calculations.

## 2.6. Compact Models

For the application in practical device simulation it is desirable to use compact models which do not require large computational resources. The most commonly used model to describe tunneling is the Fowler-Nordheim formula:[65]

$$J = \frac{\mathrm{q}^3 m_{\mathrm{eff}}}{8\pi m_{\mathrm{diel}} h \mathrm{q}\Phi_{\mathrm{B}}} E_{\mathrm{diel}}^2 \exp\left(-\frac{4\sqrt{2m_{\mathrm{diel}}(\mathrm{q}\Phi_{\mathrm{B}})^3}}{3\hbar \mathrm{q} E_{\mathrm{diel}}}\right). \quad (15)$$

This expression can be derived from the Tsu-Esaki formula (1) by the assumption of zero temperature, a triangular energy barrier, and equal materials on both sides of the dielectric. Thus, it is not valid for direct tunneling where the barrier is of trapezoidal shape. Furthermore, $\mathrm{q}\Phi_{\mathrm{B}}$ denotes the difference between the Fermi energy in the electrode and the conduction band edge in the dielectric, and not the conduction band offset, as often wrongly assumed.

Schuegraf and Hu derived correction terms for this expression to make it applicable to the regime of direct tunneling[66]

$$J = \frac{\mathrm{q}^3 m_{\mathrm{eff}}}{8\pi m_{\mathrm{diel}} h \mathrm{q}\Phi_{\mathrm{B}} B_1} E_{\mathrm{diel}}^2 \exp\left(-\frac{4\sqrt{2m_{\mathrm{diel}}(\mathrm{q}\Phi_{\mathrm{B}})^3} B_2}{3\hbar \mathrm{q} E_{\mathrm{diel}}}\right), \quad (16)$$

with the correction terms $B_1$ and $B_2$ given as

$$B_1 = \left(1 - \left(1 - \frac{\mathrm{q}E_{\mathrm{diel}} t_{\mathrm{diel}}}{\mathrm{q}\Phi_{\mathrm{B}}}\right)^{1/2}\right)^2, \quad (17)$$

$$B_2 = \left(1 - \left(1 - \frac{\mathrm{q}E_{\mathrm{diel}} t_{\mathrm{diel}}}{\mathrm{q}\Phi_{\mathrm{B}}}\right)^{3/2}\right). \quad (18)$$

For a triangular barrier the correction factors become $B_1 = B_2 = 1$ and the expression simplifies to (15). Note that (15) is only valid to describe tunneling between materials without work function difference since in the derivation a triangular barrier with slope equal to the Fermi energy differences divided by the dielectric thickness is assumed.[67, 68]

## 2.7. Simulation Results for MOS Transistors

### *2.7.1. Tunneling Paths*

Tunneling in an MOS transistor can be separated into a path between the gate and the channel, and a path between the gate and the the source and drain extension areas.[69] Tunneling in the source and drain extension areas can exceed tunneling in the channel by orders of magnitude. This is related to two effects: First, instead of n-p or p-n tunneling, n-n or p-p tunneling prevails. Second, the potential difference and thus the bending of the energy barrier is high. This increased tunneling current in the source and drain extension areas can be a serious problem if measurements are performed on long-channel MOSFETs to characterize their short-channel pendants, because the edge tunneling currents exceed the channel tunneling current by orders of magnitude. Furthermore, there is a fundamental difference between tunneling in MOS transistors and MOS capacitors.[67, 70] In contrast to MOS transistors, MOS capacitors which are biased in strong inversion cannot supply the amount of carriers as predicted by the tunneling model. This effect is termed *substrate-limited* tunneling, because the tunneling current is limited by the generation rate in the substrate. In the channel of an inverted MOS transistor, on the other hand, carriers can always be supplied by the source and drain contacts.

The typical shape of the gate current density in turned-off nMOS and pMOS devices is depicted in Figure 9.[18] A $SiO_2$ gate dielectric thickness of 2 nm and an acceptor/donor doping of $5 \times 10^{17}\,\mathrm{cm}^{-3}$ and polysilicon gates has been chosen. In the nMOS device the majority electron tunneling current always exceeds the hole tunneling current due to the lower electron mass and barrier height (3.2 eV instead of 4.65 eV for holes). In the pMOS capacitor, however, the majority hole tunneling exceeds electron tunneling only for negative and low positive bias. For positive bias the conduction band electron current again dominates due to its much lower barrier height.

The Tsu-Esaki model with an analytical WKB transmission coefficient is in good agreement with measured data for devices with different gate lengths and bulk doping as shown in Figure 10 for nMOS (top) and pMOS devices (bottom).[71] The simulations in this Figure have been performed using the device simulator MINIMOS-NT.[72] It can be seen that the gate current can be reproduced over a wide range of dielectric thicknesses with a single set of physical parameters. The compact tunneling models are compared in Figure 12 for an nMOS structure with 3 nm dielectric thickness. The Schuegraf model fails to describe the tunneling current density at low bias. For high bias it

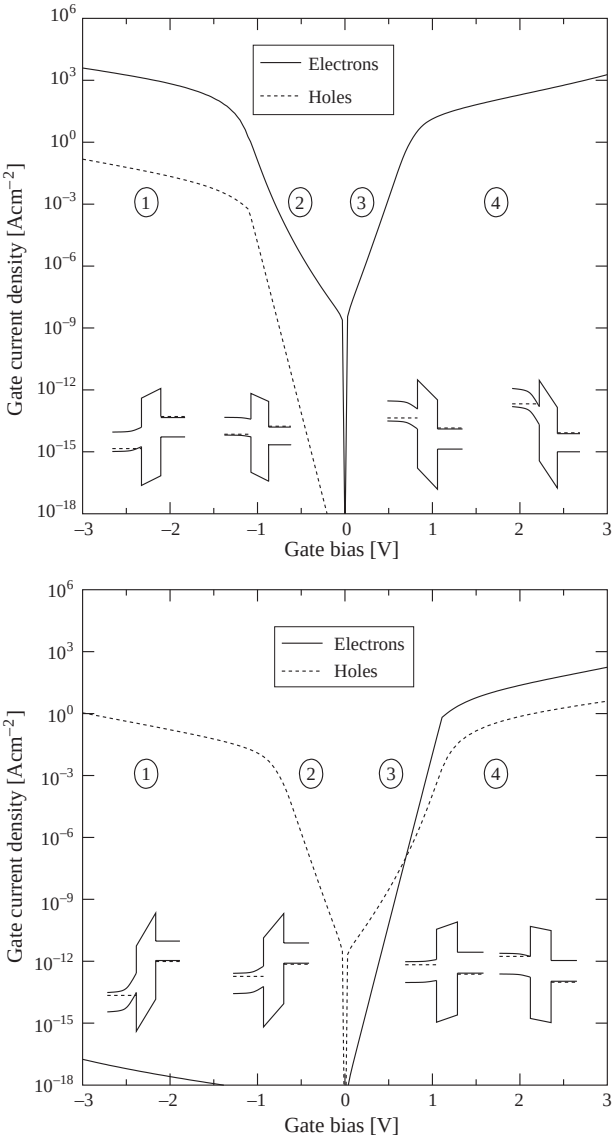


**Fig. 9.** Tunneling current components in an nMOS (top) and a pMOS (bottom) device with 2 nm dielectric thickness. The insets show the approximate shape of the band edge energies, with the gate contact located at the right side.

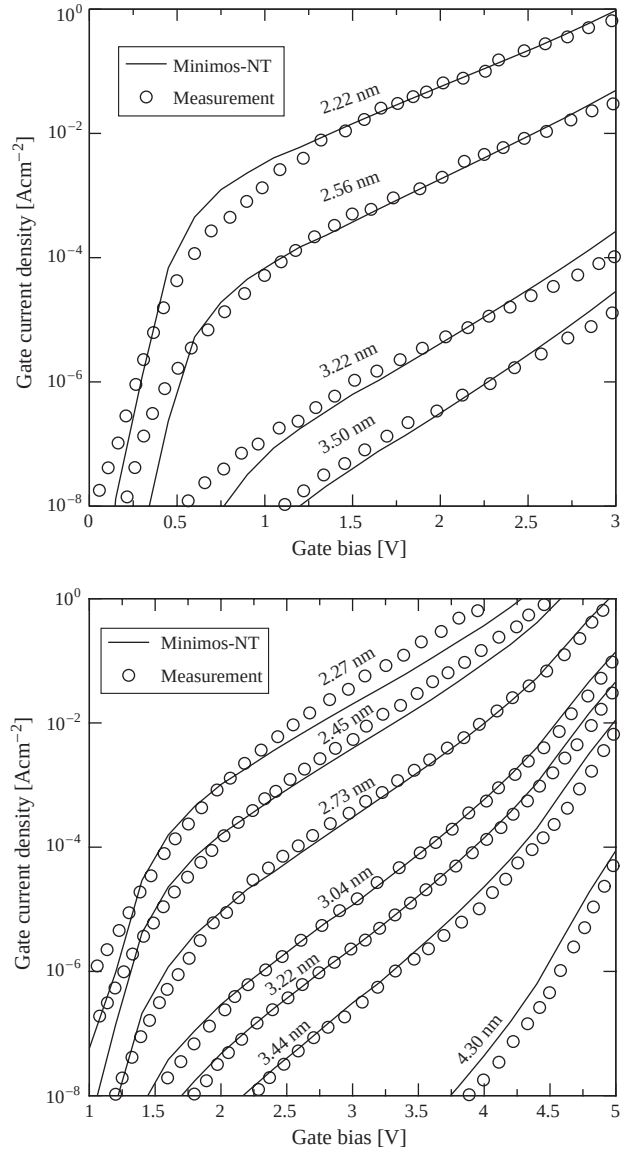





**Fig. 10.** Comparison of the gate current predicted by the Tsu-Esaki model using an analytical WKB method for the transmission coefficient with measurements of a nMOS (top) and pMOS (bottom) transistor.[71]

may be used to obtain an estimate of the gate current. The Fowler-Nordheim model totally fails for this application.

### *2.7.2. Hot Carrier Tunneling*

The distribution function in the channel of a turned-on MOS transistor heavily deviates from the shape implied by a Fermi-Dirac or Maxwellian distribution. A model for the non-Maxwellian shape of the distribution function was presented which accurately reproduced the carrier energy distribution along the channel. The Maxwellian distribution **underestimates** the distribution of high-energy electrons in the channel of turned-on devices, while the heated Maxwell **overestimates** it. To check the impact of this wrong high-energy behavior, the integrand of the Tsu-Esaki formula, namely the expression $TC(\mathscr{E})N(\mathscr{E})$ has been evaluated for a standard device, as shown in the upper part of Fig. 11, and compared to Monte Carlo results.[18, 19] The simulated device had a gate length of 100 nm and a gate dielectric thickness of 3 nm. While at low energies the difference between the non-Maxwellian distribution function (5) and the heated Maxwellian distribution (6) 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

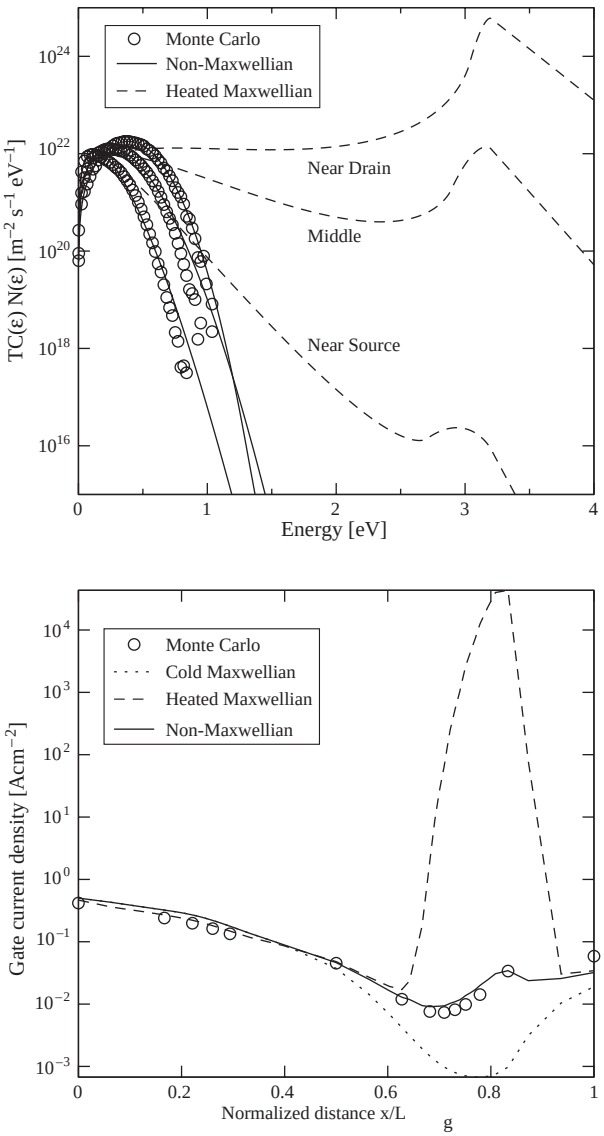


**Fig. 11.** Integrand of Tsu-Esaki's equation (left) and gate current density along the channel (right) of a MOSFET with 100 nm gate length and 3 nm gate dielectric thickness.[18,19]

non-Maxwellian shape of the distribution function, indicated by the full line, reproduces the Monte Carlo results very well.

The region of high electron temperature is confined to only a small area near the drain contact, as shown in the lower part of Figure 11, where the gate current density along the channel is compared to Monte Carlo results. At the point of the peak electron temperature, which is located at approximately $x = 0.8L_g$, the heated Maxwellian approximation overestimates the gate current density by a factor of almost $10^6$. It will therefore have a large impact on the total gate current density. The cold Maxwellian approximation underestimates the gate current density in this region, while the non-Maxwellian distribution correctly reproduces the Monte Carlo results.

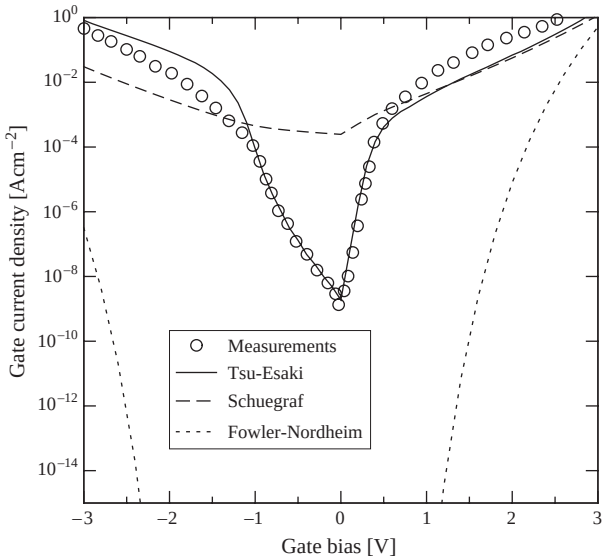


**Fig. 12.** Comparison of the compact model results with measurements of an nMOS structure.[69]

The non-Maxwellian shape yields excellent agreement, while the heated Maxwellian approximation substantially overestimates the gate current density especially near the drain region. Instead of the heated Maxwellian distribution it appears to be better to use a cold Maxwellian distribution in that regime since it leads to a comparably low underestimation of the gate current density.

### 2.8. Non-Volatile Memories Based on Layered Dielectrics

One of the most important figures of merit of a non-volatile memory cell is its $I_{on}/I_{off}$– ratio: A high on-current leads to low programming and erasing times, and a low off-current increases the retention time of the device. This ratio can be increased if, for a given device, the tunneling current in the on-state (the charging/discharging current) is increased or, in the off-state (during the retention time), decreased. With a single-layer dielectric it is not possible to tune on- and off-current independently. However, if the tunnel dielectric is replaced by a dielectric stack of varying barrier height as shown in Figure 18, it becomes possible. In this figure the device structure and the conduction band edge in the on- and off-state are shown. The device consists of a standard EEPROM structure, where the tunnel dielectric is composed of three layers. The middle layer has a higher energy barrier than the inner and outer layers. The flat-band case is indicated by the dotted lines.

In the on-state a high voltage is applied on the top contact. The middle energy barrier is strongly reduced and gives rise to a high tunneling current. If the dielectric would consist of a single layer, the peak of the energy barrier would not be reduced. Thus, the on-current is much

higher for the layered dielectric. In the off-state a low negative voltage—due to charge stored on the memory node—is applied. The middle barrier is only slightly suppressed and blocks tunneling. The off-current is only slightly lower than for a single-layer dielectric. This behavior results in a high $I_{on}/I_{off}$ ratio. A high suppression of the middle barrier in the on-state requires a low permittivity of the outer layers so that the potential drop in the outer layers is high.[73] This device design was first proposed by Capasso et al. in 1988[74] based on AlGaAs-GaAs devices and later used by several authors,[23,75] where it became popular as *crested-barrier* memory[75] or VARIOT (*varying oxide thickness* device[23]).

The gate current density of the device depicted in Figure 18 is shown as a function of the gate bias in Figure 19. A stack thickness of 5 nm was chosen. Since the middle layers must have a high band gap, only few material combinations are possible. For the simulations middle layers of $Al_2O_3$ and $SiO_2$ have been chosen, with outer layers of $Y_2O_3$, $Si_3N_4$, and $ZrO_2$. For comparison full $SiO_2$ and $Si_3N_4$ stacks have also been simulated (the dotted and dash-dotted lines). While $Y_2O_3$ shows a very high off-current, stacks with outer layers of $Si_3N_4$ or $ZrO_2$ and $Al_2O_3$ as middle layer show good ratios between the on-state (positive gate bias) current density and the off-state (negative gate bias) current density.

The important figure of merit, however, is the $I_{on}/I_{off}$-ratio. In Figure 20 the $I_{on}/I_{off}$-ratio is shown for $Si_3N_4$ and $ZrO_2$ stacks with $SiO_2$ and $Al_2O_3$ middle layers as a function of the thickness of the middle layer. Also shown is the ratio for a layer of $SiO_2$ and $Si_3N_4$ alone. It is obvious that the ratio strongly depends on the thickness of the middle layer, and both minima and maxima can be observed. Only outer layers of $Si_3N_4$ lead to a significantly increased performance as compared to full layers of $SiO_2$ or $Si_3N_4$. A middle layer thickness around 1–2 nm for the assumed 6 nm stack gives optimum performance for this application. Note however, that in these simulations no trap-assisted tunneling was assumed.

## 3. DEFECT-ASSISTED TUNNELING

Besides direct tunneling, which is a one-step tunneling processes, defects in the dielectric layer give rise to tunneling processes based on two or more steps. This tunneling component is mainly observed after writing-erasing cycles in non-volatile memory devices. It is generally assumed that this is due to traps which arise in the dielectric layer. The increased tunneling current at low bias (stress-induced leakage current or SILC) is mainly responsible for the degradation of the retention time of these devices.[76] SILC has been widely studied and modeled in MOS capacitors[77–79] and EEPROM devices.[80] This section gives a brief overview of trap-assisted tunneling models and elaborates on describes an inelastic trap-assisted tunneling model which was included in the device simulator MINIMOS-NT.

### 3.1. Model Overview

A frequently used model is the generalized trap-assisted tunneling model presented by Chang et al.[81,82] The current density reads

$$J = \mathrm{q}\int_0^{t_{\mathrm{diel}}} A N_{\mathrm{T}}(x)\frac{P_1(x)P_2(x)}{P_1(x)+P_2(x)}\,\mathrm{d}x, \tag{19}$$

where $A$ denotes a fitting constant, $N_{\mathrm{T}}(x)$ the spatial trap concentration, and $P_1$ and $P_2$ the transmission coefficients of electrons captured and emitted by traps. A similar model was used by Ghetti et al.[83]

$$J = \int_0^{t_{\mathrm{diel}}} C_{\mathrm{T}} N_{\mathrm{T}}(x)\frac{J_{\mathrm{in}}J_{\mathrm{out}}}{J_{\mathrm{in}}+J_{\mathrm{out}}}\,\mathrm{d}x, \tag{20}$$

who assumed a constant capture cross section $C_{\mathrm{T}}$ for the traps. The symbols $J_{\mathrm{in}}$ and $J_{\mathrm{out}}$ denote the capture and emission currents. Essentially the same formula was used by other authors as well.[84,85] Considerable research has been done by Ielmini et al.[86–89] who describe inelastic TAT and also take hopping conduction into account.[90,91] They derive the trap-assisted current by an integration along the dielectric thickness and energy

$$J = \int_0^{t_{\mathrm{diel}}} \mathrm{d}x \int_{\mathscr{E}_{\min}}^{\mathscr{E}_{\max}} \widetilde{J}(\mathscr{E}_{\mathrm{T}}, x)\,\mathrm{d}\mathscr{E},$$

where $\widetilde{J}$ denotes the net current flowing through the dielectric, given as the difference between capture and emission currents through the left and right side of the dielectric

$$\widetilde{J}(\mathscr{E}_{\mathrm{T}}, x) = J_{\mathrm{cl}} - J_{\mathrm{el}} = J_{\mathrm{er}} - J_{\mathrm{cr}} = \mathrm{q}N'_{\mathrm{T}}W_{\mathrm{c}}\left(1 - \frac{f_{\mathrm{T}}(\mathscr{E}_{\mathrm{T}}, x)}{f_{\mathrm{l}}(\mathscr{E}_{\mathrm{T}}, x)}\right),$$

where $f_{\mathrm{T}}$ is the trap occupancy, $\mathscr{E}_{\mathrm{T}}$ the trap energy, $W_{\mathrm{c}}$ the capture rate, and $f_{\mathrm{l}}$ the energy distribution function at the left interface. The symbol $N'_{\mathrm{T}}$ denotes the trap concentration in space and energy. Ielmini further develops the model to include transient effects and notes that in this case, the net difference between current from the left and right interfaces equals the change in the trap occupancy multiplied by the trap charge

$$(J_{\mathrm{cl}} - J_{\mathrm{el}}) + (J_{\mathrm{cr}} - J_{\mathrm{er}}) = \mathrm{q}N_{\mathrm{T}}\frac{\partial f_{\mathrm{T}}}{\partial t} \tag{21}$$

### 3.2. Inelastic Multiphonon-Emission Trap-Assisted Tunneling

Experimental evidence has been reported that SILC is caused by inelastic trap-assisted tunnel transitions.[76,92–96] A detailed model for inelastic, trap-assisted tunneling by

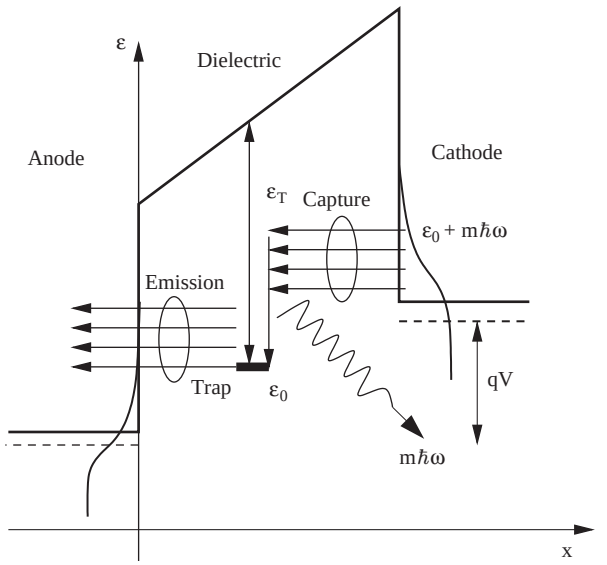


**Fig. 13.** Trap-assisted tunneling transition by inelastic phonon emission. Electrons are captured from the cathode, relax to the trap energy level $\mathscr{E}_0$ by the emission of phonons, and are emitted to the anode.[97]

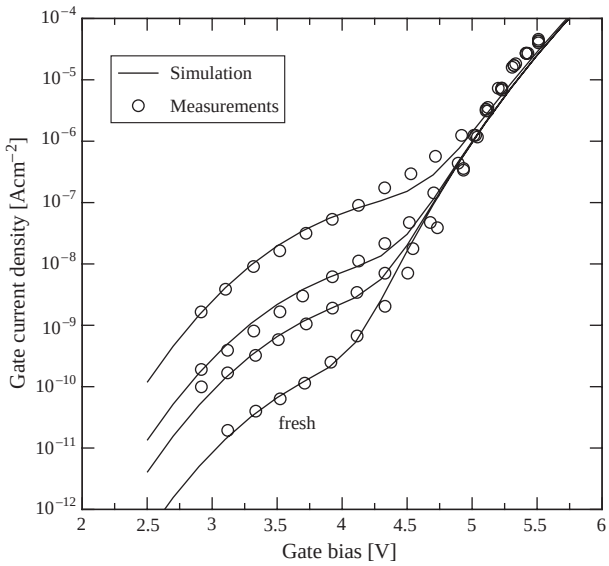


**Fig. 14.** Gate current density for different stress times[93] for $t_{\text{diel}} = 5.5$ nm. The model parameters are $\mathscr{E}_T = 2.7$ eV, $\hbar w = 20$ meV, and $N_T = 9.0 \times 10^{17}$ cm$^{-3}$, $1.0 \times 10^{17}$ cm$^{-3}$, $3.0 \times 10^{16}$ cm$^{-3}$, and $3.0 \times 10^{15}$ cm$^{-3}$ (from top to bottom).

means of multiphonon emission has been presented by Herrmann and Schenk[97] and modified versions have been used by other authors as well.[38, 98–100] The trap-assisted tunneling process is modeled via inelastic phonon-assisted transitions as shown in Figure 13.[97, 99, 101] Electrons are captured from the cathode, relax to the energy of the trap $\mathscr{E}_0$ by phonon emission with energy $m\hbar\omega$, and are emitted to the anode. The trap-assisted tunneling current is found by integration over the dielectric thickness

$$J_t = q \int_0^{t_{\text{diel}}} \frac{N_T(x)}{\tau_c(x) + \tau_e(x)} dx, \tag{22}$$

where $N_T(x)$ is the trap concentration and $\tau_c(x)$ and $\tau_e(x)$ denote the capture and emission times calculated from

$$\tau_c^{-1}(z) = \int_{\mathscr{E}_0}^{\infty} c_n(\mathscr{E}, x) T_l(\mathscr{E}) f_l(\mathscr{E}) \, d\mathscr{E} \tag{23}$$

$$\tau_e^{-1}(z) = \int_{\mathscr{E}_0}^{\infty} e_n(\mathscr{E}, x) T_r(\mathscr{E})(1 - f_r(\mathscr{E})) \, d\mathscr{E}. \tag{24}$$

In these expressions, $c_n$ and $e_n$ denote the capture and emission rates, computed as

$$c_n(\mathscr{E}, x) = c_0 \sum_m L_m \delta(\mathscr{E} - \mathscr{E}_m) \tag{25}$$

$$e_n(\mathscr{E}, x) = c_0 \exp\left( -\frac{\mathscr{E} - \mathscr{E}_T}{k_B T_L} \right) \sum_m L_m \delta(\mathscr{E} - \mathscr{E}_m) \tag{26}$$

with $c_0 = (4\pi)^2 r_T^2 (\hbar\Theta_0)^3 / (\hbar \mathscr{E}_{g,SiO_2})$ and $(\hbar\Theta_0) = (q^2 \hbar^2 F^2 / (2\, m))^{1/3}$. The symbols $f_l$ and $f_r$ the Fermi distributions, $T_l$ and $T_r$ the transmission coefficients from the left and right side of the dielectric, $F$ the electric field in the dielectric, and $\mathscr{E}_{g,SiO2}$ the band gap of $SiO_2$. A constant trap radius $r_T$ is assumed. The transmission coefficients were evaluated by a numerical WKB method, which yields reasonable accuracy for single-layer dielectrics. This model has been implemented in the device simulator MINIMOS-NT. Figure 14 shows a comparison with experimental data for MOS capacitors,[93] where the transition from the trap-assisted tunneling regime at low bias to the Fowler-Nordheim tunneling regime at high bias is clearly visible.

### 3.3. Transient Trap Charging

To predict the transient behavior of fast switching processes, the charging and discharging dynamics of the traps must be considered. The concentration of occupied traps at position $x$ and time $t$ is generally described by the rate equation

$$N_T(x) \frac{df_T(x,t)}{dt} = N_T(x) \frac{1 - f_T(x,t)}{\tau_c(x,t)} - N_T(x) \frac{f_T(x,t)}{\tau_e(x,t)}$$

where $\tau_c$ and $\tau_e$ describe the capture and emission time of the trap. For the stationary case, the time derivative on the left-hand side is zero:

$$\frac{1 - f_T(x,t)}{\tau_c(x,t)} = \frac{f_T(x,t)}{\tau_e(x,t)} = R(x) \tag{27}$$

From (27) and the incremental gate current density $dj(x) = qR(x)N_T(x)dx$, expression (22) can be derived.[97] For the transient case, the time constants must be evaluated in each time step. The occupancy function can be calculated iteratively by $f_T(x, t_i) = A_i + B_i f_T(x, t_{i-1})$ where $A_i$ and $B_i$ depend on the capture and emission times at the time step $t_i$ by[100]

$$A_i = \frac{\tau_c^{-1}(z, t_i)\Delta t_i}{1 + C_i}, \quad B_i = \frac{1 - C_i}{1 + C_i}, \quad C_i = \frac{\tau_m^{-1}(z, t_i)\Delta t_i}{2}.$$

In these expressions $\Delta t_i = t_i - t_{i-1}$ and $t_i$ denote the discretized time steps, and $\tau_m^{-1} = \tau_c^{-1} + \tau_e^{-1}$. Once the time-dependent occupancy function in the dielectric is known, the tunnel current through an interface at time $t_i$ is

$$J_{l,r}(t_i) = q \int_0^{t_{diel}} N_T(x) \tau_{l,r}^{-1}(x, t_i)\, dx \tag{28}$$

where l,r denotes the considered interface (left or right) and the time constants $\tau_l$ and $\tau_r$ are calculated from

$$\tau_{l,r}^{-1}(x, t_i) = \tau_{cl,r}^{-1}(x, t_i) - f_T(x, t_i)[\tau_{cl,r}^{-1}(x, t_i) + \tau_{el,r}^{-1}(x, t_i)]$$

with the respective values of the capture and emission times to the left and right interface $\tau_{cl,r}$ and $\tau_{el,r}$. Note that the current through the two interfaces is, in general, not equal. Only after the trap charging processes are finished, the capture and emission currents at the interfaces are in equilibrium.

By comparison with the step response of MOS capacitors, this model can be used to characterize the trap concentration, energy, and trap radius $r_T$. As an example, Figure 17 shows the step response of two pMOS capacitors with $ZrO_2$ dielectrics fabricated using MOCVD (metal-organic chemical vapor deposition) and afterwards annealed under different ambient conditions.[5] The gate voltage is first fixed at a value of 2.5 V to achieve a steady initial trap occupation and is then turned off. The resulting transient gate current peak exceeds the static gate current by orders of magnitude. Especially for the oxide annealed in forming gas atmosphere, the gate current decays very slowly with a time constant in the order of a second. This may be caused by a different trap distribution in the oxide or even different trap energy levels which lead to a different time constant for the discharging process.[102] The measurements can be fitted assuming the trap concentrations indicated in the Figure.

### 3.4. Degradation Modeling

Several models have been proposed to describe the trap generation process which is responsible for the gradual degradation of the dielectric layer in non-volatile memories over time.[103] One of the most frequently encountered models is the anode hole injection (AHI) model, where the tunneling electrons cause impact ionization of holes in the substrate which are injected back into the oxide.[104, 105] Other models such as the anode hydrogen release (AHR) model[106] assume that electrons injected into the substrate have enough energy to release hydrogen ions present at the Si-$SiO_2$ interface. However, it has been shown that MOS devices annealed in deuterium still show similar breakdown characteristics, which makes the AHR model questionable.[107] A further model is the thermochemical model proposed by McPherson et al.,[108] which describes the generation of traps in the dielectric due to the presence of a strong electric field which breaks up weak bonds. However, a comprehensive and commonly accepted model is still lacking.

In accordance with Ghetti[103] we distinguish three processes which happen sequentially and finally trigger breakdown. Starting from a fresh dielectric layer with a low trap concentration, the direct tunneling current gives rise to the creation of neutral defects. Contrary to,[104] trap generation is based on the injected charge alone, taking into account all tunneling components. The generated defects cause trap-assisted tunneling, leading to two effects. First, some of the existing traps become occupied by electrons, which changes the threshold voltage of the device. Second, new defects are created in the dielectric layer. The location of the traps is assumed to be random with a uniform distribution within the layer, while a constant energy level and a specific charge state (positive or negative) is assumed. Finally, if a conductive path through the dielectric is formed, a localized breakdown occurs and the current density increases according to the conductivity of the dielectric layer.

While the neutral defects cause trap-assisted tunneling and gate leakage, only the occupied traps lead to a shift of the threshold voltage. This is modeled by an additional space charge $\rho(x) = Q_T N_T(x) f_T(x)$ in the Poisson equation, where $f_T$ denotes the trap occupancy and $Q_T$ the trap charge state. Note that the assumption of phonon-assisted tunneling implies that, depending on the bias conditions, only a fraction of the traps in the dielectric layer is really occupied.[100] The neutral defects create percolation paths in the dielectric, which eventually connect the gate with the substrate.[109] In MINIMOS-NT the traps are placed randomly, and the defect concentration $N_T$ is assumed to be proportional to the total injected charge per area $Q_i$ via $N_T = CQ_i^{\alpha}$, as proposed by Degraeve et al.,[110] who found values of $C = 5.3 \times 10^{-19}\,cm^{-1.88}As^{-0.56}$ and $\alpha = 0.56$ for dielectric thicknesses between 7.3 and 13.8 nm. As soon as a percolation path through the dielectric is created, the dielectric layer loses its insulating behavior and the current suddenly increases. Figure 15 shows a cross section of the gate dielectric layer where the dark spots mark traps. The corresponding gate current density is shown in Figure 16 as a function of time for different gate voltages assuming an initial trap concentration of $10^{16}\,cm^{-3}$. The time-to-breakdown strongly decreases and the gate leakage strongly increases with higher gate bias. After breakdown the gate current density can no more be described by a tunneling process. Measurements indicate that the gate current

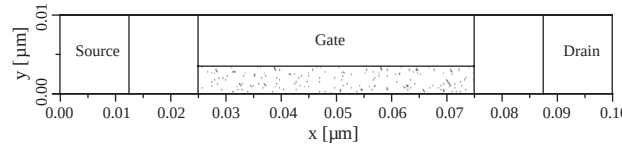


**Fig. 15.** Two-dimensional cut through the dielectric layer simulated with MINIMOS-NT and showing the random trap placement (darks spots).

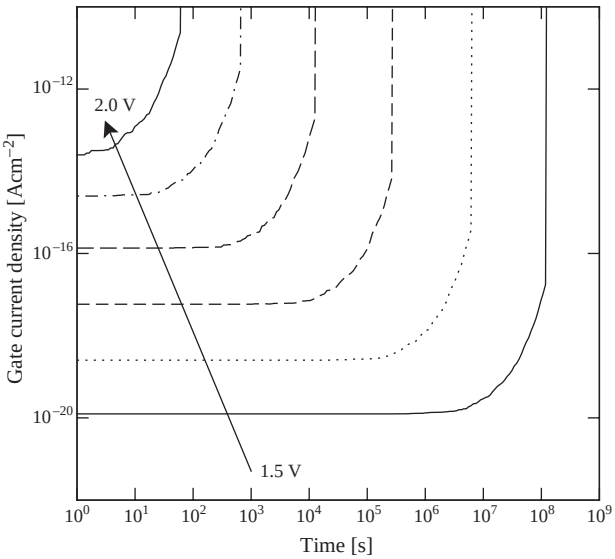


**Fig. 16.** Dielectric breakdown of a 3 nm $SiO_2$ layer as a function of gate bias.

after breakdown can be described by a point contact conduction model.[111] In this model, the gate current is related to the gate voltage by a simple power law $I = KV_G^p$, where the parameter $K$ reflects the size of the breakdown spot, and the parameter $p$ is in the range of 2–5.[112–114] Miranda et al.[112] noted that the values of $p$ and $K$ are statistically correlated: An introduction of the area prefactor $K$ comes with a reduction of the slope $p$. However, no physically sound model is available to describe this behavior.

## 4. MODEL COMPARISON

This paper outlined a number of tunneling models useful for the simulation of tunneling at the device simulation level. For practical applications, however, it is often not

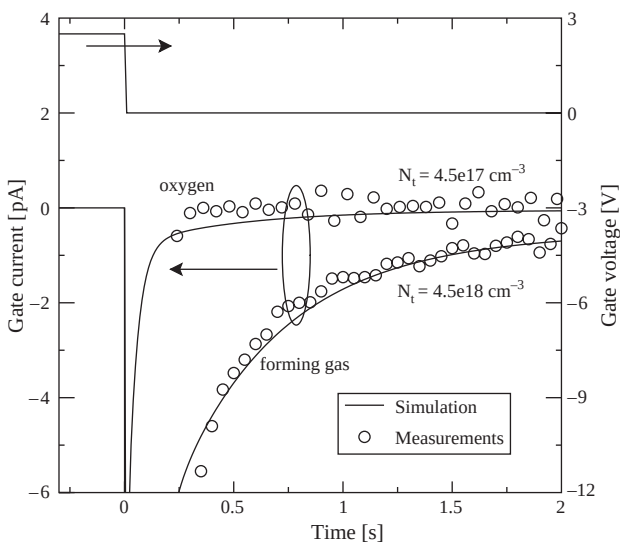


**Fig. 17.** Transient trap charging currents for a $ZrO_2$ layer fabricated by MOCVD and annealed under different ambient atmospheres.[5, 102]

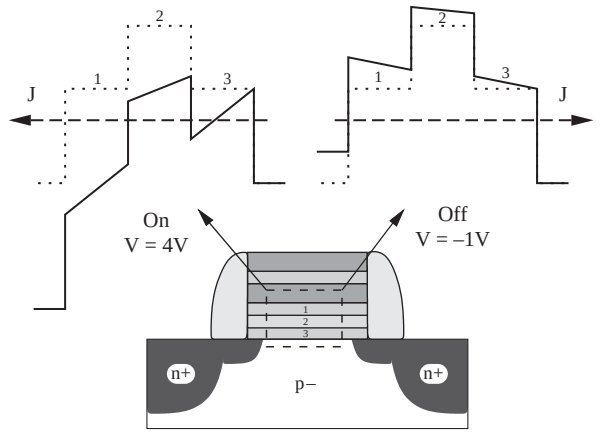


**Fig. 18.** Device structure and operating principle of a non-volatile memory based on crested barriers.[75]

clear which model to select for the application at hand. Therefore, Table III summarizes the main model features and also gives the approximate computational effort. The following points can be concluded:

- Especially the Fowler-Nordheim and Schuegraf models have a very low computational effort since they are compact models. However, they do not correctly reproduce the device physics and can only be used after careful calibration.
- The Tsu-Esaki formula with the analytical WKB or Gundlach method for the transmission coefficient combines moderate computational effort with reasonable accuracy. This approach can be used for the simulation of tunneling in devices with single-layer dielectrics.
- The inelastic TAT model allows simulation of all effects related with traps in the dielectric and poses only moderate computational effort. This model can be used

**Table III.** A hierarchy of tunneling models and their properties: Fowler-Nordheim model (FN), Schuegraf-model (SM), Tsu-Esaki model with analytic WKB transmission coefficient (TA), Tsu-Esaki model with Gundlach transmission coefficient (TG), Tsu-Esaki model with numeric WKB transmission coefficient (TN), Tsu-Esaki model with transfer-matrix method (TT), Tsu-Esaki model with QTB method (TQ), TA…Inelastic trap-assisted tunneling model.

| | FN | SM | TA | TG | TN | TT | TQ | TA |
|---|---|---|---|---|---|---|---|---|
| FN tunneling | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Direct tunneling | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| EVB tunneling process | | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| QM current oscillations | | | | ✓ | | ✓ | ✓ | |
| Dielectric stacks | | | | | ✓ | ✓ | ✓ | |
| Trap-assisted tunneling | | | | | | | | ✓ |
| Trap occupancy modeling | | | | | | | | ✓ |
| Transient TAT | | | | | | | | ✓ |
| Computational effort | low | low | | | high | high | high | |

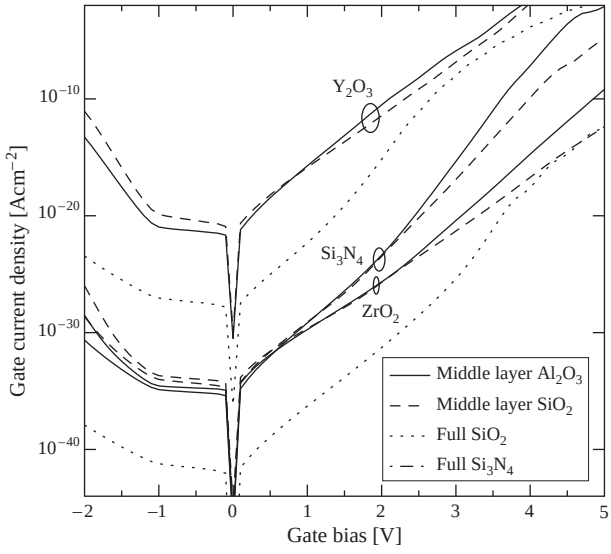


**Fig. 19.** Gate current density as a function of the gate bias for different materials of the middle layer, compared to full $SiO_2$ and $Si_3N_4$ layers.

for the simulation of leakage in EEPROMs or trap-rich dielectric devices.

- The Tsu-Esaki model with the numerical WKB, transfer-matrix, or QTB method to calculate the transmission coefficient represents the most accurate method usable for the simulation of tunneling through dielectric stacks, however, with high computational effort. If one is also interested in the wave functions, the transfer-matrix method should be used with care to avoid numerical overflow.

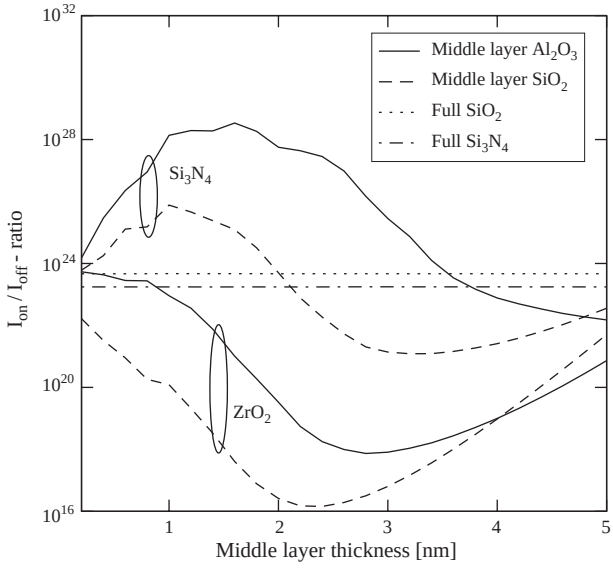


**Fig. 20.** Ratio between the on-current and the off-current as a function of the middle layer thickness for different materials of the outer layers ($Si_3N_4$ and $ZrO_2$) and middle layers ($Al_2O_3$ and $SiO_2$), compared to the resulting current density using full layers of $SiO_2$ and $Si_3N_4$.

- If tunneling in turned-on devices is studied, the correct shape of the energy distribution function in the channel must be taken into account using a supply function based on a non-Maxwellian distribution function, such as (29) or (30). Using a cold Maxwellian distribution instead leads to a underestimation of the gate current density.

## 5. SUMMARY AND CONCLUSION

We presented a hierarchy of tunneling models for semiconductor device simulation. Higher-order transport models are found suitable for the description of hot-carrier tunneling, where the correct modeling of the carrier distribution in energy is crucial. Common methods to estimate the transmission coefficient of energy barriers have been reviewed, and an overview of advantages and shortcomings of the different methods was given. For cold electron tunneling (turned-off devices) and strong channel doping, gate leakage is dominated by quasi-bound state tunneling which must be taken into account in addition—and not instead of—the conventional Tsu-Esaki formula. The correct representation of the distribution function is crucial for tunneling in turned-on devices, where the use of a heated Maxwellian distribution severely overestimates the resulting current density. To describe gate dielectric degradation we propose to link an inelastic trap-assisted tunneling model to the occurrence of dielectric wearout and breakdown phenomena in dielectrics. This method also accounts for fast transient charging and discharging processes.

Although these models represent the state-of-the-art at the device simulation level, open questions remain. These comprise the use of a constant effective mass in the dielectric layer, which contradicts *ab-initio* studies, the controversial issue of image force correction, and the modeling of high-$k$ insulator reliability issues, which are still not fully understood.

## APPENDIX A: SUPPLY FUNCTION FOR NON-MAXWELLIAN DISTRIBUTION

With expression (5) for the distribution function and the assumption of a Fermi-Dirac distribution in the polysilicon gate, the supply function (2) becomes

$$N(\mathscr{E}) = A_1 \frac{\mathscr{E}_{\mathrm{ref}}}{b} \Gamma_{\mathrm{i}}\left(\frac{1}{b}, \left(\frac{\mathscr{E}}{\mathscr{E}_{\mathrm{ref}}}\right)^b\right) - A_2 \mathrm{k_B} T_{\mathrm{L}} \ln\left(1 + \exp\left(-\frac{\mathscr{E} + \Delta\mathscr{E}_{\mathrm{c}}}{\mathrm{k_B} T_{\mathrm{L}}}\right)\right), \quad (29)$$

where $\Gamma_{\mathrm{i}}(\alpha, \beta)$ denotes the incomplete gamma function

$$\Gamma_{\mathrm{i}}(x, y) = \int_y^\infty \exp(-\alpha)\alpha^{x-1}\,\mathrm{d}\alpha.$$

In (29) the explicit value of the Fermi energy was replaced by the shift of the two conduction band edges $\Delta\mathscr{E}_{\mathrm{c}}$. Using

the accurate shape of the distribution (6), the expression for the supply function becomes

$$N(\mathscr{E}) = A_1 \frac{\mathscr{E}_{\mathrm{ref}}}{b}\Gamma_{\mathrm{i}}\left(\frac{1}{b}, \left(\frac{\mathscr{E}}{\mathscr{E}_{\mathrm{ref}}}\right)^b\right) A_1 c \mathrm{k_B} T_2 \exp\left(-\frac{\mathscr{E}}{\mathrm{k_B} T_{\mathrm{L}}}\right) - A_2 \mathrm{k_B} T_{\mathrm{L}} \ln\left(1+\exp\left(-\frac{\mathscr{E}+\Delta\mathscr{E}_{\mathrm{c}}}{\mathrm{k_B} T_{\mathrm{L}}}\right)\right) \quad (30)$$

for a Fermi-Dirac distribution in the polysilicon gate. These expressions can be used instead of (3) to account for hot-carrier tunneling.

## APPENDIX B: NORMALIZATION OF THE DISTRIBUTION FUNCTION

When implementing the analytical expressions for the distribution function and the supply function into a device simulator it is necessary to assure consistency: the carrier concentration defined by the analytical distribution function must match the carrier concentration from the transport model used. Therefore, the normalization prefactor $A$ has to be evaluated from

$$n = \langle 1 \rangle = \frac{1}{4\pi^3}\int f(\mathbf{k})\mathrm{d}^3k. \quad (31)$$

This equation can be transformed to spherical coordinates using $k = (k_x^2+k_y^2+k_z^2)^{1/2}$

$$n = \frac{1}{4\pi^3}\int_{-\pi}^{\pi}\mathrm{d}\alpha\int_0^{\pi}\sin\theta\,\mathrm{d}\theta\int_0^{\infty} f(k)k^2\,\mathrm{d}k. \quad (32)$$

For a parabolic dispersion relation we have $\mathrm{d}k = m_{\mathrm{eff}}/k\hbar^2\mathrm{d}\mathscr{E}$ which finally leads to

$$n = \int_0^{\infty} f(\mathscr{E})\frac{4\pi\sqrt{2m_{\mathrm{eff}}^3}}{h^3}\sqrt{\mathscr{E}}\,\mathrm{d}\mathscr{E}, \quad (33)$$

where the integration is performed from the conduction band edge $\mathscr{E}_{\mathrm{c}} = 0$. For a Maxwellian or heated Maxwellian distribution (expression (4)), the normalization constant evaluates to

$$A = \frac{nh^3}{4\pi(\mathrm{k_B}T_\nu)^{3/2}\Gamma\left(\frac{3}{2}\right)\sqrt{2m_{\mathrm{eff}}^3}} \quad (34)$$

where $T_\nu$ is either the lattice temperature (for the assumption of a Maxwellian distribution) or the carrier temperature (for the assumption of a heated Maxwellian distribution). Using the non-Maxwellian distribution (5) the normalization constant evaluates to

$$A = \frac{nh^3 b}{4\pi\mathscr{E}_{\mathrm{ref}}^{3/2}\Gamma\left(\frac{3}{2b}\right)\sqrt{2m_{\mathrm{eff}}^3}}, \quad (35)$$

while for expression (6) it is

$$A = \frac{nh^3}{4\pi\left(\frac{\mathscr{E}_{\mathrm{ref}}^{1/2}}{b}\Gamma\left(\frac{3}{2b}\right) + c(\mathrm{k_B}T_{\mathrm{L}})^{3/2}\Gamma\left(\frac{3}{2}\right)\right)\sqrt{2m_{\mathrm{eff}}^3}}. \quad (36)$$

## APPENDIX C: THE TRANSFER-MATRIX METHOD

If an arbitrary potential barrier is segmented into N regions with constant potentials the wave function in each region can be written as the sum of an incident and a reflected wave[115] $\Psi_j(x) = A_j\exp(\imath k_j x) + B_j\exp(-\imath k_j x)$ with the wave number $k_j = \sqrt{2m_j(\mathscr{E}-W_j)}/\hbar$. The wave amplitudes $A_j$, $B_j$, the carrier mass $m_j$, and the potential energy $W_j$ are assumed constant for each region $j$. With interface conditions for continuity of the wave function and its derivative at each layer interface, the transmitted wave of a layer relates to the incident wave by a complex transfer matrix:

$$\begin{pmatrix} A_j \\ B_j \end{pmatrix} = \underline{T}_j \begin{pmatrix} A_{j-1} \\ B_{j-1} \end{pmatrix} \quad 2 \le j \le \mathrm{N}. \quad (37)$$

The transfer matrices are of the form

$$\underline{T}_j = \frac{1}{2}\begin{pmatrix} \left(1+\frac{k_{j-1}}{k_j}\right)\gamma^{-k_j} & \left(1-\frac{k_{j-1}}{k_j}\right)\gamma^{-k_j} \\ \left(1-\frac{k_{j-1}}{k_j}\right)\gamma^{k_j} & \left(1+\frac{k_{j-1}}{k_j}\right)\gamma^{k_j} \end{pmatrix} \times \begin{pmatrix} \gamma^{k_{j-1}} & 0 \\ 0 & \gamma^{-k_{j-1}} \end{pmatrix} \quad 2 \le j \le \mathrm{N}, \quad (38)$$

with the phase factor $\gamma = \exp(\imath\Delta(j-2))$. The transmitted wave in Region N can then be calculated from the incident wave by subsequent multiplication of transfer matrices:

$$\begin{pmatrix} A_{\mathrm{N}} \\ B_{\mathrm{N}} \end{pmatrix} = \prod_{j=2..\mathrm{N}} \underline{T}_j \begin{pmatrix} A_1 \\ B_1 \end{pmatrix}. \quad (39)$$

If it is assumed that there is no reflected wave in Region N and the amplitude of the incident wave is unity, (39) simplifies to

$$\begin{pmatrix} A_{\mathrm{N}} \\ 0 \end{pmatrix} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}\begin{pmatrix} 1 \\ B_1 \end{pmatrix}, \quad (40)$$

and the transmission coefficient can be calculated from (7). Note that the straightforward calculation of $A_{\mathrm{N}}$ from[3]

$$A_{\mathrm{N}} = T_{11} - T_{12}\frac{T_{21}}{T_{22}} \quad (41)$$

may lead to erroneous results due to the subtraction of numbers which have been derived by subsequent matrix multiplications. Instead, it can be shown that $\det \underline{T} = T_{11}T_{22} - T_{12}T_{21} = 1$, and therefore the amplitude of the transmitted wave is simply $A_{\mathrm{N}} = 1/T_{22}$.

## APPENDIX D: THE QUANTUM-TRANSMITTING BOUNDARY METHOD

An efficient method to solve the Schrödinger equation with open boundary conditions has been proposed by Frensley.[31] The method is based on the tight-binding

quantum-transmitting boundary method introduced by Lent.[30] Using a simple finite-difference approximation on an equidistant grid with an effective mass $m_j$ and a grid spacing $\Delta x$, the stationary one-dimensional Schrödinger equation reads

$$-\frac{\hbar^2}{2\cdot m}\frac{\Psi(i-1)-2\Psi(i)+\Psi(i+1)}{\Delta x^2}+(W(i)-\mathscr{E})\Psi(i)=0 \tag{42}$$

Assuming a grid spacing $X_i = i\Delta x$, the wave function at position $x_i$ consists of an incident and a reflected part, with amplitudes $I(i)$ and $R(i)$[33]

$$\begin{aligned}\Psi(i) &= I(i)\cdot\exp(\imath\cdot k_x(i)\cdot i\cdot\Delta x)\\ &\quad +R(i)\cdot\exp(-\imath\cdot k_x(i)\cdot i\cdot\Delta x)\end{aligned} \tag{43}$$

At position $x_n$ it is assumed that there is no reflected wave. The transmitting boundary conditions relate the wave functions outside of the simulation domain $\Psi(-1)$ and $\Psi(n+1)$ to the wave functions inside the simulation domain:

$$\begin{aligned}&\Psi(n) = I(n)\cdot\exp(\imath\cdot k_x(n)\cdot n\cdot\Delta x)\\ &\qquad\rightarrow \Psi(n+1) = I(n)\cdot\exp(\imath\cdot k_x(n)\cdot(n+1)\cdot\Delta x)\\ &\Psi(0) = I(0)+R(0)\\ &\qquad\rightarrow \Psi(-1) = I(0)\cdot\exp(-\imath\cdot k_x(0)\cdot\Delta x)\\ &\qquad +R(0)\exp(\imath\cdot k_x(0)\cdot\Delta x)\end{aligned}$$

Assuming the transmitted wave amplitude with $I(n)=1$ allows to calculate the values of the wave function $\Psi(n-1)$, $\Psi(n-2)$, ... recursively via

$$\Psi(i-1)=\left(2+\frac{2\cdot m\cdot\Delta x^2}{\hbar^2}(W(i)-\mathscr{E})\right)\cdot\Psi(i)-\Psi(i+1) \tag{44}$$

Finally, the values of $\Psi(0)$ and $\Psi(-1)$ determine the amplitude of the incoming wave $I(0)$, and the transmission coefficient is calculated as $TC = |I(n)/I(0)|^2$. Note that the quantum transmitting boundary method overcomes the numerical problems usually encountered in transfer-matrix calculations.[24] It can thus safely be used for the evaluation of the transmission coefficient of arbitrary stacked dielectric layers.

**Acknowledgments:** The support of Hans Kosina, Tibor Grasser, Francisco Jiménez-Molinos, and Stefan Harasek is gratefully acknowledged.

## References


1. R. Bez, E. Camerlenghi, A. Modelli, and A. Visconti, *Proc. IEEE* 91, 489 **(2003)**.
2. W.-C. Lee and C. Hu, *IEEE Trans. Electron Devices* 48, 1366 **(2001)**.
3. R. Tsu and L. Esaki, *Appl. Phys. Lett.* 22, 562 **(1973)**.
4. A. Gehring, H. Kosina, and S. Selberherr, *J.Computational Electronics* 2, 219 **(2003)**.
5. S. Harasek, H. D. Wanzenböck, and E. Bertagnolli, *J. Vac. Sci. Technol. A* 21, 653 **(2003)**.
6. M. Houssa, M. Tuominen, M. Naili, V. Afanas'ev, A. Stesmans, S. Haukka, and M. M. Heyns, *J. Appl. Phys.* 87, 8615 **(2000)**.
7. A. Kumar, M. V. Fischetti, T. H. Ning, and E. Gusev, *J. Appl. Phys.* 94, 1728 **(2003)**.
8. C. B. Duke, *Tunneling in Solids*, Academic Press **(1969)**.
9. Khairurrijal, W. Mizubayashi, S. Miyazaki, and M. Hirose, *J. Appl. Phys.* 87, 3000 **(2000)**.
10. M. Lundstrom and Z. Ren, *IEEE Trans. Electron Devices* 49, 133 **(2002)**.
11. D. Cassi and B. Riccó, *IEEE Trans. Electron Devices* 37, 1514 **(1990)**.
12. K.-I. Sonoda, M. Yamaji, K. Taniguchi, C. Hamaguchi, and S. T. Dunham, *J. Appl. Phys.* 80, 5444 **(1996)**.
13. T. Grasser, H. Kosina, C. Heitzinger, and S. Selberherr, *J. Appl. Phys.* 91, 3869 **(2002)**.
14. L. Selmi, A. Ghetti, R. Bez, and E. Sangiorgi, *Microelectronic Engineering* 36, 293 **(1997)**.
15. T. Grasser, H. Kosina, and S. Selberherr, in *Proc. Intl. Conf. on Simulation of Semiconductor Processes and Devices* **(2001)**, pp. 46–49.
16. T. Grasser, H. Kosina, and S. Selberherr, *J. Appl. Phys.* 90, 6165 **(2001)**.
17. A. Abramo and C. Fiegna, *J. Appl. Phys.* 80, 889 **(1996)**.
18. A. Gehring, T. Grasser, H. Kosina, and S. Selberherr, *J. Appl. Phys.* 92, 6019 **(2002)**.
19. A. Gehring, T. Grasser, H. Kosina, and S. Selberherr, *Electron. Lett.* 39, 691 **(2003)**.
20. A. Gehring and S. Selberherr, *IEEE Trans. Dev. Mat. Rel.* 4, 306 **(2004)**.
21. S. Gasiorowicz, *Quantum Physics*, John Wiley & Sons **(1995)**.
22. K. H. Gundlach, *Solid-State Electron.* 9, 949 **(1966)**.
23. B. Govoreanu, P. Blomme, M. Rosmeulen, J. V. Houdt, and K. D. Meyer, *IEEE Electron Device Lett.* 24, 99 **(2003)**.
24. G. Wachutka, *Physical Review B* 34, 8512 **(1986)**.
25. D. Y. K. Ko and J. C. Inkson, *Physical Review B* 38, 9945 **(1988)**.
26. D. Z. Y. Ting, E. T. Yu, and T. C. McGill, *Physical Review B* 45, 3583 **(1992)**.
27. T. Usuki, M. Saito, M. Takatsu, R. A. Kiehl, and N. Yokoyama, *Physical Review B* 52, 8244 **(1995)**.
28. B. A. Biegel, *Dissertation*, Stanford University **(1997)**.
29. F. Heinz, *Dissertation*, ETH Zürich **(2004)**.
30. C. S. Lent and D. J. Kirkner, *J. Appl. Phys.* 67, 6353 **(1990)**.
31. W. R. Frensley, *Superlattices & Microstructures* 11, 347 **(1992)**.
32. L. F. Register, U. Ravaioli, and K. Hess, *J. Appl. Phys.* 69, 7153 **(1991)**.
33. U. Ravaioli, Numerical Methods for the Solution of Schrödinger Equation for Ballistic Transport, Presentation at the 2002 School on Computational Material Science, University of Illinois, Urbana Champaign, http://www.mcc.uiuc.edu/SummerSchool02/.
34. E. H. Rhoderick and R. H. Williams, *Metal-Semiconductor Contacts*, Oxford University Press, Oxford **(1988)**.
35. W. Franz, in *Handbuch der Physik*, edited by S. Flügge, Springer, Berlin **(1956)**, Vol. XVII, p. 155.
36. M. V. Fischetti, S. E. Laux, and E. Crabbé, *J. Appl. Phys.* 78, 1058 **(1995)**.
37. M. Kleefstra and G. C. Herman, *J. Appl. Phys.* 51, 4923 **(1980)**.
38. F. Jiménez-Molinos, F. Gámiz, A. Palma, P. Cartujo, and J. A. Lopez-Villanueva, *J. Appl. Phys.* 91, 5116 **(2002)**.
39. G. Yang, K. Chin, and R. Marcus, *IEEE Trans. Electron Devices* 38, 2373 **(1991)**.
40. A. Schenk and G. Heiser, *J. Appl. Phys.* 81, 7900 **(1997)**.
41. A. Schenk, *Advanced Physical Models for Silicon Device Simulation*, Springer **(1998)**.
42. C. Fiegna, E. Sangiorgi, and L. Selmi, *IEEE Trans. Electron Devices* 40, 2018 **(1993)**.

**43.** Z. A. Weinberg, *J. Appl. Phys.* 53, 5052 **(1982)**.
**44.** W.-Y. Quan, D. M. Kim, and M. K. Cho, *J. Appl. Phys.* 92, 3724 **(2002)**.
**45.** L. Larcher, A. Paccagnella, and G. Ghidini, *IEEE Trans. Electron Devices* 48, 271 **(2001)**.
**46.** B. Majkusiak, *IEEE Trans. Electron Devices* 37, 1087 **(1990)**.
**47.** F. Stern, *Physical Review B* 5, 4891 **(1972)**.
**48.** W. Magnus and W. Schoenmaker, *Microelectronics Reliability* 41, 31 **(2001)**.
**49.** E. Cassan, *J. Appl. Phys.* 87, 7931 **(2000)**.
**50.** R. Clerc, A. Spinelli, G. Ghibaudo, and G. Pananakakis, *J. Appl. Phys.* 91, 1400 **(2002)**.
**51.** *DESSIS 9.5 User's Manual*, Integrated Systems Engineering, **2004**.
**52.** *MEDICI 2002.4.0 User's Manual*, Synopsys, Mountain View, CA, **2003**.
**53.** *ATLAS User's Manual*, Silvaco, Santa Clara, CA, **2004**.
**54.** A. Dalla Serra, A. Abramo, P. Palestri, L. Selmi, and F. Widdershoven, *IEEE Trans. Electron Devices* 48, 1811 **(2001)**.
**55.** A. Gehring and S. Selberherr, in *Proc. Intl. Conf. on Simulation of Semiconductor Processes and Devices*, München **(2004)**, pp. 25–28.
**56.** J. Maserjian, *J. Vac. Sci. Technol.* 11, 996 **(1974)**.
**57.** R. Clerc, *Dissertation*, Institut National Polytechnique de Grenoble, **2001**.
**58.** M. Av-Ron, M. Shatzkes, T. H. DiStefano, and R. A. Gdula, *J. Appl. Phys.* 52, 2897 **(1981)**.
**59.** S. H. Lo, D. A. Buchanan, Y. Taur, and W. Wang, *IEEE Trans. Electron Devices* 18, 209 **(1997)**.
**60.** L. F. Register, E. Rosenbaum, and K. Yang, *Appl. Phys. Lett.* 74, 457 **(1999)**.
**61.** R. Ludeke, E. Cartier, and A. Schenk, *Appl. Phys. Lett.* 75, 1407 **(1999)**.
**62.** M. Städele, B. R. Tuttle, and K. Hess, *J. Appl. Phys.* 89, 348 **(2001)**.
**63.** M. Städele, B. Fischer, B. R. Tuttle, and K. Hess, *Solid-State Electron.* 46, 1027 **(2002)**.
**64.** M. Städele, F. Sacconi, A. D. Carlo, and P. Lugli, *J. Appl. Phys.* 93, 2681 **(2003)**.
**65.** M. Lenzlinger and E. H. Snow, *J. Appl. Phys.* 40, 278 **(1969)**.
**66.** K. F. Schuegraf, C. C. King, and C. Hu, in *Proc. Symposium on VLSI Technology* **(1992)**, pp. 18–19.
**67.** J. P. Shiely, *Dissertation*, Duke University, **1999**.
**68.** A. Gehring, *Dissertation*, Technische Universität Wien **(2003)**.
**69.** J. Cai and C.-T. Sah, *J. Appl. Phys.* 89, 2272 **(2001)**.
**70.** H. Z. Massoud and J. P. Shiely, *Microelectronic Engineering* 36, 263 **(1997)**.
**71.** S. H. Lo, D. A. Buchanan, and Y. Taur, *IBM J. Res. Dev.* 43, 327 **(1999)**.
**72.** *MINIMOS-NT 2.1 User's Guide*, Institut für Mikroelektronik, Technische Universität Wien, Austria **(2004)**.
**73.** J. D. Casperson, L. D. Bell, and H. A. Atwater, *J. Appl. Phys.* 92, 261 **(2002)**.
**74.** F. Capasso, F. Beltram, R. J. Malik, and J. F. Walker, *IEEE Electron Device Lett.* 9, 377 **(1988)**.
**75.** K. K. Likharev, *Appl. Phys. Lett.* 73, 2137 **(1998)**.
**76.** S. Aritome, R. Shirota, G. Hemink, T. Endoh, and F. Masuoka, *Proc. IEEE* 81, 776 **(1993)**.
**77.** B. Riccó, G. Gozzi, and M. Lanzoni, *IEEE Trans. Electron Devices* 45, 1554 **(1998)**.
**78.** K. Sakakibara, N. Ajika, K. Eikyu, K. Ishikawa, and H. Miyoshi, *IEEE Trans. Electron Devices* 44, 1002 **(1997)**.
**79.** A. Ghetti, E. Sangiorgi, J. Bude, T. W. Sorsch, and G. Weber, *IEEE Trans. Electron Devices* 47, 2358 **(2000)**.
**80.** C.-M. Yih, Z.-H. Ho, M.-S. Liang, and S. S. Chung, *IEEE Trans. Electron Devices* 48, 300 **(2001)**.
**81.** W. J. Chang, M. P. Houng, and Y. H. Wang, *J. Appl. Phys.* 89, 6285 **(2001)**.
**82.** W. J. Chang, M. P. Houng, and Y. H. Wang, *J. Appl. Phys.* 90, 5171 **(2001)**.
**83.** A. Ghetti, A. Hamad, P. J. Silverman, H. Vaidya, and N. Zhao, in *Proc. Intl. Conf. on Simulation of Semiconductor Processes and Devices*, IEEE Press, Piscataway, NJ **(1999)**, pp. 239–242.
**84.** M. Lenski, T. Endoh, and F. Masuoka, *J. Appl. Phys.* 88, 5238 **(2000)**.
**85.** L. Larcher, A. Paccagnella, and G. Ghidini, *IEEE Trans. Electron Devices* 48, 285 **(2001)**.
**86.** D. Ielmini, A. S. Spinelli, M. A. Rigamonti, and A. L. Lacaita, *IEEE Trans. Electron Devices* 47, 1258 **(2000)**.
**87.** D. Ielmini, A. S. Spinelli, M. A. Rigamonti, and A. L. Lacaita, *IEEE Trans. Electron Devices* 47, 1266 **(2000)**.
**88.** D. Ielmini, A. S. Spinelli, A. L. Lacaita, A. Martinelli, and G. Ghidini, *Solid-State Electron.* 45, 1361 **(2001)**.
**89.** D. Ielmini, A. S. Spinelli, A. L. Lacaita, and G. Ghidini, *Solid-State Electron.* 46, 417 **(2002)**.
**90.** D. Ielmini, A. S. Spinelli, A. L. Lacaita, and A. Modelli, *Microelectronic Engineering* 59, 189 **(2001)**.
**91.** D. Ielmini, A. S. Spinelli, A. L. Lacaita, and A. Modelli, *Solid-State Electron.* 46, 1749 **(2002)**.
**92.** R. Moazzami and C. Hu, in *Proc. Intl. Electron Devices Meeting*, IEEE Press, Piscataway, NJ (1992), pp. 139–142.
**93.** E. Rosenbaum and L. F. Register, *IEEE Trans. Electron Devices* 44, 317 **(1997)**.
**94.** S.-I. Takagi, N. Yasuda, and A. Toriumi, *IEEE J. Solid-State Circuits* 46, 348 **(1999)**.
**95.** R. Rofan and C. Hu, *IEEE Electron Device Lett.* 12, 632 **(1991)**.
**96.** J. Wu, L. F. Register, and E. Rosenbaum, in *Proc. Intl. Reliability Physics Symposium* **(1999)**, pp. 389–395.
**97.** M. Herrmann and A. Schenk, *J. Appl. Phys.* 77, 4522 **(1995)**.
**98.** A. Palma, A. Godoy, J. A. Jimenez-Tejada, J. E. Carceller, and J. A. Lopez-Villanueva, *Physical Review B* 56, 9565 **(1997)**.
**99.** F. Jiménez-Molinos, A. Palma, F. Gámiz, J. Banqueri, and J. A. Lopez-Villanueva, *J. Appl. Phys.* 90, 3396 **(2001)**.
**100.** A. Gehring, F. Jiménez-Molinos, H. Kosina, A. Palma, F. Gámiz, and S. Selberherr, *Microelectron. Reliab.* 43, 1495 **(2003)**.
**101.** L. Larcher, *IEEE Trans. Electron Devices* 50, 1246 **(2003)**.
**102.** A. Gehring, S. Harasek, E. Bertagnolli, and S. Selberherr, in *Proc. European Solid-State Device Research Conf.*, edited by J. Franca and P. Freitas, Frontier Group **(2003)**, pp. 473–476.
**103.** A. Ghetti, in *Predictive Simulation of Semiconductor Processing*, edited by E. W. J. Dabrowski, Springer **(2004)**, pp. 201–258.
**104.** M. A. B. Weir, J. Bude, P. Silverman, and A. Ghetti, *Microelectronic Engineering* 89, 137 **(2001)**.
**105.** D. J. DiMaria and J. H. Stathis, *J. Appl. Phys.* 89, 5015 **(2001)**.
**106.** J. H. Stathis, *IEEE Trans. Device and Materials Reliability* 1, 43 **(2001)**.
**107.** J. Wu, E. Rosenbaum, B. MacDonald, E. Li, J. Tao, B. Tracy, and P. Fang, in *Proc. Intl. Reliability Physics Symposium* **(2000)**, pp. 27–32.
**108.** W. McPherson, R. B. Khamankar, and A. Shanware, *J. Appl. Phys.* 88, 5351 **(2000)**.
**109.** J. H. Stathis, IBM *J. Res. Dev.* 46, 265 **(2002)**.
**110.** R. Degraeve, G. Groeseneken, R. Bellens, J. L. Ogier, M. Depas, P. J. Roussel, and H. E. Maes, *IEEE Trans. Electron Devices* 45, 904 **(1998)**.
**111.** J. Suñé, E. Miranda, M. Nafría, and X. Aymerich, in *Proc. Intl. Electron Devices Meeting* **(1998)**, pp. 191–194.
**112.** E. Miranda, J. Suñé, R. Rodríguez, M. Nafría, and X. Aymerich, *IEEE Electron Device Lett.* 20, 265 **(1999)**.
**113.** J. H. Stathis, B. P. Linder, R. Rodríguez, and S. Lombardo, *Microelectron. Reliab.* 43, 1353 **(2003)**.
**114.** R. Rodríguez, J. H. Stathis, B. P. Linder, R. V. Joshi, and C. T. Chuang, *Microelectron. Reliab.* 43, 1439 **(2003)**.
**115.** D. K. Ferry and S. M. Goodnick, *Transport in Nanostructures*, Cambridge University Press, Cambridge **(1997)**.

**116.** M. LeRoy, E. Lheurette, O. Vanbesien, and D. Lippens, *J. Appl. Phys.* 93, 2966 **(2003)**.
**117.** C. M. Osburn, I. Kim, S. K. Han, I. De, K. F. Yee, S. Gannavaram, S. J. Lee, C.-H. Lee, Z. J. Luo, W. Zhu, J. R. Hauser, D.-L. Kwong, G. Lucovsky, T. P. Ma, and M. C. Öztürk, *IBM J. Res. Dev.* 46, 299 **(2002)**.
**118.** G. D. Wilk, R. M. Wallace, and J. M. Anthony, *J. Appl. Phys.* 89, 5243 **(2001)**.
**119.** J. Robertson, *J. Vac. Sci. Technol.* 18, 1785 **(2000)**.
**120.** H.-S. P. Wong, *IBM J. Res. Dev.* 46, 133 **(2002)**.
**121.** J. Zhang, J. S. Yuan, Y. Ma, and A. S. Oates, *Solid-State Electron.* 44, 2165 **(2000)**.
**122.** Z. A. Weinberg, *Solid-State Electron.* 20, 11 **(1977)**.
**123.** M. Depas, B. Vermeire, P. W. Mertens, R. L. V. Meirhaeghe, and M. M. Heyns, *Solid-State Electron.* 38, 1465 **(1995)**.
**124.** F. Rana, S. Tiwari, and D. A. Buchanan, *Appl. Phys. Lett.* 69, 1104 **(1996)**.
**125.** A. Ghetti, *Microelectronic Engineering* 59, 127 **(2001)**.

Received: 11 January 2005. Accepted: 24 January 2005.