# Modeling of Tunneling Current and Gate Dielectric Reliability for Nonvolatile Memory Devices

Andreas Gehring, *Member, IEEE,* and Siegfried Selberherr, *Fellow, IEEE*

***Abstract*—We present a hierarchy of tunneling models suitable for the two- and three-dimensional simulation of logic and nonvolatile semiconductor memory devices. The crucial modeling topics are comprehensively discussed, namely, 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 nonvolatile memory devices.**

***Index Terms*—Device simulation, gate dielectric breakdown, nonvolatile memory reliability, tunneling.**

## I. Introduction

FOR the prediction of performance and for the optimization of nonvolatile memory devices, accurate simulation of quantum-mechanical tunneling effects has always been of paramount interest [1]. The application area of tunneling models ranges from the prediction of gate leakage in MOS transistors, the evaluation of gate stacks for advanced high-$\kappa$ gate insulator materials, and the optimization of programming and erasing times in nonvolatile semiconductor memory cells, to the study of source-drain tunneling. As shown in the silicon-dielectric-silicon structure in Fig. 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), and QBS (quasi-bound state) tunneling processes. We denote as *direct tunneling* all processes which are not defect-assisted. In Fig. 1, 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 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. 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]. 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.

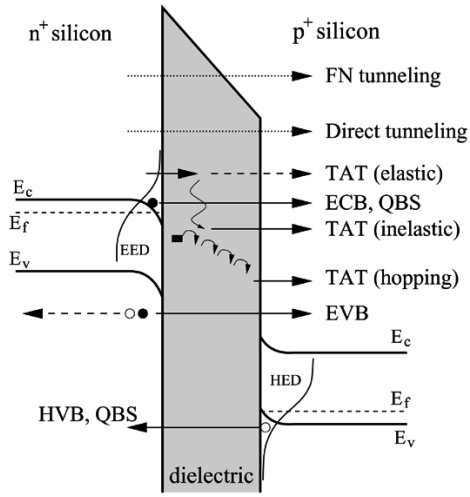



Fig. 1. Tunneling processes in a MOS structure. Direct tunneling processes (ECB, EVB, and HVB) are covered in Section II, while Section III deals with TAT transitions. Bound and quasi-bound states are studied in Section II.C.

Shrinking of gate dielectric thicknesses demands the use of alternative gate dielectrics such as $ZrO_2$, which often suffer from high defect densities [4]. Current transport through such dielectrics by means of defect-assisted tunneling has been studied intensely [4]–[6]. Furthermore, the gate dielectric reliability becomes a crucial issue not only for nonvolatile memories but also for logic applications. It is often assumed that 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 II, 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

Manuscript received June 24, 2004; revised August 18, 2004.

The authors are with the Institute for Microelectronics, Technical University Vienna, A-1040 Vienna, Austria (e-mail: gehring@iue.tuwien.ac.at).

Digital Object Identifier 10.1109/TDMR.2004.836727

and quasi-bound states as encountered in the inversion layer of a MOSFET is outlined. Typical results for MOS transistors are presented and compared with compact models. We present an example nonvolatile memory application utilizing a layered dielectric to allow independent tuning of on- and off-state currents. Section III 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 wrap up the main findings and give directions for further research.

## II. Direct Tunneling

The most prominent and almost exclusively used expression to describe direct tunneling transitions has been developed by Duke [7] and used by Tsu and Esaki to describe tunneling through a one-dimensional (1-D) superlattice [3]. It is commonly known as the Tsu–Esaki expression. The current density reads

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

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

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

The total energy $\mathcal{E}$ is the sum of a transversal component parallel to the Si–$SiO_2$ interface $\mathcal{E}_\rho$ and a transversal component $\mathcal{E}_x$. 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 $m_{3\mathrm{D}} = 2m_{\mathrm{t}} + 4\sqrt{m_{\mathrm{t}} m_{\mathrm{l}}}$, which, for (100) silicon with $m_{\mathrm{l}} = 0.92 m_0$ and $m_{\mathrm{t}} = 0.19 m_0$ equals $2.052 m_0$, and the electron mass in the dielectric $m_{\mathrm{diel}}$ [8].

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 isotropicity, the supply function evaluates to

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

where $\mathcal{E}_{\mathrm{F},1}$ and $\mathcal{E}_{\mathrm{F},2}$ 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 [9]. 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.

### A. 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., [10], [11]. The topic is of high importance, because the assumption of a cold Maxwellian distribution function underestimates the high-energy tail of the EED near the drain region [12]. 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 nonparabolic bands to check the validity of this approximation. Fig. 2 shows the contour lines of the heated Maxwellian EED in comparison to Monte Carlo results for a MOSFET with a gate length of $L_{\mathrm{g}} = 180$ nm at $V_{\mathrm{DS}} = V_{\mathrm{GS}} = 1$ V. The electron temperature was calculated in a post-processing step as $T_n = 2\langle\mathcal{E}\rangle/3k_B$.

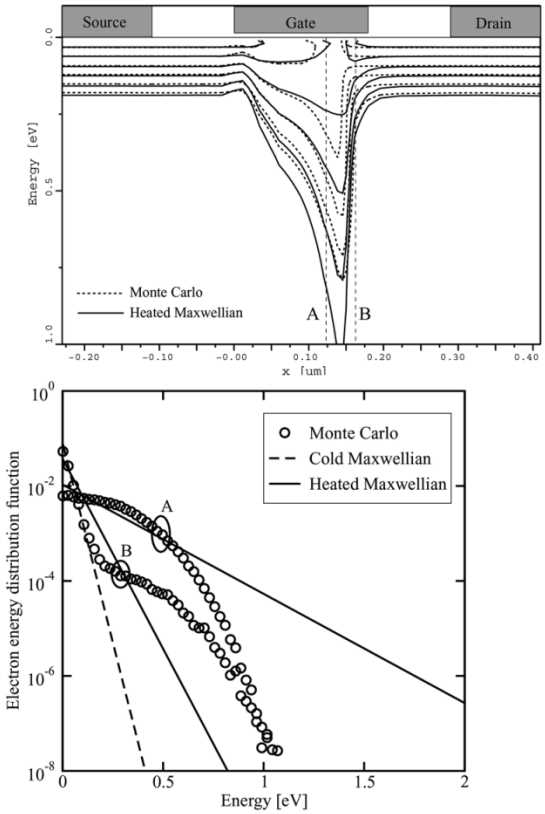



Fig. 2. Comparison of the heated Maxwellian distribution (full lines) with the results from a Monte Carlo simulation (dotted lines) in a turned-on 180 nm MOSFET. Neighboring lines differ by a factor of 10. The distributions at point A and B are compared with a cold Maxwellian in the lower figure.

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 $V_{\mathrm{GS}} > V_{\mathrm{DS}}$, the peak electric field in the channel is reduced, and the heated Maxwellian approximation delivers more reasonable results [13].

A quite generalized approach for the EED has been proposed by Grasser *et al.* [14]

$$f(\mathcal{E}) = A \exp\left( -\left( \frac{\mathcal{E}}{\mathcal{E}_{\mathrm{ref}}} \right)^b \right). \qquad (4)$$

In this expression the values of $\mathcal{E}_{\mathrm{ref}}$ and $b$ are mapped to the solution variables $T_n$ and $\beta_n$ of a six moments transport model [15]. The symbol $\beta_n$ denotes the normalized kurtosis of the distribution function ($\beta_n = 1$ for a Maxwellian distribution).

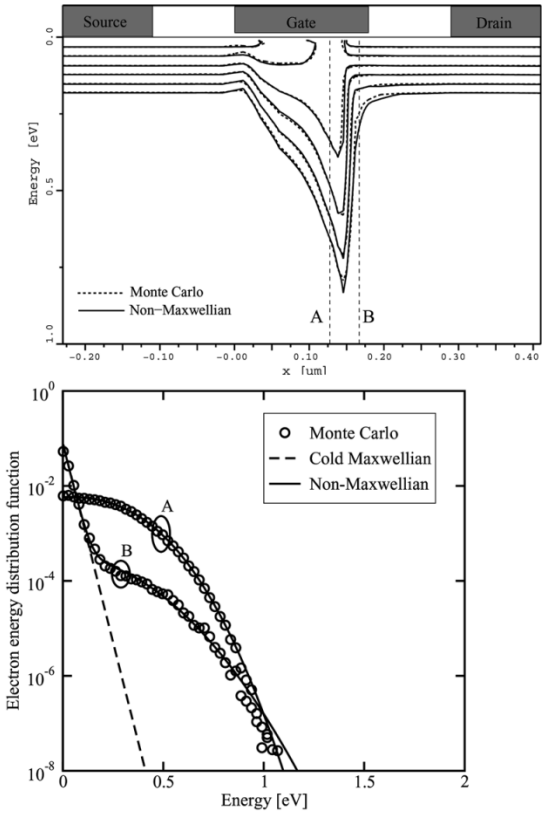



Fig. 3. Comparison of the non-Maxwellian distribution (full lines) with the results from a Monte Carlo simulation (dotted lines) in a turned-on 180 nm MOSFET. Neighboring lines differ by a factor of 10. The distributions at point A and B are compared with a cold Maxwellian in the lower figure.

Expression (4) 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 [12], [16]. Expression (4) does not explicitly account for this cold-carrier population. Therefore, when (4) is normalized to the actual carrier concentration, the high-energy tail is heavily overestimated [17], [18].

A distribution function accounting for this effect was proposed by Sonoda *et al.* [11], and an improved model has been suggested by Grasser *et al.* [12]:

$$f(\mathcal{E}) = A\left(\exp\left(-\left(\frac{\mathcal{E}}{\mathcal{E}_{\mathrm{ref}}}\right)^{b}\right) + c\exp\left(-\frac{\mathcal{E}}{k_B T_{\mathrm{L}}}\right)\right). \quad (5)$$

Here the pool of cold carriers in the drain region is correctly modeled by an additional cold Maxwellian subpopulation. The values of $\mathcal{E}_{\mathrm{ref}}$, $b$, and $c$ are again derived from the solution variables of a six moments transport model. Fig. 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 (4) and (5) are given in Appendix I. To check the impact of the distribution function, the integrand of the Tsu–Esaki formula, namely the expression $\mathrm{TC}(\mathcal{E})N(\mathcal{E})$,

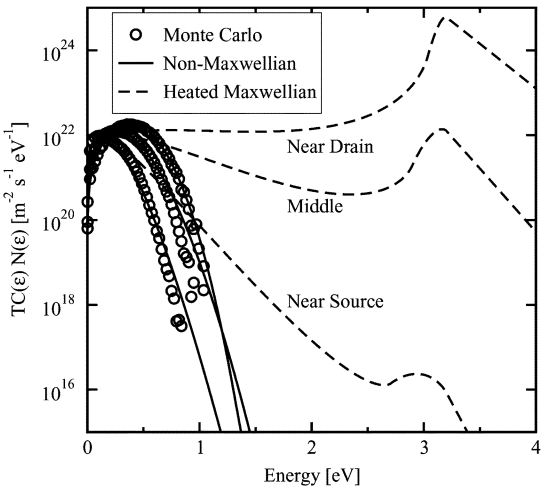



Fig. 4. Integrand of Tsu–Esaki's equation for a MOSFET with 100 nm gate length and 3 nm gate dielectric thickness at $V_{\mathrm{GS}} = V_{\mathrm{DS}} = 1$ V applying different models for the distribution function.

has been evaluated as shown in Fig. 4, and compared to post-processed Monte Carlo results. While at low energies the difference between the non-Maxwellian distribution function (5) and the heated Maxwellian distribution is negligible, the incremental gate current density is 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.

### *B. 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

$$j = \frac{\hbar}{2\imath m} \cdot (\Psi^* \cdot \boldsymbol{\nabla}\Psi - \boldsymbol{\nabla}\Psi^* \cdot \Psi) \quad (6)$$

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 be found by solving the stationary 1-D Schrödinger equation in the barrier region, which can be achieved using different numerical methods, such as the commonly applied Wentzel–Kramers–Brillouin (WKB) approximation or Gundlach's method [19]. Modern nonvolatile memories often rely on nonlinear energy barriers to increase the device performance [20]. The WKB method, however, does not account for wave function reflections, 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 II). However, several authors have noted numerical

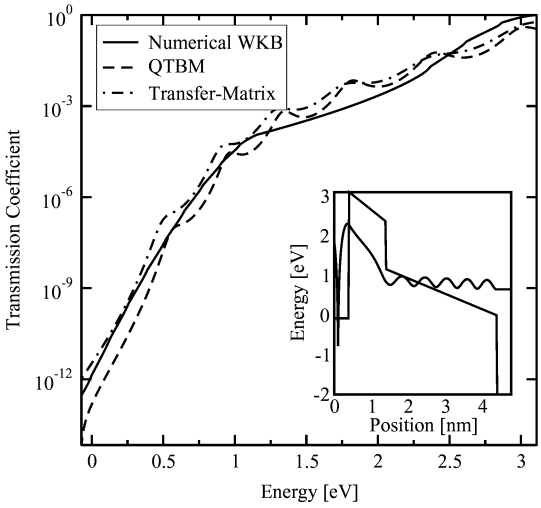



Fig. 5. The transmission coefficient using different methods for a dielectric consisting of two layers. The shape of the energy barrier and the wave function at 2.8 eV is shown in the inset.

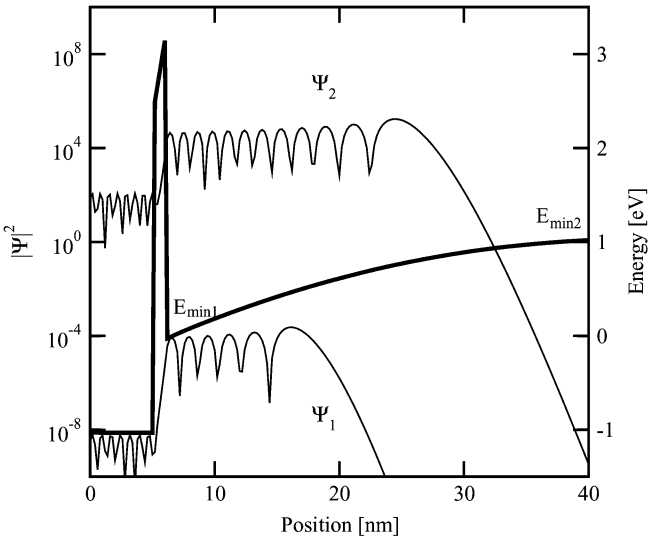



Fig. 6. Conduction band profile and two quasi-bound state wave functions. Quasi-bound state tunneling must be evaluated for $\mathcal{E}_{\min,1} < \mathcal{E} < \mathcal{E}_{\min,2}$, while the Tsu–Esaki expression must be used for $\mathcal{E} > \mathcal{E}_{\min,2}$.

problems in 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 [21]–[26].

An alternative method to compute the transmission coefficient is based on the quantum transmitting boundary (QTB) method [27], [28]. 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 1-D calculations, as it is usually the case for gate dielectric tunneling, a fast recursive solution procedure has been proposed by Ravaioli [29].

Fig. 5 shows the transmission coefficient for the different methods for a nonlinear 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. To the 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.

### C. 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 [30]. 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 [31]. 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 Fig. 6, which shows the conduction band edge and two quasi-bound state wave functions.

To account for tunneling current from both free (3-D) and quasi-bound (2-D) states, the Tsu–Esaki equation must be replaced by

$$\begin{aligned} J &= J_{\mathrm{2D}} + J_{\mathrm{3D}} \\ &= \frac{k_B T q}{\pi \hbar^2} \sum_{i,\nu} \frac{g_\nu m_{||}}{\tau_\nu(\mathcal{E}_{\nu,i}(m_q))} \ln\left(1 + \exp\left(\frac{\mathcal{E}_{\mathrm{F}} - \mathcal{E}_{\nu,i}}{k_B T}\right)\right) \\ &\quad + \frac{4\pi q m_{\mathrm{3D}}}{h^3} \int_{\mathcal{E}_{\min,2}}^{\mathcal{E}_{\max}} \mathrm{TC}(\mathcal{E}_x, m_{\mathrm{diel}}) N(\mathcal{E}_x)\, d\mathcal{E}_x \end{aligned} \tag{7}$$

where the symbols $g_\nu$, $m_{||}$, and $m_q$ denote the valley degeneracy, parallel, and quantization masses ($g = 2$: $m_{||} = m_{\mathrm{t}}$, $m_q = m_{\mathrm{l}}$, and $g = 4$: $m_{||} = \sqrt{m_{\mathrm{l}} m_{\mathrm{t}}}$, $m_q = m_{\mathrm{t}}$), and $\tau_\nu(\mathcal{E}_{\nu,i})$ is the life time of the quasi-bound state $\mathcal{E}_{\nu,i}$. 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 [32], the FWHM value of the reflection coefficient itself [33], or from the imaginary parts of the complex eigenvalues [28]. 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 entirely and use only the Tsu–Esaki formula (1) instead [34]–[36]. This formula, however, cannot reproduce the QBS tunneling component as shown in Fig. 7, where the QBS current $(J_{\mathrm{2D}})$ is compared to the continuum current $(J_{\mathrm{3D}})$.

The dotted lines indicate the continuum current $(J_{\mathrm{3D}})$ for $\mathcal{E}_{\min,2}$ as lower integration level (cf. Fig. 6), which is negligible

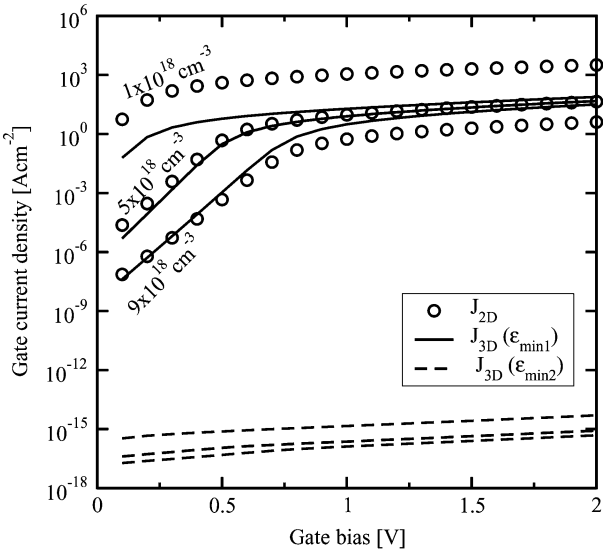



Fig. 7. Current density for different bulk doping and oxide thickness using only quasi-bound state tunneling ($J_{2D}$) and the Tsu–Esaki expression with $\mathcal{E}_{\min,1}$ ($J_{3D}(\mathcal{E}_{\min,1})$) or $\mathcal{E}_{\min,2}$ ($J_{3D}(\mathcal{E}_{\min,2})$) as lower integration level.

for this case. The full lines show $J_{3D}$ using $\mathcal{E}_{\min,1}$ 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 (7) and calculate the life times from the quasi-classical approach

$$\tau_\nu(\mathcal{E}_{\nu,i}) = \int_0^x \frac{\sqrt{2m_\nu/(\mathcal{E}_{\nu,i} - \mathcal{E}_c(\xi))}}{\mathrm{TC}(\mathcal{E}_{\nu,i})}\, d\xi \tag{8}$$

with $\mathcal{E}_c(x) = \mathcal{E}_{\nu,i}$ [37]. Furthermore, we keep the conventional shape of the Tsu–Esaki formula using $\mathcal{E}_{\min,2}$ as the lower integration level. To further reduce the computation time, the eigenvalues of the triangular well approximation

$$\mathcal{E}_{\nu,i} = -z_i \left(\frac{\hbar^2}{2m_\nu}\right)^{1/3} E^{2/3} \tag{9}$$

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 [38].

### D. 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.

*1) Barrier Heights:* 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 tradeoff 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) 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 $0.3m_0$–$0.5m_0$ for electrons and $0.32m_0$–$0.77m_0$ for holes. Note that for the assumption of a Franz-type dispersion relation [39], effective electron masses in the range of $0.41m_0$ to $0.61m_0$ have been found [40]–[44]. In the simulator MINIMOS-NT values of $m_{\mathrm{diel,e}} = 0.5m_0$ and $m_{\mathrm{diel,h}} = 0.8m_0$ have been applied, in accordance to the device simulator Dessis [34].

TABLE I
DIELECTRIC PERMITTIVITY, BAND GAP, AND BAND EDGE OFFSETS OF DIELECTRIC MATERIALS, TAKEN FROM [55], [99]–[104]

| | $\kappa/\kappa_0$ (1) | $\mathcal{E}_g$ (eV) | $\Delta\mathcal{E}_c$ (eV) | $\Delta\mathcal{E}_v$ (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 AND HOLE MASS IN $SiO_2$ [41]

| $t_{\mathrm{diel}}$ (nm) | $m_{\mathrm{diel,e}}/m_0$ (1) | $m_{\mathrm{diel,h}}/m_0$ (1) | Reference |
|---|---|---|---|
| 100 | 0.42 | | [49] |
| 100 – 12 | 0.5 | | [105] |
| 6 – 3 | 0.32 | | [106] |
| 3.5 – 1.5 | 0.5 | | [107] |
| 3.5 – 2.2 | 0.5 | | [44] |
| 6.5 – 1.56 | 0.5 | 0.42 | [108] |
| 5 – 2 | 0.437 | 0.437 | [109] |
| 3.6 – 1 | 0.4 | 0.32 | [2] |
| | 0.5 | 0.77 | [34] |

Note, however, that the assumption of a constant electron mass in the dielectric is no more justified for ultrathin $SiO_2$ layers. Here it was found both experimentally [45] and theoretically [46]–[48] by means of tight-binding simulations that the tunneling mass increases by almost 50% as the dielectric thickness is decreased down to 1 nm. The fit formula $m'_{\mathrm{diel,e}}/m_{\mathrm{diel,e}} = c + (at_{\mathrm{diel}})^{-b}$ has been proposed to describe the thickness dependence of the tunneling mass, with $m'_{\mathrm{diel,e}}$ being the corrected value and parameter values of $c = 0.706$, $a = 0.708$ nm$^{-1}$, and $b = 1.004$ for parabolic effective-mass calculations [48].

### E. Compact Models

For 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 [49]:

$$J = \frac{q^3 m_{\mathrm{eff}}}{8\pi m_{\mathrm{diel}} h q \Phi_B} E_{\mathrm{diel}}^2 \exp\left(-\frac{4\sqrt{2m_{\mathrm{diel}}(q\Phi_B)^3}}{3\hbar q E_{\mathrm{diel}}}\right). \tag{10}$$

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,

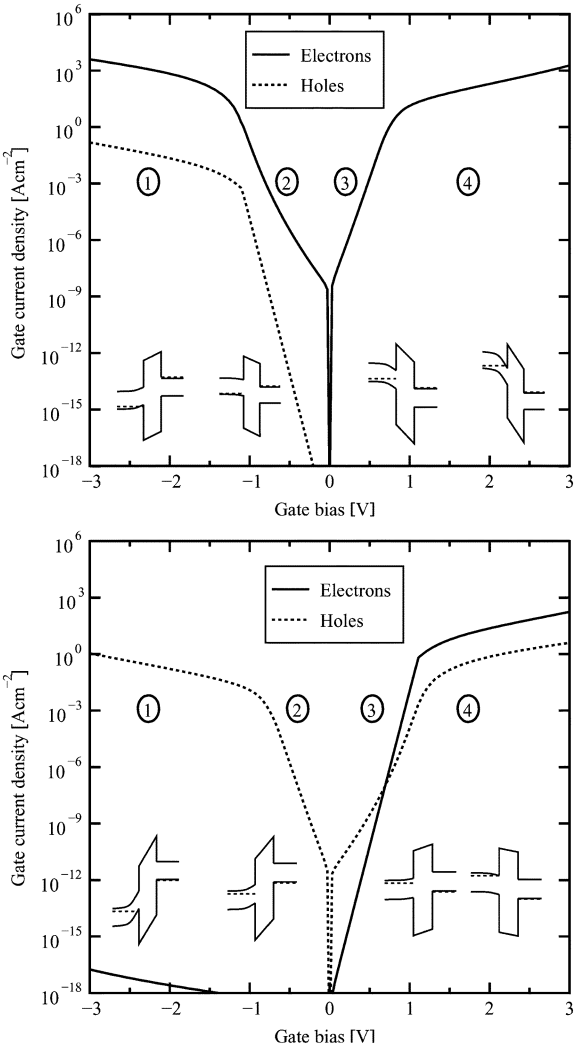



Fig. 8. Tunneling current components in an nMOS (top) and a pMOS (bottom) device with 2-nm dielectric thickness. The insets show the approximate shape of the band edge energies, with the gate contact located at the right side.

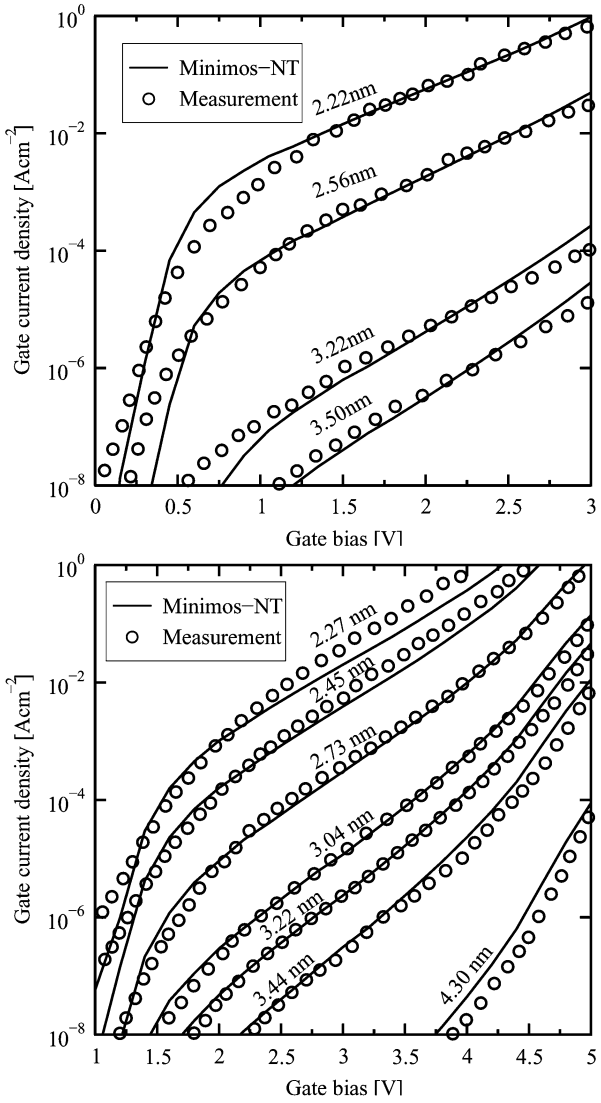



Fig. 9. Comparison of the gate current predicted by the Tsu–Esaki model using an analytical WKB method for the transmission coefficient with measurements of a nMOS (top) and pMOS (bottom) transistor [53].

it is not valid for direct tunneling where the barrier is of trapezoidal shape. Furthermore, $q\Phi_B$ 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 is often wrongly assumed.

Schuegraf and Hu derived correction terms for this expression to make it applicable to the regime of direct tunneling [50]:

$$J = \frac{q^3 m_{\text{eff}}}{8\pi m_{\text{diel}} h q \Phi_B B_1} E_{\text{diel}}^2 \exp\left(-\frac{4\sqrt{2m_{\text{diel}}(q\Phi_B)^3}B_2}{3\hbar q E_{\text{diel}}}\right) \tag{11}$$

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

$$B_1 = \left(1 - \left(1 - \frac{qE_{\text{diel}}t_{\text{diel}}}{q\Phi_B}\right)^{1/2}\right)^2 \tag{12}$$

$$B_2 = \left(1 - \left(1 - \frac{qE_{\text{diel}}t_{\text{diel}}}{q\Phi_B}\right)^{3/2}\right). \tag{13}$$

For a triangular barrier, the correction factors become $B_1 = B_2 = 1$ and the expression simplifies to (10). Note that (10) 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 [51], [52].

### *F. Simulation Results for MOS Transistors*

The typical shape of the gate current density in turned-off nMOS and pMOS devices is depicted in Fig. 8 [17]. A $SiO_2$ gate dielectric thickness of 2 nm and an acceptor/donor doping of $5\times 10^{17}$ 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 Fig. 9 for nMOS and pMOS devices [53]. The simulations

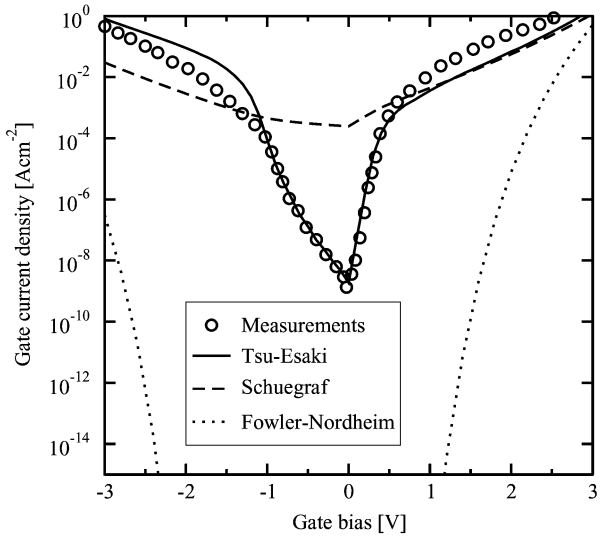



Fig. 10. Comparison of the compact model results with measurements of an nMOS structure [109].

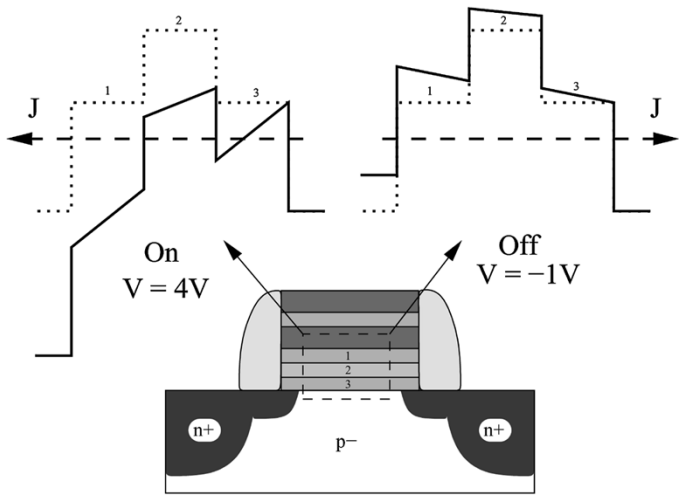



Fig. 11. Device structure and operating principle of a nonvolatile memory based on crested barriers [57].

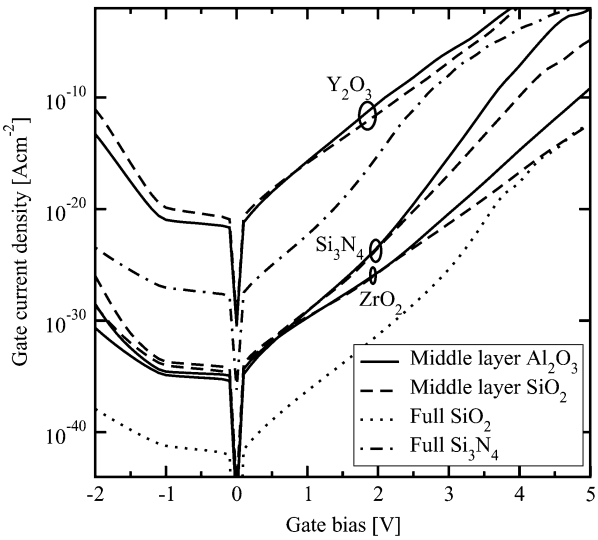



Fig. 12. Gate current density as a function of the gate bias for different materials of the middle layer, compared to full $SiO_2$ and $Si_3N_4$ layers.

in this figure have been performed using the device simulator MINIMOS-NT [54]. 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 Fig. 10 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 may be used to obtain an estimate of the gate current. The Fowler–Nordheim model totally fails for this application.

### *G. Nonvolatile Memories Based on Layered Dielectrics*

One of the most important figures of merit of a nonvolatile memory cell is its $I_{\rm on}/I_{\rm off}$ 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 the on- and off-currents independently. However, if the tunnel dielectric is replaced by a dielectric stack of varying barrier height, as shown in Fig. 11, it becomes possible. In this figure, the device structure and the conduction band edge in the on- and off-states 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 consists of a single layer, the peak of the energy barrier is not 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 $I_{\rm on}/I_{\rm off}$ 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 [55]. This device design was first proposed by Capasso *et al.* in 1988 [56] based on AlGaAs–GaAs devices and later used by several authors [20], [57], where it became popular as *crested-barrier* memory [57] or VARIOT (*varying oxide thickness* device [20]).

The gate current density of the device depicted in Fig. 11 is shown as a function of the gate bias in Fig. 12. A stack thickness of 5 nm was chosen. Since the middle layers must have a high band gap, only a 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 the middle layer show good ratios between the on-state (positive gate bias) current density and the off-state (negative gate bias) current density.

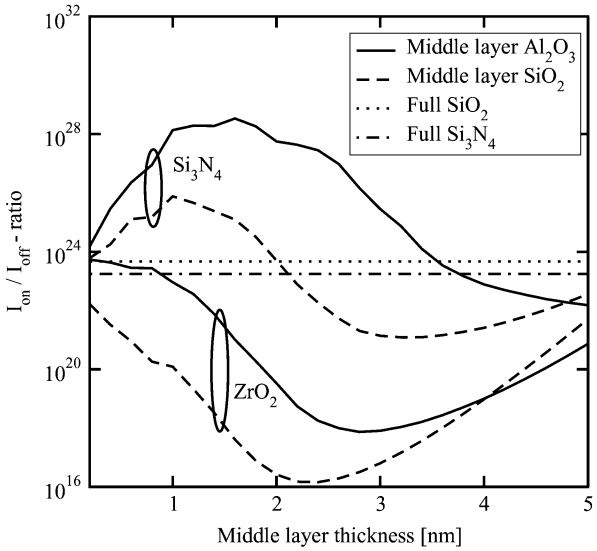



Fig. 13. Ratio between the on-current and the off-current as a function of the middle layer thickness for different materials of the outer layers ($Si_3N_4$ and $ZrO_2$) and middle layers ($Al_2O_3$ and $SiO_2$), compared to the resulting current density using full layers of $SiO_2$ and $Si_3N_4$.

The important figure of merit, however, is the $I_{\rm on}/I_{\rm off}$ ratio. In Fig. 13, the $I_{\rm on}/I_{\rm off}$ 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, trap-assisted tunneling (TAT) was neglected.

## III. 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 nonvolatile 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, SILC) is mainly responsible for the degradation of the retention time of these devices [58]. SILC has been widely studied and modeled in MOS capacitors [59]–[61] and EEPROM devices [62]. This section gives a brief overview of trap-assisted tunneling models and elaborates on an inelastic TAT model which was included in the device simulator MINIMOS-NT.

### A. Model Overview

A frequently used model is the generalized TAT model presented by Chang *et al.* [63], [64]. The current density reads

$$J = q\int_0^{t_{\rm diel}} AN_T(x)\frac{P_1(x)P_2(x)}{P_1(x)+P_2(x)}\,dx \tag{14}$$

where $A$ denotes a fitting constant, $N_T(x)$ 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.* [65]

$$J = \int_0^{t_{\rm diel}} C_T N_T(x)\frac{J_{\rm in}J_{\rm out}}{J_{\rm in}+J_{\rm out}}\,dx \tag{15}$$

who assumed a constant capture cross section $C_T$ for the traps. The symbols $J_{\rm in}$ and $J_{\rm out}$ denote the capture and emission currents. Essentially the same formula was used by other authors as well [66], [67]. Considerable research has been done by Ielmini *et al.* [68]–[71] who describe inelastic TAT and also take hopping conduction into account [72], [73]. They derive the trap-assisted current by an integration along the dielectric thickness and energy

$$J = \int_0^{t_{\rm diel}} dx \int_{\mathcal{E}_{\rm min}}^{\mathcal{E}_{\rm max}} \tilde{J}(\mathcal{E}_T,x)\,d\mathcal{E}$$

where $\tilde{J}$ 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

$$\begin{aligned}\tilde{J}(\mathcal{E}_T,x) &= J_{\rm cl}-J_{\rm el} = J_{\rm er}-J_{\rm cr}\\ &= qN_T'W_{\rm c}\left(1-\frac{f_T(\mathcal{E}_T,x)}{f_{\rm l}(\mathcal{E}_T,x)}\right)\end{aligned}$$

where $f_T$ is the trap occupancy, $\mathcal{E}_T$ the trap energy, $W_{\rm c}$ the capture rate, and $f_{\rm l}$ the energy distribution function at the left interface. The symbol $N_T'$ denotes the trap concentration in space and energy. Ielmini *et al.* further develop the model to include transient effects and note 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

$$(J_{\rm cl}-J_{\rm el})+(J_{\rm cr}-J_{\rm er}) = qN_T\frac{\partial f_T}{\partial t}. \tag{16}$$

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

Experimental evidence has been reported that SILC is caused by inelastic trap-assisted tunnel transitions [58], [74]–[78]. A detailed model for inelastic trap-assisted tunneling by means of multiphonon emission has been presented by Herrmann and Schenk [79] and modified versions have been used by other authors as well [80]–[83]. The TAT process is modeled via inelastic phonon-assisted transitions as shown in Fig. 14 [79], [81], [84].

Electrons are captured from the cathode, relax to the energy of the trap $\mathcal{E}_0$ by phonon emission with energy $m\hbar\omega$, and are emitted to the anode. The TAT current is found by integration over the dielectric thickness

$$J_{\rm t} = q\int_0^{t_{\rm diel}} \frac{N_T(x)}{\tau_{\rm c}(x)+\tau_{\rm e}(x)}\,dx \tag{17}$$

where $N_T(x)$ is the trap concentration and $\tau_{\rm c}(x)$ and $\tau_{\rm e}(x)$ denote the capture and emission times calculated from

$$\tau_{\rm c}^{-1}(z) = \int_{\mathcal{E}_0}^{\infty} c_n(\mathcal{E},x)T_{\rm l}(\mathcal{E})f_{\rm l}(\mathcal{E})\,d\mathcal{E} \tag{18}$$

$$\tau_{\rm e}^{-1}(z) = \int_{\mathcal{E}_0}^{\infty} e_n(\mathcal{E},x)T_{\rm r}(\mathcal{E})(1-f_{\rm r}(\mathcal{E}))\,d\mathcal{E}. \tag{19}$$

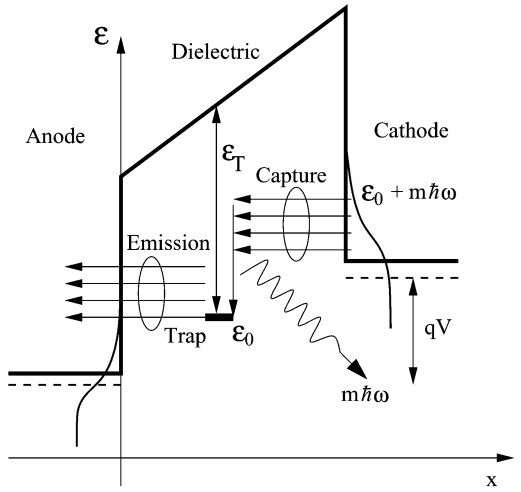



Fig. 14. Trap-assisted tunneling transition by inelastic phonon emission. Electrons are captured from the cathode, relax to the trap energy level $\mathcal{E}_0$ by the emission of phonons, and are emitted to the anode [79].

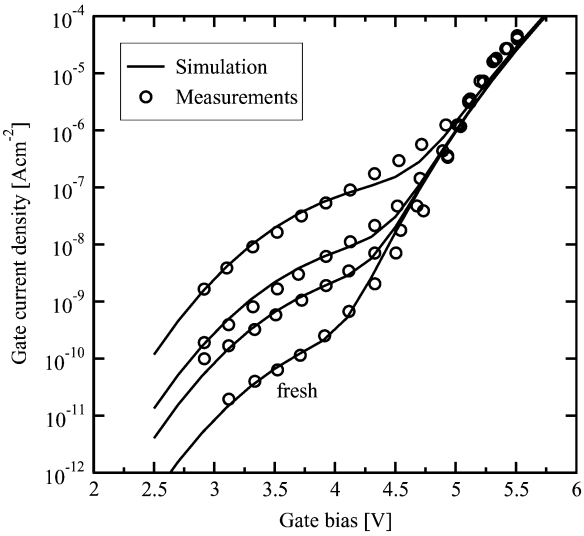



Fig. 15. Gate current density for different stress times [75] for $t_{\rm diel} = 5.5$ nm. The model parameters are $\mathcal{E}_T = 2.7$ eV, $\hbar\omega = 20$ meV, and $N_T = 9.0 \times 10^{17}$ cm$^{-3}$, $1.0 \times 10^{17}$ cm$^{-3}$, $3.0 \times 10^{16}$ cm$^{-3}$, and $3.0 \times 10^{15}$ cm$^{-3}$ (from top to bottom).

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

$$c_n(\mathcal{E},x) = c_0 \sum_m L_m \delta(\mathcal{E} - \mathcal{E}_m) \tag{20}$$

$$e_n(\mathcal{E},x) = c_0 \exp\left(-\frac{\mathcal{E}-\mathcal{E}_T}{k_B T_{\rm L}}\right) \sum_m L_m \delta(\mathcal{E} - \mathcal{E}_m) \tag{21}$$

with $c_0 = (4\pi)^2 r_T^2 (\hbar\Theta_0)^3/(\hbar\mathcal{E}_{\rm g,SiO_2})$ and $(\hbar\Theta_0) = (q^2\hbar^2F^2/(2m))^{1/3}$. The symbols $f_{\rm l}$ and $f_{\rm r}$ denote the Fermi distributions, $T_{\rm l}$ and $T_{\rm r}$ the transmission coefficients from the left and right side of the dielectric, $F$ the electric field in the dielectric, and $\mathcal{E}_{\rm g,SiO_2}$ 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. Fig. 15 shows a comparison with experimental data for MOS capacitors [75], where the transition from the TAT regime at low bias to the FN tunneling regime at high bias is clearly visible.

### C. 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

$$N_T(x)\frac{df_T(x,t)}{dt} = N_T(x)\frac{1-f_T(x,t)}{\tau_{\rm c}(x,t)} - N_T(x)\frac{f_T(x,t)}{\tau_{\rm e}(x,t)}$$

where $\tau_{\rm c}$ and $\tau_{\rm e}$ describe the capture and emission time of the trap. For the stationary case, the time derivative on the left-hand side is zero:

$$\frac{1-f_T(x,t)}{\tau_{\rm c}(x,t)} = \frac{f_T(x,t)}{\tau_{\rm e}(x,t)} = R(x). \tag{22}$$

From (22) and the incremental gate current density $dj(x) = qR(x)N_T(x)dx$, (17) can be derived [79]. 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 [83]

$$A_i = \frac{\tau_{\rm c}^{-1}(z,t_i)\Delta t_i}{1+C_i}$$

$$B_i = \frac{1-C_i}{1+C_i}$$

$$C_i = \frac{\tau_{\rm m}^{-1}(z,t_i)\Delta t_i}{2}.$$

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

$$J_{\rm l,r}(t_i) = q\int_0^{t_{\rm diel}} N_T(x)\tau_{\rm l,r}^{-1}(x,t_i)\,dx \tag{23}$$

where l, r denotes the considered interface (left or right) and the time constants $\tau_{\rm l}$ and $\tau_{\rm r}$ are calculated from

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

with the respective values of the capture and emission times to the left and right interface $\tau_{\rm cl,r}$ and $\tau_{\rm el,r}$. 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, Fig. 16 shows the step response of two pMOS capacitors with $ZrO_2$ dielectrics fabricated using metal–organic chemical vapor deposition (MOCVD) and afterwards annealed under different ambient conditions [4]. 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

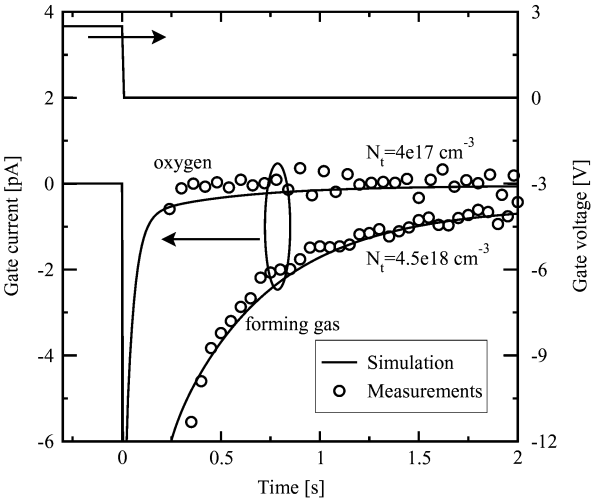



Fig. 16. Transient trap charging currents for a $ZrO_2$ layer fabricated by MOCVD and annealed under different ambient atmospheres [4], [85].

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 [85]. The measurements can be fitted assuming the trap concentrations indicated in the figure.

### *D. 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 nonvolatile memories over time [86]. 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 [87], [88]. Other models such as the anode hydrogen release (AHR) model [89] 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 [90]. A further model is the thermochemical model proposed by McPherson *et al.* [91], 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 [86], 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 [87], 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

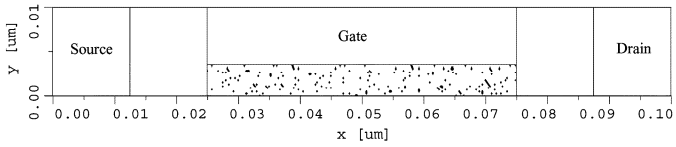



Fig. 17. Two-dimensional cut through the dielectric layer simulated with MINIMOS-NT and showing the random trap placement (darks spots).

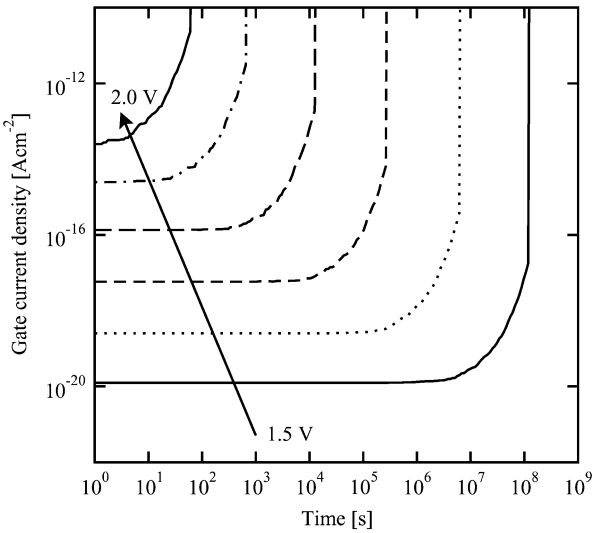



Fig. 18. Dielectric breakdown of a 3-nm $SiO_2$ layer.

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 [83]. The neutral defects create percolation paths in the dielectric, which eventually connect the gate with the substrate [92].

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 $Q_{\mathrm{i}}$ via $N_T = CQ_{\mathrm{i}}^{\alpha}$, as proposed by Degraeve *et al.* [93], 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. Fig. 17 shows a cross section of the gate dielectric layer where the dark spots mark traps. The corresponding gate current density is shown in Fig. 18 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 after breakdown can be described by a point contact conduction model [94]. In this model, the gate current is related to the gate voltage by a simple power law $I = KV_{\mathrm{G}}^{p}$, where the

parameter $K$ reflects the size of the breakdown spot, and the parameter $p$ is in the range of 2–5 [95]–[97]. Miranda *et al.* [95] 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.

## IV. Model Comparison

This paper has 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 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 (*l* for low, *m* for middle, and *h* for high). The models are abbreviated by FN (Fowler–Nordheim), SM (Schuegraf model), TA (Tsu–Esaki model with analytic WKB transmission coefficient), TG (Tsu–Esaki model with Gundlach transmission coefficient), TN (Tsu–Esaki model with numeric WKB transmission coefficient), TT (Tsu–Esaki model with transfer-matrix method), TQ (Tsu–Esaki model with QTB method), and IT (inelastic trap-assisted tunneling model). The following points can be concluded:

- The Fowler-Nordheim and Schuegraf models especially 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 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. Resonances in the transmission coefficient can only be resolved by means of the transfer-matrix or QTB methods.
- 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 (24) or (25). Using a cold Maxwellian distribution instead leads to an underestimation of the gate current density.

## V. Summary and Conclusion

We have 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 to—and not instead of—the conventional Tsu–Esaki formula. 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-$\kappa$ insulator reliability issues, which are still not fully understood.

TABLE III
A Hierarchy of Tunneling Models and Their Properties

| | FN | SM | TA | TG | TN | TT | TQ | IT |
|---|---|---|---|---|---|---|---|---|
| FN tunneling | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Direct tunneling | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| EVB tunneling | | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| QM oscillations | | | | ✓ | | ✓ | ✓ | |
| Dielectric stacks | | | | | ✓ | ✓ | ✓ | |
| Static TAT | | | | | | | | ✓ |
| Trap occupancy | | | | | | | | ✓ |
| Transient TAT | | | | | | | | ✓ |
| Effort | l | l | m | m | h | h | h | m |

## Appendix I
## Supply Function for Non-Maxwellian Distributions

With expression (4) for the distribution function and the assumption of a Fermi-Dirac distribution in the polysilicon gate, the supply function (2) becomes

$$N(\mathcal{E}) = A_1 \frac{\mathcal{E}_{\mathrm{ref}}}{b} \Gamma_{\mathrm{i}} \left( \frac{1}{b}, \left( \frac{\mathcal{E}}{\mathcal{E}_{\mathrm{ref}}} \right)^b \right) - A_2 k_B T_{\mathrm{L}} \ln \left( 1 + \exp \left( - \frac{\mathcal{E} + \Delta\mathcal{E}_{\mathrm{c}}}{k_B T_{\mathrm{L}}} \right) \right) \quad (24)$$

where $\Gamma_{\mathrm{i}}(\alpha, \beta)$ denotes the incomplete gamma function. In (24) the explicit value of the Fermi energy was replaced by the shift of the two conduction band edges $\Delta\mathcal{E}_{\mathrm{c}}$. Using the accurate shape of the distribution (5), the expression for the supply function becomes

$$N(\mathcal{E}) = A_1 \frac{\mathcal{E}_{\mathrm{ref}}}{b} \Gamma_{\mathrm{i}} \left( \frac{1}{b}, \left( \frac{\mathcal{E}}{\mathcal{E}_{\mathrm{ref}}} \right)^b \right) A_1 c k_B T_2 \exp \left( - \frac{\mathcal{E}}{k_B T_{\mathrm{L}}} \right) - A_2 k_B T_{\mathrm{L}} \ln \left( 1 + \exp \left( - \frac{\mathcal{E} + \Delta\mathcal{E}_{\mathrm{c}}}{k_B T_{\mathrm{L}}} \right) \right) \quad (25)$$

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

## APPENDIX II
## 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 [98] $\Psi_j(x) = A_j \exp(\imath k_j x) + B_j \exp(-\imath k_j x)$ with the wave number $k_j = \sqrt{2m_j(\mathcal{E} - W_j)}/\hbar$. 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:

$$\begin{pmatrix} A_j \\ B_j \end{pmatrix} = \underline{T}_j \begin{pmatrix} A_{j-1} \\ B_{j-1} \end{pmatrix}, \qquad 2 \le j \le N. \tag{26}$$

The transfer matrices are of the form

$$\underline{T}_j = \frac{1}{2} \begin{pmatrix} \alpha_j \gamma^{-k_j} & \beta_j \gamma^{-k_j} \\ \beta_j \gamma^{k_j} & \alpha_j \gamma^{k_j} \end{pmatrix} \begin{pmatrix} \gamma^{k_{j-1}} & 0 \\ 0 & \gamma^{-k_{j-1}} \end{pmatrix} \tag{27}$$

with $\alpha_j = 1 + k_{j-1}/k_j$, $\beta_j = 1 - k_{j-1}/k_j$, $2 \le j \le N$, and 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:

$$\begin{pmatrix} A_N \\ B_N \end{pmatrix} = \prod_{j=2\ldots N} \underline{T}_j \begin{pmatrix} A_1 \\ B_1 \end{pmatrix}. \tag{28}$$

If it is assumed that there is no reflected wave in Region N and the amplitude of the incident wave is unity, (28) simplifies to

$$\begin{pmatrix} A_N \\ 0 \end{pmatrix} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \begin{pmatrix} 1 \\ B_1 \end{pmatrix} \tag{29}$$

and the transmission coefficient can be calculated from (6). Note that the straightforward calculation of $A_N$ from $A_N = T_{11} - T_{12}T_{21}/T_{22}$ 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 $A_N = 1/T_{22}$.

## ACKNOWLEDGMENT

The support of H. Kosina, T. Grasser, F. Jiménez-Molinos, and S. Harasek is gratefully acknowledged. The authors also thank the anonymous reviewers for pointing out various deficiencies of the original manuscript.

## REFERENCES

[1] R. Bez, E. Camerlenghi, A. Modelli, and A. Visconti, "Introduction to flash memory," *Proc. IEEE*, vol. 91, pp. 489–502, Apr. 2003.

[2] W.-C. Lee and C. Hu, "Modeling CMOS tunneling currents through ultrathin gate oxide due to conduction- and valence-band electron and hole tunneling," *IEEE Trans. Electron Devices*, vol. 48, pp. 1366–1373, July 2001.

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

[4] S. Harasek, H. D. Wanzenböck, and E. Bertagnolli, "Compositional and electrical properties of zirconium dioxide thin films chemically deposited on silicon," *J. Vac. Sci. Technol. A*, vol. 21, no. 3, pp. 653–659, 2003.

[5] M. Houssa, M. Tuominen, M. Naili, V. Afanas'ev, A. Stesmans, S. Haukka, and M. M. Heyns, "Trap-assisted tunneling in high permittivity gate dielectric stacks," *J. Appl. Phys.*, vol. 87, no. 12, pp. 8615–8620, 2000.

[6] A. Kumar, M. V. Fischetti, T. H. Ning, and E. Gusev, "Hot-carrier charge trapping and trap generation in $HfO_2$ and $Al_2O_3$ field-effect transistors," *J. Appl. Phys.*, vol. 94, no. 3, pp. 1728–1737, 2003.

[7] C. B. Duke, *Tunneling in Solids*. New York: Academic, 1969.

[8] Khairurrijal, W. Mizubayashi, S. Miyazaki, and M. Hirose, "Analytic model of direct tunnel current through ultrathin gate oxides," *J. Appl. Phys.*, vol. 87, no. 6, pp. 3000–3005, 2000.

[9] M. Lundstrom and Z. Ren, "Essential physics of carrier transport in nanoscale MOSFETs," *IEEE Trans. Electron Devices*, vol. 49, pp. 133–141, Jan. 2002.

[10] D. Cassi and B. Riccó, "An analytical model of the energy distribution of hot electrons," *IEEE Trans. Electron Devices*, vol. 37, pp. 1514–1521, June 1990.

[11] K.-I. Sonoda, M. Yamaji, K. Taniguchi, C. Hamaguchi, and S. T. Dunham, "Moment expansion approach to calculate impact ionization rate in submicron silicon devices," *J. Appl. Phys.*, vol. 80, no. 9, pp. 5444–5448, 1996.

[12] T. Grasser, H. Kosina, C. Heitzinger, and S. Selberherr, "Characterization of the hot electron distribution function using six moments," *J. Appl. Phys.*, vol. 91, no. 6, pp. 3869–3879, 2002.

[13] L. Selmi, A. Ghetti, R. Bez, and E. Sangiorgi, "Trade-offs between tunneling and hot-carrier injection in short channel floating gate MOSFETs," *Microelectron. Eng.*, vol. 36, no. 1–4, pp. 293–296, 1997.

[14] T. Grasser, H. Kosina, and S. Selberherr, "An impact ionization model including non-Maxwellian and nonparabolicity effects," in *Proc. Int. Conf. Simulation of Semiconductor Processes and Devices*, 2001, pp. 46–49.

[15] ——, "Influence of the distribution function shape and the band structure on impact ionization modeling," *J. Appl. Phys.*, vol. 90, no. 12, pp. 6165–6171, 2001.

[16] A. Abramo and C. Fiegna, "Electron energy distributions in silicon structures at low applied voltages and high electric fields," *J. Appl. Phys.*, vol. 80, no. 2, pp. 889–893, 1996.

[17] A. Gehring, T. Grasser, H. Kosina, and S. Selberherr, "Simulation of hot-electron oxide tunneling current based on a non-Maxwellian electron energy distribution function," *J. Appl. Phys.*, vol. 92, no. 10, pp. 6019–6027, 2002.

[18] ——, "Energy transport gate current model accounting for non-Maxwellian energy distribution," *Electron. Lett.*, vol. 39, no. 8, pp. 691–692, 2003.

[19] K. H. Gundlach, "Zur Berechnung des Tunnelstroms durch eine trapezförmige Potentialstufe," *Solid-State Electron.*, vol. 9, pp. 949–957, 1966.

[20] B. Govoreanu, P. Blomme, M. Rosmeulen, J. Van Houdt, and K. De Meyer, "VARIOT: A novel multilayer tunnel barrier concept for low-voltage nonvolatile memory devices," *IEEE Electron Device Lett.*, vol. 24, pp. 99–101, Feb. 2003.

[21] G. Wachutka, "New layer method for the investigation of the electronic properties of two-dimensional periodic spatial structures: First applications to copper and aluminum," *Phys. Rev. B*, vol. 34, no. 12, pp. 8512–8527, 1986.

[22] D. Y. K. Ko and J. C. Inkson, "Matrix method for tunneling in heterostructures: Resonant tunneling in multilayer systems," *Phys. Rev. B*, vol. 38, no. 14, pp. 9945–9951, 1988.

[23] D. Z. Y. Ting, E. T. Yu, and T. C. McGill, "Multiband treatment of quantum transport in interband tunnel devices," *Phys. Rev. B*, vol. 45, no. 7, pp. 3583–3592, 1992.

[24] T. Usuki, M. Saito, M. Takatsu, R. A. Kiehl, and N. Yokoyama, "Numerical analysis of ballistic-electron transport in magnetic fields by using a quantum point contact and a quantum wire," *Phys. Rev. B*, vol. 52, no. 11, pp. 8244–8258, 1995.

[25] B. A. Biegel, "Quantum electronic device simulation," dissertation, Stanford Univ., Stanford, CA, 1997.

[26] F. Heinz, "Simulation approaches for nanoscale semiconductor devices," dissertation, ETH Zürich, Switzerland, 2004.

[27] C. S. Lent and D. J. Kirkner, "The quantum transmitting boundary method," *J. Appl. Phys.*, vol. 67, no. 10, pp. 6353–6359, 1990.

[28] W. R. Frensley, "Numerical evaluation of resonant states," *Superlattices and Microstructures*, vol. 11, no. 3, pp. 347–350, 1992.

[29] U. Ravaioli, "Numerical methods for the solution of Schrödinger equation for ballistic transport," presented at the 2002 School on Computational Material Science, Univ. Illinois at Urbana-Champaign, http://www.mcc.uiuc.edu/SummerSchool02/.

[30] F. Stern, "Self-consistent results for n-type Si inversion layers," *Phys. Rev. B*, vol. 5, no. 12, pp. 4891–4899, 1972.

[31] W. Magnus and W. Schoenmaker, "On the calculation of gate tunneling currents in ultra-thin metal-insulator-semiconductor capacitors," *Microelectron. Reliabil.*, vol. 41, no. 1, pp. 31–35, 2001.

[32] E. Cassan, "On the reduction of direct tunneling leakage through ultrathin gate oxides by a one-dimensional Schrödinger-Poisson solver," *J. Appl. Phys.*, vol. 87, no. 11, pp. 7931–7939, 2000.

[33] R. Clerc, A. Spinelli, G. Ghibaudo, and G. Pananakakis, "Theory of direct tunneling current in metal-oxide-semiconductor structures," *J. Appl. Phys.*, vol. 91, no. 3, pp. 1400–1409, 2002.

[34] *DESSIS 9.5 User's Manual*, Integrated Systems Engineering (ISE), Mountain View, CA, 2004.

[35] *MEDICI 2002.4.0 User's Manual*, Synopsys, Mountain View, CA, 2003.

[36] *ATLAS User's Manual*, Silvaco, Santa Clara, CA, 2004.

[37] A. D. Serra, A. Abramo, P. Palestri, L. Selmi, and F. Widdershoven, "Closed- and open-boundary models for gate-current calculation in n-MOSFETs," *IEEE Trans. Electron Devices*, vol. 48, pp. 1811–1815, Aug. 2001.

[38] A. Gehring and S. Selberherr, "On the calculation of quasi-bound states and their impact on direct tunneling in CMOS devices," in *Proc. Int. Conf. Simulation of Semiconductor Processes and Devices*, 2004.

[39] W. Franz, *Handbuch der Physik*. Berlin, Germany: Springer, 1956, vol. XVII, p. 155.

[40] J. Maserjian, "Tunneling in thin MOS structures," *J. Vac. Sci. Technol.*, vol. 11, no. 6, pp. 996–1003, 1974.

[41] R. Clerc, "Etude des Effets Quantiques dans les Composants CMOS a Oxydes de Grille Ultra Minces—Modelization et Caracterization," dissertation, Institut National Polytechnique de Grenoble, France, 2001.

[42] M. Av-Ron, M. Shatzkes, T. H. DiStefano, and R. A. Gdula, "Electron tunneling at Al-$SiO_2$ interfaces," *J. Appl. Phys.*, vol. 52, no. 4, pp. 2897–2908, 1981.

[43] S. H. Lo, D. A. Buchanan, Y. Taur, and W. Wang, "Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFETs," *IEEE Trans. Electron Devices*, vol. 18, pp. 209–211, May 1997.

[44] L. F. Register, E. Rosenbaum, and K. Yang, "Analytic model for direct tunneling current in polycristalline silicon-gate metal-oxide-semiconductor devices," *Appl. Phys. Lett.*, vol. 74, no. 3, pp. 457–459, 1999.

[45] R. Ludeke, E. Cartier, and A. Schenk, "Determination of the energy-dependent conduction band mass in $SiO_2$," *Appl. Phys. Lett.*, vol. 75, no. 10, pp. 1407–1409, 1999.

[46] M. Städele, B. R. Tuttle, and K. Hess, "Tunneling through ultrathin $SiO_2$ gate oxides from microscopic models," *J. Appl. Phys.*, vol. 89, no. 1, pp. 348–363, 2001.

[47] M. Städele, B. Fischer, B. R. Tuttle, and K. Hess, "Resonant electron tunneling through defects in ultrathin $SiO_2$ gate oxides in MOSFETs," *Solid-State Electron.*, vol. 46, no. 7, pp. 1027–1032, 2002.

[48] M. Städele, F. Sacconi, A. Di Carlo, and P. Lugli, "Enhancement of the effective tunnel mass in ultrathin silicon dioxide layers," *J. Appl. Phys.*, vol. 93, no. 5, pp. 2681–2690, 2003.

[49] M. Lenzlinger and E. H. Snow, "Fowler-Nordheim tunneling into thermally grown $SiO_2$," *J. Appl. Phys.*, vol. 40, no. 1, pp. 278–283, 1969.

[50] K. F. Schuegraf, C. C. King, and C. Hu, "Ultra-thin silicon dioxide leakage current and scaling limit," in *Symp. VLSI Technol. Tech. Dig.*, 1992, pp. 18–19.

[51] J. P. Shiely, "Simulation of tunneling in MOS devices," dissertation, Duke Univ., Durham, NC, 1999.

[52] A. Gehring, "Simulation of tunneling in semiconductor devices," dissertation, Technische Universität Wien, Austria, 2003.

[53] S. H. Lo, D. A. Buchanan, and Y. Taur, "Modeling and characterization of quantization, polysilicon depletion, and direct tunneling effects in MOSFET's with ultrathin oxides," *IBM J. Res. Dev.*, vol. 43, no. 3, pp. 327–337, 1999.

[54] *MINIMOS-NT 2.1 User's Guide*, Institut für Mikroelektronik, Technische Universität Wien, Austria, 2004.

[55] J. D. Casperson, L. D. Bell, and H. A. Atwater, "Materials issues for layered tunnel barrier structures," *J. Appl. Phys.*, vol. 92, no. 1, pp. 261–267, 2002.

[56] F. Capasso, F. Beltram, R. J. Malik, and J. F. Walker, "New floating-gate AlGaAs/GaAs memory devices with graded-gap electron injector and long retention times," *IEEE Electron Device Lett.*, vol. 9, pp. 377–379, Aug. 1988.

[57] K. K. Likharev, "Layered tunnel barriers for nonvolatile memory devices," *Appl. Phys. Lett.*, vol. 73, no. 15, pp. 2137–2139, 1998.

[58] S. Aritome, R. Shirota, G. Hemink, T. Endoh, and F. Masuoka, "Reliability issues of flash memory cells," *Proc. IEEE*, vol. 81, pp. 776–788, May 1993.

[59] B. Riccó, G. Gozzi, and M. Lanzoni, "Modeling and simulation of stress-induced leakage current in ultrathin $SiO_2$ Films," *IEEE Trans. Electron Devices*, vol. 45, pp. 1554–1560, July 1998.

[60] K. Sakakibara, N. Ajika, K. Eikyu, K. Ishikawa, and H. Miyoshi, "A quantitative analysis of time-decay reproducible stress-induced leakage current in $SiO_2$ films," *IEEE Trans. Electron Devices*, vol. 44, pp. 1002–1008, June 1997.

[61] A. Ghetti, E. Sangiorgi, J. Bude, T. W. Sorsch, and G. Weber, "Tunneling into interface states as reliability monitor for ultrathin oxides," *IEEE Trans. Electron Devices*, vol. 47, pp. 2358–2365, Dec. 2000.

[62] C.-M. Yih, Z.-H. Ho, M.-S. Liang, and S. S. Chung, "Characterization of hot-hole injection induced SILC and related disturbs in flash memories," *IEEE Trans. Electron Devices*, vol. 48, pp. 300–306, Feb. 2001.

[63] W. J. Chang, M. P. Houng, and Y. H. Wang, "Simulation of stress-induced leakage current in silicon dioxides: A modified trap-assisted tunneling model considering Gaussian-distributed traps and electron energy loss," *J. Appl. Phys.*, vol. 89, no. 11, pp. 6285–6293, 2001.

[64] ——, "Electrical properties and modeling of ultrathin impurity-doped silicon dioxides," *J. Appl. Phys.*, vol. 90, no. 10, pp. 5171–5179, 2001.

[65] A. Ghetti, A. Hamad, P. J. Silverman, H. Vaidya, and N. Zhao, "Self-consistent simulation of quantization effects and tunneling current in ultra-thin gate oxide MOS devices," in *Proc. Int. Conf. Simulation of Semiconductor Processes and Devices*, 1999, pp. 239–242.

[66] M. Lenski, T. Endoh, and F. Masuoka, "Analytical modeling of stress-induced leakage currents in 5.1–9.6 nm-thick silicon-dioxide films based on two-step inelastic trap-assisted tunneling," *J. Appl. Phys.*, vol. 88, no. 9, pp. 5238–5245, 2000.

[67] L. Larcher, A. Paccagnella, and G. Ghidini, "A model of the stress induced leakage current in gate oxides," *IEEE Trans. Electron Devices*, vol. 48, pp. 285–288, Feb. 2001.

[68] D. Ielmini, A. S. Spinelli, M. A. Rigamonti, and A. L. Lacaita, "Modeling of SILC based on electron and hole tunneling—Part I: Transient effects," *IEEE Trans. Electron Devices*, vol. 47, pp. 1258–1265, June 2000.

[69] ——, "Modeling of SILC based on electron and hole tunneling—Part II: Steady-state," *IEEE Trans. Electron Devices*, vol. 47, pp. 1266–1272, June 2000.

[70] D. Ielmini, A. S. Spinelli, A. L. Lacaita, A. Martinelli, and G. Ghidini, "A recombination- and trap-assisted tunneling model for stress-induced leakage current," *Solid-State Electron.*, vol. 45, no. 8, pp. 1361–1369, 2001.

[71] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and G. Ghidini, "Modeling of stress-induced leakage current and impact ionization in MOS devices," *Solid-State Electron.*, vol. 46, no. 3, pp. 417–422, 2002.

[72] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and A. Modelli, "A new two-trap tunneling model for the anomalous stress-induced leakage current (SILC) in flash memories," *Microelectron. Eng.*, vol. 59, no. 1–4, pp. 189–195, 2001.

[73] ——, "Modeling of anomalous SILC in flash memories based on tunneling at multiple defects," *Solid-State Electron.*, vol. 46, no. 11, pp. 1749–1756, 2002.

[74] R. Moazzami and C. Hu, "Stress-induced current in thin silicon dioxide films," in *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, 1992, pp. 139–142.

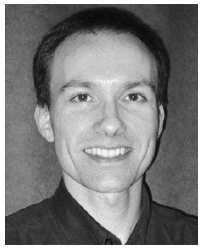
[75] E. Rosenbaum and L. F. Register, "Mechanism of stress-induced leakage current in MOS capacitors," *IEEE Trans. Electron Devices*, vol. 44, pp. 317–323, Feb. 1997.

[76] S.-I. Takagi, N. Yasuda, and A. Toriumi, "A new I-V model for stress-induced leakage current including inelastic tunneling," *IEEE J. Solid-State Circuits*, vol. 46, pp. 348–354, Feb. 1999.

[77] R. Rofan and C. Hu, "Stress-induced oxide leakage," *IEEE Electron Device Lett.*, vol. 12, pp. 632–634, Nov. 1991.

[78] J. Wu, L. F. Register, and E. Rosenbaum, "Trap-assisted tunneling current through ultra-thin oxide," in *Proc. Int. Reliability Physics Symp.*, 1999, pp. 389–395.

[79] M. Herrmann and A. Schenk, "Field and high-temperature dependence of the long term charge loss in erasable programmable read only memories: Measurements and modeling," *J. Appl. Phys.*, vol. 77, no. 9, pp. 4522–4540, 1995.
[80] A. Palma, A. Godoy, J. A. Jimenez-Tejada, J. E. Carceller, and J. A. Lopez-Villanueva, "Quantum two-dimensional calculation of time constants of random telegraph signals in metal-oxide-semiconductor structures," *Phys. Rev. B*, vol. 56, no. 15, pp. 9565–9574, 1997.
[81] F. Jiménez-Molinos, A. Palma, F. Gámiz, J. Banqueri, and J. A. Lopez-Villanueva, "Physical model for trap-assisted inelastic tunneling in metal-oxide-semiconductor structures," *J. Appl. Phys.*, vol. 90, no. 7, pp. 3396–3404, 2001.
[82] F. Jiménez-Molinos, F. Gámiz, A. Palma, P. Cartujo, and J. A. Lopez-Villanueva, "Direct and trap-assisted elastic tunneling through ultrathin gate oxides," *J. Appl. Phys.*, vol. 91, no. 8, pp. 5116–5124, 2002.
[83] A. Gehring, F. Jiménez-Molinos, H. Kosina, A. Palma, F. Gámiz, and S. Selberherr, "Modeling of retention time degradation due to inelastic trap-assisted tunneling in EEPROM devices," *Microelectron. Reliabil.*, vol. 43, no. 9–11, pp. 1495–1500, 2003.
[84] L. Larcher, "Statistical simulation of leakage currents in MOS and flash memory devices with a new multiphonon trap-assisted tunneling model," *IEEE Trans. Electron Devices*, vol. 50, pp. 1246–1253, May 2003.
[85] A. Gehring, S. Harasek, E. Bertagnolli, and S. Selberherr, "Evaluation of $ZrO_2$ gate dielectrics for advanced CMOS devices," in *Proc. Eur. Solid-State Device Research Conf.*, J. Franca and P. Freitas, Eds., 2003, pp. 473–476.
[86] A. Ghetti, *Gate Oxide Reliability: Physical and Computational Models*. New York: Springer, 2004, pp. 201–258.
[87] M. Alam, B. Weir, J. Bude, P. Silverman, and A. Ghetti, "A computational model for oxide breakdown: Theory and experiments," *Microelectron. Eng.*, vol. 89, no. 1–4, pp. 137–147, 2001.
[88] D. J. DiMaria and J. H. Stathis, "Anode hole injection, defect generation, and breakdown in ultrathin silicon dioxide films," *J. Appl. Phys.*, vol. 89, no. 9, pp. 5015–5024, 2001.
[89] J. H. Stathis, "Physical and predictive models of ultrathin oxide reliability in CMOS devices and circuits," *IEEE Trans. Device Mater. Reliabil.*, vol. 1, pp. 43–59, Mar. 2001.
[90] J. Wu, E. Rosenbaum, B. MacDonald, E. Li, J. Tao, B. Tracy, and P. Fang, "Anode hole injection versus hydrogen release: The mechanism for gate oxide breakdown," in *Proc. Int. Reliability Physics Symp.*, 2000, pp. 27–32.
[91] W. McPherson, R. B. Khamankar, and A. Shanware, "Complementary model for intrinsic time-dependent dielectric breakdown in $SiO_2$ dielectrics," *J. Appl. Phys.*, vol. 88, no. 9, pp. 5351–5359, 2000.
[92] J. H. Stathis, "Reliability limits for the gate insulator in CMOS technology," *IBM J. Res. Dev.*, vol. 46, no. 2/3, pp. 265–286, 2002.
[93] R. Degraeve, G. Groeseneken, R. Bellens, J. L. Ogier, M. Depas, P. J. Roussel, and H. E. Maes, "New insights in the relation between electron trap generation and the statistical properties of oxide breakdown," *IEEE Trans. Electron Devices*, vol. 45, pp. 904–911, Apr. 1998.
[94] J. Suñé, E. Miranda, M. Nafría, and X. Aymerich, "Point contact conduction at the oxide breakdown of MOS devices," in *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, 1998, pp. 191–194.
[95] E. Miranda, J. Suñé, R. Rodríguez, M. Nafría, and X. Aymerich, "A function-fit model for the soft breakdown failure mode," *IEEE Electron Device Lett.*, vol. 20, pp. 265–267, June 1999.
[96] J. H. Stathis, B. P. Linder, R. Rodríguez, and S. Lombardo, "Reliability of ultra-thin oxides in CMOS circuits," *Microelectron. Reliabil.*, vol. 43, no. 9–11, pp. 1353–1360, 2003.
[97] R. Rodríguez, J. H. Stathis, B. P. Linder, R. V. Joshi, and C. T. Chuang, "Influence and model of gate oxide breakdown on CMOS inverters," *Microelectron. Reliabil.*, vol. 43, no. 9–11, pp. 1439–1444, 2003.
[98] D. K. Ferry and S. M. Goodnick, *Transport in Nanostructures*. Cambridge, U.K.: Cambridge University Press, 1997.
[99] M. LeRoy, E. Lheurette, O. Vanbesien, and D. Lippens, "Wave-mechanical calculations of leakage current through stacked dielectrics for nanotransistor metal-oxide-semiconductor design," *J. Appl. Phys.*, vol. 93, no. 5, pp. 2966–2971, 2003.
[100] 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, "Vertically scaled MOSFET gate stacks and junctions: How far are we likely to go?," *IBM J. Res. Dev.*, vol. 46, no. 2/3, pp. 299–315, 2002.
[101] G. D. Wilk, R. M. Wallace, and J. M. Anthony, "High-k gate dielectrics: Current status and materials properties considerations," *J. Appl. Phys.*, vol. 89, no. 10, pp. 5243–5275, 2001.
[102] J. Robertson, "Band offsets of wide-bandgap oxides and implications for future electronic devices," *J. Vac. Sci. Technol.*, vol. 18, no. 3, pp. 1785–1791, 2000.
[103] H.-S. P. Wong, "Beyond the conventional transistor," *IBM J. Res. Dev.*, vol. 46, no. 2/3, pp. 133–168, 2002.
[104] J. Zhang, J. S. Yuan, Y. Ma, and A. S. Oates, "Design optimization of stacked layer dielectrics for minimum gate leakage currents," *Solid-State Electron.*, vol. 44, no. 12, pp. 2165–2170, 2000.
[105] Z. A. Weinberg, "Tunneling of electrons from Si into thermally grown $SiO_2$," *Solid-State Electron.*, vol. 20, no. 1, pp. 11–18, 1977.
[106] M. Depas, B. Vermeire, P. W. Mertens, R. L. Van Meirhaeghe, and M. M. Heyns, "Determination of tunneling parameters in ultra-thin oxide layer Poly-Si/$SiO_2$/Si structures," *Solid-State Electron.*, vol. 38, no. 8, pp. 1465–1471, 1995.
[107] F. Rana, S. Tiwari, and D. A. Buchanan, "Self-consistent modeling of accumulation layers and tunneling currents through very thin oxides," *Appl. Phys. Lett.*, vol. 69, no. 8, pp. 1104–1106, 1996.
[108] A. Ghetti, "Characterization and modeling of the tunneling current in Si-$SiO_2$-Si structures with ultra-thin oxide layer," *Microelectron. Eng.*, vol. 59, no. 1–4, pp. 127–136, 2001.
[109] J. Cai and C.-T. Sah, "Gate tunneling currents in ultrathin oxide metal-oxide-silicon transistors," *J. Appl. Phys.*, vol. 89, no. 4, pp. 2272–2285, 2001.



**Andreas Gehring** (M'04) 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, CA, in summer 2003. His scientific interests include the modeling of quantum effects for device simulation, the simulation of tunneling, and gate dielectric reliability issues.

**Siegfried Selberherr** (F'93) 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.

He has held 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 research topics are modeling and simulation of problems for microelectronics engineering.