# Gate Current Modeling for MOSFETs

A. Gehring and S. Selberherr
Institute for Microelectronics, TU Vienna, A-1040 Vienna, Austria
{gehring|selberherr}@iue.tuwien.ac.at

*Abstract*— **The topic of gate current modeling has been of strong interest in recent years, and with the accelerating pace of device miniaturization it is becoming more and important. We present a survey of tunneling models describing carrier transport through insulating layers for semiconductor device simulation. The crucial topics are particularly discussed, namely, models for the energy distribution function, the transmission coefficient for single and layered dielectrics, defect-assisted tunneling and its relation to dielectric degradation and breakdown, and the influence of quasi-bound states in the inversion layer. The models are compared to measurements.**

## I. Introduction

For the prediction of device performance in state-of-the-art semiconductor devices the simulation of quantum-mechanical tunneling effects is of increasing importance. The application area of such models ranges from the prediction of gate leakage in MOS transistors, the evaluation of gate stacks for advanced high-$\kappa$ gate insulator materials, the optimization of programming and erasing times in non-volatile semiconductor memory cells up to the study of source-drain tunneling. 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 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 [1]. The second weakness lies in the estimation of the transmission coefficient by the WKB or Gundlach method. Energy barriers which are not of triangular or trapezoidal shape are not treated correctly by these models. To accurately describe tunneling in such cases, Schrödinger's equation must be solved. This is usually achieved using the transfer-matrix method [2]. This method, however, is numerically stable only for layer thicknesses up to a few nanometers. We therefore propose to use the quantum transmitting boundary method instead [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, strongly overestimates the tunneling current in this case.

In the silicon-dielectric-silicon structure sketched in Fig. 1 a variety of tunneling processes can be identified [4]. 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

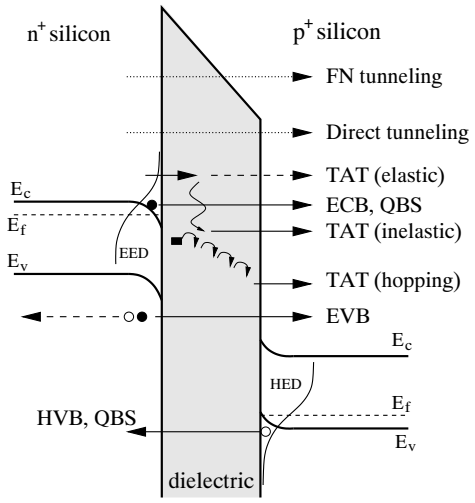


Fig. 1: Tunneling processes in a MOS structure. Direct tunneling processes (ECB, EVB, and HVB) are covered in Section II, while Section III deals with TAT transitions. Bound and quasi-bound states are studied in Section IV.

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.

The paper is structured as follows. In Section II the theory of direct tunneling mechanisms with emphasis on the modeling of the distribution function and the transmission coefficient is outlined. Section III outlines a set of models which can be used to describe the effects of defect-assisted tunneling based on inelastic phonon-assisted transitions, such as dielectric degradation and breakdown. Finally, Section IV describes the calculation of tunneling in the presence of bound and quasi-bound states as encountered in the inversion layer of a MOSFET. A conclusion wraps up the main findings and gives 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 [5] and used by Tsu and Esaki to describe tunneling through a one-dimensional superlattice [2]. It is commonly

0-7803-8777-5/04/$20.00 © 2004 IEEE

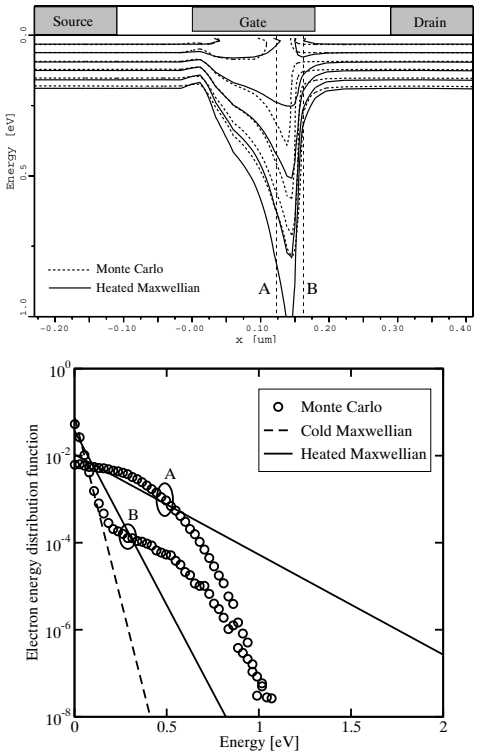


Fig. 2: Comparison of the heated Maxwellian distribution (full lines) with the results from a Monte Carlo simulation (dotted lines) in a turned-on 180 nm MOSFET. Neighboring lines differ by a factor of 10. The distributions at point A and B are compared with a cold Maxwellian in the lower figure.

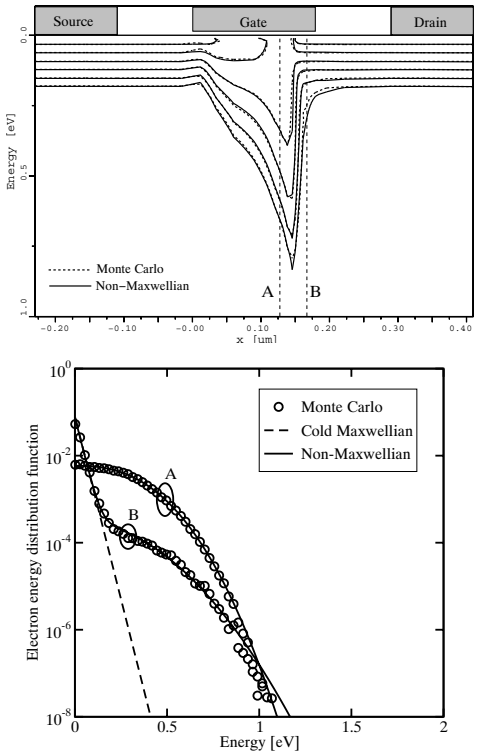


Fig. 3: Comparison of the non-Maxwellian distribution (full lines) with the results from a Monte Carlo simulation (dotted lines) in a turned-on 180 nm MOSFET. Neighboring lines differ by a factor of 10. The distributions at point A and B are compared with a cold Maxwellian in the lower figure.

known as Tsu-Esaki expression. The current density reads

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

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

$$N(\mathcal{E}_x) = \int_0^{\infty} \left(f_1(\mathcal{E}) - f_2(\mathcal{E})\right) \mathrm{d}\mathcal{E}_\rho \ . \tag{2}$$

In these expressions 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. 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 the supply function evaluates to

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

where $\mathcal{E}_{\mathrm{F},1}$ and $\mathcal{E}_{\mathrm{F},2}$ denote the Fermi energies at the semiconductor-oxide interfaces. However, the assumption of a Fermi-Dirac distribution is not valid in the channel of a turned-on submicron MOSFET. Advanced models for the distribution function are necessary.

### 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. [6, 7]. The topic is of high importance because the assumption of a cold Maxwellian distribution function

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

where $T_{\mathrm{L}}$ denotes the lattice temperature and $A$ a normalization constant, underestimates the high-energy tail of the EED near the drain region. The straightforward approach is to use a heated Maxwellian distribution function where the lattice temperature $T_{\mathrm{L}}$ is simply replaced by the electron temperature $T_n$. We applied a Monte Carlo simulator employing analytical non-parabolic bands to check the validity of 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 heated Maxwellian distribution (full lines) yields only poor agreement with the Monte Carlo results (dashed lines). Particularly the high-energy tail near the drain side of the channel is heavily overestimated by the heated Maxwellian

model. A quite generalized approach for the EED has been proposed by Grasser *et al.* [8]

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

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 [9]. 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 because it does not account for the population of cold carriers coming from the drain. A distribution function accounting for this effect was proposed by Sonoda *et al.* [7], and an improved model has been suggested by Grasser *et al.* [1]:

$$f(\mathcal{E}) = A\left(\exp\left(-\left(\frac{\mathcal{E}}{\mathcal{E}_{\mathrm{ref}}}\right)^b\right) + c\exp\left(-\frac{\mathcal{E}}{\mathrm{k_B}T_{\mathrm{L}}}\right)\right). \tag{6}$$

Here the pool of cold carriers in the drain region is correctly modeled by an additional cold Maxwellian subpopulation. The values of $\mathcal{E}_{\mathrm{ref}}$, $b$, and $c$ are again derived from the solution variables of a six moments transport model [1]. 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. With the generalized distribution (5) in the channel and a Maxwellian EED in the poly gate, the supply function (2) becomes [10]

$$N(\mathcal{E}_{\mathrm{t}}) = A_1 \frac{a}{b}\,\Gamma_{\mathrm{i}}\left[\frac{1}{b}, \left(\frac{\mathcal{E}_{\mathrm{t}}}{a}\right)^b\right] - N_2 \exp\left[-\frac{\mathcal{E}_{\mathrm{t}} + \Delta\mathcal{E}_{\mathrm{C}}}{\mathrm{k_B}T_{\mathrm{L}}}\right] \tag{7}$$

where $N_2 = A_2 \mathrm{k_B} T_{\mathrm{L}}$ and $\Gamma_{\mathrm{i}}(\alpha, \beta)$ denotes the incomplete gamma function.

### *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

$$j = \frac{\hbar}{2im} \cdot (\Psi^* \cdot \nabla\Psi - \nabla\Psi^* \cdot \Psi) \tag{8}$$

where $\Psi$ is the wave function, $m$ the carrier mass, and $i = \sqrt{-1}$. 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 one-dimensional Schrödinger equation in the barrier region. This can be achieved using different numerical methods, such as the commonly applied Wentzel-Kramers-Brillouin (WKB) approximation or Gundlach's method [11] which is accurate for triangular and trapezoidal barriers. A more general approach is the transfer-matrix method [2] the basic principle of which is the approximation of an arbitrary-shaped energy barrier by a series of barriers with constant or linear potential. Since the wave function for such barriers can easily be calculated, the transfer matrix can be derived by a number of subsequent matrix computations. From the transfer matrix, the transmission coefficient can be calculated. However, the main shortcoming of the method is that it becomes numerically instable for thick barriers which is due to the multiplication of exponentially growing and decaying states, leading to rounding errors which eventually exceed the amplitude of the wave function itself [12]. An alternative method to compute the transmission coefficient is based on the quantum transmitting boundary method [3, 13]. The method uses a finite-difference approximation of Schrödinger's equation with open boundary conditions. This results in a complex-valued linear equation system for the unknown values of the wave amplitudes. The method is easy to implement, fast, and more robust than the transfer-matrix method. Fig. 4 shows the transmission coefficient for the described methods for a triangular energy barrier (top) and a non-linear energy barrier (bottom). The inset shows the energy barrier and the values of $|\Psi|^2$ for an energy of 2.8 eV on a logarithmic scale. The dotted lines refer to the constant-potential transfer-matrix method. In the top figure the numerical instability of the transfer-matrix method leads to an increasing transmission coefficient for energies below 1 eV. These numerical problems occur for both the constant-

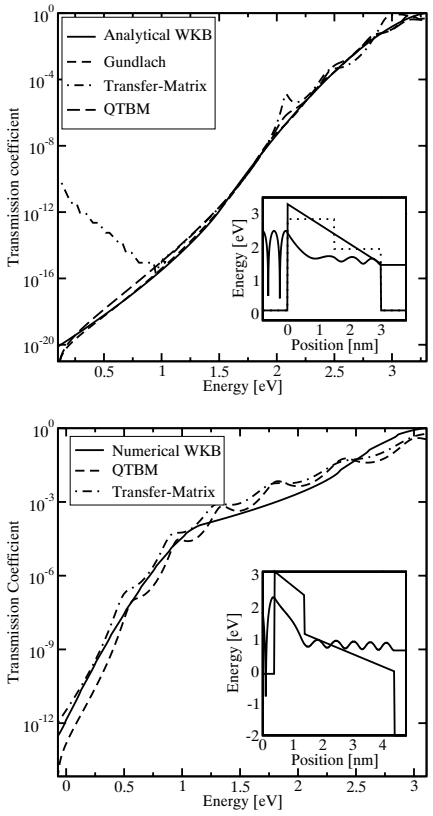


Fig. 4: The transmission coefficient using different methods for a dielectric consisting of a single layer (left) and for a dielectric consisting of two layers (right). The shape of the energy barrier and the wave function at 2.8 eV is shown in the inset.

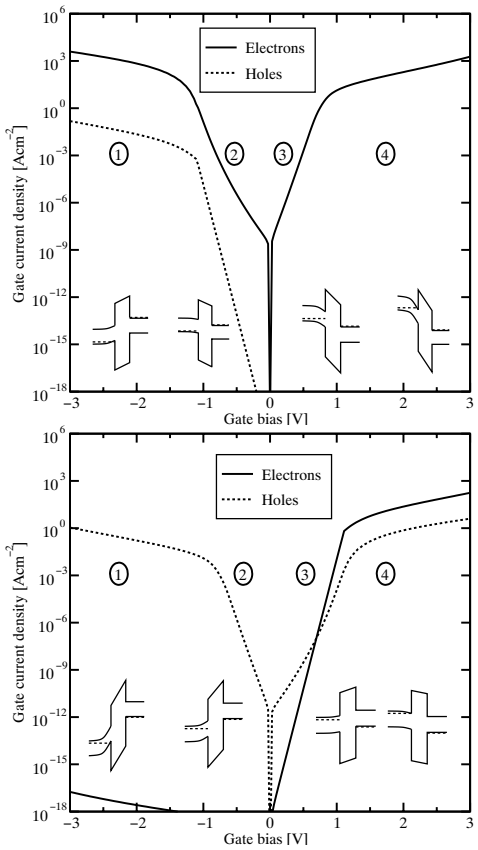


Fig. 5: Tunneling current components in an nMOS (top) and a pMOS (bottom). The insets show the approximate shape of the band edge energies, with the gate contact located at the right side.

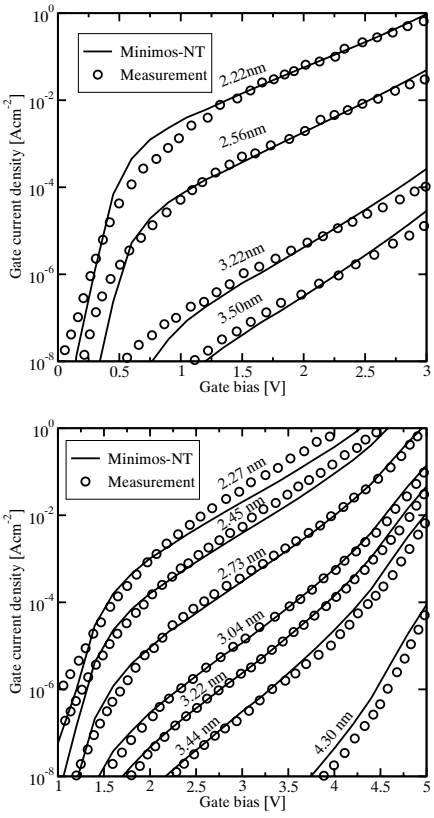


Fig. 6: Comparison of the gate current predicted by the Tsu-Esaki model with measurements of a nMOS (left) and pMOS (right) transistor [16].

potential and the linear-potential approaches. The Gundlach and analytical WKB methods deliver similar results for the triangular barrier. For the stacked dielectric shown in the right figure, the analytical WKB and Gundlach methods cannot be used. The numerical WKB, transfer-matrix, and QTB methods deliver similar results, however, the WKB method does not resolve oscillations in the transmission coefficient. The transmitting boundary method delivers the same results as the Gundlach method which provides an accurate analytical solution in this case. It therefore promises to be a reliable method for the estimation of the transmission coefficient of high-$\kappa$ gate stacks [14].

### *C. Typical Results for MOS Transistors*

The typical shape of the gate current density in turned-off nMOS and pMOS devices is depicted in Fig. 5 [15]. A $SiO_2$ gate dielectric thickness of 2 nm and an acceptor or donor doping of $5 \times 10^{17}\,\mathrm{cm}^{-3}$ and polysilicon gates was 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. 6 for nMOS (top) and pMOS devices (bottom) [16]. The simulations in this figure have been performed using the device simulator MINIMOS-NT [17]. It can be seen that the gate current density can be reproduced over a wide range of dielectric thicknesses with a single set of physical parameters. Note, however, that the assumption of a constant electron mass in the dielectric may no more be justified for ultrathin $SiO_2$ layers but must be replaced by an energy-dependent mass [18].

## III. DEFECT-ASSISTED TUNNELING

Shrinking of gate dielectric thicknesses demands the use of alternative gate dielectrics such as $ZrO_2$. These materials, however, suffer from high defect densities [19]. Therefore, gate dielectric reliability becomes a crucial issue not only for non-volatile memories but also for logic applications. While the current transport through high-$\kappa$ dielectric layers by defect-assisted tunneling has been studied intensely [20], modeling of dielectric breakdown has been investigated only

recently [21]. 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.

### A. Leakage, Wearout, and Breakdown

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. These defects cause trap-assisted tunneling, leading to two effects. First, some of the existing traps become occupied by electrons, which degrades 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, as shown in Fig. 7 for a three-dimensional simulation. 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.

The defects give rise to additional trap-assisted tunneling which is modeled via inelastic phonon-assisted transitions [22, 23]. Fig. 8 shows the basic trap-assisted tunneling process through the gate dielectric. 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 trap-assisted tunneling current is found by integration over the dielectric thickness

$$J_\mathrm{t} = \mathrm{q} \int_0^{t_\mathrm{diel}} \frac{N_\mathrm{T}(x)}{\tau_\mathrm{c}(x) + \tau_\mathrm{e}(x)} \mathrm{d}x, \tag{9}$$

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

$$\tau_\mathrm{c}^{-1}(z) = \int_{\mathcal{E}_0}^{\infty} c_n(\mathcal{E}, x) T_\mathrm{l}(\mathcal{E}) f_\mathrm{l}(\mathcal{E})\, \mathrm{d}\mathcal{E} \tag{10}$$

$$\tau_\mathrm{e}^{-1}(z) = \int_{\mathcal{E}_0}^{\infty} e_n(\mathcal{E}, x) T_\mathrm{r}(\mathcal{E})(1 - f_\mathrm{r}(\mathcal{E}))\, \mathrm{d}\mathcal{E}\ . \tag{11}$$

In these expressions, $c_n$ and $e_n$ denote the capture and emission rates, $f_l$ and $f_r$ the Fermi distributions, and $T_l$ and $T_r$ the transmission coefficients from the left and right side of the dielectric, respectively. The capture and emission processes are described by their respective probabilities which can be calculated by assuming constant [22] or energy-dependent capture cross sections [23], and the transmission coefficients were evaluated by a numerical WKB method. Fig. 9 shows a comparison with experimental data for MOS capacitors [24], where the transition from the trap-assisted tunneling regime at low bias to the Fowler-Nordheim tunneling regime at high bias is clearly visible. The symbol $S$ denotes the Huang-Rhys factor which characterizes the electron-phonon interaction [23].

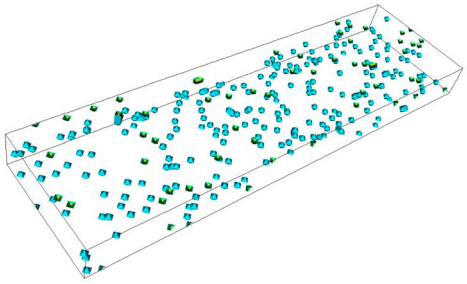

Fig. 7: Random trap distribution in a MOSFET dielectric layer simulated by MINIMOS-NT.

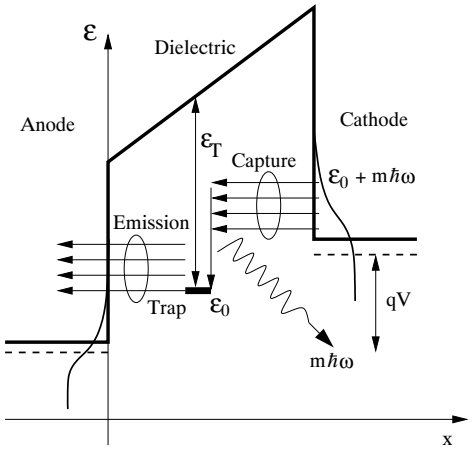


Fig. 8: Trap-assisted tunneling transition by inelastic phonon emission. Electrons are captured from the cathode, relax to the trap energy level $\mathcal{E}_0$ by the emission of phonons, and are emitted to the anode.

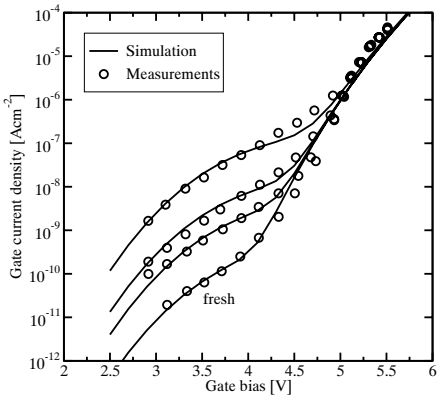


Fig. 9: Gate current density for different stress times [24] for $t_\mathrm{diel}$=5.5 nm. The model parameters are $\mathcal{E}_\mathrm{T}$=2.7 eV, $S\hbar\omega$=1.3 eV, and $N_\mathrm{T}$= $9.0\times10^{17}$ cm$^{-3}$, $1.0\times10^{17}$ cm$^{-3}$, $3.0\times10^{16}$ cm$^{-3}$, and $3.0\times10^{15}$ cm$^{-3}$ (from top to bottom).

While the neutral defects cause trap-assisted tunneling and gate leakage, only the occupied traps lead to threshold voltage degradation and wearout of the gate dielectric. This is modeled by an additional space charge $\rho(x) = Q_\mathrm{T} N_\mathrm{T}(x) f_\mathrm{T}(x)$ in the Poisson equation, where $f_\mathrm{T}$ denotes the trap occupancy and $Q_\mathrm{T}$ 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 [25].

The neutral defects create percolation paths in the dielectric, which eventually connect the gate with the substrate [21]. In MINIMOS-NT the traps are placed randomly, and the defect concentration $N_{\mathrm{T}}$ is assumed to be proportional to the total injected charge $Q_{\mathrm{i}}$ via

$$N_{\mathrm{T}} = CQ_{\mathrm{i}}^{\alpha} \ , \tag{12}$$

as proposed by Degraeve *et al.* [26], who found values of $C = 5.3 \times 10^{-19}\,\mathrm{cm}^{-1.88}\mathrm{As}^{-0.56}$ 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. The gate current density is shown in Fig. 10 for a 3 nm layer of $SiO_2$ as a function of time for different gate voltages assuming an initial trap concentration of $10^{16}\,\mathrm{cm}^{-3}$. The time-to-breakdown strongly decreases and the gate leakage strongly increases with higher gate bias.

However, the gate current density after breakdown can no more be described by a tunneling process. Measurements indicate that the gate current after breakdown 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 $2 - 5$ [27].

### *B. Transient Trap Charging*

To predict the transient behavior of fast switching processes, the charging and decharging 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

$$N_{\mathrm{T}}(x)\frac{\mathrm{d}f_{\mathrm{T}}(x,t)}{\mathrm{d}t} = N_{\mathrm{T}}(x)\frac{1-f_{\mathrm{T}}(x,t)}{\tau_{\mathrm{c}}(x,t)} - N_{\mathrm{T}}(x)\frac{f_{\mathrm{T}}(x,t)}{\tau_{\mathrm{e}}(x,t)}$$

where $\tau_{\mathrm{c}}$ and $\tau_{\mathrm{e}}$ describe the capture and emission time of the trap. For the stationary case, the time derivative on the left-hand side is zero and expression (9) can be derived, while for the transient case, the time constants must be evaluated in each time step. The occupancy function can be calculated iteratively by $f_{\mathrm{T}}(x,t_i) = A_i + B_i f_{\mathrm{T}}(x,t_{i-1})$ where $A_i$ and $B_i$ depend on the capture and emission times at the time step $t_i$ by [25]

$$A_i = \frac{\tau_{\mathrm{c}}^{-1}(z,t_i)\Delta t_i}{1+C_i}, \quad B_i = \frac{1-C_i}{1+C_i}, \quad C_i = \frac{\tau_{\mathrm{m}}^{-1}(z,t_i)\Delta t_i}{2} \ .$$

In these expressions $\Delta t_i = t_i - t_{i-1}$ and $t_i$ denote the discretized time steps. Once the time-dependent occupancy function in the dielectric is known, the tunnel current through one of the interfaces at time $t_i$ is

$$J_{\mathrm{l,r}}(t_i) = \mathrm{q}\int_0^{t_{\mathrm{diel}}} N_{\mathrm{T}}(x)\tau_{\mathrm{l,r}}^{-1}(x,t_i)\,\mathrm{d}x \tag{13}$$

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

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

with the respective values of the capture and emission times to the left and right interface $\tau_{\mathrm{cl,r}}$ and $\tau_{\mathrm{el,r}}$. 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. This model can be applied to the characterization of traps in the dielectric layer. Fig. 11 shows the step response of two MOS capacitors with $ZrO_2$ dielectrics annealed in reducing oxidizing conditions [19]. 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 decharging process [28]. The measurements can be fitted assuming the trap concentrations indicated in the figure.

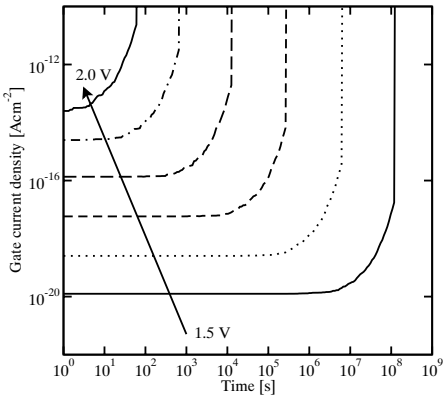


Fig. 10: Dielectric breakdown for a 3 nm $SiO_2$ layer as a function of gate bias.

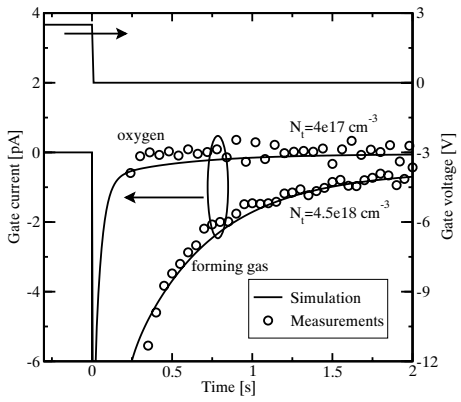


Fig. 11: Transient trap charging currents for a $ZrO_2$ layer.

## IV. 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 [29]. If it is assumed that the wave function does not penetrate into the gate, discrete energy levels can be identified. However, it cannot be assumed that electrons tunnel from these energies, since for the derivation of the levels it was assumed that there is no wave function penetration into the dielectric. This leads to the *paradox* which was addressed by Magnus and Schoenmaker [30]: How can a bound state, which has vanishing current density, lead to tunneling current? 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. Bound states are formed at energies for which the wave function decays to zero at both sides. Quasi-bound states (QBS) have closed boundary conditions at one side and open boundary conditions at the other side. Free states, finally, are states which do not decay at any side. The Tsu-Esaki equation (1) must therefore be replaced by a formula accounting for quasi-bound state and continuum tunneling

$$J = \frac{\mathrm{k_B}T\mathrm{q}}{\pi\hbar^2}\sum_{i,\nu}\frac{g_\nu m_\parallel}{\tau_\nu(\mathcal{E}_{\nu,i})}\ln\left(1+\exp\left(\frac{\mathcal{E}_\mathrm{F}-\mathcal{E}_{\nu,i}}{\mathrm{k_B}T}\right)\right) + \frac{4\pi\mathrm{q}m_\mathrm{dos}}{h^3}\int_{\mathcal{E}_\mathrm{min}}^{\mathcal{E}_\mathrm{max}} TC(\mathcal{E}_x, m_\mathrm{diel})N(\mathcal{E}_x)\,\mathrm{d}\mathcal{E}_x \ , \qquad (14)$$

where the symbols $g_\nu$ and $m_\parallel$ denote the valley degeneracy and parallel masses ($g_\nu = 2$ for $m_\parallel = m_\mathrm{t}$ and $g_\nu = 4$ for $m_\parallel = \sqrt{m_\mathrm{l} m_\mathrm{t}}$), and $\tau_\nu(\mathcal{E}_{\nu,i})$ is the life time of the quasi-bound state $\mathcal{E}_{\nu,i}$. The eigenvalues $\mathcal{E}_{\nu,i}$ in (14) are the real parts of the complex eigenvalue problem.

Fig. 12 shows the resulting wave function for two specific eigenvalues. 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 [31], the FWHM value of the reflection coefficient itself [32], or from the imaginary parts of the complex eigenvalues [3]. However, these methods are computationally demanding and therefore hardly suitable for implementation in a device simulator. Conventional device simulation packages even neglect the QBS tunneling component at all and use only the Tsu-Esaki formula (1) instead. This formula, however, cannot reproduce the QBS tunneling component as shown in Fig. 13, where the QBS current ($J_\mathrm{2D}$) is compared to the continuum current ($J_\mathrm{3D}$). Setting the lower integration level to $\mathcal{E}_{\mathrm{min},2}$ shows that for cold-electron tunneling, the QBS tunneling component is the dominant mechanism. Using $\mathcal{E}_{\mathrm{min},1}$ as lower integration level makes the calibration for different substrate doping necessary to reproduce the QBS results. We propose to use the quasi-classical approach where the life time is calculated from

$$\tau_\nu(\mathcal{E}_{\nu,i}) = \int_0^x \frac{\sqrt{2m_\nu/(\mathcal{E}_{\nu,i}-\mathcal{E}_\mathrm{c}(\xi))}}{TC(\mathcal{E}_{\nu,i})}\,\mathrm{d}\xi \ , \qquad (15)$$

with $\mathcal{E}_\mathrm{c}(x) = \mathcal{E}_{\nu,i}$ [33]. Furthermore, we keep the conventional shape of the Tsu-Esaki formula using $\mathcal{E}_{\mathrm{min},2}$ as lower integration level. To further reduce the computation time, the QBS tunneling current can be calculated based on the eigenvalues of the triangular well approximation $\mathcal{E}_{\nu,i} = -z_i(\hbar^2/2m_\nu)^{1/3}E^{2/3}$ with $z_i$ being the zeros of the Airy function and $E$ the electric field, 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. Thus, an easy and stable formula for the evaluation of quantum and continuum tunneling in CMOS devices is achieved [34].

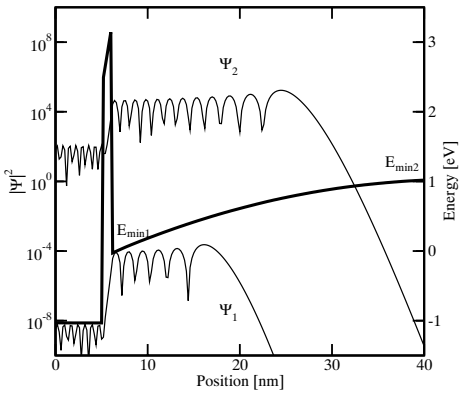


Fig. 12: The conduction band profile and two quasi-bound state wave functions.

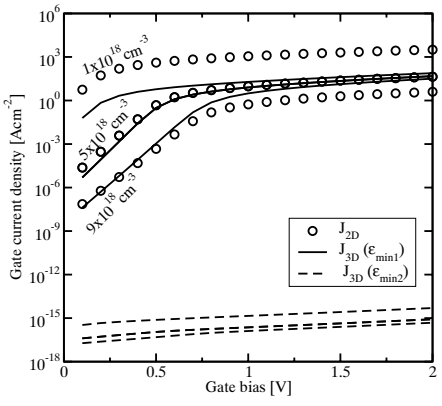


Fig. 13: Current density for different bulk doping and oxide thickness calculated by different approaches.

## V. 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 only the transmitting boundary method remained for the reliable evaluation of dielectric stacks. Furthermore, we propose to link an inelastic trap-assisted tunneling model to the occurrence of dielectric wearout and breakdown phenomena in high-$\kappa$ dielectric materials, also allowing the modeling of fast transient charging and decharging processes. Finally, the emergence of quasi-bound states in inverted MOSFETs was discussed. This requires a modification of the Tsu-Esaki formula, and we recommend a method where the life times are calculated based on the eigenvalues of the closed-boundary triangular well approximation. 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 which is still in its beginnings.

## Acknowledgment

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

## References

[1] 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.

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

[3] W. R. Frensley, "Numerical Evaluation of Resonant States," *Superlattices & Microstructures*, vol. 11, no. 3, pp. 347–350, 1992.

[4] 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, no. 7, pp. 1366–1373, 2001.

[5] C. B. Duke, *Tunneling in Solids*, Academic Press, 1969.

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

[7] 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.

[8] T. Grasser, H. Kosina, and S. Selberherr, "An Impact Ionization Model Including Non-Maxwellian and Non-Parabolicity Effects," in *Proc. Simulation of Semiconductor Processes and Devices*, 2001, pp. 46–49.

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

[10] A. Gehring, T. Grasser, H. Kosina, and S. Selberherr, "Energy Transport Gate Current Model Accounting for Non-Maxwellian Energy Distribution," *Electron.Lett.*, vol. 39, no. 8, pp. 691–692, 2003.

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

[12] 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," *Physical Review B*, vol. 52, no. 11, pp. 8244–8258, 1995.

[13] C. S. Lent and D. J. Kirkner, "The Quantum Transmitting Boundary Method," *J.Appl.Phys.*, vol. 67, no. 10, pp. 6353–6359, 1990.

[14] A. Gehring, H. Kosina, and S. Selberherr, "Analysis of Gate Dielectric Stacks Using the Transmitting Boundary Method," *J.Computational Electronics*, vol. 2, no. 2–4, pp. 219–223, 2003.

[15] 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.

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

[17] Institut für Mikroelektronik, Technische Universität Wien, Austria, *MINIMOS-NT User's Guide*, 2002.

[18] M. Städele, B. Fischer, B. R. Tuttle, and K. Hess, "Resonant Electron Tunneling through Defects in Ultrathin SiO2 Gate Oxides in MOSFETs," *Solid-State Electron.*, vol. 46, no. 7, pp. 1027–1032, 2002.

[19] 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.

[20] 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, no. 5, pp. 1246–1253, 2003.

[21] 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.

[22] 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.

[23] 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.

[24] E. Rosenbaum and L. F. Register, "Mechanism of Stress-Induced Leakage Current in MOS Capacitors," *IEEE Trans.Electron Devices*, vol. 44, no. 2, pp. 317–323, 1997.

[25] 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.Reliab.*, vol. 43, no. 9-11, pp. 1495–1500, 2003.

[26] 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, no. 4, pp. 904–911, 1998.

[27] J. H. Stathis, B. P. Linder, R. Rodríguez, and S. Lombardo, "Reliability of Ultra-Thin Oxides in CMOS Circuits," *Microelectron.Reliab.*, vol. 43, no. 9-11, pp. 1353–1360, 2003.

[28] A. Gehring, S. Harasek, E. Bertagnolli, and S. Selberherr, "Evaluation of $ZrO_2$ Gate Dielectrics for Advanced CMOS Devices," in *Proc. European Solid-State Device Research Conf.*, Estoril, Portugal, 2003, pp. 473–476.

[29] F. Stern, "Self-Consistent Results for n-Type Si Inversion Layers," *Physical Review B*, vol. 5, no. 12, pp. 4891–4899, 1972.

[30] W. Magnus and W. Schoenmaker, "On the Calculation of Gate Tunneling Currents in Ultra-Thin Metal-Insulator-Semiconductor Capacitors," *Microelectronics Reliability*, vol. 41, no. 1, pp. 31–35, 2001.

[31] 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.

[32] 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.

[33] A. Dalla 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, no. 8, pp. 1811–1815, 2001.

[34] A. Gehring and S. Selberherr, "On the Calculation of Quasi-Bound States and Their Impact on Direct Tunneling in CMOS Devices," in *Proc. Simulation of Semiconductor Processes and Devices*, 2004.