CHAPTER 9

Tunneling Models for Semiconductor Device Simulation

Andreas Gehring, Siegfried Selberherr
Institute for Microelectronics, Technical University Vienna, Vienna, Austria

CONTENTS

1. INTRODUCTION

The increasing demand for higher computing power, smaller dimensions, and lower power consumption of electronic devices leads to a pressing need to downscale semiconductor components. This process has already led to length scales where the electrical device characteristics are dominated by quantum-mechanical effects. One of the most interesting of these effects is the quantum-mechanical tunneling of charge carriers through classically forbidden regions.

It is therefore necessary to account for tunneling effects in the design of semiconductor devices. Several models of varying complexity and accuracy can be derived to describe

ISBN: 1-58883-052-7

Handbook of Theoretical and Computational Nanotechnology
Edited by Michael Rieth and Wolfram Schommers
Volume 10: Pages (469–543)

the tunneling current density in semiconductor devices. The models depend on two central quantities; namely, the supply function, which describes the supply of available electrons, and the transmission coefficient, which describes the probability that an electron can tunnel through the barrier. The supply function is determined by the energy distribution of the electrons. In equilibrium, this distribution can be approximated by a Maxwellian distribution. However, the electric field in miniaturized devices is so high that non-Maxwellian models have to be considered to describe accurately the shape of the distribution function and especially the shape of the high-energy tail of the distribution.

To calculate the transmission coefficient of a dielectric layer, Schrodinger's equation must be solved. One of the most frequently used methods is the Wentzel–Kramers–Brillouin (WKB) approximation, which, however, does not reproduce transmission coefficient oscillations as observed in thin gate dielectrics. To describe accurately tunneling through dielectric stacks, it is necessary to resolve the effects of wave function interference. This can be achieved using the transfer-matrix method with either constant or linear potential segments. However, this method is prone to numerical instabilities. A more promising approach is the quantum transmitting boundary method, which allows a stable and reliable evaluation of the transmission coefficient.

Unlike what is assumed in idealized models, dielectric layers are not ideal insulators. Caused by electric stress or processing conditions, defects arise in the dielectric that give rise to trap-assisted tunneling. This results in increased tunneling current at low bias, which is referred to as SILC (stress-induced leakage current). The trap-assisted tunneling process is caused by inelastic transitions of carriers supported by the emission of phonons. As this is a transient process, it is necessary to account for the creation and annihilation of traps in the dielectric based on the rate equation of the traps.

All these effects are discussed in Section 2, which treats the theory of tunneling in semiconductors. This comprises modeling of the supply function, the transmission coefficient, and trap-assisted tunneling. In Section **3,** several applications are presented. The general-purpose device simulator Minimos-NT is used for the simulation of gate leakage currents in metal-oxide-semiconductor (MOS) capacitors and MOSFETs (MOS field-effect transistors). Emphasis is put on modeling of the different tunneling paths in MOS transistors and on the evaluation of alternative high-Kdielectric materials. Furthermore, several NVM (nonvolatile memory) devices such as electrically erasable programmable read-only memory (EEPROM) devices, trap-rich dielectric, or multi-barrier tunneling based devices are investigated.

2. THEORY OF TUNNELING

This section outlines the theory of quantum-mechanical tunneling in semiconductor devices. Different tunneling mechanisms, such as direct-, Fowler–Nordheim, and trap-assisted tunneling are covered. As a first step, the Tsu–Esaki model is derived. The derivation of the supply function and the transmission coefficient is described in detail. Tunneling from quasi-bound states and compact tunneling models is covered as well. The section continues with the description of trap-assisted tunneling and discusses some of the most frequently used models.

2.1. Tunneling Mechanisms

In the silicon-dielectric-silicon structure sketched in Fig. 1, a variety of tunneling processes can be identified. Considering the shape of the energy barrier alone, 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), and TAT (trap-assisted tunneling) processes. The EVB process is caused by electrons tunneling from the valence band to the conduction band. It thus creates free carriers at both sides of the dielectric, which, for MOS transistors, gives rise to increased substrate current. The TAT process can either be elastic, which means that the energy of the carrier is conserved, or inelastic, where the carrier loses energy due to the emission of phonons. Furthermore, in dielectrics with a very high defect density, hopping conduction via multiple defects may occur.

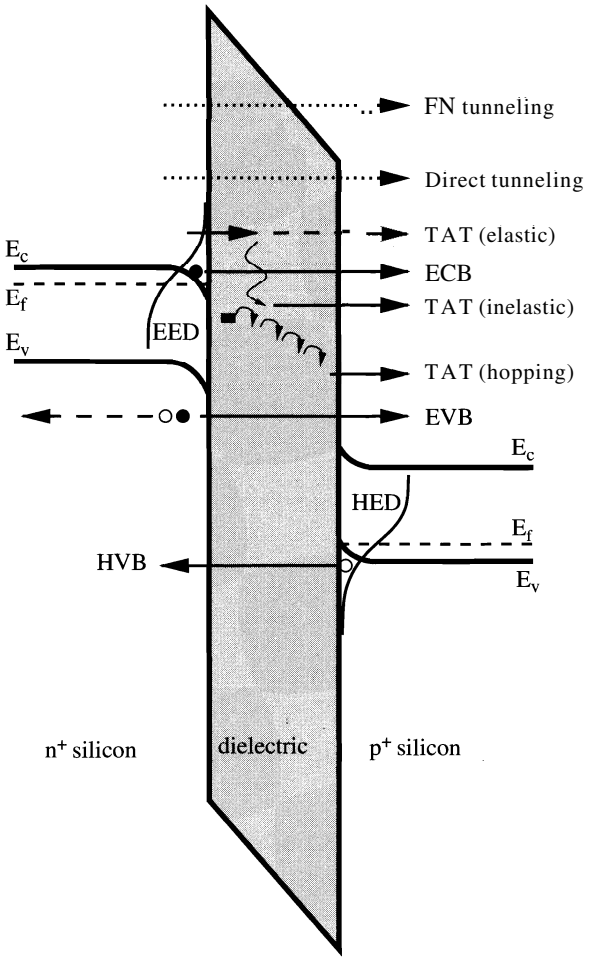


Figure 1. Schematic of the tunneling processes in a silicon-dielectric-silicon structure. The different tunneling processes are indicated by arrows and are described in the text. The abbreviations EED and HED denote the electron and hole energy distribution function.

2.2. The Tsu–Esaki Model

The processes ECB and HVB shown in Fig. 1 can be investigated considering an energy barrier as shown in Fig. 2. Two semiconductor or metal regions are separated by an energy barrier with barrier height $q\Phi_B$ measured from the Fermi energy to the conduction band edge of the insulating layer. Electrons tunnel from Electrode 1 to Electrode 2. The distribution functions at both sides of the barrier are indicated in the figure.

In the derivation, the following assumptions are made:

- Effective-mass approximation: The different masses corresponding to the band structure of the considered material are lumped into a single value for the effective mass. This is denoted by m_{eff} in the electrodes and m_{diel} in the dielectric layer.
- Parabolic bands: The dispersion relation in semiconductors is approximated by

$$\mathscr{E} = \frac{\hbar^2 \mathbf{k}^2}{2m_{\text{eff}}} = \frac{\hbar^2 (k_x^2 + k_y^2 + k_z^2)}{2m_{\text{eff}}} \tag{1}$$

  with the wave vector $\mathrm{k} = k_x \mathbf{e}_x + k_y \mathbf{e}_y + k_z \mathbf{e}_z$.
- Conservation of parallel momentum: Only transitions in the x-direction are considered; the parallel wave vector $\mathbf{k}_\rho = (k_y \mathbf{e}_y + k_z \mathbf{e}_z)$ is not altered by the tunneling process.

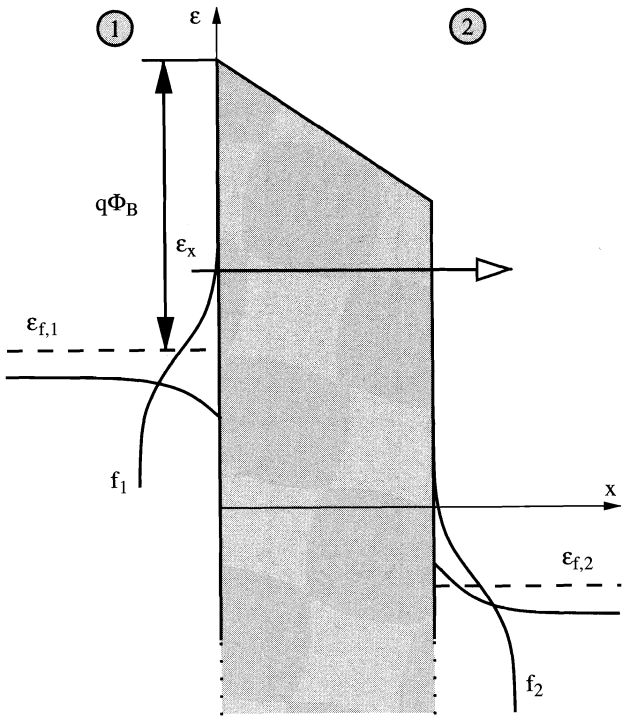


Figure 2. Schematic of an energy barrier with two electrodes that can be used to describe the ECB or HVB processes.

The net tunneling current density from Electrode *1* to Electrode 2 can be written as the net difference between current flowing from Side *1* to Side 2 and vice versa [1, 2]

$$J = J_{1\to 2} - J_{2\to 1} \tag{2}$$

The current density through the two interfaces depends on the perpendicular component of the wave vector k_x, the transmission coefficient TC, the perpendicular velocity v_x, the density of states g, and the distribution function at both sides of the barrier:

$$\begin{aligned} dJ_{1\to 2} &= \mathrm{q}TC(k_x)v_x g_1(k_x)f_1(\mathscr{E})[1 - f_2(\mathscr{E})]\,dk_x \\ dJ_{2\to 1} &= \mathrm{q}TC(k_x)v_x g_2(k_x)f_2(\mathscr{E})[1 - f_1(\mathscr{E})]\,dk_x \end{aligned} \tag{3}$$

In this expression, it is assumed that the transmission coefficient only depends on the momentum perpendicular to the interface. The density of k_x states $g(k_x)$ is

$$g(k_x) = \int_0^\infty \int_0^\infty g(k_x, k_y, k_z)\,dk_y\,dk_z \tag{4}$$

where $g(k_x, k,, k,)$ denotes the three-dimensional density of states in the momentum space. Considering the quantized wave vector components within a cube of side length L

$$\Delta k_x = \frac{2\pi}{L} \quad \Delta k = \frac{2\pi}{L} \quad Ak, = \frac{2\pi}{L} \tag{5}$$

yields for the density of states within the cube

$$g(k_x, k_y, k_z) = 2\frac{1}{\Delta k_x \Delta k_y \Delta k_z}\frac{1}{L^3} = \frac{1}{4\pi^3} \tag{6}$$

where the factor 2 stems from spin degeneracy. For the parabolic dispersion relation (1), the velocity and energy components in tunneling direction obey

$$v_x = \frac{1}{\hbar}\frac{\partial \mathscr{E}}{\partial k_x} = \frac{\hbar k_x}{m_{\mathrm{eff}}} \quad \mathscr{E}_x = \frac{\hbar^2 k_x^2}{2m_{\mathrm{eff}}} \quad v_x\,dk_x = \frac{1}{\hbar}d\mathscr{E}_x \tag{7}$$

Hence, expressions (3) become

$$\begin{aligned} dJ_{1\to 2} &= \frac{\mathrm{q}}{4\pi^3\hbar} TC(\mathscr{E}_x)\, d\mathscr{E}_x \int_0^\infty \int_0^\infty f_1(\mathscr{E})[1 - f_2(\mathscr{E})]\, dk_y\, dk_z \\ dJ_{2\to 1} &= \frac{\mathrm{q}}{4\pi^3\hbar} TC(\mathscr{E}_x)\, d\mathscr{E}_x \int_0^\infty \int_0^\infty f_2(\mathscr{E})[1 - f_1(\mathscr{E})]\, dk_y\, dk_z \end{aligned} \tag{8}$$

Using polar coordinates for the parallel wave vector components

$$\begin{aligned} k_\rho &= \sqrt{k_y^2 + k_z^2} \quad k_y = k_\rho \cos(\gamma) \\ \gamma &= \arctan\left(\frac{k_z}{k_y}\right) \quad k_z = k_\rho \sin(\gamma) \end{aligned} \tag{9}$$

the current density evaluates to

$$\begin{aligned} J_{1\to 2} &= \frac{4\pi m_{\mathrm{eff}}\mathrm{q}}{h^3} \int_{\mathscr{E}_{\min}}^{\mathscr{E}_{\max}} TC(\mathscr{E}_x)\, d\mathscr{E}_x \int_0^\infty f_1(\mathscr{E})[1 - f_2(\mathscr{E})]\, d\mathscr{E}_\rho \\ J_{2\to 1} &= \frac{4\pi m_{\mathrm{eff}}\mathrm{q}}{h^3} \int_{\mathscr{E}_{\min}}^{\mathscr{E}_{\max}} TC(\mathscr{E}_x)\, d\mathscr{E}_x \int_0^\infty f_2(\mathscr{E})[1 - f_1(\mathscr{E})]\, d\mathscr{E}_\rho \end{aligned} \tag{10}$$

In these expressions, the total energy $\mathscr{E}$ has been split into a longitudinal part $\mathscr{E}_\rho$ and a transversal part $\mathscr{E}_x$

$$\mathscr{E}_\rho = \frac{\hbar^2(k_y^2 + k_z^2)}{2m_{\mathrm{eff}}} = \frac{\hbar^2 k_\rho^2}{2m_{\mathrm{eff}}} \quad \mathscr{E}_x = \frac{\hbar^2 k_x^2}{2m_{\mathrm{eff}}} \tag{11}$$

Evaluating the difference $J = J_{1\to 2} - J_{2\to 1}$, the net current through the interface equals

$$J = \frac{4\pi m_{\mathrm{eff}}\mathrm{q}}{h^3} \int_{\mathscr{E}_{\min}}^{\mathscr{E}_{\max}} TC(\mathscr{E}_x)\, d\mathscr{E}_x \int_0^\infty (f_1(\mathscr{E}) - f_2(\mathscr{E}))\, d\mathscr{E}_\rho \tag{12}$$

This expression is usually written as an integral over the product of two independent parts, which only depend on the energy perpendicular to the interface: the transmission coefficient $TC(\mathscr{E}_x)$ and the supply function $N(\mathscr{E}_x)$:

$$J = \frac{4\pi m_{\mathrm{eff}}\mathrm{q}}{h^3} \int_{\mathscr{E}_{\min}}^{\mathscr{E}_{\max}} TC(\mathscr{E}_x) N(\mathscr{E}_x)\, d\mathscr{E}_x \tag{13}$$

which is the expression known as Tsu–Esaki formula. This model has been proposed by Duke [3] and was used by Tsu and Esaki for the modeling of tunneling current in resonant tunneling devices [4]. The values of $\mathscr{E}_{\min}$ and $\mathscr{E}_{\max}$ depend on the considered tunneling process:

- Electrons tunneling from the conduction band (ECB): $\mathscr{E}_{\min}$ is the highest conduction band edge of the two electrodes; $\mathscr{E}_{\max}$ is the highest conduction band edge of the dielectric.
- Holes tunneling from the valence band (HVB): $\mathscr{E}_{\min}$ is the absolute value of the lowest valence band edge of the electrodes; $\mathscr{E}_{\max}$ is the absolute value of the lowest valence band edge of the dielectric. The sign of the integration must be changed.
- Electrons tunneling from the valence band (EVB): $\mathscr{E}_{\min}$ is the lowest conduction band edge of the two electrodes; $\mathscr{E}_{\max}$ the highest valence band edge of the two electrodrc It must be checked if $\mathscr{E}_{\min} < \mathscr{E}_{\max}$.

The next sections concentrate on the calculation of the supply function and the transmission coefficient.

2.3. Supply Function Modeling

The supply function describes the difference in the supply of carriers at the interfaces of the dielectric layer. Following (12), it is given as

$$N(\mathscr{E}_x) = \int_0^\infty (f_1(\mathscr{E}) - f_2(\mathscr{E}))\, d\mathscr{E}_\rho \tag{14}$$

where f_1 and f_2 denote the energy distribution functions near the interfaces. Because the exact shape of these distributions is usually not known, approximative shapes are commonly used. Furthermore, it is assumed that the distributions are isotropic.

2.3.1. Fermi–Dirac Distribution

In equilibrium, the energy distribution function of electrons or holes is given by the Fermi–Dirac statistics

$$f(\mathscr{E}) = \frac{1}{1 + \exp\left(\frac{\mathscr{E} - \mathscr{E}_F}{k_B T}\right)} \tag{15}$$

which can be derived from statistical thermodynamics [5]. Separating the longitudinal and transversal energy components $\mathscr{E} = \mathscr{E}_x + \mathscr{E}_\rho$ and splitting the integral in (14) $N(\mathscr{E}_x) = \xi_1(\mathscr{E}_x) - \xi_2(\mathscr{E}_x)$, the values of ξ_1 and ξ_2 become

$$\xi_i = \int_0^\infty f_i(\mathscr{E})\, d\mathscr{E}_\rho = \int_0^\infty \frac{1}{1 + \exp\left(\frac{\mathscr{E}_x + \mathscr{E}_\rho - \mathscr{E}_{F,i}}{k_B T}\right)} d\mathscr{E}_\rho \quad i = 1, 2 \tag{16}$$

This expression can be integrated analytically using

$$\int \frac{dx}{1 + \exp(x)} = \ln\left(\frac{1}{1 + \exp(-x)}\right) + C \tag{17}$$

so expression (16) evaluates to

$$\xi_i = k_B T \ln\left[1 + \exp\left(-\frac{\mathscr{E}_x - \mathscr{E}_{F,i}}{k_B T}\right)\right] \quad i = 1, 2 \tag{18}$$

and the total supply function (14) becomes

$$N(\mathscr{E}_x) = k_B T \ln\left(\frac{1 + \exp\left(-\frac{\mathscr{E}_x - \mathscr{E}_{F,1}}{k_B T}\right)}{1 + \exp\left(-\frac{\mathscr{E}_x - \mathscr{E}_{F,2}}{k_B T}\right)}\right) \tag{19}$$

2.3.2. Maxwell–Boltzmann Distribution

For nondegenerate semiconductors, the Fermi energy is located below the conduction band edge. Therefore, $\mathscr{E}_{min} - \mathscr{E}_F \gg k_B T$ holds in expression (13), and the Fermi–Dirac distribution (15) can be approximated by a Maxwell–Boltzmann (or Maxwellian) distribution

$$f(\mathscr{E}) = \exp\left(\frac{\mathscr{E}_F - \mathscr{E}}{k_B T}\right) \tag{20}$$

Using this expression, ξ in (14) becomes

$$\xi_i = \int_0^\infty f_i(\mathscr{E})\, d\mathscr{E}_\rho = \int_0^\infty \exp\left(-\frac{\mathscr{E}_x + \mathscr{E}_\rho - \mathscr{E}_{F,i}}{k_B T}\right) d\mathscr{E}_\rho \quad i = 1, 2 \tag{21}$$

which evaluates to

$$\xi_i = k_B T \exp\left(-\frac{\mathscr{E}_x - \mathscr{E}_{F,i}}{k_B T}\right) \quad i = 1, 2 \tag{22}$$

and yields a supply function of

$$N(\mathscr{E}_x) = k_B T \left[\exp\left(-\frac{\mathscr{E}_x - \mathscr{E}_{F,1}}{k_B T}\right) - \exp\left(-\frac{\mathscr{E}_x - \mathscr{E}_{F,2}}{k_B T}\right)\right] \tag{23}$$

2.3.3. Non-Maxwellian Distributions

The Fermi–Dirac or Maxwell–Boltzmann distribution functions are frequently used to describe the distribution of carriers in equilibrium, because they are the solution of Boltzmann's transport equation for the case of zero electric field. In the channel region of a MOSFET, however, the energy distribution deviates from the ideal shape implied by expressions (15) or (20). Carriers gain energy by the electric field in the channel, and they experience scattering events. Models to describe the distribution function of such hot carriers have been studied by numerous authors [6–8]. One possibility to describe the distribution of hot carriers is to use a heated Maxwellian distribution function

$$\mathrm{f}(\mathscr{E}) = A \exp\left(-\frac{\mathscr{E}}{\mathrm{k_B} T_n}\right) \tag{24}$$

where T_n denotes the electron temperature and A is a normalization constant. The validity of this approach, however, is limited. Figure 3 shows in the left part the contour lines of the heated Maxwellian distribution function at the $Si–SiO_2$ interface in comparison to Monte Carlo results. A Monte Carlo simulator employing analytical nonparabolic bands was used for this simulation for a MOSFET with a gate length of $L_g = 180$ nm and a thickness of the gate dielectric of 1.8 nm at a bias of $V_{DS} = V_{GS} = 1V$. It is evident that the heated Maxwellian distribution (full lines) yields only poor agreement with the Monte Carlo results (dashed lines). The distribution function at two points near the middle of the channel (point A) and near the drain contact (point B) are shown in the right part of this figure. Particularly, the high-energy tail in the middle of the channel is heavily overestimated by the heated Maxwellian model. This is unsatisfactory, because a correct description of the high-energy tail is crucial for the evaluation of hot-carrier injection at the drain side used for programming and erasing of EEPROM devices.

To obtain a better prediction of hot-carrier effects, Cassi and Riccó presented an expression to account for the non-Maxwellian shape of the electron energy distribution function [6]

$$\mathrm{f}(\mathscr{E}) = A \exp\left(-\frac{\chi \mathscr{E}^3}{E^{1.5}}\right) \tag{25}$$

with χ as fitting parameter and E being the local electric field in the channel. This local-field dependence was soon questioned by other authors such as Fiegna et al. [9], who replaced the electric field with an effective field calculated from the average electron energy to model the EEPROM writing process. Hasnat et al. used a similar form for the distribution function [10]

$$f(\mathscr{E}) = A \exp\left(-\frac{\mathscr{E}^{\xi}}{\eta(\mathrm{k_B} T_n)^{\nu}}\right) \tag{26}$$

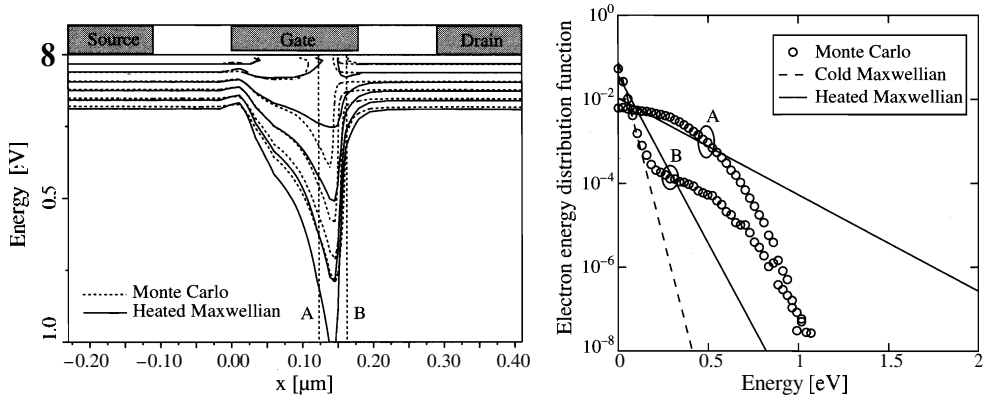


Figure 3. Comparison of the heated Maxwellian distribution (full lines) with the results from a Monte Carlo simulation (dotted lines) in a turned-on 180-nm MOSFET [150]. Neighboring lines differ by a factor of 10. The distributions at point A and B are compared with a cold Maxwellian distribution in the right figure.

They obtained values of $\xi = 1.3$, $\eta = 0.265$, and $\nu = 0.75$ by fitting simulation results to measured gate currents. However, these values fail to describe the shape of the distribution function along the channel when compared to Monte Carlo results [11]. A quite generalized approach to describe the shape of the electron energy distribution (EED) has been proposed by Grasser et al.

$$f(\mathscr{E}) = A \exp\left[-\left(\frac{\mathscr{E}}{\mathscr{E}_{\text{ref}}}\right)^b\right] \tag{27}$$

In this expression, the values of $\mathscr{E}_{\text{ref}}$ and b are mapped to the solution variables T_n and β_n of a six moments transport model [12]. Expression (27) has been shown to reproduce appropriately 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. This was already visible in the right part of Fig. 3 near the drain side of the channel: The distribution consists of a cold Maxwellian, a high-energy tail, and a second cold Maxwellian at higher energies. Expression (27) cannot reproduce the low-energy Maxwellian. A distribution function accounting for the cold carrier population near the drain contact was proposed by Sonoda et al. [8], and an improved model has been suggested by Grasser et al. [11]:

$$f(\mathscr{E}) = A\left\{\exp\left[-\left(\frac{\mathscr{E}}{\mathscr{E}_{\text{ref}}}\right)^b\right] + c\exp\left(-\frac{\mathscr{E}}{k_B T_L}\right)\right\} \tag{28}$$

Here, the pool of cold carriers in the drain region is correctly modeled by an additional cold Maxwellian subpopulation. The values of $\mathscr{E}_{\text{ref}}$, b, and c are again derived from the solution variables of a six moments transport model [11]. Figure 4 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.

This model for the distribution function, however, requires calculation of the third even moment of the distribution function: the kurtosis β_n. As an approximation, β_n can be calculated by an expression obtained for a bulk semiconductor where a fixed relationship between β_n, T_n, and the lattice temperature T_L exists:

$$\beta_{\text{Bulk}}(T_n) = \frac{T_L^2}{T_n^2} + 2\frac{\tau_\beta}{\tau_\mathscr{E}}\frac{\mu_S}{\mu_n}\left(1 - \frac{T_L}{T_n}\right) \tag{29}$$

In this expression, $\tau_\mathscr{E}$, τ_β, μ_n, and μ_S are the energy relaxation time, the kurtosis relaxation time, the electron mobility, and the energy flux mobility, respectively. The value of

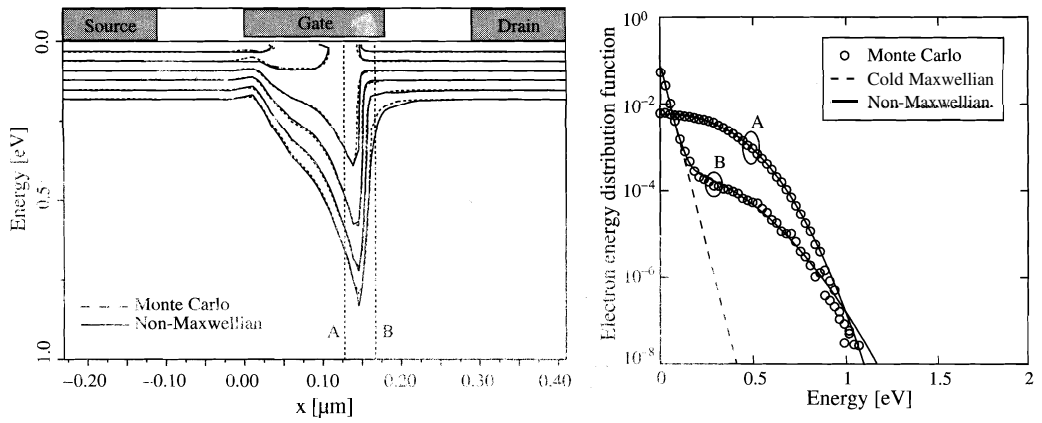


Figure 4. Comparison of the non-Maxwellian distribution (full lines) with the results from a Monte Carlo simulation (dotted lines) in a turned-on 180-nm MOSFET [150]. Neighboring lines differ by a factor of 10. The distributions at point A and B are compared with a coid Maxwellian distribution in the right figure.

$\tau_\beta \mu_S / \tau_\mathscr{E} \mu_n$ can be approximated by a fit to Monte Carlo data [11]. Estimating the kurtosis from (29), the distribution (27) can be used within the energy-transport or hydrodynamic model. For a parabolic band structure, the expressions

$$T_n = \frac{2}{3} \frac{\Gamma\left(\frac{5}{2b}\right)}{\Gamma\left(\frac{3}{2b}\right)} \frac{\mathscr{E}_{\text{ref}}}{\text{k}_\text{B}} \tag{30}$$

$$\beta_n = \frac{3}{5} \frac{\Gamma\left(\frac{3}{2b}\right)\Gamma\left(\frac{7}{2b}\right)}{\Gamma\left(\frac{5}{2b}\right)^2} \tag{31}$$

are found [12], where $\Gamma(x)$ denotes the Gamma function

$$\Gamma(x) = \int_0^\infty \exp(-\alpha)\alpha^{x-1}\, d\alpha \tag{32}$$

Though (30) can easily be inverted to obtain $\mathscr{E}_{\text{ref}}(T_n)$, the inversion of (31) to find $b(T_n)$ at $\beta_n(b) = \beta_{\text{Bulk}}(T_n)$ cannot be given in a closed form. Instead, a fit expression

$$b(T_n) = 1 + b_0\left(1 - \frac{T_\text{L}}{T_n}\right)^{b_1} + b_2\left(1 - \frac{T_\text{L}}{T_n}\right)^{b_3} \tag{33}$$

with the parameters $b_0 = 38.82$, $b_1 = 101.11$, $b_2 = 3.40$, and $b_3 = 12.93$ can be used. Using $\mathscr{E}_{\text{ref}}(T_n)$ and $b(T_n)$, the Monte Carlo distribution can be approximated without knowledge of β_n. Figure 5 shows simulation results for a 500-nm MOSFET using the heated Maxwellian distribution (24), the non-Maxwellian distribution (28), and the non-Maxwellian distribution (27) using (30) and (33) to calculate the values of $\mathscr{E}_{\text{ref}}$ and b. It can be seen that the fit to the results from Monte Carlo simulations is good. However, the emerging population of cold carriers near the drain end of the channel leads to a significant error in the shape of the distribution at low energy. This is important for certain processes, whereas in the case of tunneling, the high-energy tail is more crucial.

With expression (27) for the distribution function and the assumption of a Fermi–Dirac distribution in the polysilicon gate, the supply function (14) becomes

$$N(\mathscr{E}) = A_1 \frac{\mathscr{E}_{\text{ref}}}{b} \Gamma i\left[\frac{1}{b}, \left(\frac{\mathscr{E}}{\mathscr{E}_{\text{ref}}}\right)^b\right] - A_2 \text{k}_\text{B} T_\text{L} \ln\left[1 + \exp\left(-\frac{\mathscr{E} + \Delta\mathscr{E}_\text{c}}{\text{k}_\text{B} T_\text{L}}\right)\right] \tag{34}$$

where $\Gamma_i(\alpha, \beta)$ denotes the incomplete gamma function

$$\Gamma_i(x, y) = \int_y^\infty \exp(-\alpha)\alpha^{x-1}\, d\alpha \tag{35}$$

In (34), the explicit value of the Fermi energy was replaced by the shift of the two conduction band edges $\Delta\mathscr{E}_\text{c}$. Assuming a Maxwellian distribution in the polysilicon gate, the supply function can be further simplified to

$$N(\mathscr{E}) = A_1 \frac{\mathscr{E}_{\text{ref}}}{b} \Gamma_i\left[\frac{1}{b}, \left(\frac{\mathscr{E}}{\mathscr{E}_{\text{ref}}}\right)^b\right] - A_2 \text{k}_\text{B} T_\text{L} \exp\left(-\frac{\mathscr{E} + \Delta\mathscr{E}_\text{c}}{\text{k}_\text{B} T_\text{L}}\right) \tag{36}$$

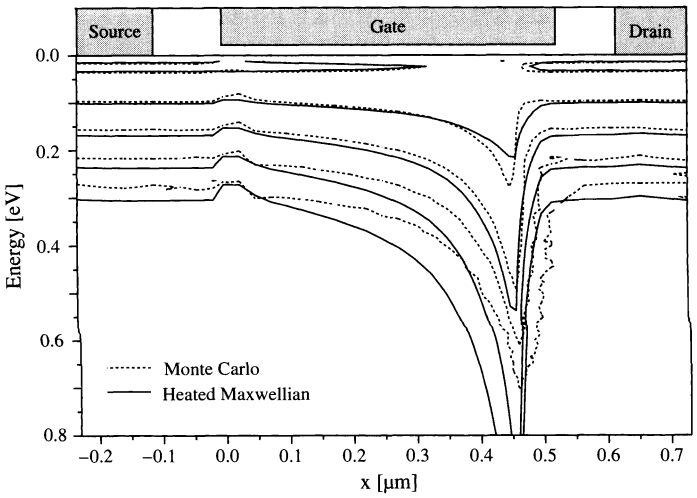
Using the accurate shape of the distribution (28), the expressions for the supply function become

$$N(\mathscr{E}) = A_1 \frac{\mathscr{E}_{\text{ref}}}{b} \Gamma_i\left[\frac{1}{b}, \left(\frac{\mathscr{E}}{\mathscr{E}_{\text{ref}}}\right)^b\right] + A_1 c \text{k}_\text{B} T_2 \exp\left(-\frac{\mathscr{E}}{\text{k}_\text{B} T_\text{L}}\right) - A_2 \text{k}_\text{B} T_\text{L} \ln\left[1 + \exp\left(-\frac{\mathscr{E} + \Delta\mathscr{E}_\text{c}}{\text{k}_\text{B} T_\text{L}}\right)\right] \tag{37}$$

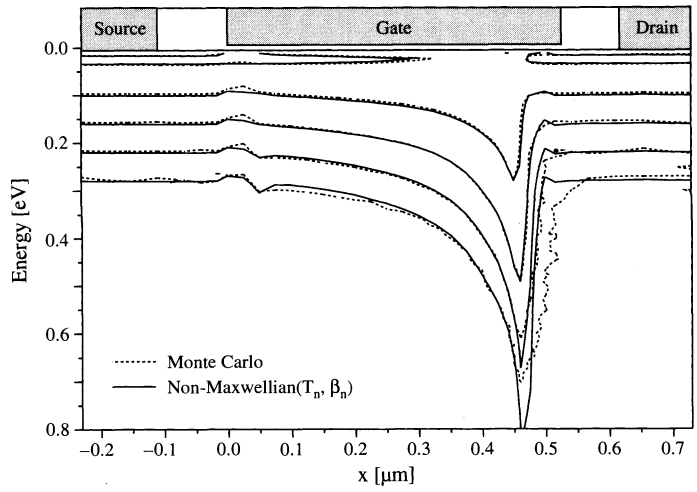
for a Fermi–Dirac distribution, and

$$N(\mathscr{E}) = A_1 \frac{\mathscr{E}_{\text{ref}}}{b} \Gamma_i\left[\frac{1}{b}, \left(\frac{\mathscr{E}}{\mathscr{E}_{\text{ref}}}\right)^b\right] + A_1 c \text{k}_\text{B} T_2 \exp\left(-\frac{\mathscr{E}}{\text{k}_\text{B} T_\text{L}}\right) - A_2 \text{k}_\text{B} T_\text{L} \exp\left(-\frac{\mathscr{E} + \Delta\mathscr{E}_\text{c}}{\text{k}_\text{B} T_\text{L}}\right) \tag{38}$$

assuming a Maxwellian distribution in the polysilicon gate.

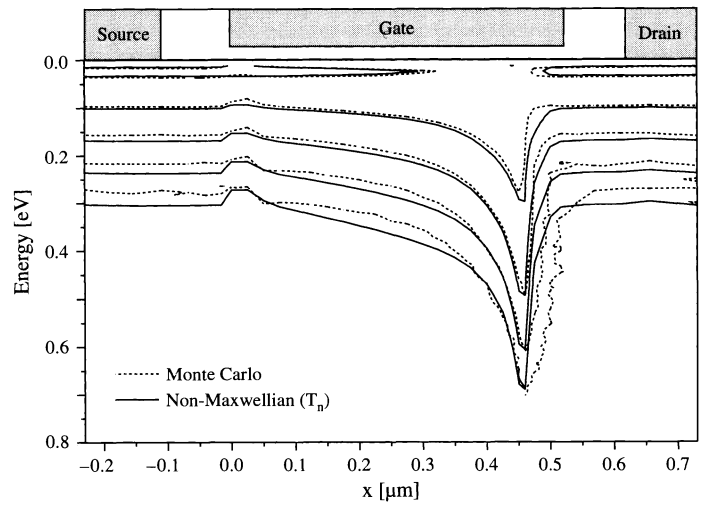


$$f(\mathscr{E}) = A\exp\left(-\frac{\mathscr{E}}{k_B T_n}\right)$$



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

$\mathscr{E}_{\text{ref}}$, b, and c derived from n, T_n, and β_n.



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

$\mathscr{E}_{\text{ref}}$ and b derived from n and T_n.

Figure 5. Different expressions for the energy distribution function in a 500-nm MOSFET compared with Monte Carlo results [152].

2.3.4. Normalization

When implementing the analytical expressions for the distribution function and the supply function into a device simulator, it is necessary to assure consistency: the carrier concentration defined by the analytical distribution function must match the carrier concentration

from the transport model used. Therefore, the normalization prefactor A has to be evaluated from

$$n = \langle 1 \rangle = \frac{1}{4\pi^3} \int f(\mathbf{k})\, d^3k \tag{39}$$

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

$$n = \frac{1}{4\pi^3} \int_{-\pi}^{\pi} d\alpha \int_0^{\pi} \sin\theta\, d\theta \int_0^{\infty} f(k) k^2\, dk \tag{40}$$

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

$$n = \int_0^{\infty} f(\mathscr{E}) \frac{4\pi\sqrt{2m_{\mathrm{eff}}^3}}{h^3} \sqrt{\mathscr{E}}\, d\mathscr{E} \tag{41}$$

where the integration is performed from the conduction band edge $\mathscr{E}_c = 0$. For a Maxwellian or heated Maxwellian distribution [expressions (20) or (24)], the normalization constant evaluates to

$$A = \frac{nh^3}{4\pi(\mathrm{k_B}T_\nu)^{3/2}\Gamma\left(\frac{3}{2}\right)\sqrt{2m_{\mathrm{eff}}^3}} \tag{42}$$

where T_ν is either the lattice temperature (for the assumption of a Maxwellian distribution) or the carrier temperature (for the assumption of a heated Maxwellian distribution). Using the non-Maxwellian distribution (27), the normalization constant evaluates to

$$A = \frac{nh^3 b}{4\pi\mathscr{E}_{\mathrm{ref}}^{3/2}\Gamma\left(\frac{3}{2b}\right)\sqrt{2m_{\mathrm{eff}}^3}} \tag{43}$$

whereas for expression (28), it is

$$A = \frac{nh^3}{4\pi\left[\frac{\mathscr{E}_{\mathrm{ref}}^{1/2}}{b}\Gamma\left(\frac{3}{2b}\right) + c(\mathrm{k_B}T_{\mathrm{L}})^{3/2}\Gamma\left(\frac{3}{2}\right)\right]\sqrt{2m_{\mathrm{eff}}^3}} \tag{44}$$

2.4. The Energy Barrier

For the calculation of the transmission coefficient, it is necessary to take the shape of the energy barrier into account. Electrons tunnel from a semiconductor or metal segment through a dielectric layer to another semiconductor or metal segment. Thus, the band diagram of a MOS capacitor has to be investigated. Furthermore, the image force, which leads to a reduction of both the electron and hole energy barrier for thin dielectrics, will be described in this section.

2.4.1. The Metal-Oxide-Semiconductor Capacitor

Figure **6** shows the band diagram and the electrostatic potential in a metal-oxide-semiconductor structure for different voltages at the metal contact [13–15]. A central quantity is the work function, which is defined as the energy required to extract an electron from the Fermi energy to the vacuum level. The work function of the semiconductor is

$$\mathrm{q}\Phi_{\mathrm{S}} = \mathrm{q}\chi_{\mathrm{S}} + \mathscr{E}_{\mathrm{g}} - \mathscr{E}_i + \mathscr{E}_{\mathrm{v}} + \mathrm{q}\Phi_{\mathrm{f}} \tag{45}$$

where χ_s denotes the electron affinity of the semiconductor. The work function difference between the work function in the metal $\mathrm{q}\Phi_{\mathrm{M}}$ and the work function in the semiconductor $\mathrm{q}\Phi_{\mathrm{S}}$ is

$$\mathrm{q}\Phi_{\mathrm{MS}} = \mathrm{q}\Phi_{\mathrm{M}} - \mathrm{q}\Phi_{\mathrm{S}} \tag{46}$$

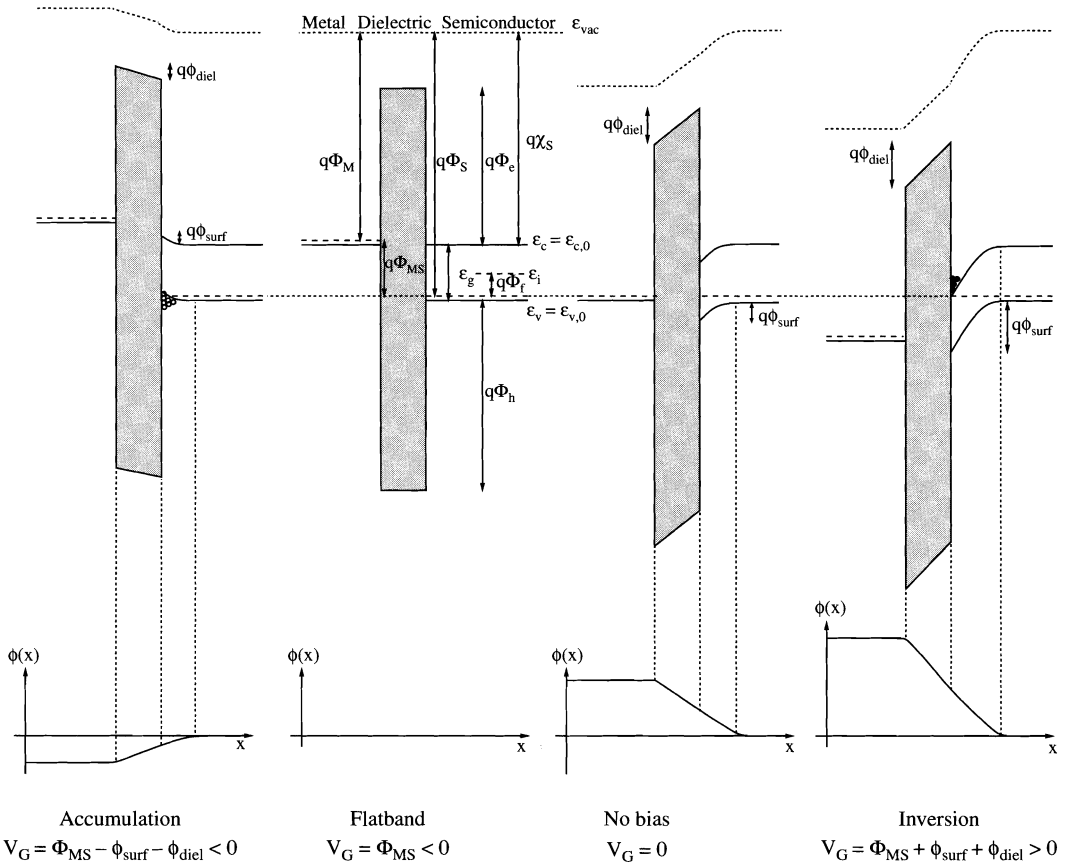


Figure 6. Band diagram and electrostatic potential in an *n*MOS structure (negative work function difference) in accumulation, under flatband condition, without bias, and under inversion condition.

The values of Φ_M and χ_S depend on the material, as shown in Table 1 [5, 16, 17]. However, the actual value of the work function of a metal deposited on SiO_2 is not exactly the same as that of the metal in vacuum [17].

As long as Boltzmann statistics can be applied, the Fermi potential Φ_f depends on the doping concentration of the semiconductor in the following way:

$$p\text{-type: } \Phi_f = \frac{k_B T}{q} \ln\left(\frac{N_A}{n_i}\right) > 0, \tag{47}$$

$$n\text{-type: } \Phi_f = -\frac{k_B T}{q} \ln\left(\frac{N_D}{n_i}\right) < 0 \tag{48}$$

Table 1. Electron affinity of various semiconductors (left), work function and the radius of the Fermi sphere of various metals (right) [18, 197].

Semiconductor	χ_S (V)	Metal	$q\Phi_M$ (eV)	k_f (nm^{-1})
Si	4.05	Al	4.28	17.52
Ge	4.00	Pt	5.65	
GaAs	4.07	W	4.63	
GaP	3.80	Mg	3.66	13.74
GaSb	4.06	Ag	4.30	12.04
InAs	4.90	Au	4.80	12.06
InP	4.38	Cu	4.25	13.61
InSb	4.59	Cr	4.50	

where N_A, N_D, and n_i denote the acceptor, donor, and intrinsic concentrations, respectively. The concentration-independent part of (46) is labeled Φ'_{MS}:

$$q\Phi'_{MS} = q\Phi_M - q\chi_S - \mathscr{E}_g + \mathscr{E}_i - \mathscr{E}_v \tag{49}$$

The voltage that has to be applied to achieve flat bands is denoted the flatband voltage. If we deviate from this voltage, a space charge region forms near the interface between the dielectric and the semiconductor. The total potential drop across this space charge region is the surface potential ϕ_{surf}. Due to this potential, all energy levels in the conduction and valence bands are shifted by a constant amount, therefore

$$\begin{aligned} \mathscr{E}_c(x) &= \mathscr{E}_{c,0} - q\phi(x) \\ \mathscr{E}_v(x) &= \mathscr{E}_{v,0} - q\phi(x) \end{aligned} \tag{50}$$

where $\mathscr{E}_{c,0}$ and $\mathscr{E}_{v,0}$ are the conduction and valence bands in the flatband case. Note that in the flatband case $\phi(x) = 0$ in the whole structure.

In metals, the Fermi energy is located at a higher energy level than the conduction band. The difference between the conduction band edge in the metal and the Fermi energy in the metal can be calculated considering the free-electron theory of metals, which assumes that the metal electrons are unaffected by their metallic ions. The sphere of radius k_f (the Fermi wave vector) contains all occupied levels and determines the electron concentration

$$k_f = \sqrt[3]{3\pi^2 n} \tag{51}$$

The values of the metal work function and k_f for various metals are summarized in the right part of Table 1 [18]. The value of $\mathscr{E}_F - \mathscr{E}_c$ can then be calculated from the carrier concentration assuming a parabolic dispersion relation.

At the semiconductor side, the height of the energy barrier is given by $q\Phi_e$ for electrons and $q\Phi_h$ for holes. Note that in the derivation of the Tsu–Esaki formula, the barrier height $q\Phi_B$, which denotes the energetic difference between the Fermi energy and the band edge in the dielectric, is used. Depending on the considered tunneling process, $q\Phi_B$ must be calculated from $q\Phi_e$ or $q\Phi_h$.

2.4.2. Image Force Correction

When an electron approaches a dielectric layer, it induces a positive charge on the interface that acts like an image charge within the layer. This effect leads to a reduction of the barrier height for both electrons and holes [19–21]: The conduction band bends downward and the valence band bends upward, respectively. To account for this effect, the band edge energies (50) must be modified

$$\begin{aligned} \mathscr{E}_c(x) &= \mathscr{E}_{c,0} - q\phi(x) + \mathscr{E}_{image}(x) \\ \mathscr{E}_v(x) &= \mathscr{E}_{v,0} - q\phi(x) + \mathscr{E}_{image}(x) \end{aligned} \tag{52}$$

where the image force correction in the dielectric with thickness t_{diel} can be calculated as [22]

$$\mathscr{E}_{image}(x) = -\frac{q^2}{16\pi\kappa_{diel}} \sum_j^{\infty} (k_1 k_2)^j \left(\frac{k_1}{|x| + jt_{diel}} + \frac{k_2}{(j+1)t_{diel} - |x|} + \frac{2k_1 k_2}{(j+1)t_{diel}} \right) \tag{53}$$

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

$$k_1 = \frac{\kappa_{diel} - \kappa_{si}}{\kappa_{diel} + \kappa_{si}} \quad k_2 = \frac{\kappa_{diel} - \kappa_{metal}}{\kappa_{diel} + \kappa_{metal}} = -1 \tag{54}$$

Here, k_2 accounts for the interface between the insulator and the metal and evaluates to -1.

In the semiconductor, the band edge energies are also altered

$$\mathscr{E}_{\text{image}}(x) = -\frac{q^2}{16\pi\kappa_{si}} \sum_j^{\infty} (k_1 k_2)^j \left(\frac{-k_1}{|x| + jt_{\text{diel}}} + \frac{k_2}{(j+1)t_{\text{diel}} + |x|} \right) \tag{55}$$

In practice, it is sufficient to evaluate the sums in (53) and (55) up to $j = 11$ [23]. Figure 7 shows the band edge energies in an MOS structure for a dielectric layer with a thickness of 2 nm and different dielectric permittivities for an applied bias of 0 V (left) and 2 V (right). A lower dielectric permittivity leads to a stronger band bending due to the image force and therefore strongly influences the transmission coefficient.

However, there is still some uncertainty if the image force has to be considered for tunneling calculations. Though it is used in some works [23–26], others neglect it or report only minor influence on the results [27–31]. For rigorous investigations, however, it is necessary to include it in the simulations. This, however, raises the need for a high spatial resolution along the dielectric. Simple models like the analytical Wentzel–Kramers–Brillouin (WKB) formula or the Gundlach formula are not valid for this case, as described in the following sections. It may therefore be justified to account for the image force barrier lowering by correction factors.

2.5. Transmission Coefficient Modeling

Now that the shape of the energy barrier has been treated, the calculation of the quantum-mechanical transmission coefficient can be investigated. The transmission coefficient TC is defined as the ratio of the quantum-mechanical current density

$$\mathbf{J}(\mathbf{r}) = \frac{\imath\hbar q}{2m}(\Psi\nabla\Psi^* - \Psi^*\nabla\Psi) \tag{56}$$

due to an incident wave in Region 1 and a transmitted wave in Region N; see Fig. 8. The assumption of plane waves in both regions

$$\begin{aligned} \Psi_1(x) &= A_1 \exp(\imath k_1 x) \\ \Psi_N(x) &= A_N \exp(\imath k_N x) \end{aligned} \tag{57}$$

leads to the transmission coefficient

$$TC = \frac{J_N}{J_1} = \frac{k_1 m_1}{k_N m_N} \frac{|A_N|^2}{|A_1|^2} \tag{58}$$

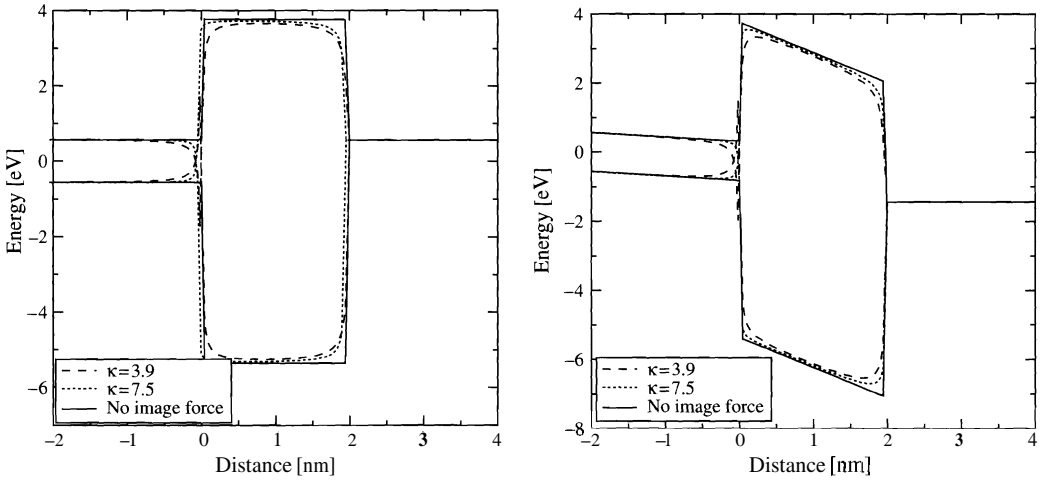


Figure 7. Effect of the image force in an *n*MOS device with a dielectric thickness of 2 nm at a gate bias of 0 V (left) and 2 V (right).

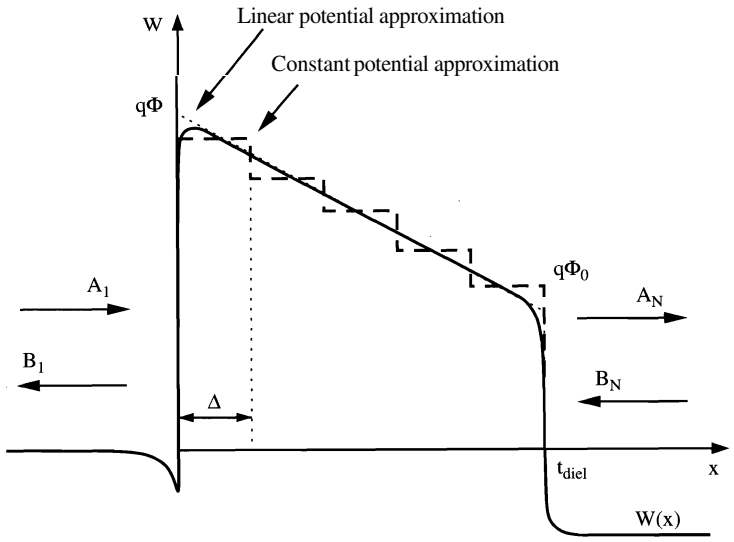


Figure 8. Schematic of an energy barrier of a single-layer dielectric. The potential energy $W(x)$ may either be the conduction band or the valence band energy, depending on the tunneling process. The linear and constant potential approximations refer to the transfer-matrix method described in Section 2.5.3.

Note that the quantum-mechanical current density (56) is equal in Region 1 and Region N. Considering only the incident wave in Region 1 and the transmitted wave in Region N allows definition of a transmission coefficient TC ≤ 1. The wave function amplitudes A, and A_N can be found by solving the stationary Schrodinger equation [32]

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + W(\mathbf{r})\right]\Psi(\mathbf{r}) = \mathscr{E}\Psi(\mathbf{r}) \tag{59}$$

where $W(\mathbf{r})$ is an external potential energy, in the barrier region. This can be achieved by various methods. The Wentzel–Kramers–Brillouin approximation can be applied either analytically for a linear barrier or numerically for arbitrary barriers. Gundlach's method can be used for a single linear energy barrier, whereas the transfer-matrix and quantum transmitting boundary methods are applicable for arbitrary-shaped barriers. The transfer-matrix method can be applied using either constant or linear potential segments as shown in Fig. 8. The different methods will be described in this section, and a brief comparison at the end summarizes their advantages and shortcomings.

2.5.1. The Wentzel–Kramers–Brillouin Approximation

The Wentzel–Kramers–Brillouin approximation is one of the most frequently applied approximations to solve Schrodinger's equation [31, 33, 34]. Starting from the time-independent Schrodinger equation (59), the one-dimensional case reads

$$\left[-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + W(x) - \mathscr{E}\right]\Psi(x) = 0 \tag{60}$$

If the following *Ansatz* is used for the wave function

$$\Psi(\mathrm{x}) = R(x)\exp\left(\imath\frac{S(x)}{\hbar}\right) \tag{61}$$

the equations

$$\frac{d^2R}{dx^2} - \frac{R}{\hbar^2}\left(\frac{dS}{dx}\right)^2 + \frac{2m[\mathscr{E} - W(x)]}{\hbar^2}R = 0 \tag{62}$$

and

$$R\frac{d^2S}{dx^2} + 2\frac{dR}{dx}\frac{dS}{dx} = 0 \tag{63}$$

for the real and imaginary part of (60) can be found. Equation (63) can be solved by

$$\frac{dS}{dx} = \frac{C}{R^2} \tag{64}$$

where C is a constant. With (64), Eq. (62) becomes

$$\frac{1}{R}\frac{d^2R}{dx^2} - \frac{1}{\hbar^2}\left(\frac{dS}{dx}\right)^2 + \frac{2m[\mathscr{E} - W(x)]}{\hbar^2} = 0 \tag{65}$$

With the approximation

$$\frac{1}{R}\frac{d^2R}{dx^2} \ll \frac{1}{\hbar^2}\left(\frac{dS}{dx}\right)^2 \tag{66}$$

we can write

$$S(x) \approx \int \sqrt{2m[\mathscr{E} - W(x)]}\,dx \tag{67}$$

and the wave function $\Psi(x)$ becomes

$$\Psi(x) = R(x)\exp\left(\frac{\imath}{\hbar}\int \sqrt{2m[\mathscr{E} - W(x)]}\,dx\right) \tag{68}$$

Now we consider an energy barrier between the classical turning points x_1 and x_2 with an incoming wave Ψ_1 and a transmitted wave Ψ_2, and $x_2 > x_1$

$$\begin{aligned}\Psi_1(x \le x_1) &\sim \exp\left(\frac{\imath}{\hbar}\int_{-\infty}^{x_1} \sqrt{2m(\mathscr{E} - W(x'))}\,dx'\right)\\ \Psi_2(x \ge x_2) &\sim \exp\left(\frac{\imath}{\hbar}\int_{-\infty}^{x_2} \sqrt{2m(\mathscr{E} - W(x'))}\,dx'\right)\end{aligned} \tag{69}$$

The transmission probability $TC(\mathscr{E})$ is proportional to $|\Psi_2(x_2)/\Psi_1(x_1)|^2$:

$$\begin{aligned}TC &= \left|\frac{\exp\left(\frac{\imath}{\hbar}\int_{-\infty}^{x_2}\sqrt{2m[\mathscr{E} - W(x')]}\,dx'\right)}{\exp\left(\frac{\imath}{\hbar}\int_{-\infty}^{x_1}\sqrt{2m[\mathscr{E} - W(x')]}\,dx'\right)}\right|^2 = \left|\exp\left(\frac{\imath}{\hbar}\int_{x_1}^{x_2}\sqrt{2m[\mathscr{E} - W(x')]}\,dx'\right)\right|^2\\ &= \exp\left(-\frac{2}{\hbar}\int_{x_1}^{x_2}\sqrt{2m[W(x') - \mathscr{E}]}\,dx'\right)\end{aligned} \tag{70}$$

This expression can be evaluated for arbitrary barriers. In Ref. [33], however, it is shown that the WKB approximation is only valid for

$$m\hbar\frac{dW(x)}{dx} \ll \sqrt{|2m[W(x) - \mathscr{E}]|^3} \tag{71}$$

This inequality is fulfilled for points where the variation of the energy barrier is small. The WKB approximation is therefore not valid in the close vicinity of the classical turning points.

The WKB approximation is often used for tunneling simulations and has been implemented in device simulators [1, 35, 36]. For a linear energy barrier, the numerical calculation of the integral in (70) can be avoided. Still, it is necessary to distinguish between regions where direct or Fowler–Nordheim tunneling takes place. For the direct tunneling regime, $\mathscr{E} < q\Phi_0$ holds (see Fig. 8). Therefore, the transmission coefficient

$$TC(\mathscr{E}) = \exp\left(-\frac{2}{\hbar}\int_0^{t_{\text{diel}}}\sqrt{2m_{\text{diel}}(q\Phi - qE_{\text{diel}}x - \mathscr{E})}\,dx\right) \tag{72}$$

evaluates to

$$TC(\mathscr{E}) = \exp\left\{-4\frac{\sqrt{2m_{\text{diel}}}}{3\hbar q E_{\text{diel}}}[(q\Phi - \mathscr{E})^{3/2} - (q\Phi_0 - \mathscr{E})^{3/2}]\right\} \tag{73}$$

with E_{diel} being the electric field defined as $V_{\text{diel}}/t_{\text{diel}}$ and m_{diel} the electron mass in the dielectric. The symbols Φ and Φ_0 denote the upper and lower barrier heights as shown in Fig. 8. The value of Φ_0 is calculated assuming a linear potential in the barrier

$$\Phi_0 = \Phi - E_{\text{diel}} t_{\text{diel}} \tag{74}$$

For the Fowler–Nordheim tunneling regime it holds $\mathscr{E} > q\Phi_0$, and therefore with x_1 defined by $q\Phi - qE_{\text{diel}}x_1 = \mathscr{E}$, the transmission coefficient

$$TC(\mathscr{E}) = \exp\left(-\frac{2}{\hbar}\int_0^{x_1}\sqrt{2m_{\text{diel}}(q\Phi - qE_{\text{diel}}x - \mathscr{E})}\,dx\right) \tag{75}$$

evaluates to

$$TC(\mathscr{E}) = \exp\left[-4\frac{\sqrt{2m_{\text{diel}}}}{3\hbar q E_{\text{diel}}}(q\Phi - \mathscr{E})^{3/2}\right] \tag{76}$$

The WKB tunneling coefficient is frequently multiplied by an oscillating prefactor to reproduce Fowler–Nordheim-induced oscillations [37–41]. However, because no wave function interference is taken into account, the general validity of this method is questionable.

2.5.2. The Gundlach Method

The Gundlach method [42] provides an analytical solution of Schrödinger's equation for a linear energy barrier. The one-dimensional time-independent Schrödinger equation in this case reads

$$\frac{d^2}{dx^2}\Psi(x) + \frac{2m}{\hbar^2}[\mathscr{E} - W(x)]\Psi(x) = 0 \tag{77}$$

with the linear potential energy $W(x)$ between the points x_0 and x_1, $W_0 = W(x_0)$, and $W_1 = W(x_1)$,

$$W(x) = W_0 + (x - x_0)\frac{W_1 - W_0}{x_1 - x_0} \tag{78}$$

for $x_0 < x < x_1$. Using the abbreviations

$$\begin{aligned} l &= -\left(\frac{\hbar^2}{2m}\frac{x_1 - x_0}{W_1 - W_0}\right)^{1/3} \\ \lambda &= -\left(\frac{2m}{\hbar^2}\right)^{1/3}\left(\frac{x_1 - x_0}{W_1 - W_0}\right)^{2/3}\left(\mathscr{E} - W_0 + x_0\frac{W_1 - W_0}{x_1 - x_0}\right) \end{aligned} \tag{79}$$

and $u(x) = \lambda - x/l$, expression (77) turns into

$$\frac{d^2}{dx^2}\Psi(x) - \frac{1}{l^2}u(x)\Psi(x) = 0 \tag{80}$$

With

$$\frac{d^2}{dx^2}\Psi(x) = \frac{d}{du}\frac{du}{dx}\left\{\frac{d}{du}\frac{du}{dx}\Psi[u(x)]\right\} = \frac{1}{l^2}\frac{d^2}{du^2}\Psi[u(x)] \tag{81}$$

Schrödinger's equation evolves into the Airy differential equation

$$\frac{d^2}{du^2}\Psi[u(x)] - u(x)\Psi[u(x)] = 0 \tag{82}$$

The solutions of this differential equation are the Airy functions $\mathrm{Ai}[u(x)]$ and $\mathrm{Bi}[u(x)]$ [43], which are depicted in Fig. 9 together with their derivatives. The wave functions consist of linear superpositions of these Airy functions

$$\Psi(x) = A\mathrm{Ai}[u(x)] + B\mathrm{Bi}[u(x)] \tag{83}$$

where the function $u(x)$ is given as

$$u(x) = -\left(\frac{2m}{\hbar^2}\right)^{1/3}\left(\frac{x_1 - x_0}{W_1 - W_0}\right)^{2/3}[\mathscr{E} - W(x)] \tag{84}$$

Assuming a constant electron mass in the dielectric, Gundlach derives an expression for the transmission coefficient [42]

$$TC = \frac{k_n}{k_1}\frac{4}{\pi^2}\left[\left(\frac{z'}{k_1}A + \frac{k_n}{z'}B\right)^2 + \left(\frac{k_n}{k_1}C + D\right)^2\right]^{-1} \tag{85}$$

where the abbreviations

$$A = \mathrm{Ai}'(z_0)\mathrm{Bi}'(z_s) - \mathrm{Ai}'(z_s)\mathrm{Bi}'(z_0) \tag{86}$$

$$B = \mathrm{Ai}(z_0)\mathrm{Bi}(z_s) - \mathrm{Ai}(z_s)\mathrm{Bi}(z_0) \tag{87}$$

$$C = \mathrm{Ai}(z_s)\mathrm{Bi}'(z_0) - \mathrm{Ai}'(z_0)\mathrm{Bi}(z_s) \tag{88}$$

$$D = \mathrm{Ai}(z_0)\mathrm{Bi}'(z_s) - \mathrm{Ai}'(z_s)\mathrm{Bi}(z_0) \tag{89}$$

have been used, and the symbols z_0, z_s, and z' are given by

$$z_0 = (\mathrm{q}\Phi_0 - \mathscr{E})\left(\frac{at_{\mathrm{diel}}}{2\mathrm{q}(\Phi - \Phi_0)}\right)^{2/3} \quad z_s = (\mathrm{q}\Phi - \mathscr{E})\left(\frac{at_{\mathrm{diel}}}{2\mathrm{q}(\Phi - \Phi_0)}\right)^{2/3} \tag{90}$$

and

$$z' = -\left(\frac{a^2}{4}\frac{\mathrm{q}\Phi - \mathrm{q}\Phi_0}{t_{\mathrm{diel}}}\right)^{1/3} \quad a = \frac{2}{\hbar}\sqrt{2m_{\mathrm{diel}}} \tag{91}$$

The symbols $\mathrm{q}\Phi$ and $\mathrm{q}\Phi_0$ denote the two edges of the energy barrier as shown in Fig. 8. The Gundlach method is frequently used in the literature [25, 44] and implemented in device simulators. Numerical problems may occur for flat barriers ($\Phi \approx \Phi_0$) due to the exponential increase of the Airy functions Bi and Bi′ for positive arguments. In practical implementations, the values of z_0 and z_s have been bounded to values below ≈200 to avoid floating point overflow.

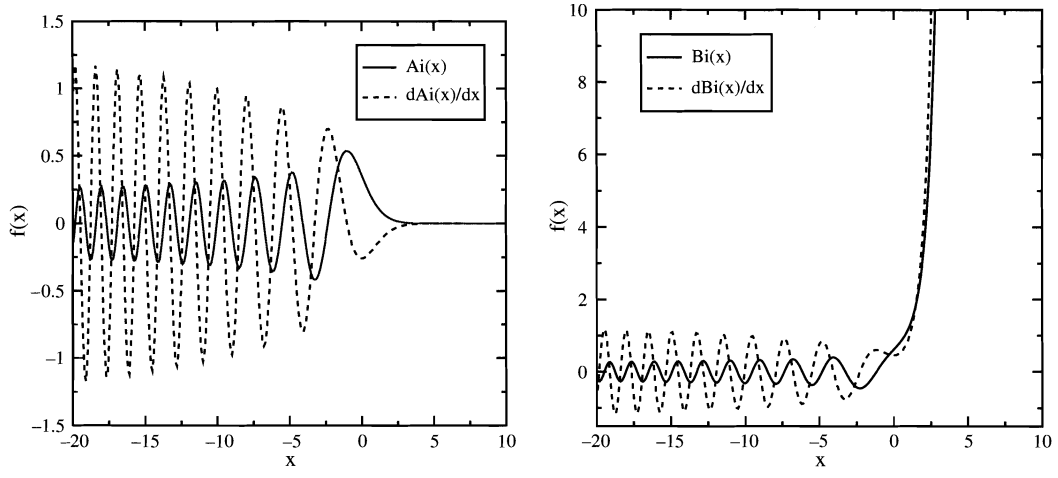


Figure 9. The Airy functions Ai and Bi and their derivatives.

2.5.3. The Transfer-Matrix Method

The use of the transfer-matrix (TM) method for the calculation of the transmission coefficient of energy barriers is based on the work of Tsu and Esaki on electron tunneling through one-dimensional super lattices [4]. It has been used by numerous authors to describe tunneling processes in semiconductor devices [45–49]. The basic principle of the transfer-matrix method is the approximation of an arbitrary-shaped energy barrier by a series of piece-wise constant or piece-wise linear functions. Because the wave function in such barriers can easily be calculated, the total transfer matrix can be derived by a number of subsequent matrix computations. From the transfer matrix, the transmission coefficient can easily be derived.

2.5.3.1. Piecewise-Constant Potential If an arbitrary potential barrier is segmented into N regions with constant potentials (see Fig. 8), the wave function in each region can be written as the sum of an incident and a reflected wave [50] $\Psi_j(x) = A_j \exp(\imath k_j x) + \mathrm{B}, \exp(-\imath k_j x)$ with the wave number k, $= \sqrt{2m_j(\mathscr{E} - W_j)}/\hbar$. The wave amplitudes A,, B,, the carrier mass rn,, and the potential energy W, are assumed constant for each region j. With the interface conditions for energy and momentum conservation

$$\Psi_j(x^-) = \Psi_{j+1}(x^+) \tag{92}$$

$$\frac{1}{m_j}\frac{d\Psi_j(x^-)}{dx} = \frac{1}{m_{j+1}}\frac{d\Psi_{j+1}(x^+)}{dx} \tag{93}$$

the outgoing wave of a layer relates to the incident wave by a complex transfer matrix:

$$\begin{pmatrix} A_j \\ B_j \end{pmatrix} = \underline{T}_j \begin{pmatrix} A_{j-1} \\ B_{j-1} \end{pmatrix} \quad 2 \le j \le \mathrm{N} \tag{94}$$

The transfer matrices are of the form

$$\underline{T}_j = \frac{1}{2}\begin{pmatrix} \left(1 + \frac{k_{j-1}}{k_j}\right)\gamma^{-k_j} & \left(1 - \frac{k_{j-1}}{k_j}\right)\gamma^{-k_j} \\ \left(1 - \frac{k_{j-1}}{k_j}\right)\gamma^{k_j} & \left(1 + \frac{k_{j-1}}{k_j}\right)\gamma^{k_j} \end{pmatrix} \begin{pmatrix} \gamma^{k_{j-1}} & 0 \\ 0 & \gamma^{-k_{j-1}} \end{pmatrix} \quad 2 \le j \le \mathrm{N} \tag{95}$$

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

$$\begin{pmatrix} A_\mathrm{N} \\ B_\mathrm{N} \end{pmatrix} = \prod_{j=2..\mathrm{N}} \underline{T}_j \begin{pmatrix} A_1 \\ B_1 \end{pmatrix} \tag{96}$$

If it is assumed that there is no reflected wave in Region N and the amplitude of the incident wave is unity, (96) simplifies to

$$\begin{pmatrix} A_\mathrm{N} \\ 0 \end{pmatrix} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \begin{pmatrix} 1 \\ B_1 \end{pmatrix} \tag{97}$$

and the transmission coefficient can be calculated from (58). The transfer-matrix method based on constant potential segments has the obvious shortcoming that, for practical barriers, the accuracy of the resulting matrix strongly depends on the chosen resolution. A more rigorous approach is to use linear potential segments.

2.5.3.2. Piecewise-Linear Potential A general barrier may consist of several segments with linear potential sandwiched between contact segments where the potential is constant, as depicted in Fig. 10. The wave functions within these four regions can be written as [confer (83) and (84) for a linear potential]

$$\Psi_1(x) = \mathrm{A},\, \exp(\imath k_1 x) + \mathrm{B},\, \exp(-\imath k_1 x) \tag{98}$$

$$\Psi_2(x) = A_2 \mathrm{Ai}[u_2(x)] + B_2 \mathrm{Bi}[u_2(x)] \tag{99}$$

$$\Psi_3(x) = A_3 \mathrm{Ai}[u_3(x)] + B_3 \mathrm{Bi}[u_3(x)] \tag{100}$$

$$\Psi_4(x) = A_4 \exp(\imath k_4 x) + B_4 \exp(-\imath k_4 x) \tag{101}$$

with $u(x)$ from (84) and the x-independent derivative

$$u' = \frac{du(x)}{dx} = -\left(\frac{2m}{\hbar^2}\right)^{1/3} \left(\frac{W_2 - W_1}{x_2 - x_1}\right)^{1/3} \tag{102}$$

The conditions for continuity of the wave functions and their derivatives yield the following equation system, where abbreviations for the left and right value of $u(x)$ in a layer $\overleftarrow{u}_j = u_j(l_{j-2})$, $\vec{u}_j = u_j(l_{j-1})$, and their derivatives u'_j for $2 \le \mathrm{j} \le \mathrm{N} - 1$ have been used.

$$\begin{aligned}
A_1 \exp(\imath k_1 l_0) + B_1 \exp(-\imath k_1 l_0) &= A_2 \mathrm{Ai}(\overleftarrow{u}_2) + B_2 \mathrm{Bi}(\overleftarrow{u}_2) \\
A_1 \imath k_1 \exp(ik_1 l_0) - B_1 \imath k_1 \exp(-ik_1 l_0) &= A_2 \mathrm{Ai}'(\overleftarrow{u}_2) u'_2 + B_2 \mathrm{Bi}'(\overleftarrow{u}_2) u'_2 \\
A_2 \mathrm{Ai}(\vec{u}_2) + B_2 \mathrm{Bi}(\vec{u}_2) &= A_3 \mathrm{Ai}(\overleftarrow{u}_3) + B_3 \mathrm{Bi}(\overleftarrow{u}_3) \\
A_2 \mathrm{Ai}'(\vec{u}_2) u'_2 + B_2 \mathrm{Bi}'(\vec{u}_2) u'_2 &= A_3 \mathrm{Ai}'(\overleftarrow{u}_3) u'_3 + B_3 \mathrm{Bi}'(\overleftarrow{u}_3) u'_3 \\
A_3 \mathrm{Ai}(\vec{u}_3) + B_3 \mathrm{Bi}(\vec{u}_3) &= A_4 \exp(\imath l_2 k_4) + B_4 \exp(-\imath l_2 k_4) \\
A_3 \mathrm{Ai}'(\vec{u}_3) u'_3 + B_3 \mathrm{Bi}'(\vec{u}_3) u'_3 &= A_4 \imath k_4 \exp(\imath l_2 k_4) - B_4 \imath k_4 \exp(-\imath l_2 k_4)
\end{aligned} \tag{103}$$

The transfer matrices between adjacent layers are again calculated from (94). Using the first two equations of (103) and the Wronskian [43]

$$\mathrm{Wr}\{\mathrm{Ai}(z), \mathrm{Bi}(z)\} = \mathrm{Ai}(z)\mathrm{Bi}'(z) - \mathrm{Ai}'(z)\mathrm{Bi}(z) = \pi^{-1} \tag{104}$$

the matrix $\underline{T}_1$ can be simplified to

$$\underline{T}_1 = \pi \begin{pmatrix} \exp(\imath k_1 l_0)\left(\mathrm{Bi}'(\overleftarrow{u}_2) - \mathrm{Bi}(\overleftarrow{u}_2)\dfrac{\imath k_1}{u'_2}\right) & \exp(-\imath k_1 l_0)\left(\mathrm{Bi}'(\overleftarrow{u}_2) + \mathrm{Bi}(\overleftarrow{u}_2)\dfrac{\imath k_1}{u'_2}\right) \\ \exp(\imath k_1 l_0)\left(-\mathrm{Ai}'(\overleftarrow{u}_2) + \mathrm{Ai}(\overleftarrow{u}_2)\dfrac{\imath k_1}{u'_2}\right) & \exp(-\imath k_1 l_0)\left(-\mathrm{Ai}'(\overleftarrow{u}_2) - \mathrm{Ai}(\overleftarrow{u}_2)\dfrac{\imath k_1}{u'_2}\right) \end{pmatrix}$$

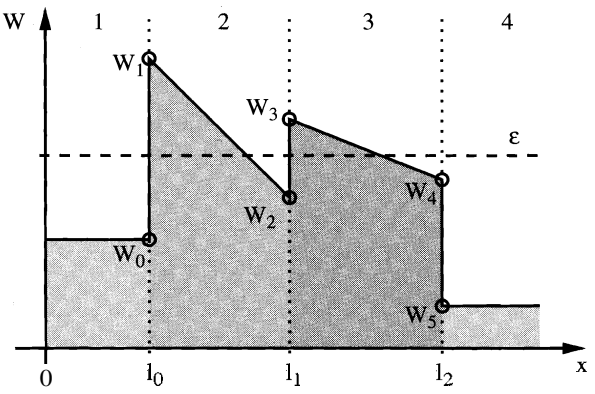


Figure 10. An energy barrier consisting of constant and linear potential segments.

Using the next two lines of (103) yields

$$\underline{T}_2 = \pi \begin{pmatrix} \mathrm{Ai}(\vec{u}_2)\mathrm{Bi}'(\overleftarrow{u}_3) - \frac{u_2'}{u_3'}\mathrm{Bi}(\overleftarrow{u}_3)\mathrm{Ai}'(\vec{u}_2) & \mathrm{Bi}(\vec{u}_2)\mathrm{Bi}'(\overleftarrow{u}_3) - \frac{u_2'}{u_3'}\mathrm{Bi}(\overleftarrow{u}_3)\mathrm{Bi}'(\vec{u}_2) \\ \frac{u_2'}{u_3'}\mathrm{Ai}(\overleftarrow{u}_3)\mathrm{Ai}'(\vec{u}_2) - \mathrm{Ai}(\vec{u}_2)\mathrm{Ai}'(\overleftarrow{u}_3) & \frac{u_2'}{u_3'}\mathrm{Ai}(\overleftarrow{u}_3)\mathrm{Bi}'(\vec{u}_2) - \mathrm{Bi}(\vec{u}_2)\mathrm{Ai}'(\overleftarrow{u}_3) \end{pmatrix}$$

and the last two equations yield with the phase factor $y = \exp(\imath l_2 k_4)$

$$\underline{T}_3 = \frac{1}{2} \begin{pmatrix} \mathrm{Ai}(\vec{u}_3)\gamma^{-1} + \frac{u_3'}{\imath k_4}\mathrm{Ai}'(\vec{u}_3)\gamma^{-1} & \mathrm{Bi}(\vec{u}_3)\gamma^{-1} + \frac{u_3'}{\imath k_4}\mathrm{Bi}'(\vec{u}_3)\gamma^{-1} \\ \mathrm{Ai}(\vec{u}_3)\gamma - \frac{u_3'}{\imath k_4}\mathrm{Ai}'(\vec{u}_3)\gamma & \mathrm{Bi}(\vec{u}_3)\gamma - \frac{u_3'}{\imath k_4}\mathrm{Bi}'(\vec{u}_3)\gamma \end{pmatrix}$$

Though being more accurate than the constant potential approach, this method is computationally more expensive. This drawback, however, is offset by the fact that a lower resolution and thus fewer matrix multiplications are necessary to resolve an energy barrier consisting of linear potential segments.

Simulations using the transfer-matrix method have been reported by several authors [51–54]. Others compared the constant and linear potential approaches and found the constant potential method more feasible for device simulation [55]. The main advantage of the linear-potential transfer-matrix method is that for linear potential segments, the accuracy does not depend on the resolution as it does for the constant-potential transfer-matrix method. However, the evaluation of the Airy functions must be carefully implemented to avoid overflow.

Although the transfer-matrix method for constant or linear potential segments is intuitively easy to understand and implement, the main shortcoming of the method is that it becomes numerically instable for thick barriers. This has been observed by several authors [55–59]. The reason for the numerical problems is that during the matrix multiplications, exponentially growing and decaying states have to be multiplied, leading to rounding errors that eventually exceed the amplitude of the wave function itself for thick barriers.

These problems have been overcome by a further segmentation of the barrier into slices with more accurate transfer matrices [56], the use of scattering matrices instead of transfer matrices [57], iterative methods [58], or by simply setting the transfer matrix entries to zero if the decay factor $\sum k_j x_j$ exceeds a certain value of about 20 [55]. In the next section, a method will be presented that avoids this problem and allows a fast and reliable transmission coefficient estimation.

2.5.4. The Quantum Transmitting Boundary Method

An alternative method to solve the Schrodinger equation has been proposed by Frensley and Einspruch [60], which is based on the tight-binding quantum transmitting boundary method (QTBM) introduced by Lent [61]. It has been used to simulate electron transport in resonant tunneling diodes [59]. The method is based on the finite-difference approximation of the stationary one-dimensional Schrodinger equation (77) on an equidistant grid with an effective mass m_j and a grid spacing A

$$\underline{H}\Psi_j = -s_{j-1}\Psi_{j-1} + d_j\Psi_j - s_{j+1}\Psi_{j+1} = \mathscr{E}\Psi_j \tag{105}$$

where $s_j = \hbar^2/(2m_j\Delta^2)$ and $d_j = \hbar^2/(m_j\Delta^2) + W_j$. For the evaluation of the transmission coefficient, it is necessary to assume open boundary conditions. They are introduced by writing the wave functions at the boundaries of the simulation domain as

$$\Psi_1 = a_1 + b_1 \tag{106}$$

$$\Psi_N = a_N + b_N \tag{107}$$

and relate them to the wave functions outside of the simulation domain by

$$\Psi_0 = a_1 \exp(-\imath k_1 \Delta) + b_1 \exp(\imath k_1 \Delta), \tag{108}$$

$$\Psi_{N+1} = a_N \exp(-\imath k_N \Delta) + b_N \exp(\imath k_N \Delta) \tag{109}$$

This introduces four unknowns and two equations into the system. Setting

$$a_1 = \zeta_1 \Psi_0 + \xi_1 \Psi_1 \tag{110}$$

$$a_N = \zeta_N \Psi_{N+1} + \xi_N \Psi_N \tag{111}$$

eliminates the unknown values of b_1 and b_N and gives a linear system for the $N+2$ complex values Ψ_j

$$\begin{pmatrix} \zeta_1 & \xi_1 & & & & & \\ -s_1 & d_1 - \mathscr{E} & -s_2 & & & & \\ & -s_2 & d_1 - \mathscr{E} & -s_3 & & & \\ & & & \cdots & & & \\ & & & & -s_N & d_N - \mathscr{E} & -s_N \\ & & & & & \xi_N & \zeta_N \end{pmatrix} \begin{pmatrix} \Psi_0 \\ \Psi_1 \\ \Psi_2 \\ \cdots \\ \Psi_N \\ \Psi_{N+1} \end{pmatrix} = \begin{pmatrix} a_1 \\ 0 \\ 0 \\ \cdots \\ 0 \\ a_N \end{pmatrix} \tag{112}$$

Setting $a_i = 1$ and $a_{j \neq i} = 0$ yields the values of the wave function in the whole simulation domain for an incident wave from the left side as in the transfer-matrix method. The method is easy to implement, fast, and more robust than the transfer-matrix method. A further advantage of this method is its suitability for two- and three-dimensional problems. It directly yields the values of the open-boundary wave functions, which can be used to estimate the carrier concentration in the dielectric. Note that the QTBM is closely linked with the nonequilibrium Green's function formalism (NEGF): The matrix in expression (112) is the inverse of the retarded Green's function for an open system without scattering. However, the values of ζ and ξ are complex, so the matrix admits complex eigenvalues and complex solving routines are necessary.

2.5.5. Comparison

Figure 11 shows the transmission coefficient for the described methods for a triangular energy barrier (left) and a two-step nonlinear energy barrier (right). 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 left 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-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.

It can be concluded that for a single-layer dielectric, the analytical WKB method yields reasonable accuracy as compared to the other, computationally more expensive methods. For stacked dielectrics, however, only the numerical WKB, transfer-matrix, or QTB methods can be used in the first place. Because transfer-matrix–based methods exhibit problems regarding numerical stability, only the QTBM and the numerical WKB methods remain. Because the numerical WKB method also needs a numerical integration, its advantage in terms of computational effort is not high enough to rule out the QTBM. Furthermore, if resonance effects—such as in dielectrics with quantum wells—have to be taken into account, the QTBM remains as the method of choice for a reliable transmission coefficient estimation.

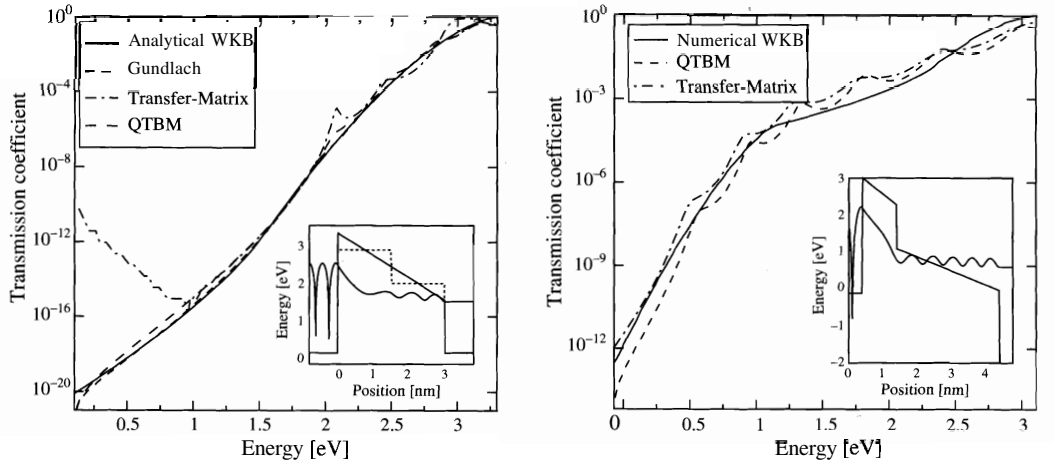


Figure 11. The transmission coefficient using different methods for a dielectric consisting of a single layer (left) and for a dielectric consisting of two layers (right) [140]. The shape of the energy barrier and the wave function at 2.8 eV is shown in the inset.

2.6. Bound and Quasi-Bound States

Up to now, it has been assumed that all energetic states in the substrate contribute to the tunneling current. However, the high doping and the high electric field in the channel leads to a quantum-mechanical quantization of carriers [62, 63]. 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, as for the derivation of the levels, it was assumed that there is no wave function penetration into the dielectric. This leads to theparadox that was addressed by Magnus and Schoenmaker [64]: How can a bound state, which has vanishing current density, lead to tunneling current?

The answer is that it cannot. 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 [65]: Bound states are formed at energies for which the wave function decays to zero at both sides of the dielectric. Quasi-bound states (QBS) have closed boundary conditions at one side and open boundary conditions at the other side. Free states, finally, are states that do not decay at any side of the dielectric layer. This is shown schematically in Fig. 12. The total tunnel current density therefore consists of

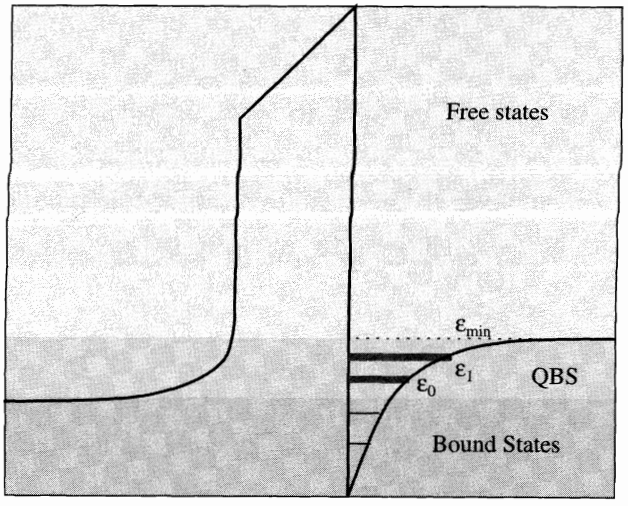


Figure 12. Free, bound, and quasi-bound states in a typical MOS inversion layer.

current from the QBS and from the free states.

$$J = \mathrm{q}\sum_i \frac{n_\nu(\mathscr{E}_i)}{\tau_\mathrm{q}(\mathscr{E}_i)} + \frac{4\pi m_\mathrm{eff}\mathrm{q}}{\hbar^3}\int_{\mathscr{E}_\mathrm{min}}^{\mathscr{E}_\mathrm{max}} TC(\mathscr{E})N(\mathscr{E})d\mathscr{E} \tag{113}$$

where the symbol $n_\nu(\mathscr{E}_i)$ denotes the two-dimensional carrier concentration [66]

$$n_\nu = g_\nu \frac{m\mathrm{k_B}T}{\pi\hbar^2}\ln\left[1+\exp\left(\frac{\mathscr{E}_\mathrm{F}-\mathscr{E}_i}{\mathrm{k_B}T}\right)\right] \tag{114}$$

the symbol g_ν is the valley degeneracy, and τ_q is the lifetime of the quasi-bound state $\mathscr{E}_i$. The lifetime is based on Gamow's theory of nuclear decay [67] and denotes the time constant with which an electron leaks through the energy barrier. Because bound and quasi-bound states are closely related, the computation of bound states will be described first.

2.6.1. Eigenvalues of a Triangular Energy Well

To first order the conduction band edge in a MOSFET inversion layer can be approximated by a linear potential (this is actually done by various authors, see Refs. [68–71]). The solution of Schrodinger's equation for a linear potential has been derived in Section 2.5.2 and consists of a linear superposition of Airy functions. If the triangular energy well is defined as

$$W(x) = W_0 + \frac{W_1 - W_0}{x_1 - x_0}x \tag{115}$$

and no wave function penetration for x <= x_0 is taken into account, the wave function for x > 0 can be written as [62]

$$\Psi(x) = A\mathrm{Ai}[u(\mathrm{x})] \tag{116}$$

$$\Psi(x_0) = A\mathrm{Ai}[u(x_0)] = 0 \tag{117}$$

Therefore, $u(x_0)$ must equal one of the zeros of the Airy function z_i:

$$u(x_0) = z_i < 0 \tag{118}$$

With $u(x)$ from expression (84), the energy eigenvalues are found as

$$\mathscr{E}_i = W_0 - z_i\left(\frac{\hbar^2}{2m}\right)^{1/3}\left(\frac{W_1 - W_0}{x_1 - x_0}\right)^{2/3} \tag{119}$$

The first five zeros of the Airy function are −2.34, −4.09, −5.52, −6.79, and −7.94. These values are often used to approximate the quantized carrier concentration in the channel of MOS devices.

For the assumption of a triangular energy well, the wave function is approximately given as (see Section 2.6.1)

$$\Psi(x) = A\mathrm{Ai}[u(x)] \tag{120}$$

with

$$u(x) = -\left(\frac{2m}{\hbar^2}\right)^{1/3}\left(\frac{x_1 - x_0}{W_1 - W_0}\right)^{2/3}(\mathscr{E} - W(x)) \tag{121}$$

The square of the wave function is a probability, therefore the normalization can be written as [62]

$$\int_0^\infty |\Psi[u(x)]|^2\,dx = 1 \tag{122}$$

$$\int_0^\infty |A\mathrm{Ai}[u(x)]|^2\,dx = 1 \tag{123}$$

$$\int_0^\infty \mathrm{Ai}^2\left\{-\left(\frac{2m}{\hbar^2}\right)^{1/3}\left(\frac{x_1 - x_0}{W_1 - W_0}\right)^{2/3}[\mathscr{E}_i - W(x)]\right\}dx = \frac{1}{A^2} \tag{124}$$

where an infinite barrier is assumed for $x < 0$. With $x_0 = 0$, $W_0 = 0$, and the electric field

$$E = \frac{W_1}{\mathrm{q}x_1} \tag{125}$$

the integral becomes

$$\int_0^\infty \mathrm{Ai}^2\left[\left(\frac{2m\mathrm{q}E}{\hbar^2}\right)^{1/3}\left(x - \frac{\mathscr{E}_i}{\mathrm{q}E}\right)\right]dx = \frac{1}{A^2}. \tag{126}$$

Substituting

$$\lambda(x) = \left(\frac{2m\mathrm{q}E}{\hbar^2}\right)^{1/3}\left(x - \frac{\mathscr{E}_i}{\mathrm{q}E}\right) \tag{127}$$

$$d\lambda(x) = \left(\frac{2m\mathrm{q}E}{\hbar^2}\right)^{1/3} dx \tag{128}$$

yields

$$\left(\frac{\hbar^2}{2m\mathrm{q}E}\right)^{1/3}\int_{\lambda(0)}^\infty \mathrm{Ai}^2[\lambda(x)]\,d\lambda(x) = \frac{1}{A^2} \tag{129}$$

Using the expression [63]

$$\int_z^\infty \mathrm{Ai}^2(x)\,dx = -z\mathrm{Ai}^2(z) + \mathrm{Ai}'^2(z) \tag{130}$$

and $\lambda(0) = \lambda_0$, the normalization constant becomes

$$A = \left(\frac{\left(\frac{2m\mathrm{q}E}{\hbar^2}\right)^{1/3}}{\mathrm{Ai}'^2(\lambda_0) - \lambda_0\mathrm{Ai}^2(\lambda_0)}\right)^{1/2} \tag{131}$$

This method can be used to get an estimate of the first few eigenvalues of the system or to find initial values for the calculation of the eigenvalues described in the next section.

2.6.2. Eigenvalues of Arbitrary Energy Wells

To calculate the eigenvalues of an arbitrary energy well, it is necessary to solve Schrödinger's equation. This can be done using the method of finite differences. It is based on a discretization of the Hamiltonian on a spatial grid and given by (105), which is repeated here for convenience

$$\underline{H}\Psi_j = -s_j\Psi_{j-1} + d_j\Psi_j - s_{j+1}\Psi_{j+1} = \mathscr{E}\Psi_j$$

Though in Section 2.5.4, a constant value of the electron mass in the simulated region was used, a discretization that allows for a position-dependent carrier mass reads

$$d_j = \frac{\hbar^2}{4\Delta^2}\left(\frac{1}{m_{j-1}} + \frac{2}{m_j} + \frac{1}{m_{j+1}}\right) + W_j \tag{132}$$

and

$$s_j = \frac{\hbar^2}{4\Delta^2}\left(\frac{1}{m_{j-1}} + \frac{1}{m_j}\right) \tag{133}$$

The system Hamiltonian is tridiagonal and, for a six-point example, can be written similar to (112) but without the entries for ζ and ξ:

$$\begin{pmatrix} d_1 & -s_2 & & \\ -s_2 & d_2 & -s_3 & \\ & -s_3 & d_3 & -s_4 \\ & & -s_4 & d_4 \end{pmatrix} \begin{pmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \end{pmatrix} = \mathscr{E} \begin{pmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \end{pmatrix} \tag{134}$$

The values Ψ_0 and Ψ_5 must be 0 in this case; that is, closed boundary conditions are assumed. The system Hamiltonian is real and symmetric, therefore all eigenvalues are real. Though this matrix equation looks similar to (112), there are important differences. Here it is necessary to solve the eigenvalue equation to get a value for $\mathscr{E}_i$ and Ψ_i. In (112), any value of $\mathscr{E}$ leads to a valid solution for Ψ_i, and the solution is obtained by solving a complex equation system.

2.6.3. The Lifetime of Quasi-Bound States

The tunneling current from quasi-bound states in (113) depends on their quantum-mechanical lifetime τ_q: In contrast to electrons in bound states, which have an infinite lifetime, electrons in quasi-bound states have a nonzero probability to tunnel through the energy barrier, thus their lifetime is finite [72–74]. This can be seen if the time evolution of the states is considered [75]

$$\Psi(t) = \Psi_0 \exp\left(-\imath \frac{\mathscr{E}_i}{\hbar} t\right) \tag{135}$$

where Ψ_0 is the initial wave function and the complex eigenenergy is

$$\mathscr{E}_i = \mathscr{E}_{\text{re}} - \imath \mathscr{E}_{\text{im}} \tag{136}$$

The time-dependent probability becomes

$$P(t) = \Psi^*(t)\Psi(t) = \Psi_0^2 \exp\left(-\frac{2\mathscr{E}_{\text{im}}}{\hbar} t\right) = \Psi_0^2 \exp\left(-\frac{t}{\tau_q}\right) \tag{137}$$

Thus, the imaginary component of the eigenenergy $\mathscr{E}$ is related to the decay time constant by

$$\tau_q = \frac{\hbar}{2\mathscr{E}_{\text{im}}} \tag{138}$$

The QBS are frequently used for tunneling current calculations [76–81]. Three methods are established to compute the lifetime of a quasi-bound state in MOS inversion layers: computing the full width half-maximum (FWHM) of the reflection coefficient resonances, using the quasi-classical formula based on the Wentzel–Kramers–Brillouin method, or from the complex eigenvalues of the non-Hermitian Hamiltonian. These methods will be described in the following.

2.6.3.1. The Reflection Coefficient Resonances A quasi-bound state forms if one of the system boundary conditions is open ($\neq 0$) and the other one is closed ($=0$). The carrier wave function is reflected at the interface; there is no transmitted wave. Using the transfer-matrix method described in Section 2.5.3, the system can be described by

$$\begin{pmatrix} A_N \\ B_N \end{pmatrix} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \begin{pmatrix} A_1 \\ B_1 \end{pmatrix} \tag{139}$$

where the wave functions are plane waves

$$\Psi_j(x) = A, \exp(\imath k_j x) + \mathrm{B}, \exp(-\imath k_j x) \tag{140}$$

However, no transmission coefficient can be defined for a quasi-bound state: The transmitted wave amplitude A_N must vanish to fulfill the assumption of closed boundary conditions. Instead, a reflection coefficient can be defined, which is

$$RC(\mathscr{E}) = \frac{B_1}{A_1} = -\frac{T_{21}}{T_{22}} \tag{141}$$

For free states, which is the kind of application investigated in Section 2.5.3, the transfer matrix is Hermitian:

$$T_{11} = T_{22}^* \tag{142}$$

$$T_{12} = T_{21}^* \tag{143}$$

It is shown in Ref. [72] that for a quasi-bound state, the transfer matrix is not Hermitian and its elements obey

$$T_{11} = T_{12}^*, \tag{144}$$

$$T_{21} = T_{22}^* \tag{145}$$

Therefore, the reflection coefficient $RC(\mathscr{E})$ can be written as

$$RC(\mathscr{E}) = \exp[\imath\Theta(\mathscr{E})] \tag{146}$$

The phase $\Theta(\mathscr{E})$ varies only weakly at energies away from the resonance energy of the QBS, whereas near the QBS the phase changes strongly. Near the complex energy levels $\mathscr{E}_i$, the derivative of the phase factor $\Theta(\mathscr{E})$ follows a Lorentzian distribution

$$\frac{d\Theta}{d\mathscr{E}} = \frac{2\mathscr{E}_i}{(\mathscr{E} - \mathscr{E}_{re})^2 + \mathscr{E}_{im}^2} \tag{147}$$

where $2\mathscr{E}_{im}$ is the FWHM value of $d\Theta/d\mathscr{E}$. Thus, by calculating the phase of the reflection coefficient as a function of energy, the lifetimes can be determined. This method has been studied intensely by Cassan et al. [66, 82]. They reported numerical difficulties in the calculation of the value of $d\Theta/d\mathscr{E}$, which is prone to numerical noise. Similar problems have been reported by other groups [83].

An alternative approach has been presented by Clerc et al., who noted that the lifetimes can also be extracted directly from the transfer matrix [49]. For a free state, $B_N = 0$ in (139), and the transmission coefficient becomes

$$TC = \left|\frac{A_N}{A_1}\right|^2 = \frac{1}{|T_{11}|^2} \tag{148}$$

For a quasi-bound state, $A_N = 0$. Therefore,

$$A_1 = T_{12} B_N \tag{149}$$

but, because $T_{11} = T_{12}^*$, the value of $|T_{11}|^{-2}$ may be evaluated as well—even if it cannot be interpreted as a transmission coefficient. The lifetime of the QBS is proportional to the resonance peak of the Lorentzian around the real component of the eigenenergy $\mathscr{E}_{re}$

$$\frac{1}{|T_{11}|^2} \propto \frac{1}{(\mathscr{E} - \mathscr{E}_{re})^2 + \frac{\hbar^2}{4\tau_q^2}} \tag{150}$$

but no derivative must be calculated this time. As an example of this method, the left part of Fig. 13 shows the shape of the conduction band edge of a MOS structure in the substrate, dielectric, and polysilicon gate. In the substrate, a triangular quantum well forms. Considering closed boundaries, eigenvalues and wave functions can be calculated. The corresponding wave functions are shown in the figure, where closed boundary conditions have been used at the boundaries of the simulation domain. Note the wave function penetration into the classically forbidden region of the dielectric layer. The eigenvalues of the quasi-bound states are located at 0.27, 0.47, 0.63, 0.76, 0.86, and 0.95 eV. The same information can be found when the value of $|T_{11}|^{-2}$ is investigated, as shown in right part of Fig. 13: Every quasi-bound state in the inversion layer manifests as a peak in the value of $|T_{11}|^{-2}$. The width of each peak is directly related to its lifetime.

2.6.3.2. The Quasi-Classical Formula The calculation of the lifetimes using the approaches shown so far is cumbersome and error-prone, as a precise value for the FWHM in regions where different QBS overlap is difficult to obtain. As an approximation, the lifetime of a QBS can be computed from the quasi-classical formula [83]

$$\tau_q = \frac{1}{TC(\mathscr{E}_i)} \int_0^{x_i} \sqrt{\frac{2m_i}{\mathscr{E}_i - \mathscr{E}_c(x)}}\, dx \tag{151}$$

where $\mathscr{E}_i$ is the resonance energy of the respective bound state and x_i the classical turning point for this energy. The transmission coefficient $TC(\mathscr{E}_i)$ can be calculated by the transfer-matrix method or any other method that solves Schrodinger's equation.

2.6.3.3. The Eigenvalues of the Non-Hermitian Hamiltonian For open-boundary conditions, the system is described by a Hamiltonian that is not Hermitian and admits complex eigenvalues. The most straightforward way to calculate the lifetimes is to find directly the complex eigenvalues of the system Hamiltonian. This, however, is not easily possible because the eigenvalue problem is nonlinear [84]: The values of the matrix elements ζ and ξ depend on the eigenvalue $\mathscr{E}$.

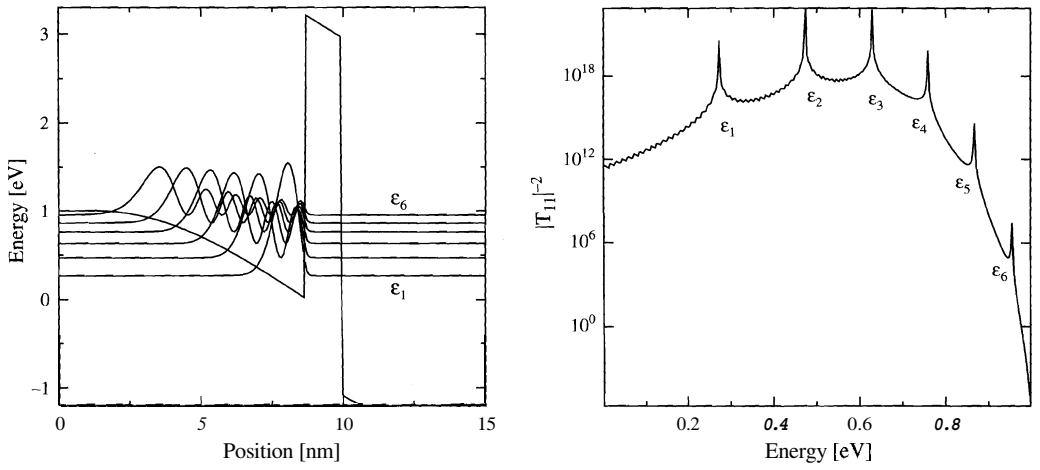


Figure 13. Wave function of quasi-bound states. Note the wave function penetration into classically forbidden regions (left). The respective value of $|T_{11}|^{-2}$ as a function of energy is shown in the right plot. The energy broadening around the poles is clearly visible.

Sophisticated methods have been developed to allow an easy solution of this matrix so that the lifetimes can be calculated [85–88]. First, the closed-boundary Hamiltonian is constructed, and the eigenvalues are calculated. In the one-dimensional case, the matrix is tridiagonal. It is shown in Ref. [89] that in this case, the LU algorithm is advantageous for the calculation of eigenvalues compared to the commonly used QR algorithm, which transforms the matrix into an upper Hessenberg matrix [90].

Then, the eigenvalues are filtered so that only the values remain that are located in the considered energy range. These values are then used as initial values for a Newton search around the closed-boundary eigenvalue [86, 88]. This is motivated by the fact that for $\mathscr{E}_i$ being an eigenvalue of $\underline{H}$, the determinant

$$m(\mathscr{E}_i) = \det(\underline{H} - \mathscr{E}_i \underline{I}) = 0 \tag{152}$$

must be zero. To find the roots of this equation, a Newton search around the closed-boundary eigenvalues $\mathscr{E}_i$ is used

$$\mathscr{E}_{i,j+1} = \mathscr{E}_{i,j} - \frac{m(\mathscr{E}_{i,j})}{m'(\mathscr{E}_{i,j})} \tag{153}$$

where $m'(\mathscr{E})$ denotes the derivative of the determinant

$$m'(\mathscr{E}) = \frac{dm(\mathscr{E})}{d\mathscr{E}} \tag{154}$$

For a tridiagonal matrix, it is possible to find an analytical expression for $m'(\mathscr{E})$ [91, 92]. For general situations, however, the derivative can only be found numerically by

$$m'(\mathscr{E}_i) \approx \frac{m(\mathscr{E}_i + \Delta\mathscr{E}/2) - m(\mathscr{E}_i - \Delta\mathscr{E}/2)}{\Delta\mathscr{E}} \tag{155}$$

This has the advantage that it is not limited to one-dimensional problems but can be applied to any shape of the Hamiltonian.

The complex eigenvalues have been used to calculate the lifetimes of the structure shown in the left part of Fig. 13. The complex energies and lifetimes found are shown in Table 2 and agree with the values found using the method based on the evaluation of the reflection-coefficient.

2.7. Compact Tunneling Models

The above-presented models for the calculation of tunneling currents require a considerable computational effort. However, for practical device simulation, it is desirable to use compact models that do not require large computational resources. That may be necessary for a quick estimation of the dielectric thickness from IV data or to predict the impact of gate leakage on the performance of CMOS circuits [93–98]. The most frequently used model to describe tunneling is the Fowler–Nordheim formula [99]. The Tsu–Esaki expression *(12)* for the tunnel current density reads

$$J = \frac{4\pi q m_{\text{eff}}}{h^3} \int_{\mathscr{E}_{\min}}^{\mathscr{E}_{\max}} TC(\mathscr{E}_x)\, d\mathscr{E}_x \int_0^{\infty} [f_1(\mathscr{E}) - f_2(\mathscr{E})]\, d\mathscr{E}_\rho \tag{156}$$

Table 2. Eigenvalues found by using a resonance-finding algorithm based on the determinant of the open-boundary Hamiltonian.

$\mathscr{E}_i$	$\mathscr{E}_{\text{re}}$ (eV)	$\mathscr{E}_{\text{im}}$ (eV)	τ_q (s)
1	0.2695	1.503×10^{-20}	4.376×10^{4}
2	0.4695	1.830×10^{-19}	3.594×10^{3}
3	0.6256	5.285×10^{-15}	1.244×10^{-1}
4	0.7549	2.794×10^{-11}	2.354×10^{-4}
5	0.8629	4.231×10^{-8}	1.555×10^{-8}
6	0.9503	2.005×10^{-5}	3.281×10^{-11}

where the total energy is split into a longitudinal and a transversal energy

$$\mathscr{E} = \mathscr{E}_x + \mathscr{E}_\rho \tag{157}$$

The goal is to find a simple approximation of (156) that avoids numerical integration. As a first approximation, $T \to 0$ is assumed [1]. This allows replacement of the Fermi function $f(x)$ by the step function

$$\begin{aligned} f_1(\mathscr{E}) = f(\mathscr{E} - \mathscr{E}_{\mathrm{F},1}) &= \begin{cases} 1 & \text{for } \mathscr{E} \le \mathscr{E}_{\mathrm{F},1} \\ 0 & \text{for } \mathscr{E} > \mathscr{E}_{\mathrm{F},1} \end{cases} \\ f_2(\mathscr{E}) = f(\mathscr{E} - \mathscr{E}_{\mathrm{F},2}) &= \begin{cases} 1 & \text{for } \mathscr{E} \le \mathscr{E}_{\mathrm{F},2} \\ 0 & \text{for } \mathscr{E} > \mathscr{E}_{\mathrm{F},2} \end{cases} \end{aligned} \tag{158}$$

Without loss of generality, it can be assumed that $\mathscr{E}_{\mathrm{F},1} > \mathscr{E}_{\mathrm{F},2}$ (see Fig. 14). The innermost integral can then be evaluated analytically for three distinct regions

$$\begin{aligned} \int_0^\infty [f(\mathscr{E} - \mathscr{E}_{\mathrm{F},1}) - f(\mathscr{E} - \mathscr{E}_{\mathrm{F},2})]\, d\mathscr{E}_\rho &= \mathscr{E}_{\mathrm{F},1} - \mathscr{E}_{\mathrm{F},2} \quad \text{for } \mathscr{E}_x \le \mathscr{E}_{\mathrm{F},2} \\ &= \mathscr{E}_{\mathrm{F},1} - \mathscr{E}_x \quad \text{for } \mathscr{E}_{\mathrm{F},2} \le \mathscr{E}_x \le \mathscr{E}_{\mathrm{F},1} \\ &= 0 \quad \text{for } \mathscr{E}_x > \mathscr{E}_{\mathrm{F},1} \end{aligned} \tag{159}$$

This leads to the following expression for the current density:

$$J = \frac{4\pi q m_{\mathrm{eff}}}{h^3} \left(\underbrace{\int_{-\infty}^{\mathscr{E}_{\mathrm{F},2}} TC(\mathscr{E}_x)(\mathscr{E}_{\mathrm{F},1} - \mathscr{E}_{\mathrm{F},2})\, d\mathscr{E}_x}_{\approx 0} + \int_{\mathscr{E}_{\mathrm{F},2}}^{\mathscr{E}_{\mathrm{F},1}} TC(\mathscr{E}_x)(\mathscr{E}_{\mathrm{F},1} - \mathscr{E}_x)\, d\mathscr{E}_x \right) \tag{160}$$

The left integral represents tunneling current from electron states that are low in energy and face a high energy barrier. Hence, as a second approximation, the left integral is neglected. Still, it is necessary to insert an expression for the transmission coefficient in the right integral. For a single-layer dielectric, two shapes are possible: triangular and trapezoidal. First, the formula will be derived assuming a triangular shape.

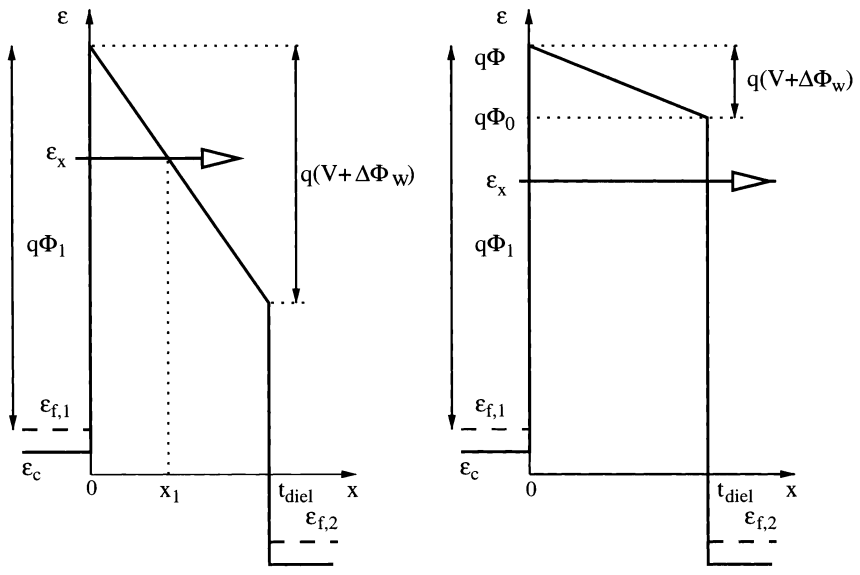


Figure 14. Schematic of an energy barrier in the Fowler–Nordheim tunneling (left) and direct tunneling (right) regime.

2.7.1. Original Fowler–Nordheim Formula

The original Fowler–Nordheim formula assumes a triangular shape of the energy barrier. This is motivated by the fact that only tunneling at strong electric fields was studied. The WKB approximation (70) for the transmission coefficient reads

$$TC(\mathscr{E}_x) = \exp\left(-\frac{2}{\hbar}\int_0^{x_1}\sqrt{2m_{\text{diel}}(\mathscr{E}_c - \mathscr{E}_x)}\,dx\right) \tag{161}$$

The classical turning point x_1 is (see the left part of Fig. 14)

$$x_1 = \frac{\mathscr{E}_{F,1} + q\Phi_1 - \mathscr{E}_x}{qE_{\text{diel}}} \tag{162}$$

and the dielectric conduction band edge for a triangular barrier

$$\mathscr{E}_c(x) = \mathscr{E}_{F,1} + q\Phi_1 - qE_{\text{diel}}x \tag{163}$$

where the electric field in the dielectric E_{diel} is caused by the different Fermi levels and the work function difference $\Delta\Phi_W$:

$$E_{\text{diel}} = \frac{\mathscr{E}_{F,1} - \mathscr{E}_{F,2} + q\Delta\Phi_W}{qt_{\text{diel}}} \tag{164}$$

The third approximation is to assume equal materials for both electrodes, so that $\Delta\Phi_W = 0$. The WKB-based transmission coefficient can then be applied and yields

$$TC(\mathscr{E}_x) = \exp\left(-2\frac{\sqrt{2m_{\text{diel}}}}{\hbar}\int_0^{x_1}\sqrt{\mathscr{E}_{F,1} + q\Phi_1 - qE_{\text{diel}}x - \mathscr{E}_x}\,dx\right) \tag{165}$$

$$= \exp\left[4\frac{\sqrt{2m_{\text{diel}}}}{3\hbar qE_{\text{diel}}}(\mathscr{E}_{F,1} + q\Phi_1 - qE_{\text{diel}}x - \mathscr{E}_x)^{3/2}\Big|_0^{x_1}\right] \tag{166}$$

$$= \exp\left[4\frac{\sqrt{2m_{\text{diel}}}}{3\hbar qE_{\text{diel}}}(-\mathscr{E}_{F,1} - q\Phi_1 + \mathscr{E}_x)^{3/2}\right] \tag{167}$$

$$= \exp\left\{-4\frac{\sqrt{2m_{\text{diel}}}}{3\hbar qE_{\text{diel}}}[q\Phi_1 - (\mathscr{E}_x - \mathscr{E}_{F,1})]^{3/2}\right\} \tag{168}$$

Using this expression in (160), the current density becomes

$$J = \frac{4\pi qm_{\text{eff}}}{h^3}\int_{\mathscr{E}_{F,2}}^{\mathscr{E}_{F,1}}\exp\left\{-\frac{4\sqrt{2m_{\text{diel}}}}{3\hbar qE_{\text{diel}}}[q\Phi_1 - (\mathscr{E}_x - \mathscr{E}_{F,1})]^{3/2}\right\}(\mathscr{E}_{F,1} - \mathscr{E}_x)\,d\mathscr{E}_x \tag{169}$$

This integral cannot be solved analytically. Hence, the fourth approximation is to expand the square root into a first-order Taylor series around $q\Phi_1$:

$$[q\Phi_1 - (\mathscr{E}_x - \mathscr{E}_{F,1})]^{3/2} \approx (q\Phi_1)^{3/2} + \frac{3}{2}(\mathscr{E}_x - \mathscr{E}_{F,1})(q\Phi_1)^{1/2} \tag{170}$$

Inserting this expression into (169) and setting $\epsilon = \mathscr{E}_x - \mathscr{E}_F$, yields

$$J = \frac{4\pi qm_{\text{eff}}}{h^3}\exp\left[-\frac{4\sqrt{2m_{\text{diel}}}}{3\hbar qE_{\text{diel}}}(q\Phi_1)^{3/2}\right]\int_{\mathscr{E}_{F,2}-\mathscr{E}_{F,1}}^{0}\exp\left[\frac{2\sqrt{2m_{\text{diel}}}}{\hbar qE_{\text{diel}}}(q\Phi_1)^{1/2}\epsilon\right]\epsilon\,d\epsilon \tag{171}$$

With

$$\int \epsilon\exp(\lambda\epsilon)\,d\epsilon = \frac{1}{\lambda^2}\exp(\lambda\epsilon)(\lambda\epsilon - 1) \tag{172}$$

and

$$a = -\frac{4\sqrt{2m_{\text{diel}}}}{3\hbar q E_{\text{diel}}}(q\Phi_1)^{3/2}, \quad \lambda = \frac{2\sqrt{2m_{\text{diel}}}}{\hbar q E_{\text{diel}}}(q\Phi_1)^{1/2} \tag{173}$$

the current density becomes

$$J = \frac{4\pi q m_{\text{eff}}}{h^3}\exp(a)\int_{\mathscr{E}_{\text{F},2}-\mathscr{E}_{\text{F},1}}^{0}\exp(\lambda\epsilon)\epsilon\, d\epsilon \tag{174}$$

$$= \frac{4\pi q m_{\text{eff}}}{h^3}\exp(a)\frac{1}{\lambda^2}\exp[\lambda(\mathscr{E}_{\text{F},2}-\mathscr{E}_{\text{F},1})][\lambda(\mathscr{E}_{\text{F},2}-\mathscr{E}_{\text{F},1})-1] \tag{175}$$

The fifth assumption is now that $\mathscr{E}_{\text{F},1} \gg \mathscr{E}_{\text{F},2}$, leading to

$$J = \frac{4\pi q m_{\text{eff}}}{h^3}\exp(a)\frac{1}{\lambda^2} \tag{176}$$

or

$$J = \frac{q^3 m_{\text{eff}}}{8\pi m_{\text{diel}} h q\Phi_1}E_{\text{diel}}^2\exp\left(-\frac{4\sqrt{2m_{\text{diel}}(q\Phi_1)^3}}{3\hbar q E_{\text{diel}}}\right) \tag{177}$$

which is the equation commonly known as the Fowler–Nordheim formula [100]. Note that there is a difference between the effective electron mass in the electrode (m_{eff}) and the effective electron mass in the dielectric (m_{diel}).

2.7.2. Correction for Direct Tunneling

The equation derived above is only valid for triangular barriers; that is, the case of high applied voltages. In Ref. [101], Schuegraf proposed a correction to the Fowler–Nordheim formula to account for tunneling in the direct tunneling regime. In this case, the transmission coefficient is

$$TC(\mathscr{E}) = \exp\left(-\frac{2}{\hbar}\int_0^{t_{\text{diel}}}\sqrt{2m_{\text{diel}}(\mathscr{E}_c-\mathscr{E}_x)}\,dx\right) \tag{178}$$

where t_{diel} is the dielectric thickness. The conduction band edge is again approximated by a linear shape

$$\mathscr{E}_c(x) = \mathscr{E}_{\text{F},1} + q\Phi_1 - qE_{\text{diel}}x \tag{179}$$

The band edges $q\Phi$ and $q\Phi_0$ are given by (see the right part of Fig. 14)

$$q\Phi = \mathscr{E}_{\text{F},1} + q\Phi_1 \tag{180}$$

$$q\Phi_0 = \mathscr{E}_{\text{F},1} + q\Phi_1 - qE_{\text{diel}}t_{\text{diel}} \tag{181}$$

As for the triangular energy barrier, it is assumed that the electrodes have equal work functions: $\Delta\Phi_W = 0$. Using these expressions, the transmission coefficient becomes

$$TC(\mathscr{E}_x) = \exp\left(-2\frac{\sqrt{2m_{\text{diel}}}}{\hbar}\int_0^{t_{\text{diel}}}\sqrt{q\Phi - qE_{\text{diel}}x - \mathscr{E}_x}\,dx\right) \tag{182}$$

$$= \exp\left[4\frac{\sqrt{2m_{\text{diel}}}}{3\hbar q E_{\text{diel}}}(q\Phi - qE_{\text{diel}}x - \mathscr{E}_x)^{3/2}\Big|_0^{t_{\text{diel}}}\right] \tag{183}$$

$$= \exp\left\{-4\frac{\sqrt{2m_{\text{diel}}}}{3\hbar q E_{\text{diel}}}[(q\Phi - \mathscr{E}_x)^{3/2} - (q\Phi_0 - \mathscr{E}_x)^{3/2}]\right\} \tag{184}$$

The exponent can be approximated using a first-order Taylor series expansion around $q\Phi_1$ and $q\Phi_1 - qE_{diel}t_{diel}$, respectively:

$$(q\Phi - \mathscr{E}_x)^{3/2} = (\mathscr{E}_{F,1} + q\Phi_1 - \mathscr{E}_x)^{3/2} \tag{185}$$

$$= [q\Phi_1 - (\mathscr{E}_x - \mathscr{E}_{F,1})]^{3/2} \tag{186}$$

$$\approx (q\Phi_1)^{3/2} + \frac{3}{2}(\mathscr{E}_x - \mathscr{E}_{F,1})(q\Phi_1)^{1/2} \tag{187}$$

$$(q\Phi_0 - \mathscr{E}_x)^{3/2} = (\mathscr{E}_{F,1} + q\Phi_1 - qE_{diel}t_{diel} - \mathscr{E}_x)^{3/2} \tag{188}$$

$$= [(q\Phi_1 - qE_{diel}t_{diel}) - (\mathscr{E}_x - \mathscr{E}_{F,1})]^{3/2} \tag{189}$$

$$\approx (q\Phi_1 - qE_{diel}t_{diel})^{3/2} + \frac{3}{2}(\mathscr{E}_x - \mathscr{E}_{F,1})(q\Phi_1 - qE_{diel}t_{diel})^{1/2} \tag{190}$$

With the temporary variable η

$$\eta = (q\Phi - \mathscr{E}_x)^{3/2} - (q\Phi_0 - \mathscr{E}_x)^{3/2}$$

$$\approx -(q\Phi_1 - qE_{diel}t_{diel})^{3/2} + (q\Phi_1)^{3/2} - \frac{3}{2}(\mathscr{E}_x - \mathscr{E}_{F,1})[(q\Phi_1)^{1/2} - q\Phi_1 - qE_{diel}t_{diel})^{1/2}] \tag{191}$$

the tunnel current density becomes

$$J = \frac{4\pi q m_{eff}}{h^3} \int_{\mathscr{E}_{F,2}}^{\mathscr{E}_{F,1}} TC(\mathscr{E}_x)(\mathscr{E}_{F,1} - \mathscr{E}_x)\, d\mathscr{E}_x \tag{192}$$

$$\approx \frac{4\pi q m_{eff}}{h^3} \int_{\mathscr{E}_{F,2}}^{\mathscr{E}_{F,1}} \exp\left(-4\frac{\sqrt{2m_{diel}}}{3\hbar q E_{diel}}\eta\right)(\mathscr{E}_{F,1} - \mathscr{E}_x)\, d\mathscr{E}_x \tag{193}$$

With the abbreviations

$$a = \frac{4\pi q m_{eff}}{h^3} \tag{194}$$

$$b = -\frac{4\sqrt{2m_{diel}}}{3\hbar q E_{diel}}[(q\Phi_1)^{3/2} - (q\Phi_1 - qE_{diel}t_{diel})^{3/2}] \tag{195}$$

$$c = -\frac{2\sqrt{2m_{diel}}}{\hbar q E_{diel}}[(q\Phi_1)^{1/2} - (q\Phi_1 - qE_{diel}t_{diel})^{1/2}] \tag{196}$$

the tunnel current density can be written as

$$J = a\exp(b) \int_{\mathscr{E}_{F,1}}^{\mathscr{E}_{F,2}} \exp[c(\mathscr{E}_x - \mathscr{E}_{F,1})](\mathscr{E}_{F,1} - \mathscr{E}_x)\, d\mathscr{E}_x \tag{197}$$

With $\epsilon = \mathscr{E}_x - \mathscr{E}_{F,1}$ this yields

$$J = -a\exp(b) \int_{\mathscr{E}_{F,2} - \mathscr{E}_{F,1}}^{0} \exp(c\epsilon)\epsilon\, d\epsilon \tag{198}$$

Using (172), this integral becomes

$$J = \frac{a\exp(b)}{c^2}\{1 - \exp[-c(\mathscr{E}_{F,1} - \mathscr{E}_{F,2})][1 + c(\mathscr{E}_{F,1} - \mathscr{E}_{F,2})]\} \tag{199}$$

which, for $\mathscr{E}_{F,1} \gg \mathscr{E}_{F,2}$, simplifies to

$$J = \frac{a\exp(b)}{c^2} \tag{200}$$

or, inserting the expressions for a, b, and c

$$J = \frac{q^3 m_{eff}}{8\pi h m_{diel}[(q\Phi_1)^{1/2} - (q\Phi_1 - qV_{diel})^{1/2}]^2} E_{diel}^2 \times \exp\left\{-\frac{4\sqrt{2m_{diel}}}{3\hbar q E_{diel}}[(q\Phi_1)^{3/2} - (q\Phi_1 - qV_{diel})^{3/2}]\right\} \quad (201)$$

which is the equation used in Refs. [101, 102]. In some publications, the equation is rewritten to make it more similar to the Fowler–Nordheim formula:

$$J = \frac{q^3 m_{eff}}{8\pi m_{diel} h q\Phi_1 B_1} E_{diel}^2 \exp\left(-\frac{4\sqrt{2m_{diel}(q\Phi_1)^3}B_2}{3\hbar q E_{diel}}\right) \quad (202)$$

with the additional correction terms B_1, B_2 given as

$$B_1 = \left[1 - \left(1 - \frac{qE_{diel}t_{diel}}{q\Phi_1}\right)^{1/2}\right]^2$$
$$B_2 = \left[1 - \left(1 - \frac{qE_{diel}t_{diel}}{q\Phi_1}\right)^{3/2}\right] \quad (203)$$

For a triangular barrier, the correction factors become $B, = B, = 1,$ and the expression simplifies to (177). Note that using these equations, the minimum tunneling current occurs for $E_{diel} = 0$ Vim, which, for a work function difference $\neq 0$, does not occur at the minimum applied bias.

2.8. Trap-Assisted Tunneling

Besides direct or Fowler–Nordheim tunneling, which are 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 electrically erasable programmable read-only-memories (EEPROMs). It is therefore assumed that traps arise in the dielectric layer due to the repeated high-voltage stress. The increased tunneling current at low bias is called stress-induced leakage current (SILC) and is mainly responsible for the degradation of the retention time of nonvolatile memory devices [103]. It is now generally accepted that it is caused by inelastic trap-assisted tunnel transitions and that the traps are created by the electric high-field stress during the writing and erasing processes [103–108]. SILC has widely been studied and modeled in MOS capacitors [109–111] and EEPROM devices [112].

This section gives a brief overview of trap-assisted tunneling models, describes two frequently encountered models (Chang's and Ielmini's model), and elaborates on a sophisticated model originally proposed by Jimenez et al. The adaption of this model to allow its inclusion in device simulators is described in some detail.

2.8.1. Model Overview

Numerous models have been presented to describe trap-assisted tunneling in the gate dielectric of MOS devices. These models usually share the equation for the current density, which is given by an integration along the gate dielectric [113]:

$$J = q\int_0^{t_{diel}} \frac{N_T(x)}{\tau_c(x) + \tau_e(x)} dx \quad (204)$$

In this expression, N_T denotes the trap concentration, and τ_c and τ_e denote the capture and emission times of the considered trap. Because both processes—capture and emission—must happen in sequence, they both determine the current density. However, differences exist in how the capture and emission times are calculated. Some models use constant capture and emission cross sections to calculate the respective times. Another important point is the

distribution in space, where the traps are usually assumed to follow a Gaussian distribution. The distribution in energy is also crucial. Commonly, it is either assumed that traps have a Gaussian distribution in energy or that they are located at a certain energy level below the dielectric conduction band. The assumption of a discrete energy level for specific trap types is backed by spectroscopic analyses [114]. Additionally, the tunneling process can either be elastic, where the energy of the tunneling electron is conserved, or inelastic, where the energy of the tunneling electron changes. Recent studies and experiments have shown strong evidence for the tunneling process being inelastic [115–117].

2.8.1.1. Chang's Model A frequently used model is the generalized trap-assisted tunneling model presented by Chang et al. [118, 119]. The current density reads

$$J = \mathrm{q}\int_0^{t_{\mathrm{diel}}} A N_{\mathrm{T}}(x)\frac{P_1(x)P_2(x)}{P_1(x)+P_2(x)}\,dx \tag{205}$$

where A denotes a fitting constant, $N_{\mathrm{T}}(x)$ the spatial trap concentration, and P_1 and P_2 the transmission coefficients of electrons captured and emitted by traps. Using $\tau_{\mathrm{c}} \sim P_1/P_2$ and $\tau_e \sim P_2/P_1$, this expression reduces to (204). A similar model was used by Ghetti et al. [76]

$$J = \int_0^{t_{\mathrm{diel}}} C_{\mathrm{T}} N_{\mathrm{T}}(x)\frac{J_{\mathrm{in}}J_{\mathrm{out}}}{J_{\mathrm{in}}+J_{\mathrm{out}}}\,dx \tag{206}$$

who assumed a constant capture cross section C_{T} for the traps. The symbols J_{in} and J_{out} denote the capture and emission currents. Essentially the same formula was used by other authors as well [116, 120].

2.8.1.2. Ielmini's Model Considerable research has been done by Ielmini et al. [121–124], who describe inelastic TAT and also take hopping conduction into account [125, 126]. They derive the trap-assisted current by an integration along the dielectric thickness and energy

$$J = \int_0^{t_{\mathrm{diel}}} dx \int_{\mathscr{E}_{\mathrm{min}}}^{\mathscr{E}_{\mathrm{max}}} \tilde{J}(\mathscr{E}_{\mathrm{T}}, x)\,d\mathscr{E} \tag{207}$$

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

$$\tilde{J}(\mathscr{E}_{\mathrm{T}}, x) = J_{\mathrm{cl}} - J_{\mathrm{el}} = J_{\mathrm{er}} - J_{\mathrm{cr}} = \mathrm{q}N'_{\mathrm{T}}W_{\mathrm{c}}\left(1 - \frac{f_{\mathrm{T}}(\mathscr{E}_{\mathrm{T}}, x)}{f_{\mathrm{l}}(\mathscr{E}_{\mathrm{T}}, x)}\right) \tag{208}$$

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

$$(J_{\mathrm{cl}} - J_{\mathrm{el}}) + (J_{\mathrm{cr}} - J_{\mathrm{er}}) = \mathrm{q}N_{\mathrm{T}}\frac{\partial f_{\mathrm{T}}}{\partial t} \tag{209}$$

an observation that will be revisited in Section 2.8.2.4. The model assumes a constant capture cross section.

2.8.1.3. Compact Trap-Assisted Tunneling Models For application in circuit simulators or to catch a quick glimpse at the effects of trap-assisted tunneling, compact models are required. A frequently used expression is based on the work of Ricco et al. [109]. They describe the trapping- and detrapping processes by

$$J_{\mathrm{TAT}} = JC_{\mathrm{T}}TC_1(N_{\mathrm{T}} - n_{\mathrm{T}}) = \mathrm{q}\nu n_{\mathrm{T}}TC_2 \tag{210}$$

where J is the supply current density at the interface, C_{T} the capture cross section, TC_1 and TC_2 the transmission coefficients from the left and right sides of the dielectric to the trap, n_{T} the concentration of trapped electrons that is smaller than or equal to the trap concentration N_{T}, and ν their escape frequency. The highest contribution comes from traps that have $TC_1 \approx TC_2$; therefore, the trap-assisted tunnel current becomes

$$J_{\mathrm{TAT}} = \mathrm{q}\nu n_{\mathrm{T}}TC = \mathrm{q}\nu C_{\mathrm{T}}N_{\mathrm{T}}\frac{J}{JC_{\mathrm{T}} + \mathrm{q}\nu}TC \tag{211}$$

A modified version of this expression was used by Ghetti et al. [111, 127]. Other more or less empirical trap-assisted tunneling models based on SILC measurements are presented in Ref. [128]. These comprise hopping conduction

$$J = C_1 E_{\mathrm{diel}} \exp\left(-\frac{\mathrm{q}\Phi_{\mathrm{a}}}{\mathrm{k_B}T}\right) \tag{212}$$

where Φ_{a} is an activation potential, and the frequently applied Poole–Frenkel tunneling formula [128–134]. This model describes the emission of trapped electrons and reads

$$J = AE_{\mathrm{diel}} \exp\left(-\frac{\mathscr{E}_{\mathrm{T}}}{\mathrm{k_B}T}\right) \exp\left(\frac{\mathrm{q}}{\mathrm{k_B}T}\sqrt{\frac{\mathrm{q}E_{\mathrm{diel}}}{\pi\kappa_0 r^2}}\right) \tag{213}$$

where r is the refractive index of the dielectric, $\mathscr{E}_{\mathrm{T}}$ is the difference between the conduction band in the dielectric and the trap energy, and the coefficient A depends on the trap concentration. The main motivation to use this expression is that the trap-assisted gate current density was found to be a linear function of the square root of the dielectric field, in contrast to the Fowler–Nordheim tunneling current, which is a linear function of the dielectric field. Note, however, that no trapping-detrapping considerations enter this equation.

2.8.2. The Model of Jiménez et al.

A model for trap-assisted inelastic tunneling has been developed by Jiménez et al. [135]. Their model is based on the theory of nonradiative capture and emission of electrons by multiphonon processes [136]. The main difference to the models described before is that it does not require constant capture cross sections as fitting parameters but calculates them for each trap based on the trap energy level and the shape of the energy barrier.

2.8.2.1. Capture and Emission Probabilities The tunneling model is based on a two-step tunneling process via traps in the dielectric that incorporates energy loss by phonon emission [135]. Figure 15 shows the basic two-step process of an electron tunneling from a region with higher Fermi energy (the cathode) to a region with lower Fermi energy (the anode). To avoid integration in energy, the initial electron energy is assumed to be located at the average kinetic energy, which, for the parabolic dispersion relation (1) and the Maxwellian distribution (20), is

$$\frac{\langle\mathscr{E}\rangle}{\langle 1\rangle} = \frac{\int_0^\infty \mathscr{E}f(\mathscr{E})g(\mathscr{E})\,d\mathscr{E}}{\int_0^\infty f(\mathscr{E})g(\mathscr{E})\,d\mathscr{E}} = \frac{\int_0^\infty \mathscr{E}^{3/2}\exp\left(-\frac{\mathscr{E}}{\mathrm{k_B}T}\right)d\mathscr{E}}{\int_0^\infty \mathscr{E}^{1/2}\exp\left(-\frac{\mathscr{E}}{\mathrm{k_B}T}\right)d\mathscr{E}} = \frac{3}{2}\mathrm{k_B}T \tag{214}$$

During the capture process (W_{c}), the difference in total energy between the initial and final state is released by means of phonon emission ($\hbar\omega$). An electron captured by a trap can then be emitted into the anode (W_{e}).

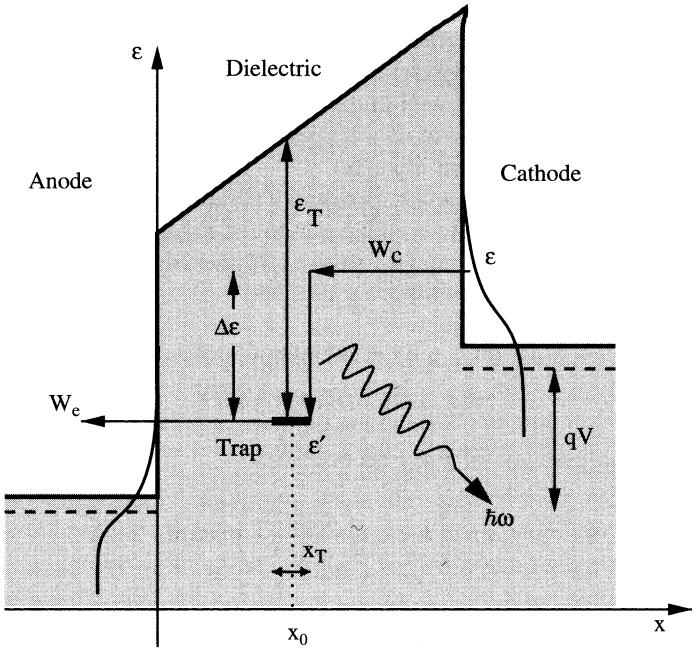


Figure 15. The trap-assisted tunneling process [135].

The rate with which an electron with energy $\mathscr{E}$ is captured by a trap located at position x and energy $\mathscr{E}'$ is given by [137]

$$W_c(x, \mathscr{E}', \mathscr{E}) = \frac{\pi}{\hbar^2 \omega} |V_e|^2 S \left(1 - \frac{P}{S}\right)^2 I_P(\xi) \exp\left[-(2f_P + 1)S + \frac{\Delta\mathscr{E}}{2k_B T}\right] \tag{215}$$

Here, S is the Huang–Rhys factor, which characterizes the electron–phonon interaction [138]; $\hbar\omega$ is the energy of the phonons involved in the transitions, $\Delta\mathscr{E} = \mathscr{E} - \mathscr{E}'$; and $P = \Delta\mathscr{E}/\hbar\omega$ is the number of phonons emitted due to this energy difference. In the simulations, the value of $S\hbar\omega$ was used as fitting parameter.

The population of phonons f_P is given by the Bose–Einstein statistics

$$f_P = \left[\exp\left(\frac{\hbar\omega}{k_B T}\right) - 1\right]^{-1} \tag{216}$$

The function $I_P(\xi)$ is the modified Bessel function of order P, with

$$\xi = 2S\sqrt{f_P(f_P + 1)} \tag{217}$$

The term $|V_e|^2$ in (215) denotes the transition matrix element, which is calculated by an integration over the trap cube [136]

$$|V_e|^2 = 5\pi S(\hbar\omega)^2 \frac{\hbar^2}{2m_{diel}\mathscr{E}_T} \int_{x_0 - x_T/2}^{x_0 + x_T/2} |\Psi(x)|^2 \, dx \tag{218}$$

In this expression, x_T denotes the side length of the trap cube, estimated as

$$x_T = \frac{\hbar}{\sqrt{2m_{diel}\mathscr{E}_T}} \left(\frac{4\pi}{3}\right)^{1/3} \tag{219}$$

The symbol $\mathscr{E}_T$ denotes the energy difference between the trap energy and the barrier conduction band edge as shown in Fig. 15. For the emission of electrons from the trap to the anode, elastic tunneling is assumed. Hence, the probability of emission to the anode is equal to the probability of capture from the anode, which is calculated from (215).

The numerical evaluation of (218) requires the calculation of the wave functions in the dielectric layer, which, however, degrades the computational efficiency of a multipurpose device simulator where simulation speed is crucial. To avoid this, the barriers have been transformed to take advantage of the well-known solutions for constant potentials. Two cases must be distinguished; namely, the case of a trapezoidal barrier and the case of a triangular barrier. The two cases are depicted in Fig. 16.

For capture processes and for emission processes where the electron faces a trapezoidal barrier, the barrier is transformed into a step function of height equal to the potential at the middle point between x = 0 and x = x_0 ($\mathscr{E}_m$ in the left part of Fig. 16), x_0 being the position of the trap inside the dielectric. Assuming

$$\begin{aligned} \Psi(x \le 0) &= A\sin(k_1 x + \alpha) \\ \Psi(x > 0) &= B\exp(-k_2 x) \end{aligned} \tag{220}$$

the wave function at the position of the trap becomes

$$\Psi(x) = A\sin\left[\arctan\left(\frac{m_{\text{diel}}}{m_{\text{eff}}}\frac{k_1}{k_2}\right)\right]\exp(-k_2 x) \tag{221}$$

where m_{diel} and m_{eff} are the electron masses in the dielectric and the neighboring electrode, respectively. The wave numbers are given by

$$\begin{aligned} k_1 &= \frac{1}{\hbar}\sqrt{2m_{\text{eff}}(\mathscr{E} - \mathscr{E}_c)} \\ k_2 &= \frac{1}{\hbar}\sqrt{2m_{\text{diel}}(\mathscr{E}_m - \mathscr{E})} \end{aligned} \tag{222}$$

For emission processes in which the barrier is triangular (the electron energy is above the dielectric conduction band at some point between the trap and the anode), two regions in the dielectric must be distinguished. The first one, between the interface at x = 0 and the point x = x_{FN} (see the right part of Fig. 16) has the height $\mathscr{E}_{FN}$. The height of the approximated barrier in the other region is then the value of the barrier, $\mathscr{E}_m$, in the middle point between x = x_{FN} and the position of the trap x = x_0. With this new barrier and the assumptions for the wave functions in the three regions

$$\Psi(x \le 0) = A\sin(k_1 x + \text{a},) \tag{223}$$

$$\Psi(0 < \text{x} \le \text{x}_{\text{m}}) = B\sin(k_2 x + \text{a},) \tag{224}$$

$$\Psi(x_{FN} < \text{x} \le x_0) = \text{C}\exp[-k_3(\text{x} - x_{FN})] \tag{225}$$

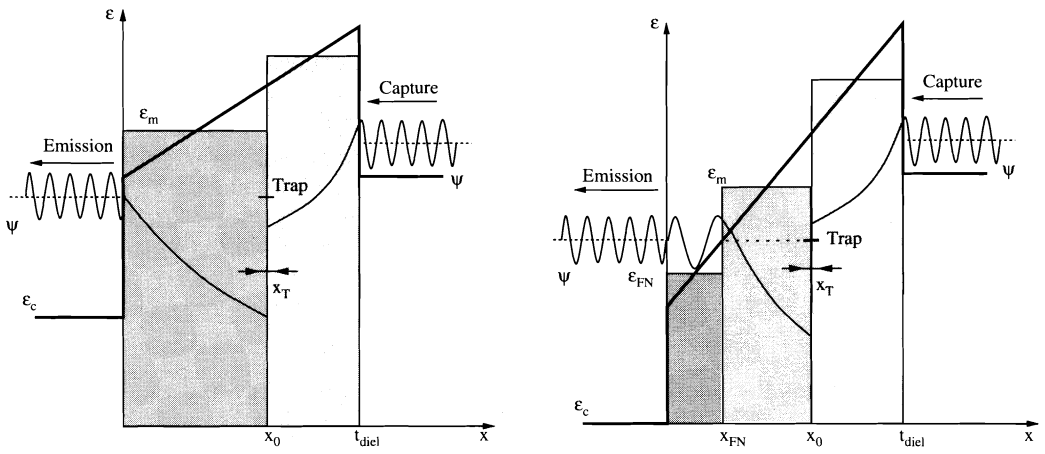


Figure 16. The approximate shape of the barrier in the direct (left) and Fowler–Nordheim regime (right) [171].

the wave function at the position of the trap becomes

$$\Psi(x) = A\frac{\sin\alpha_1}{\sin\alpha_2}\sin(k_2 x_{\mathrm{FN}} + \alpha_2)\exp[-k_3(x - x_{\mathrm{FN}})] \tag{226}$$

with the symbols

$$\begin{aligned}\alpha_1 &= \arctan\left(\frac{k_1}{k_2}\tan\alpha_2\right)\\ \alpha_2 &= \arctan\left(\frac{k_2}{k_3}\right) - k_2 x_{\mathrm{FN}}\end{aligned} \tag{227}$$

The corresponding wave numbers are given as

$$\begin{aligned}k_1 &= \frac{1}{\hbar}\sqrt{2m_{\mathrm{eff}}(\mathscr{E} - \mathscr{E}_{\mathrm{c}})}\\ k_2 &= \frac{1}{\hbar}\sqrt{2m_{\mathrm{diel}}(\mathscr{E} - \mathscr{E}_{\mathrm{FN}})}\\ k_3 &= \frac{1}{\hbar}\sqrt{2m_{\mathrm{diel}}(\mathscr{E}_{\mathrm{m}} - \mathscr{E})}\end{aligned} \tag{228}$$

Using expression (221) and (226), the integration in (218) can be performed analytically, which allows the capture and emission probabilities to be calculated without the need for numerical integration.

2.8.2.2. Capture and Emission Times Once the capture and emission probabilities have been obtained, the corresponding times can be calculated. The inverse of the capture time is given by [135, 139]

$$\tau_{\mathrm{c}}^{-1}(x) = \int_{\mathscr{E}'}^{\infty} W_{\mathrm{c}}(x, \mathscr{E}', \mathscr{E}) g_{\mathrm{c}}(\mathscr{E}) f_{\mathrm{c}}(\mathscr{E})\, d\mathscr{E} \tag{229}$$

where $g_{\mathrm{c}}(\mathscr{E})$ denotes the two-dimensional density of states and $f_{\mathrm{c}}(\mathscr{E})$ the electron energy distribution function in the cathode. For the above-stated assumption that all electrons are captured from the same energy level $\mathscr{E}_{\mathrm{c}} + 3/2\mathrm{k_B}T$ in the cathode, this expression can be approximated by

$$\tau_{\mathrm{c}}^{-1}(x) \approx W_{\mathrm{c}}\left(x, \mathscr{E}', \mathscr{E}_{\mathrm{c}} + \frac{3}{2}\mathrm{k_B}T\right) n_{\mathrm{c}} \tag{230}$$

where n_c is the sheet carrier concentration in the cathode, which is determined by the transport model used in the device simulator. The inverse of the emission time is [135]

$$\tau_{\mathrm{e}}^{-1}(x) = \int_{-\infty}^{\mathscr{E}'} W_{\mathrm{e}}(x, \mathscr{E}', \mathscr{E}) g_{\mathrm{a}}(\mathscr{E})[1 - f_{\mathrm{a}}(\mathscr{E})]\, d\mathscr{E} \tag{231}$$

Assuming $f_{\mathrm{a}}(\mathscr{E}) \approx 0$ in the anode and elastic tunneling for the emission process ($\mathscr{E} = \mathscr{E}'$), the emission time becomes

$$\tau_{\mathrm{e}}^{-1}(x) \approx W_{\mathrm{e}}(x, \mathscr{E}', \mathscr{E}') g_{\mathrm{a}}(\mathscr{E}')\hbar\omega \tag{232}$$

where the energy loss is restricted to values less than $\hbar\omega$. To check the validity of the approximations for the wave functions, the resulting capture and emission times have been compared to results using a Schrodinger-Poisson solver for a MOS capacitor with the parameters $\mathscr{E}_{\mathrm{T}} = 2.8$ eV, $S\hbar\omega = 1.6$ eV, and a trap concentration of $N_{\mathrm{T}} = 10^{19}$ cm^{-3}. As can be seen in Fig. 17, the analytical and the numerical results are very close. Electrons are captured from the right and emitted to the left in this figure. Thus, for traps near the right side of the barrier, the capture time is very low and the emission time is very high. The oscillations in the emission time for high bias are due to the fact that in this regime, the energy barrier has a triangular shape, which gives rise to an oscillating wave function, in contrast to the decaying wave function for a trapezoidal barrier.

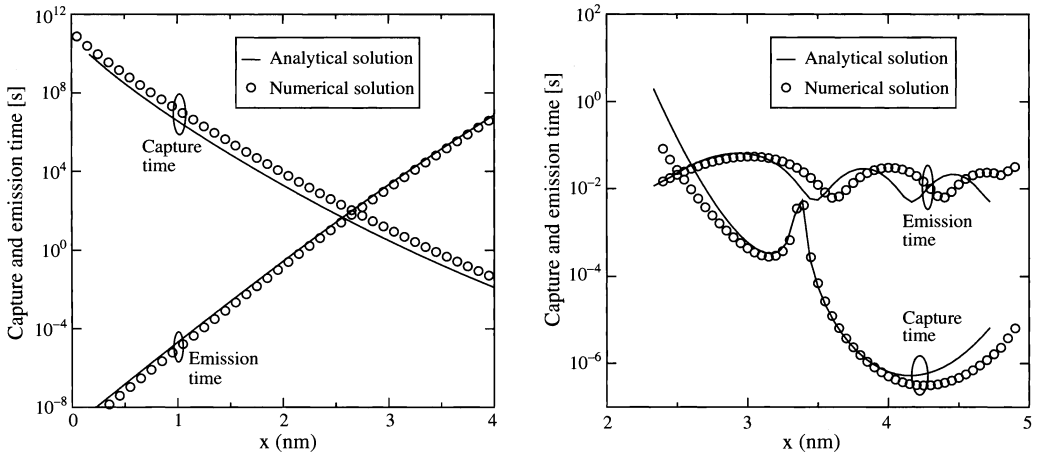


Figure 17. Comparison of the analytic solution with a numerical solution for the capture and emission times at a gate bias of 3 V (left) and 7 V (right) [171].

2.8.2.3. Steady-State Current The total steady-state tunneling current is derived as the sum of the trap-assisted tunneling current (204) and the direct tunneling current computed from the Tsu–Esaki formula (13)

$$J = J_{\mathrm{TAT}} + J_{\mathrm{Tsu\text{-}Esaki}} \tag{233}$$

Figure 18 shows the dependence of the gate current density on the model parameters $\mathscr{E}_{\mathrm{T}}$ (trap energy level) and $S\hbar\omega$ for a fixed phonon energy of $\hbar\omega = 10$ meV in an MOS capacitor. For a low trap energy level, traps are located near the conduction band edge in the dielectric, and direct tunneling prevails. With increasing trap energy level, the trap-assisted component becomes stronger and exceeds the direct tunneling current for low bias. The current density shows a peak at low bias, which is due to the alignment of the trap energy level with the cathode conduction band edge. The Huang–Rhys factor has only a minor influence on the results, as shown in the right part of Fig. 18.

2.8.2.4. Transient Current Models of trap-assisted transitions are commonly employed to calculate steady-state SILC in MOS capacitors, whereas transient SILC has hardly been studied [110, 121]. However, transient tunneling current becomes important at high switching speed where the transients of the trap charging and discharging processes may degrade

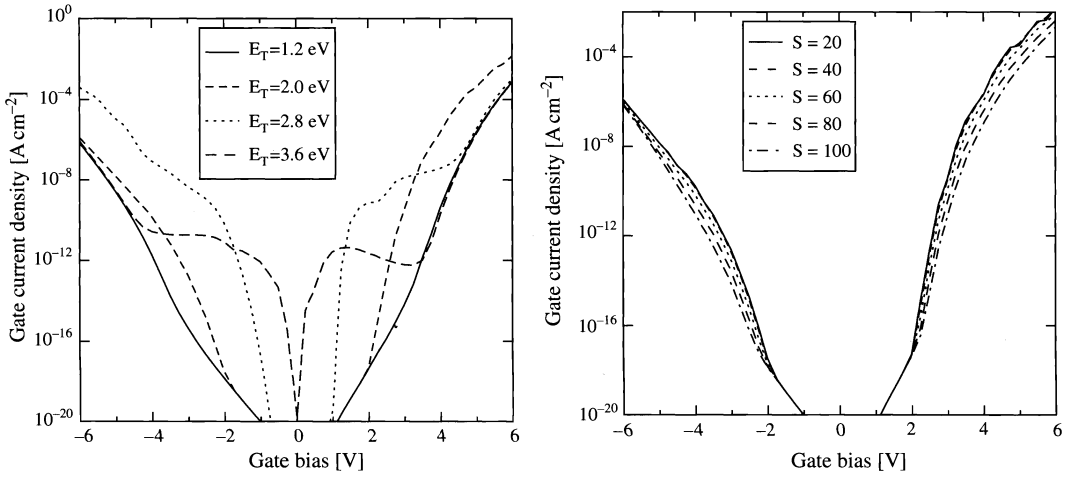


Figure 18. Dependence of the tunneling current on the trap energy level (left) and on the Huang–Rhys factor for a fixed phonon energy of 10 meV (right) [171].

signal integrity. For the calculation of transient SILC, it is necessary to calculate capture and emission times at each time step. Considering a spatial trap distribution $N_T(x)$ across the dielectric layer, the rate equation for the concentration of occupied traps at position x reads

$$N_T(x)\frac{df_T(x,t)}{dt} = N_T(x)[1 - f_T(x,t)]\tau_c^{-1}(x,t) - N_T(x)f_T(x,t)\tau_e^{-1}(x,t) \tag{234}$$

where $f_T(x, t)$ is the trap occupancy function, and $\tau_c(x, t)$ and $\tau_e(x, t)$ are the inverse capture and emission times of electrons by a trap placed at position x. In the static case, capture and emission processes are in equilibrium and $df_T(x,t)/dt = 0$. In the transient case, however, capture and emission times include transitions from the cathode and the anode

$$\begin{aligned}\tau_c^{-1}(x,t) &= \tau_{ca}^{-1}(x,t) + \tau_{cc}^{-1}(x,t)\\ \tau_e^{-1}(x,t) &= \tau_{ea}^{-1}(x,t) + \tau_{ec}^{-1}(x,t)\end{aligned} \tag{235}$$

where τ_{ca} and τ_{cc} are the capture times to the anode and to the cathode and τ_{ea} and τ_{ec} the corresponding emission times. To calculate the local trap occupancy, the differential equation (234) must be solved. If the capture and emission times τ_c^{-1} and τ_e^{-1} are constant over time, like in a discharging process with a constant potential distribution, the solution of (234) can be given in a closed form

$$f_T(x,t) = f_T(x,0)\exp\left(-\frac{t}{\tau_m(x,t)}\right) + \frac{\tau_m(x,t)}{\tau_c(x,t)}\left[1 - \exp\left(-\frac{t}{\tau_m(x,t)}\right)\right] \tag{236}$$

with $\tau_m^{-1} = \tau_c^{-1} + \tau_e^{-1}$.

A more general approach is to look at the change of the trap distribution at discrete time steps. Integration of (234) in time between t_i and t_{i+1} and changing to discrete time steps yields

$$f_T(x,t_i) - f_T(x,t_{i-1}) \approx \tau_c^{-1}(x,t_{i-1})\Delta t_i - \tau_m^{-1}(x,t_{i-1})\bar{f}_i\Delta t_i$$

where the abbreviations $\Delta t_i = t_i - t_{i-1}$ and $\bar{f}_i = [f_T(x,t_i) + f_T(x,t_{i-1})]/2$ have been used. Thus, it is possible to write the trap distribution over time in the following recursive manner:

$$f_T(x,t_i) = A_i + B_i f_T(x,t_{i-1}) \tag{237}$$

where the symbols A_i, B_i, and C_i are calculated from

$$\begin{aligned}A_i &= \frac{\tau_c^{-1}(x,t_i)\Delta t_i}{1 + C_i}\\ B_i &= \frac{1 - C_i}{1 + C_i}\\ C_i &= \frac{\tau_m^{-1}(x,t_i)\Delta t_i}{2}\end{aligned} \tag{238}$$

Once the time-dependent occupancy function in the dielectric is known, the tunnel current through each of the interfaces is

$$J_{TAT,\,Anode}(t) = q\int_0^{t_{diel}} N_T(x)\left\{\tau_{ca}^{-1}(x,t) - f_T(x,t)\left[\tau_{ca}^{-1}(x,t) + \tau_{ea}^{-1}(x,t)\right]\right\}dx \tag{239}$$

$$J_{TAT,\,Cathode}(t) = q\int_0^{t_{diel}} N_T(x)\left\{\tau_{cc}^{-1}(x,t) - f_T(x,t)\left[\tau_{cc}^{-1}(x,t) + \tau_{ec}^{-1}(x,t)\right]\right\}dx \tag{240}$$

2.9. Model Comparison

This section outlined a number of tunneling models useful for the simulation of tunneling in semiconductor devices. For practical device simulation, however, it is often not clear which model to select for the application at hand. Therefore, Table 3 summarizes the main model

Table 3. A hierarchy of tunneling models and their properties.

	Fowler–Nordheim Model	Schuegraf Model	Tsu–Esaki Analytic WKB	Tsu–Esaki Gundlach	Tsu–Esaki Numeric WKB	Tsu–Esaki Transfer-Matrix	Tsu–Esaki QTBM	Inelastic TAT	Frenkel–Poole
FN tunneling	✓	✓	✓	✓	✓	✓	✓		
Direct tunneling		✓	✓	✓	✓	✓	✓		
EVB tunneling process			✓	✓	✓	✓	✓		
QM current oscillations				✓		✓	✓		
Dielectric stacks					✓	✓	✓		
Numerical stability						—			
Trap-assisted tunneling								✓	✓
Trap occupancy modeling								✓	
Transient TAT								✓	
Computational effort	**Low**	**Low**			**High**	**High**	**High**		**Low**

Note: WKB, Wentzel–Kramers–Brillouin; QTBM, quantum transmitting boundary method; TAT, trap-assisted tunneling; FN, Fowler–Nordheim; EVB, electrons from valence band; QM, quantum mechanical.

features and also gives the approximate computational effort. The following points can be concluded [140]:

- Especially the Fowler–Nordheim, Schuegraf, and Frenkel–Poole models have a very low computational effort because 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, due to the analytical calculation of the overlap integral, poses only moderate computational effort. This model can be used for the simulation of leakage in EEPROMs or trap-rich dielectric devices (see Section 3.2.2.1).
- 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. The transfer-matrix method should be used with care due to its poor numerical stability.

3. APPLICATIONS

Gate leakage is one of the most important issues for contemporary complementary metal-oxide semiconductor (CMOS) devices. Based on the tunneling models outlined so far, two different application areas will be investigated in this section. First, gate leakage in contemporary MOS transistors will be studied and compared to measurements. Emphasis is put on the distinction between the different sources of the tunneling current; namely, the region below the gate and the region near the drain and source extensions.

Device engineers commonly rely on gate leakage measurements of turned-off devices to evaluate the power consumption of CMOS circuits. This may lead to erroneous results because for turned-on devices, hot-carrier tunneling prevails that may exceed the turned-off tunneling current. Models that are based on simplified assumptions of the carrier energy distribution function fail to predict gate leakage in such cases.

Advanced CMOS devices will use alternative dielectric materials as gate dielectrics. However, a pronounced trade-off between the height of the energy barrier and the dielectric permittivity exists. This makes the use of optimization necessary to find the optimum layer composition. Furthermore, alternative dielectrics are not ideal insulators but contain defects that give rise to trap-assisted tunneling. As a state-of-the-art example, tunneling in ZrO_2-based MOS capacitors will be studied and compared to measurements.

As a second important application area, nonvolatile memories will be studied. Unlike MOS transistors, nonvolatile memory devices represent an application where tunneling is not a spurious effect, but crucial for the device functionality. After a short review of nonvolatile memory technology, the tunneling current of conventional EEPROMs and advanced structures will be studied. In contrast to these devices, SONOS (silicon-oxide-nitride-oxide-silicon) EEPROM devices store the charge not on an isolated contact, but in a layer of trap-rich dielectric.

Recent efforts to reduce the charging time of nonvolatile memory devices resulted in multibarrier tunneling devices and EEPROMs with asymmetrically layered tunnel dielectrics. The operation of these devices will briefly be described at the end of this chapter. All simulations are performed using the device simulator Minimos-NT [141].

3.1. Tunneling in MOS Transistors

The gate leakage current in contemporary MOS transistors poses a major problem for further device scaling. This section describes simulation results of MOS transistors, outlines the effect of various device parameters, shows how to account for hot-carrier tunneling in turned-on devices, and elaborates on the use of alternative dielectric materials to replace SiO_2 as a gate dielectric. First, however, the tunneling paths in MOS transistor structures will be reviewed.

3.1.1. Tunneling Paths in MOS Transistors

Tunneling in an MOS transistor, as shown in the left part of Fig. 19, basically can be separated into a path between the gate and the channel and a path between the gate and the source and drain extension areas [142]. Tunneling in the source and drain extension areas can exceed tunneling in the channel by orders of magnitude. This is related to two effects: First, instead of n-p or p-n tunneling, n-n or p-p tunneling prevails. Second, the potential difference and thus the bending of the energy barrier is high. This increased tunneling current in the source and drain extension areas can be a serious problem if measurements are performed on long-channel MOSFETs to characterize their short-channel pendants, because the edge tunneling currents exceed the channel tunneling current by orders of magnitude. Furthermore, there is a fundamental difference between tunneling in MOS transistors and MOS capacitors [1, 143]. In contrast to MOS transistors, MOS capacitors, which are biased in strong inversion, cannot supply the amount of carriers as predicted by the tunneling model. This effect is termed *substrate-limited* tunneling, because the tunneling current is limited by the generation rate in the substrate. In the channel of an inverted MOS transistor, on the other hand, carriers can always be supplied by the source and drain contacts. This effect is depicted in the right part of Fig. 19.

3.1.2. Channel Tunneling

In this section, the effects of various device parameters on the gate leakage of MOS capacitors are studied. This is equivalent to tunneling in MOS transistors if only channel tunneling (n-p or p-n) is considered, and the source, drain, and bulk contacts are grounded. The parameters investigated are

- the doping of the polysilicon gate contact,
- the doping of the substrate,
- the thickness of the dielectric layer,
- the barrier height of the dielectric,
- the carrier mass in the dielectric,
- the dielectric permittivity, and
- the lattice temperature.

The typical shape of the gate current density in turned-off nMOS and pMOS devices is depicted in Fig. 20. A SiO_2 gate dielectric thickness of 2 nm and an acceptor or donor doping of 5×10^{17} 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 [144].

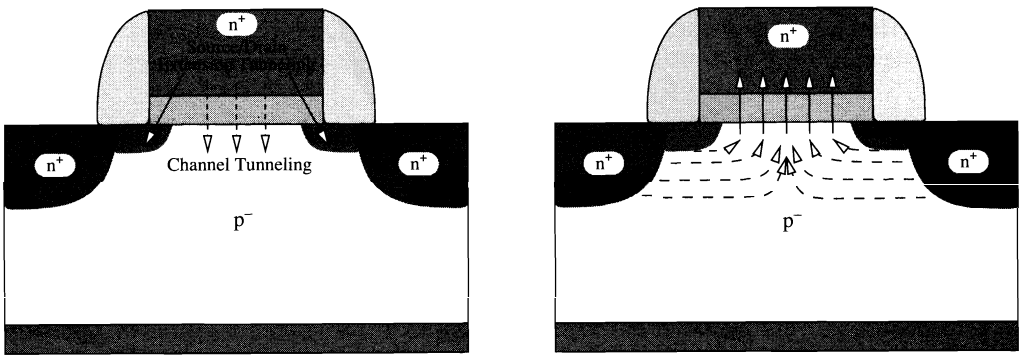


Figure 19. The different tunneling paths (channel tunneling, source and drain extension tunneling) in a MOS transistor (left). In a MOS transistor biased in inversion (right), tunneling electrons are supplied from the source and drain reservoirs, which is not possible in a MOS capacitor.

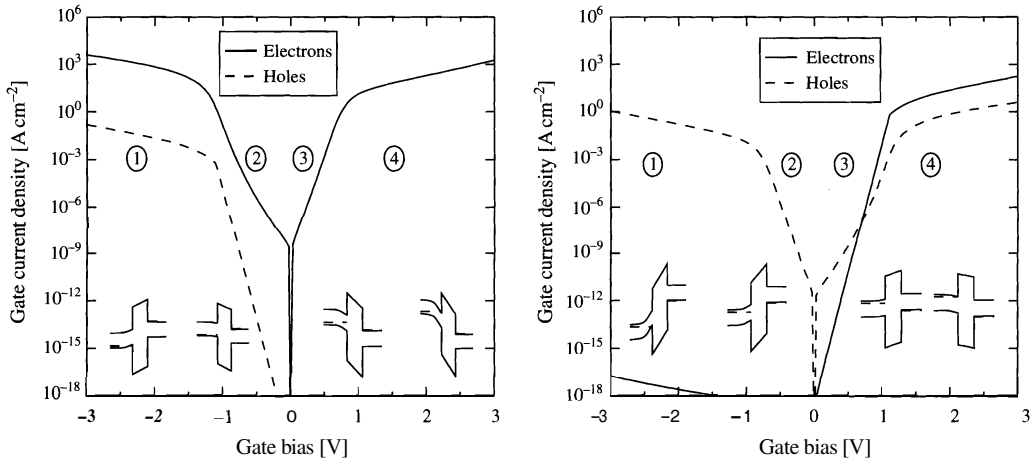


Figure 20. Channel tunneling regions in an *n*MOS (left) and a *p*MOS (right). The insets show the approximate shape of the band edge energies.

3.1.2.1. Effect of the Polysilicon Gate Doping on the Channel Tunneling Highly doped polysilicon is used as material for the gate contact to allow adjustable work functions and realize CMOS circuits. Figure 21 shows the electron and hole tunneling current density for different doping of the polysilicon gate contact. In the *n*MOS, gate leakage generally

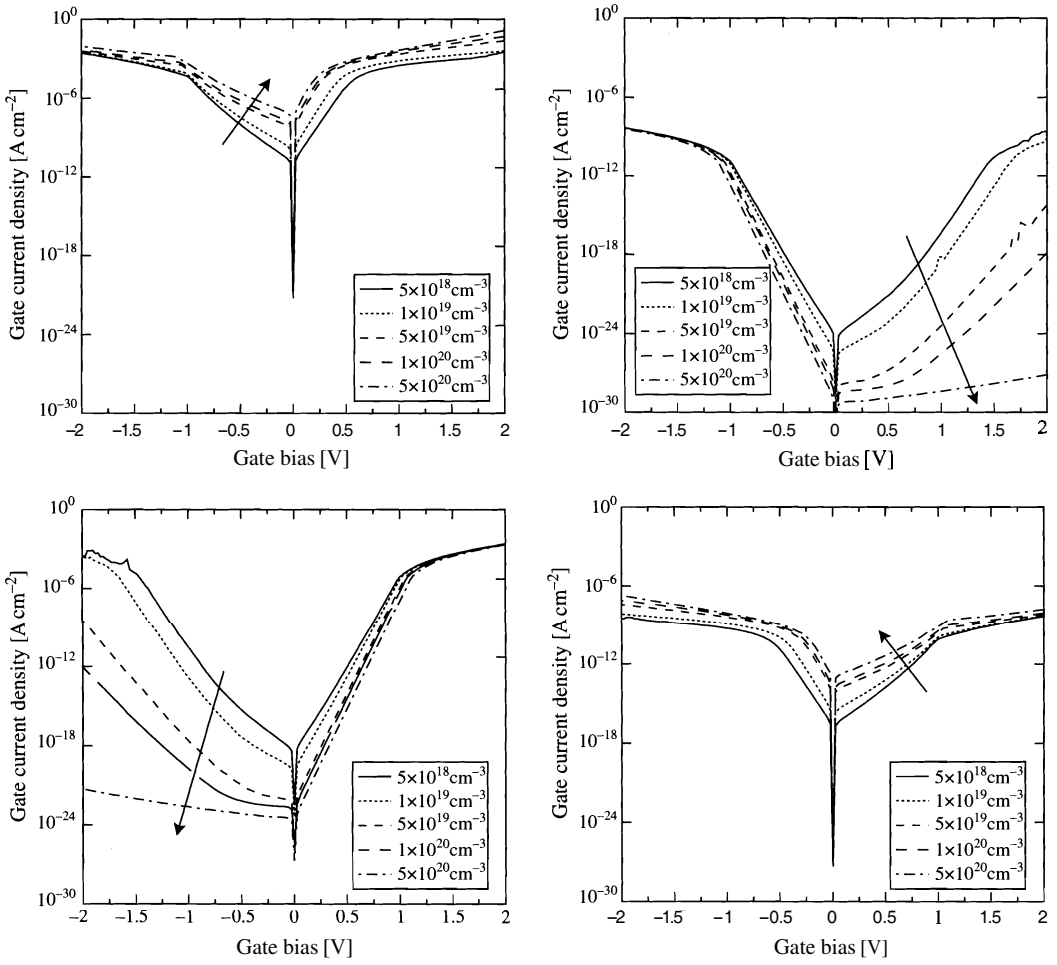


Figure 21. Electron (left) and hole (right) current density in an *n*MOS (top) and a *p*MOS (bottom) with different doping of the polysilicon gate. Substrate doping is 10^{18} cm^{-3}; dielectric thickness is 2 nm.

increases with increasing doping of the polysilicon gate because tunneling current is dominated by electrons. In the *p*MOS, a higher polysilicon doping leads to reduced electron tunneling current and increased hole tunneling current. The effect on the overall leakage depends on the doping and the gate bias.

3.1.2.2. Effect of the Substrate Doping on the Channel Tunneling Figure ***22*** shows the electron and hole tunneling current density for different doping of the substrate. With increasing substrate doping, the majority tunneling component (electrons in the *n*MOS, holes in the *p*MOS) is reduced in both the *n*MOS and *p*MOS devices, whereas the minority component increases.

3.1.2.3. Effect of the Dielectric Thickness on the Channel Tunneling The physical thickness of the dielectric has the largest impact on the gate current density, as shown in Fig. 23. Increasing the gate dielectric thickness by 0.4 nm leads to a decrease of all tunneling current components by several orders of magnitude.

3.1.2.4. Effect of the Barrier Height on the Channel Tunneling The main parameter, besides the thickness of the dielectric, influencing tunneling current is the height of the energy barrier. The influence of this parameter is depicted in Fig. 24. Different dielectric materials strongly differ in their work function difference to silicon. It must be distinguished between the barrier height for electrons and for holes. The most frequently used dielectric material SiO_2 has an electron barrier height of about 3.2 eV and a hole barrier height of

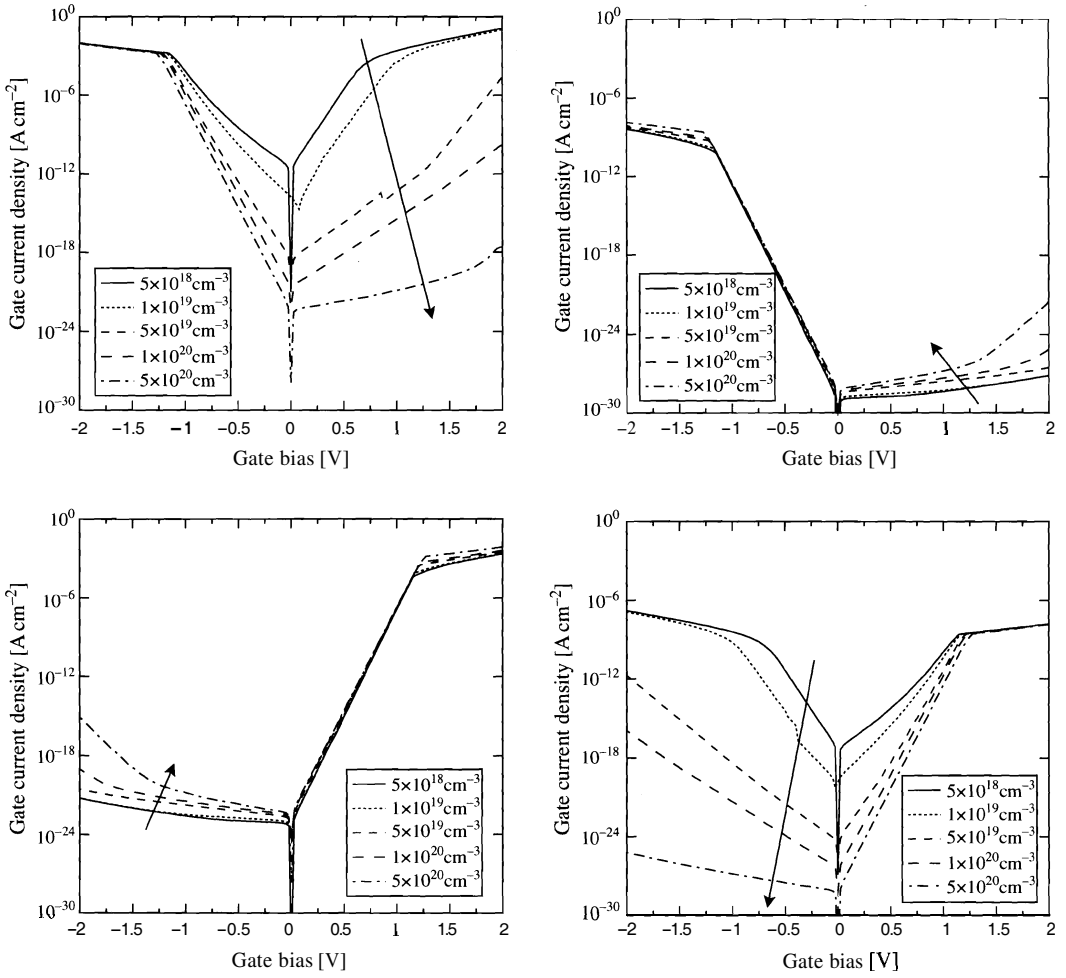


Figure 22. Electron (left) and hole (right) current density in an *n*MOS (top) and a *p*MOS (bottom) with different doping of the substrate. Gate polysilicon doping is 5×10^{20} cm^{-3}; dielectric thickness is 2 nm.

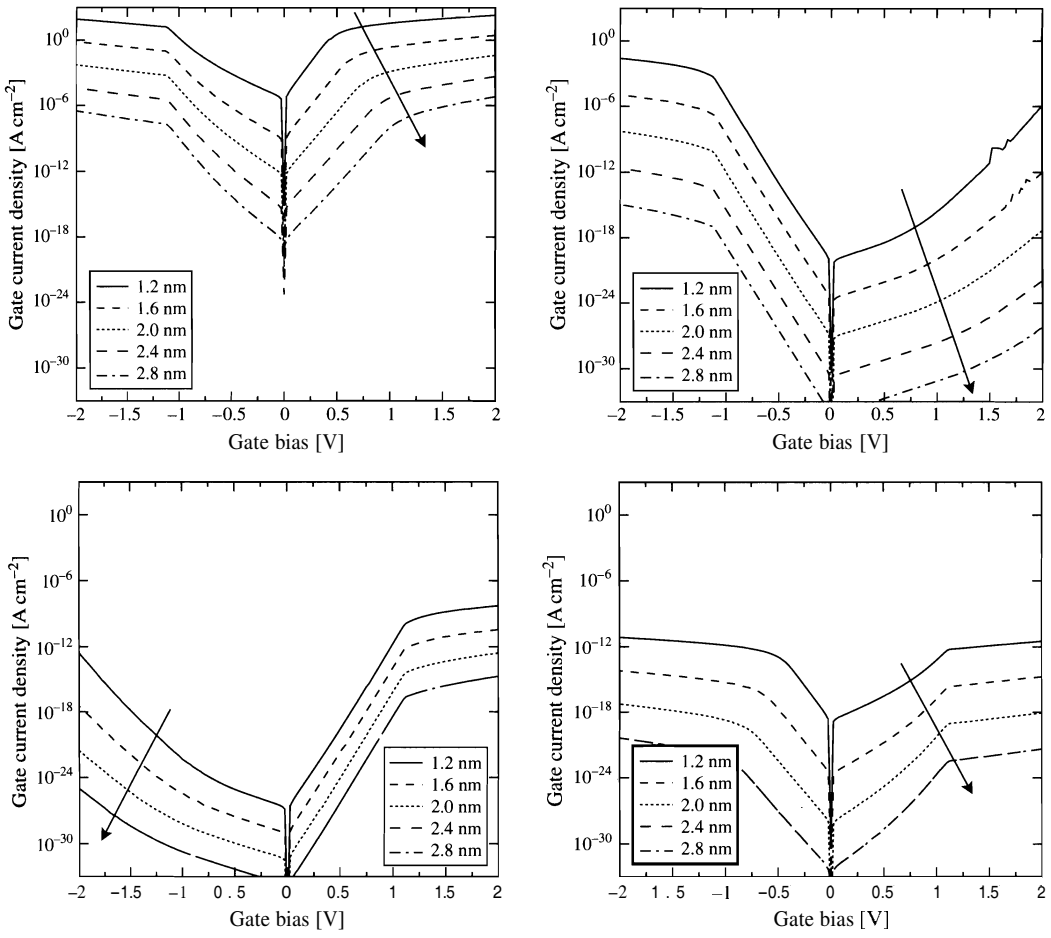


Figure 23. Electron (left) and hole (right) current density in an *n*MOS (top) and a *p*MOS (bottom) with different thickness of the dielectric layer. Gate polysilicon doping is 5×10^{20} cm^{-3}; substrate doping is 5×10^{18} cm^{-3}.

approximately 4.6 eV. The measurement of these material parameters is difficult, and values in the available literature vary widely (see Section 3.1.5).

3.1.2.5. Effect of the Carrier Mass in the Dielectric on the Channel Tunneling Being the parameter with the highest uncertainty, the electron and hole mass in the dielectric is commonly used as a fitting parameter to reproduce measurements. Its influence on the gate current density is shown in Fig. 25. An increase in the carrier mass by 0.1 m_0 leads to a reduction in the gate current density by about a factor of 10. It must, of course, be held in mind that with the approaches described so far, tunneling is described by a single value for the carrier mass. Its use as a fitting parameter may thus well be justified. Recent investigations, however, report an increase of the electron mass with reducing thickness of the dielectric layer, which is backed by measurements and tight-binding band structure calculations [145–147].

3.1.2.6. Effect of the Dielectric Permittivity on the Channel Tunneling The permittivity of the dielectric layer influences the tunneling current density in two ways: First, the shape of the energy barrier — and thus the transmission coefficient-changes. Second, the inversion charge — and thus the band edge energy — in the channel is affected. The effect of varying dielectric permittivity is shown in Fig. 26. Especially in the low-bias regime, a higher permittivity strongly increases the gate current density.

3.1.2.7. Effect of the Lattice Temperature on the Channel Tunneling The lattice temperature enters the gate tunneling current via the electron energy distribution functions in the polysilicon gate and in the channel. The transmission coefficient, being based on

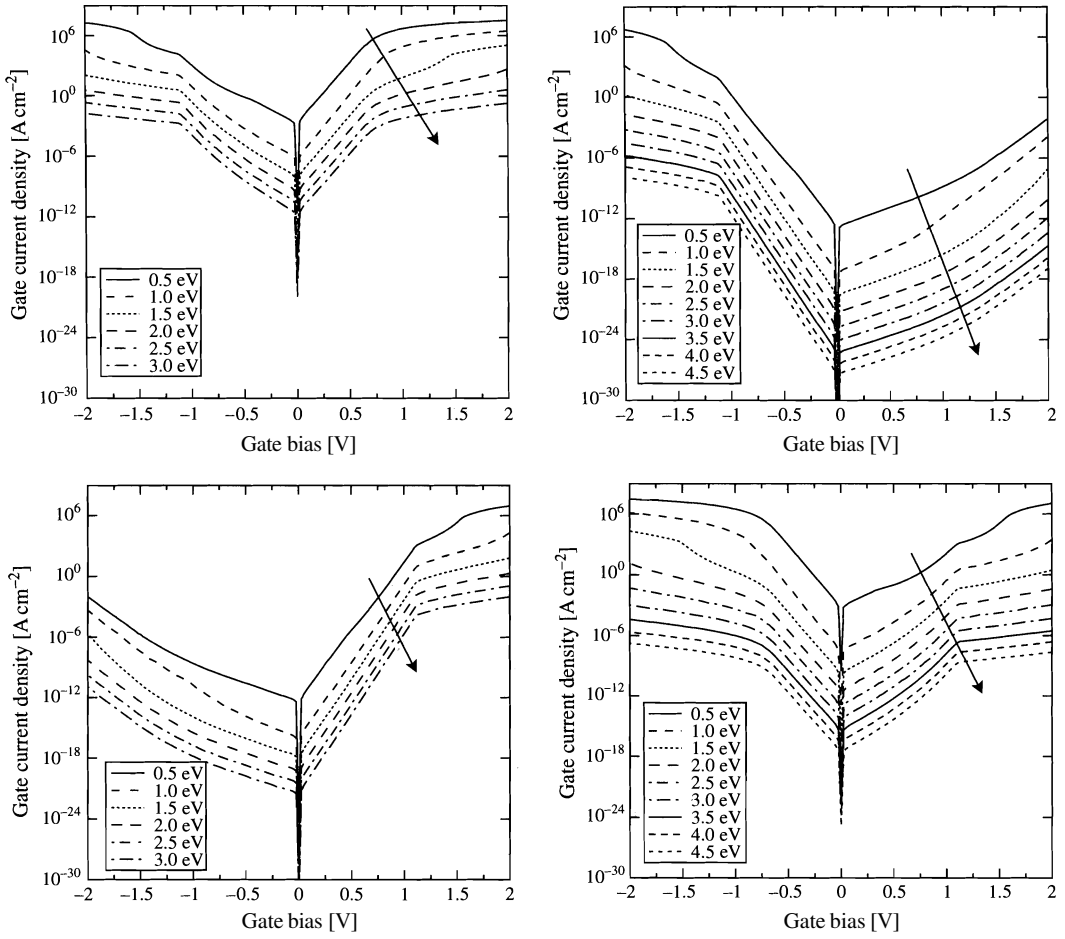


Figure 24. Effect of the electron and hole barrier height on electron tunneling current (left) and hole tunneling current (right) in an *n*MOS (top) and a *p*MOS (bottom) with 2-nm dielectric thickness, 10^{20} cm^{-3} polysilicon, and 5 x 10^{18} cm^{-3} substrate doping.

quantum-mechanical reasoning alone, is not affected by the lattice temperature. However, the supply function depends on the lattice temperature. The impact on the gate current density is shown in Fig. 27. Rising temperature increases the tunneling current density in all cases.

3.1.2.8. Comparison to Measurements Because almost all available measurements of gate leakage in MOS devices are performed on turned-off MOS transistors, a cbmparison with measurements will be given before turned-on devices are investigated in Section 3.1.4. The Tsu–Esaki model with an analytical WKB transmission coefficient is in good agreement with recently reported data for devices with different gate lengths and bulk doping [1, 142] as shown in Fig. 28 for *n*MOS (left) and *p*MOS devices (right) [148]. 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. Additional measurements have been performed on MOSFETs with a gate dielectric thickness of 1.5 nm (see the lower part of Fig. 28) and compared with the results of other simulators (UTQUANT [149] and MEDICI [150]). Under inversion condition the fit is not perfect, whereas under accumulation the measurements can be reproduced well. Note that with UTQUANT, the low-bias tunneling current cannot be reproduced, and MEDICI completely failed for the pMOS device.

3.1.2.9. Validity of Compact Models Because the computational effort for the numerical integration in Tsu–Esaki's formula or the evaluation of the quasi-bound states is numerically expensive, it is reasonable to ask if compact models can describe tunneling, at least for

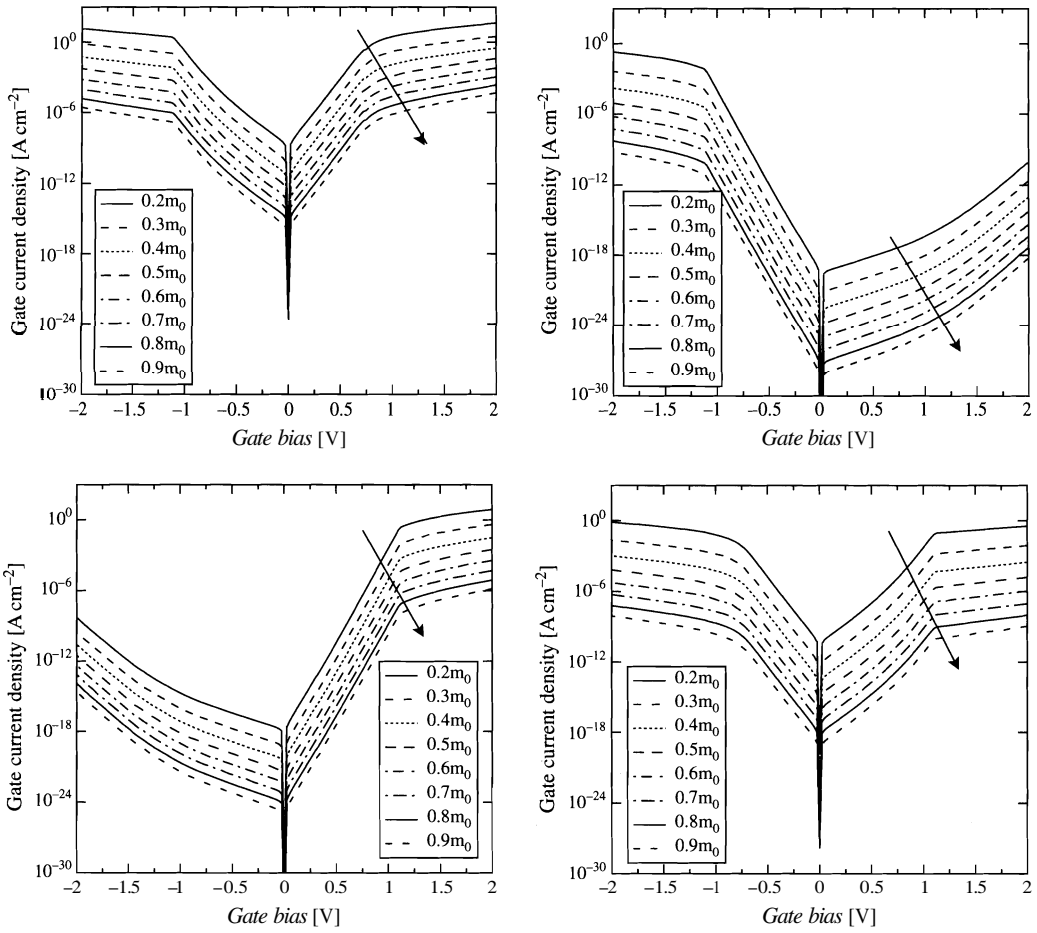


Figure 25. *Effect of the carrier mass on electron tunneling current (left) and hole tunneling current (right) in an nMOS (top) and a pMOS (bottom) with* 2-nm *dielectric thickness,* 10^{20} cm^{-3} *polysilicon, and* 5 x 10^{18} cm^{-3} *substrate doping.*

single-layer dielectrics. The compact tunneling models outlined in Section 2.7 are compared in Fig. 29 for a symmetrical metal-dielectric-metal structure (left) and for an *n*MOS structure with 3-nm dielectric thickness (right). For the metal-dielectric-metal structure, Schuegraf's model yields almost the same results as the computationally much more expensive Tsu–Esaki model. The Fowler–Nordheim model delivers correct values only for high bias. It is thus only applicable to describe high-field transport through gate dielectrics, like program and erase cycles in EEPROM devices. For the MOS structure in the right part of Fig. 29, the Schuegraf model fails to describe the tunneling current density at low bias. For high bias, however, it may be used to provide an estimation of the gate current. The Fowler–Nordheim model totally fails for this application. Furthermore, the Fowler–Nordheim model shows the minimum gate current at minimum electric field in the dielectric, and not for the minimum gate bias.

3.1.3. Source/Drain Extension Tunneling

In the following examples, the same devices as in Section 3.1.1 are investigated, but this time only the tunneling current in the source and drain extension areas (n-n or p-p) is taken into account. Because the barrier height, carrier mass, and dielectric thickness shows the same impact on the gate current density as for the case of channel tunneling, the corresponding figures are omitted.

3.1.3.1. Effect of the Polysilicon Gate Doping on the Source and Drain Extension Tunneling Figure 30 shows the effect of the doping concentration in the polysilicon gate on the extension region gate current density. Increasing the polysilicon doping leads to a

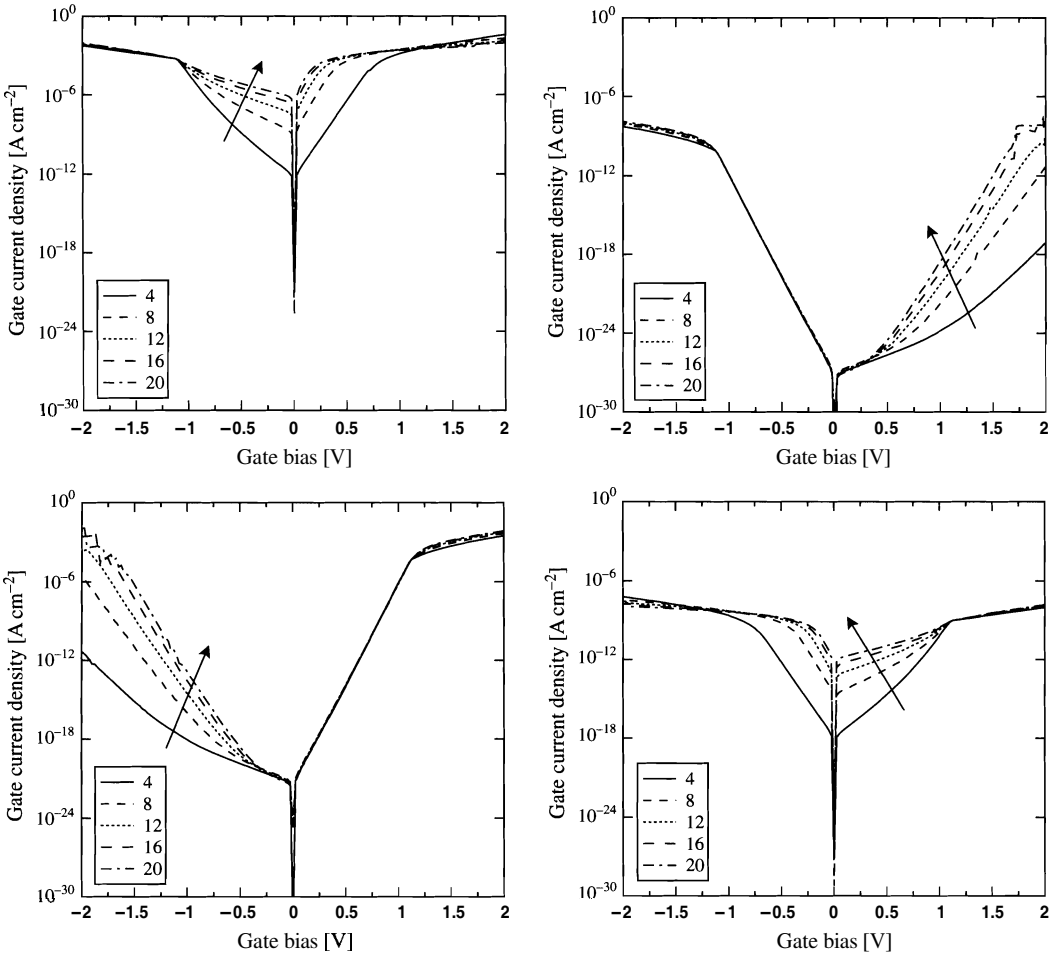


Figure *26.* Effect of the dielectric permittivity κ/κ_0 on electron tunneling current (left) and hole tunneling current (right) in an *n*MOS (top) and a *p*MOS (bottom) with 2-nm dielectric thickness, 10^{20} cm^{-3} polysilicon, and 5 x 10^{18} cm^{-3} substrate doping.

slight increase of the main tunneling component and to a strong decrease of the minority tunneling component in both nMOS and pMOS devices.

3.1.3.2. Effect of the Substrate Doping Concentration on the Source and Drain Extension Tunneling Figure ***31*** shows the effect of the substrate doping concentration on the extension region gate current density. Similar to the polysilicon gate doping, a higher substrate doping leads to increased majority and decreased minority tunneling current.

3.1.3.3. Effect of the Dielectric Permittivity on the Source and Drain Extension Tunneling Figure ***32*** shows the effect of the dielectric permittivity on the extension region gate current density. In contrast to the channel-tunneling case, the low-bias regime is not influenced by the permittivity. Furthermore, the influence on the majority tunneling current component depends on the bias: The electron tunneling component in the *n*MOS decreases for negative bias and increases for positive bias. The hole tunneling component in the *p*MOS shows exactly the inverse trend.

3.1.3.4. Effect of the Lattice Temperature on the Source and Drain Extension Tunneling Figure ***33*** shows the effect of the temperature on the extension region gate current density. Especially the minority carriers (holes in the *n*MOS, electrons in the *p*MOS) show strongly increased tunneling current with higher temperature. Unlike in the channel tunneling case, the majority tunneling component is hardly influenced by the temperature.

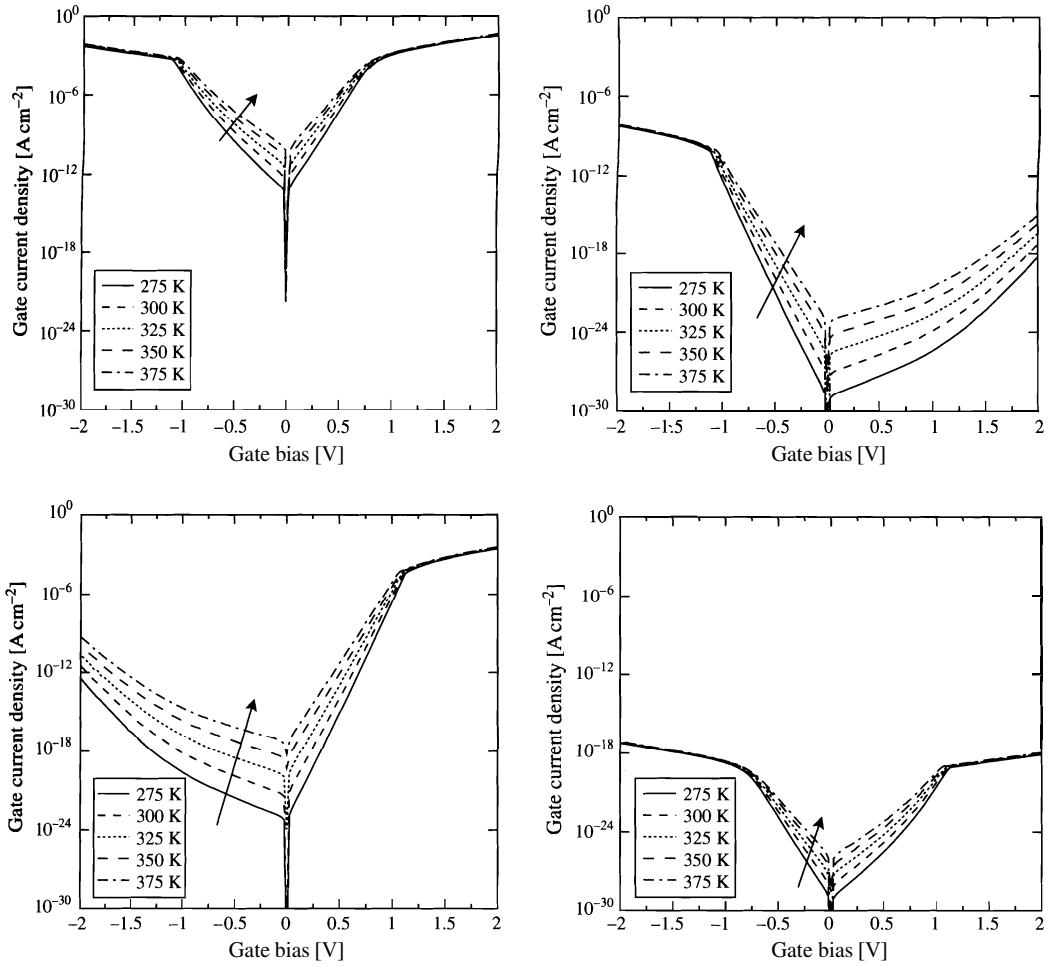


Figure 27. Effect of the lattice temperature on electron tunneling current (left) and hole tunneling current (right) in an *n*MOS (top) and a *p*MOS (bottom) with 2-nm dielectric thickness, 10^{20} cm^{-3} polysilicon, and 5 x 10^{18} cm^{-3} substrate doping.

3.1.4. Hot-Carrier Tunneling in MOS Transistors

It has been shown in Section 2.3 that the distribution function in the channel of a turned-on MOS transistor heavily deviates from the shape implied by a Fermi–Dirac or Maxwellian distribution. A model for the non-Maxwellian shape of the distribution function was presented that accurately reproduced the carrier energy distribution along the channel.

To check the impact of this wrong high-energy behavior, the integrand of the Tsu–Esaki formula, namely the expression $TC(\mathscr{E})N(\mathscr{E})$, has been evaluated for a standard device, as shown in the left part of Fig. 34, and compared to Monte Carlo results [151, 152]. The simulated device had a gate length of 100 nm and a gate dielectric thickness of 3 nm. Though at low energies, the difference between the non-Maxwellian distribution function (28) and the heated Maxwellian distribution (24) seems to be negligible, the amount of overestimation of the incremental gate current density for the heated Maxwellian distribution reaches several orders of magnitude at 1 eV and peaks when the electron energy exceeds the barrier height. This spurious effect is clearly more pronounced for points at the drain end of the channel where the electron temperature is high. The non-Maxwellian shape of the distribution function, indicated by the full line, reproduces the Monte Carlo results very well.

The region of high electron temperature is confined to only a small area near the drain contact, as shown in the right part of Fig. 34, where the gate current density along the channel is compared to Monte Carlo results. At the point of the peak electron temperature, which is located at approximately $x = 0.8L_g$, the heated Maxwellian approximation overestimates

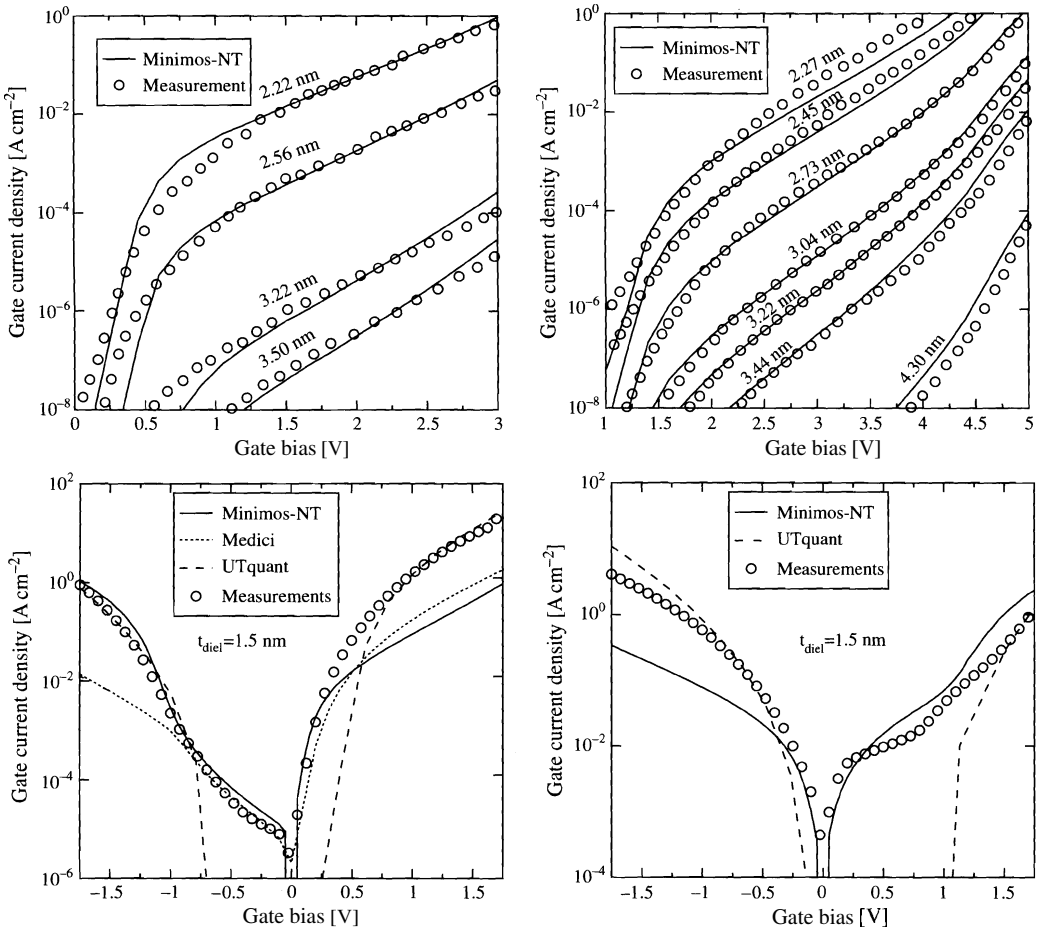


Figure 28. Comparison of simulations using different simulators with measurements of *n*MOS (left) and *p*MOS (right) devices [1, 142, 148].

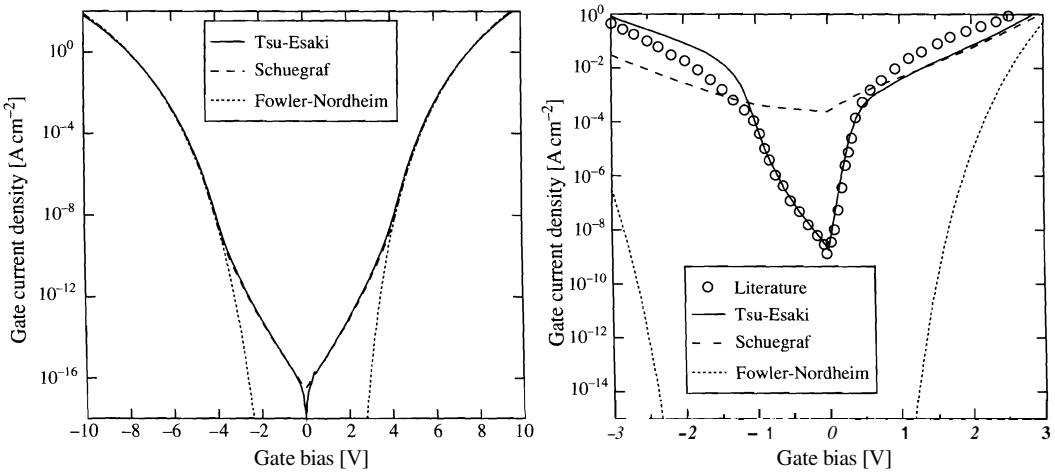


Figure 29. Compact models for a metal-dielectric-metal structure (left) and an *n*MOS structure (right; literature values from [142]).

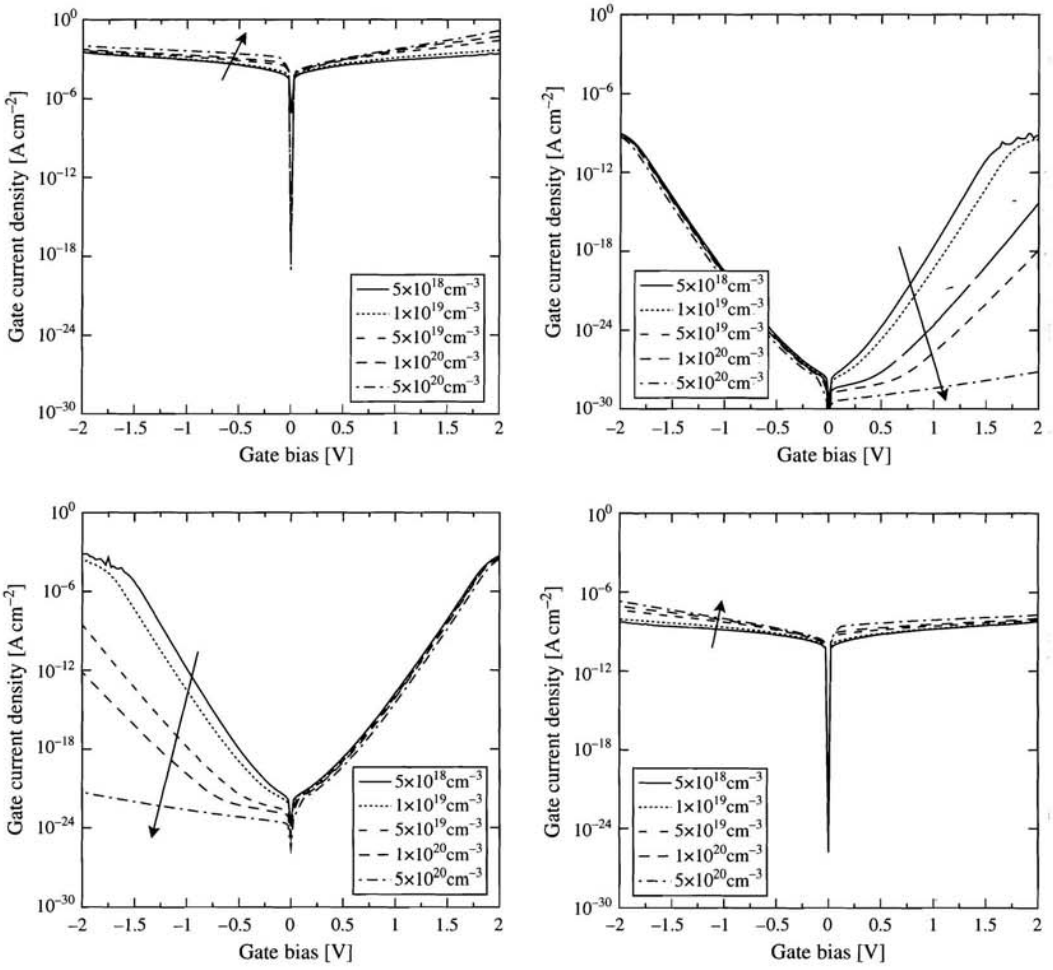


Figure 30. Effect of the polysilicon doping on the electron tunneling current (left) and the hole tunneling current (right) in the source and drain extension region of an *n*MOS (top) and a *p*MOS (bottom) with 2-nm dielectric thickness and 5×10^{18} cm^{-3} substrate doping.

the gate current density by a factor of almost 10^6. It will therefore have a large impact on the total gate current density. The cold Maxwellian approximation underestimates the gate current density in this region, whereas the non-Maxwellian distribution correctly reproduces the Monte Carlo results.

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

The effect of hot-carrier tunneling on the total gate current of the devices is shown in Fig. 35. In the left part of this figure, the gate current density for a 0.5-μm turned-on MOSFET with a dielectric thickness of 4 nm is shown as a function of the gate bias. Results from Monte Carlo simulations are also shown in this figure. For low gate voltages ($V_{GS} < V_{DS}$), the peak electric field in the channel increases with increasing gate bias. The electron temperature is high, and the heated Maxwellian approximation massively overestimates the total gate current. If the gate bias exceeds the drain-source voltage, however, the peak electric field in the channel is reduced [153]. Therefore, for $V_{GS} > V_{DS}$, the electron temperature reduces with increasing gate bias, and the heated Maxwellian approximation delivers correct results. The non-Maxwellian model (28) delivers correct results for all gate voltages.

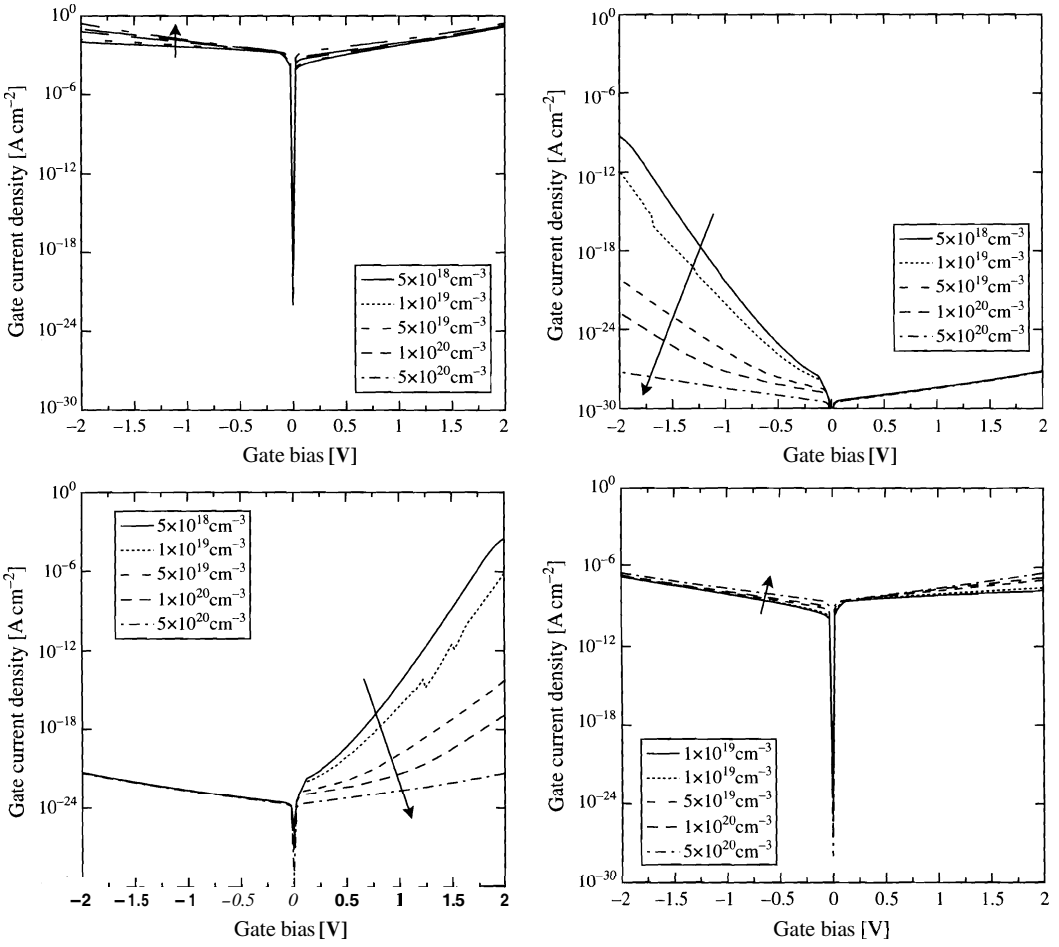


Figure 31. Effect of the substrate doping on the electron tunneling current (left) and the hole tunneling current (right) in the source and drain extension region of an *n*MOS (top) and a *p*MOS (bottom) with 2-nm dielectric thickness and 5 x 10^{20} cm^{-3} polysilicon doping.

The question remains if the hot-carrier tunneling current strongly depends on the gate length of the device. In the right part of Fig. 35, the gate current is given as a function of the gate length for different gate dielectric thicknesses (2.2 nm–3.0 nm). Again, Monte Carlo simulation results are used as reference. It can be seen that the heated Maxwellian distribution delivers correct results only for large gate lengths, whereas it totally fails for smaller devices. The use of a cold Maxwellian distribution, on the other hand, underestimates the gate current only slightly and seems to be the better choice if accurate modeling of the device physics is not that important or only a quick estimation is asked for. The non-Maxwellian model correctly reproduces the Monte Carlo results for all gate lengths and gate dielectric thicknesses.

3.1.5. Alternative Dielectrics for MOS Transistors

The further reduction of device dimensions makes the introduction of alternative dielectric materials necessary. Because none of the possible materials forms a native oxide on silicon, a thin interfacial layer of SiO_2 cannot be avoided. Thus, a two-layer band edge diagram is commonly assumed as depicted in Fig. 36 [154]. A wide variety of high-K materials can be considered as alternative dielectrics. However, several points must be considered when evaluating these materials:

(1) The dielectric permittivity κ.
(2) The barrier height for electrons $q\Phi_e$ and holes $q\Phi_h$ on silicon. These values are equivalent to the band edge offsets $\Delta\mathscr{E}_c$ and $\Delta\mathscr{E}_v$.

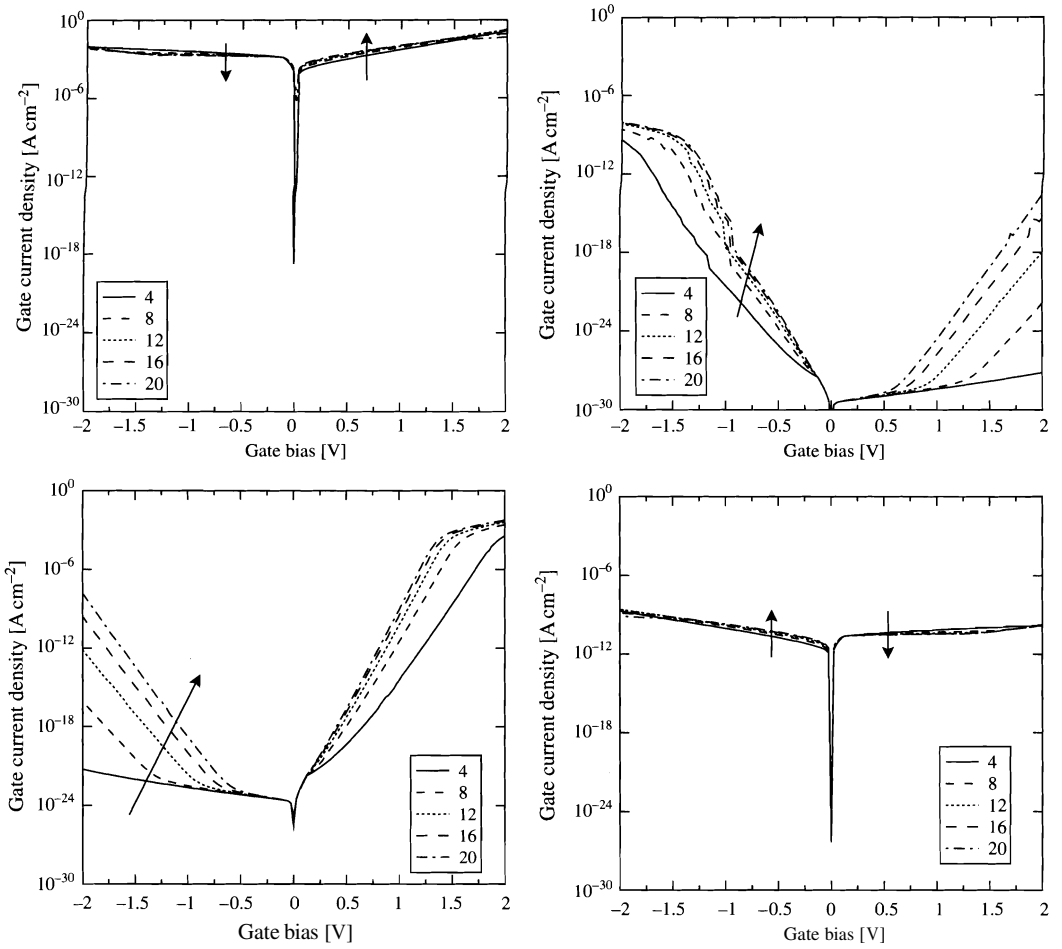


Figure 32. Effect of the dielectric permittivity κ/κ_0 on the electron tunneling current (left) and the hole tunneling current (right) in the source and drain extension region of an *n*MOS (top) and a *p*MOS (bottom) with 2-nm dielectric thickness and 5×10^{18} cm^{-3} substrate doping.

(3) The thermodynamic stability of the dielectric material on silicon: The material must withstand all following processing steps.
(4) The quality of the interfaces: High interface roughness may cause increased scattering, which reduces the carrier mobility in the channel.
(5) The trap concentration, which leads to trap-assisted tunneling.
(6) The feasibility and integrability of the deposition method in the fabrication process.

Only the permittivity, the trap concentration, and the barrier heights influence the tunneling current. When looking at the barrier height and permittivity of various dielectrics in Table 4, one notices a strong trade-off between the barrier height and the dielectric permittivity: dielectrics with a high energy barrier have a low permittivity and vice versa; see Figs. 37 and 38. Hence, optimization becomes necessary to find the optimum material.

Choosing the material parameters from Table 4, the gate current density can be computed as a function of the gate bias [155]. It is commonly assumed that an underlying layer of SiO_2 cannot be avoided—or is even deliberately introduced to achieve a lower trap density at the interface to silicon. Thus, an underlying SiO_2 layer with a thickness of 0.5 nm was assumed. The thickness of the high-K layer was adjusted so that the effective oxide thickness (EOT) remains unchanged at 1 nm. The gate current density is shown in the left part of Fig. 39 as a function of the gate bias for different material combinations. The commonly assumed limit of 1 Acm^{-2} gate leakage is also indicated. Both SiO_2 and Si_3N_4 show a much too high leakage, whereas Ta_2O_5, ZrO_2, and HfO_2 stay below 1 Acm^{-2} at $V_{GS} = 1$ V. Due to the low

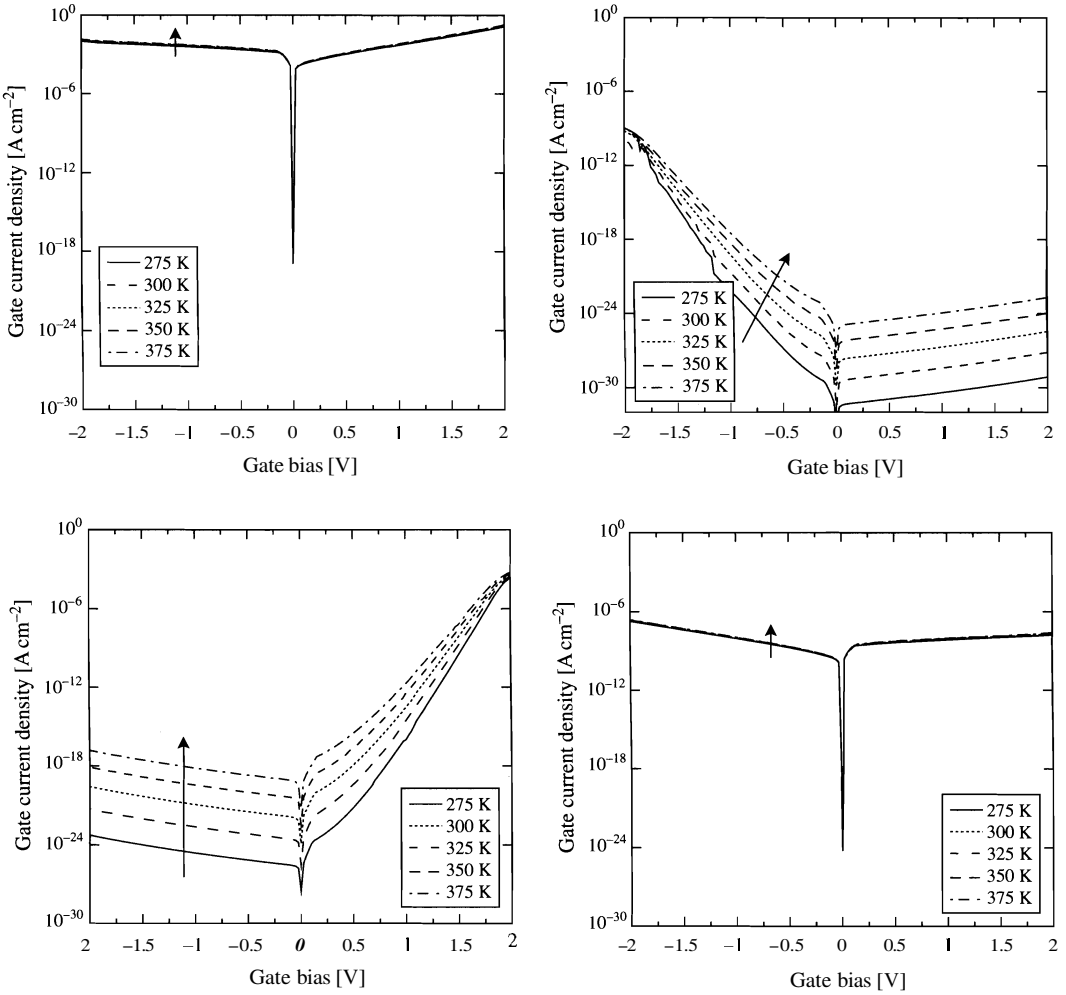


Figure 33. Effect of the lattice temperature on the electron tunneling current (left) and the hole tunneling current (right) in the source and drain extension region of an *n*MOS (top) and a *p*MOS (bottom) with 2-nm dielectric thickness and 5 x 10^{18} cm ^{3} substrate doping.

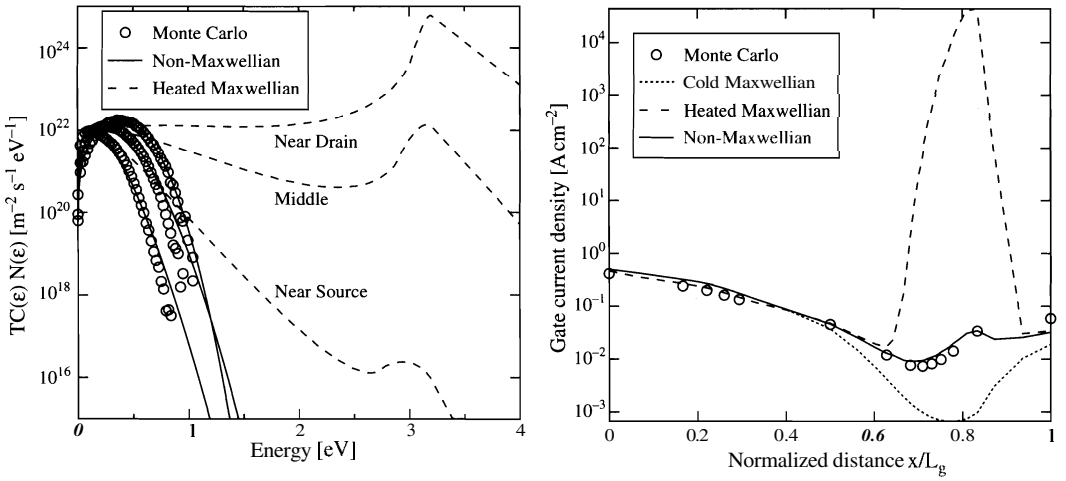


Figure 34. Integrand of Tsu–Esaki's equation (left) and gate current density along the channel (right) of a MOSFET with 100-nm gate length and 3-nm gate dielectric thickness [151, 152].

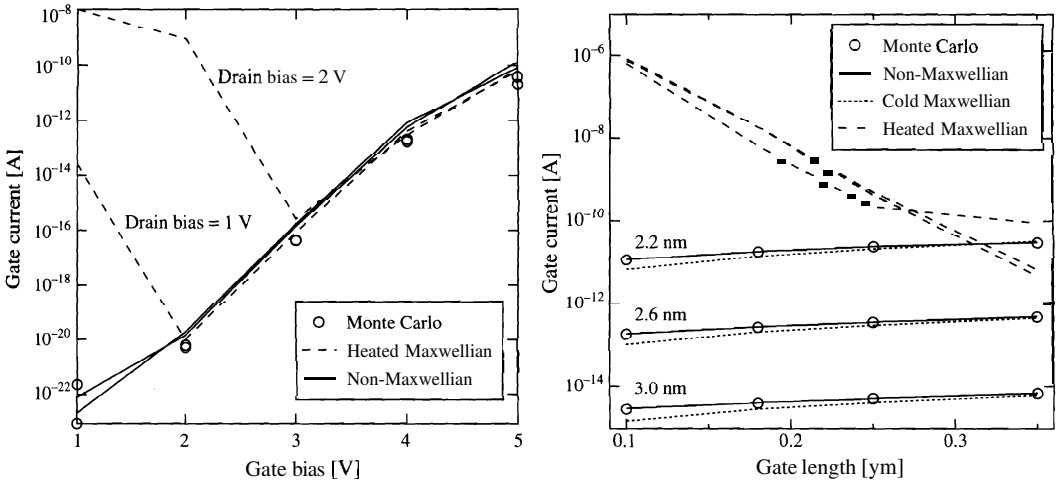


Figure 35. Gate current for different values of the gate bias (left). Dependency of the total gate current on the gate length (right) [151, 152].

conduction band offset, TiO_2 shows an especially pronounced current increase for positive gate bias.

To assess the material parameters necessary to stay below a specific maximum gate current density, the gate current has been calculated as a function of the conduction band offset and dielectric permittivity as shown in the right part of Fig. 39. Because it is often not possible to vary the thickness of the underlying SiO_2 layer, it was again fixed at 0.5 nm and the high-K thickness was adjusted to reach an EOT of 1.5 nm. The gate current density was evaluated at a fixed bias point of $V_{GS} = 1.5$ V and $V_{DS} = 0$ V. The current density decreases strongly with increasing conduction band offset. Increasing the value of the dielectric permittivity κ also strongly reduces the leakage current due to the higher physical stack thickness. However, materials with a conduction band offset below 1 eV never reach acceptable gate current densities.

It may be asked which thickness of the high-K layer is necessary to achieve a certain gate current density. In the left part of Fig. 40, the gate current density is shown for an effective oxide thickness ranging from 0.5 nm to 2.0 nm as a function of the high-K layer thickness. Again, the stack consists of an underlying 0.5 nm layer of SiO_2 and the simulations are

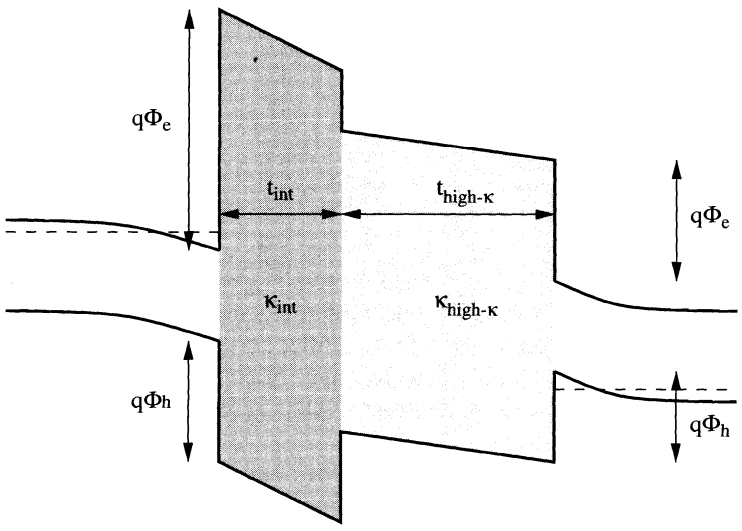


Figure 36. Schematic of a band energy diagram of a stacked dielectric consisting of a thin underlying interface layer and a thick layer of a high-κ material with higher dielectric permittivity, but lower barrier height.

Table 4. Band gap energy and conduction band offset of various dielectric materials.

	κ/κ_0 (1)	Band Gap $\mathscr{E}_g$ (eV)	Conduction Band Offset $\Delta\mathscr{E}_c$ (eV)	Valence Band Offset $\Delta\mathscr{E}_v$ (eV)	Reference
SiO_2	3.9	9.00	3.00	4.90	[198]
	3.9	9.00	3.50	4.40	[193]
	3.9	9.00	3.15	4.75	[199]
	3.9	8.90	3.20	4.60	[200]
		9.00	3.50	4.40	[201, 202]
	3.9	9.00	3.00	4.90	[41]
Si_3N_4	7.5	5.00	2.00	1.90	[198]
	7.6	5.00–5.30	2.40	1.50–1.80	[193]
	7.9	5.30	2.40	1.80	[199]
	7.0	5.10	2.00	2.00	[200]
		5.30	2.40	1.80	[201, 202]
	7.5	5.00	2.00	1.90	[41]
Ta_2O_5	25.0	4.40	1.40	1.90	[41, 198]
	23.0-25.0	4.40	0.30	3.00	[193]
	25.0	4.40	0.36	2.94	[199, 202]
	26.0	4.50	1.00-1.50	1.90–2.40	[200]
		4.40	0.36	2.94	[201]
TiO_2	40.0	3.50	1.10	1.30	[41, 198]
	39.0–110.0	3.00–3.27	0.00	1.90–1.97	[193]
	80.0-170.0	3.05	0.00	1.95	[199]
	80.0	3.50	1.20	1.20	[200]
		3.05	0.00	1.95	
	9.0	8.70	2.80	4.80	[200]
	8.0–9.0	8.8–9.00	2.78–2.80	4.92–5.10	[193]
	9.5–12.0	8.8	2.80	4.90	[199]
		8.80	2.80	4.90	[201]
	10.0	8.80	2.80	4.90	
ZrO_2	23.0	5.80	1.40	3.30	[202]
	25.0	7.80	1.40	5.30	[198, 200]
	22.0-25.0	5.00–5.80	1.40	2.50–3.30	[193]
	12.0-16.0	5.70–5.80	1.40–1.50	3.10-3.30	[199]
		5.80	2.50	2.20	[201]
HfO_2	25.0	5.70	1.50	3.10	[198, 200]
	22.0–40.0	6.00	1.50	3.50	[193]
	16.0–30.0	4.50–6.00	1.50	1.90–3.40	[199]
		6.00	1.50	3.40	[201]
	20.0	6.00	1.50	3.40	[202]
Y_2O_3	15.0	5.60	2.30	2.20	[200]
	11.3–18.0	5.50–6.00	1.30	3.10-3.60	[193]
	4.4	6.00	1.30	3.60	[201]
	15.0	6.00	2.30	2.60	[202]
$ZrSiO_4$	12.6	6.00	1.50	3.40	[193]
		4.50	0.70	2.70	[199]
	3.8	6.00	1.50	3.40	[201]
		6.00	1.50	3.40	[202]

performed at a fixed bias point of $V_{GS} = 1.5$ V and $V_{DS} = 0$ V. In this plot, the curves are only drawn for an EOT of 0.5 nm–2.0 nm, and conduction band offsets of $q\Phi_e = 1$ eV to $q\Phi_e = 3$ eV have been considered. For a conduction band offset of 1 eV, large high-κ thicknesses are necessary to reduce the leakage. Such large stacks may pose problems due to fringing fields from the drain contact, which reduce the threshold voltage of the device.

The trade-off between the dielectric permittivity and the conduction band offset gives rise to further effects as shown in the right part of Fig. 40. If the EOT has to be held at a fixed value, an increase of the SiO_2 layer thickness causes a reduced thickness of the high-κ layer. This is shown for different values of the permittivity ($\kappa = 8.0$ to $\kappa = 24.0$). So, the total stack thickness may be larger than 8 nm for $\kappa = 24$, or as small as 1.5 nm

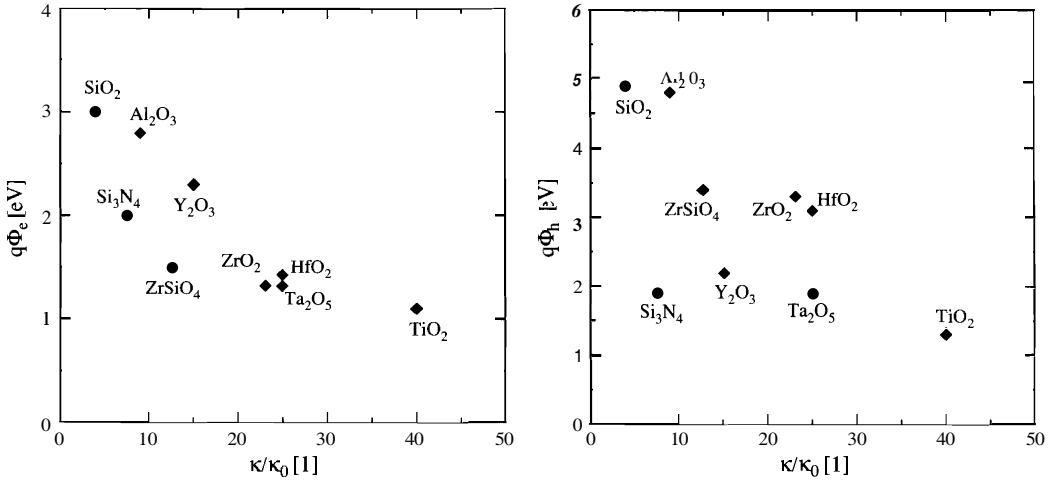


Figure 37. Trade-off between electron barrier height (left) or hole barrier height (right) and the permittivity of various dielectric materials [41, 193, 198–2021.

if only SiO_2 is used. Such a reduction of the total stack thickness, however, has no clear effect on the leakage. It may cause the gate current density at a specific bias point to stay constant, increase, or even decrease depending on the material parameters. For example, the gate leakage for a material with $\kappa = 24$ and a conduction band offset of 1 eV shows the maximum leakage at a SiO_2 layer thickness of approximately 0.8 nm. Therefore, a clear statement about the optimum thickness of the interface layer obviously depends on the material parameters.

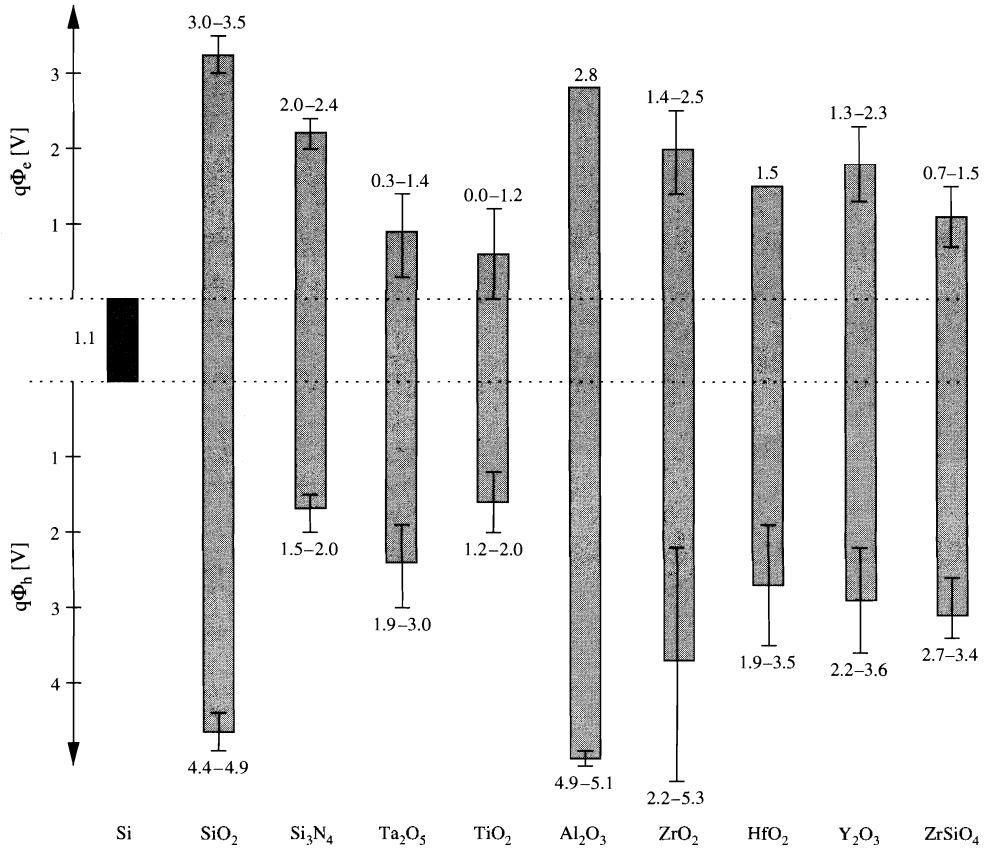


Figure 38. Conduction and valence band edges of various dielectric materials compared to silicon [41, 193, 198–202].

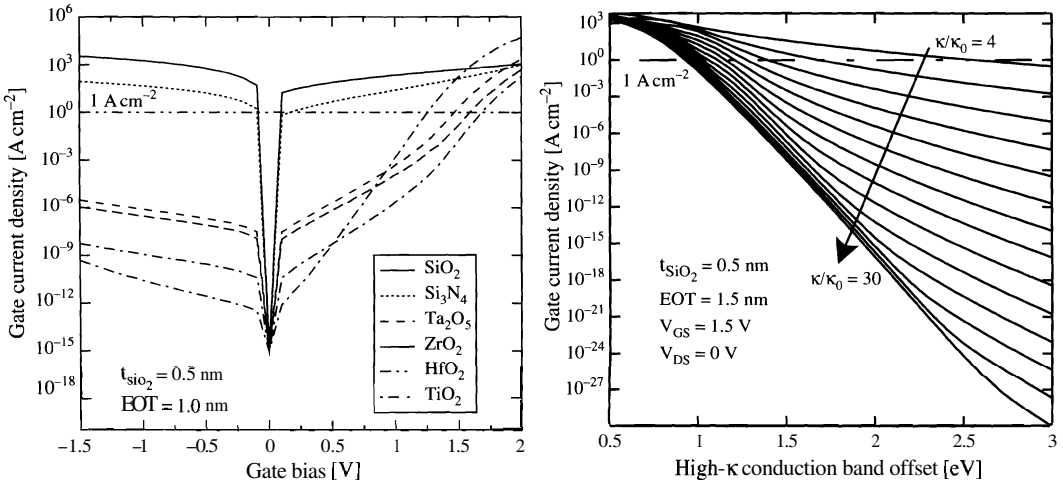


Figure 39. Gate current density as a function of the gate voltage for different materials. The dielectric stack consists of a 0.5-nm SiO_2 layer and a high-K layer with a total EOT of 1.0 nm (left). Dependence of the gate current on the high-K conduction band offset and dielectric permittivity of a stack with EOT = 1.5 nm and an 0.5-nm SiO_2 interface layer at a gate bias of 1.5 V (right) [155].

3.1.6. Trap-Assisted Tunneling in ZrO_2 Dielectrics

Because ZrO_2 offers good material parameters, it was further investigated by means of experiments, and numerous results were published [156, 157]. ZrO_2 pMOS capacitors have been fabricated by MOCVD (metal-organic chemical vapor deposition) on p-type (100) silicon wafers with an acceptor doping of 1.5 x 10^{18} cm^{-3} and Al gate electrodes [157]. The overall thicknesses of the dielectric layers have been evaluated by spectroscopic ellipsometry. Employing a dielectric permittivity of the high-K material of $\kappa/\kappa_0 = 18$, which has been found for thicker films, the comparison of optical measurements and the results of CV characterization implicates the presence of an interfacial layer with a permittivity in the range of 4 to 8. Table 5 summarizes the results of an evaluation of the thicknesses of the high-K films and interfacial layers. Also given is the effective oxide thickness EOT. The values t_{int} and $t_{high-\kappa}$ denote the thicknesses of the interface and the high-K layer.

In the left part of Fig. 41, the measured gate current is shown for the two dielectric layers with the approximate shape of the energy barrier sketched in the insets. As reference,

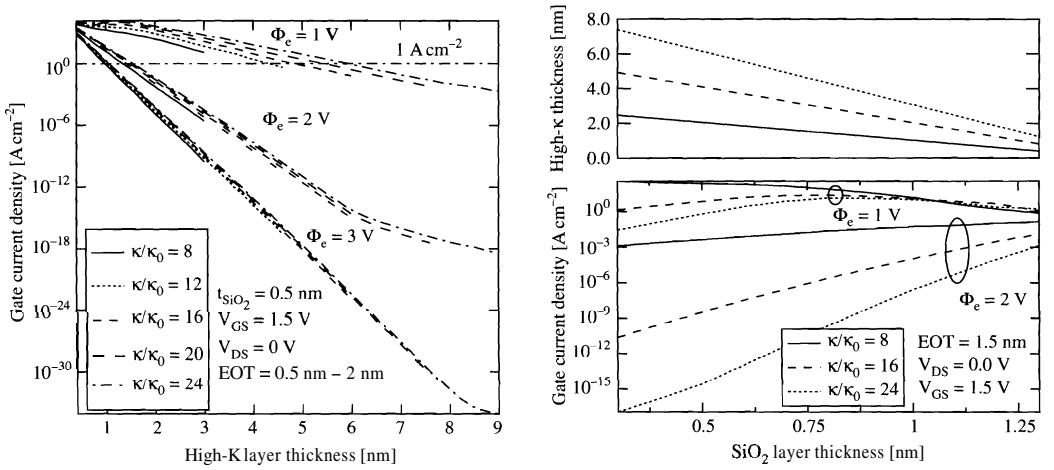


Figure 40. Dependence of the gate current on the high-κ layer thickness, conduction band offset, and permittivity of a stack with EOT = 2.0 nm and a 0.5-nm SiO_2 interface layer at a gate bias of 1.5 V (left). Dependence of the gate current on the interface layer thickness, conduction band offset, and permittivity of a stack with EOT = 1.5 nm at a gate bias of 1.5 V (right) [155].

Table 5. Layer thicknesses and effective oxide thickness of metal organic chemical vapor deposition-deposited ZrO_2 layers in nanometers, after Harasek [156].

Layer Thickness	t_{int}	$t_{high-\kappa}$	EOT
6.9	0.75–2.0	6.15–4.9	2.0
12.7	0.3–1.0	12.4–11.7	3.0

the figure also shows the gate current for a 2-nm and a 3-nm SiO_2 layer (dotted lines). As expected, the measured current density is lower than for the SiO_2 counterparts. However, the Tsu–Esaki model cannot reproduce the measurements as it yields tunneling currents orders of magnitude lower than the measurements. This indicates the presence of strong trap-assisted tunneling due to a high trap concentration in the dielectric layer. By assuming a Frenkel–Poole-like conduction through the dielectric layer, the measurements could be reproduced (full lines). Note that in previous studies [156], tunneling through ZrO_2 layers fabricated by magnetron sputtering could be reproduced without considering trap-assisted tunneling. That indicates the presence of a high trap concentration due to the MOCVD process, in contrast to the sputtering process.

To clarify the trap energy level and concentration, the step response of the MOS capacitors has been measured as shown in the right part of Fig. 41 for the 12.7-nm ZrO_2 layer annealed in reducing conditions (forming gas) and the 6.9-nm layer annealed under oxidizing conditions [158]. The gate voltage is turned off after being fixed at a value of 2.5 V, and the resulting gate current is measured over time. The transient gate current exceeds the static gate current by orders of magnitude and decays very slowly. This behavior can be explained assuming defects in the dielectric layer [159]. Using the trap-assisted tunneling model outlined in Section 2.8.2, a trap energy level of 1.3 eV below the ZrO_2 conduction band edge, a trap concentration of 4.5 x 10^{18} cm^{-3}, and an energy loss of 1.5 eV have been found. For the dielectric layer annealed under oxidizing conditions, a trap concentration of 4 x 10^{17} cm^{-3} was found.

To predict the performance of devices based on ZrO_2 dielectrics, a well-tempered MOSFET as described in Ref. [160] with an effective channel length of 50 nm has been simulated. EOT thicknesses of 2-nm and 3-nm SiO_2 and respective ZrO_2 layers have been considered. The left part of Fig. 42 depicts the conduction band edge in the channel for different gate-source voltages. It can be seen that the barrier is slightly lower for the ZrO_2 layer at $V_{GS} = 1.2$ V, whereas it is strongly reduced at $V_{GS} = 0.1$ V, which is due to the pronounced fringing fields from the drain contact.

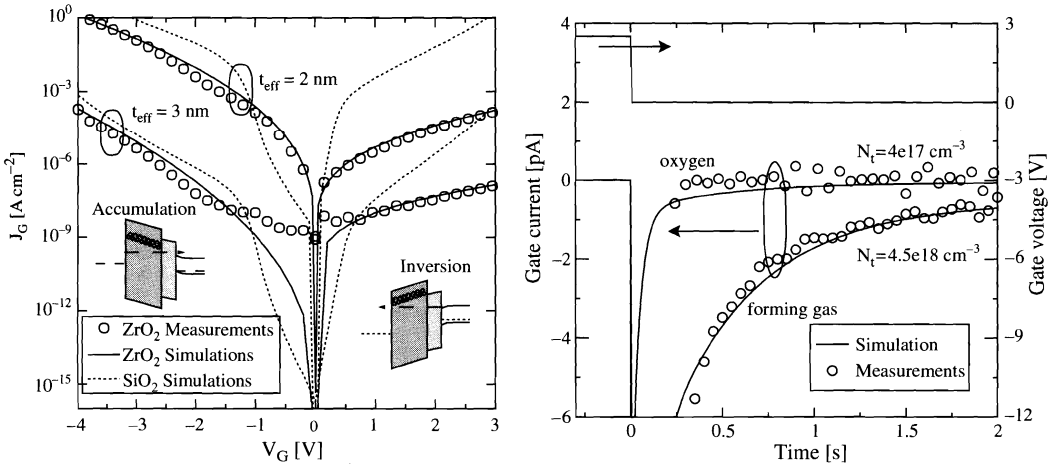


Figure 41. Stationary (left) and transient (right) gate current measurements of the ZrO_2 layers performed by Harasek [157] compared with simulations [158].

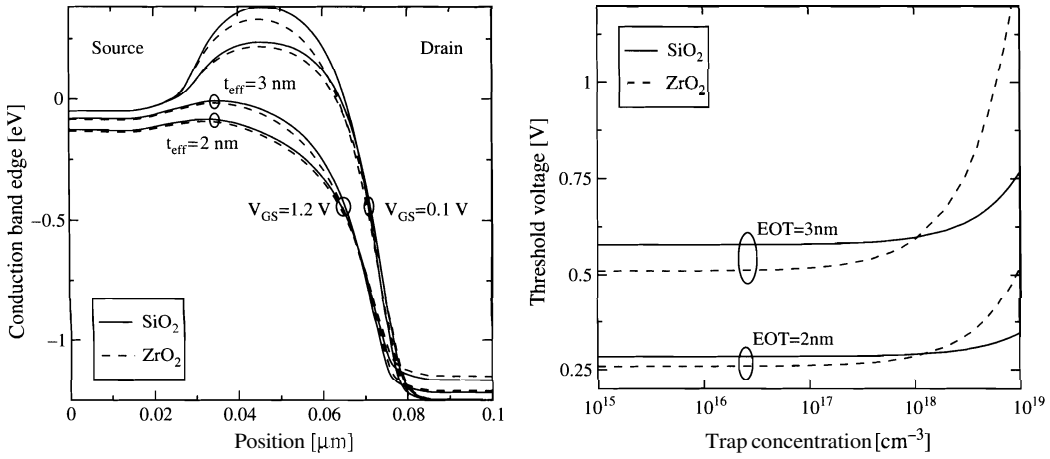


Figure 42. Well-tempered MOSFET conduction band edge along the channel for SiO_2 and ZrO_2 dielectrics (left). Influence of the dielectric trap concentration on the MOSFET threshold voltage (right) [158].

An additional topic of interest for high-K dielectrics is the influence of trapped charges in the high-K layer on the threshold voltage of the device. The trap concentration in the ZrO_2 layer was increased from 10^{15} cm^{-3} to 10^{19} cm^{-3} with full trap occupancy in the dielectric layer. It can be seen in the right part of Fig. 42 that the threshold voltage strongly increases with rising trap concentration. This effect is therefore contrary to the decrease of the threshold voltage due to fringing fields described above.

3.2. Tunneling in Nonvolatile Memory Devices

Tunneling effects are crucial not only for MOS transistors but also for nonvolatile semiconductor memory devices. In contrast to volatile memory devices, they retain the stored information without external power supply. Nonvolatile memory (NVM) devices can be read and programmed like random-access memory (RAM) devices, have a low power consumption, are mechanically robust, and offer the possibility of large-scale integration. They constitute about 10% of the total semiconductor memory market [161]. However, simulation of such devices is often carried out using simplified compact models [162–167]. For the case of stacked gate dielectrics or hot electron injection, such models do not capture the device physics and can reproduce measured data only on a fit-formula level. In this section, some examples of conventional EEPROM and alternative devices will be studied using the tunneling models described above.

3.2.1. Conventional EEPROM Devices

The basic operating principle of an EEPROM was presented by Kahng and Sze in 1967 at Bell Laboratories [168]. The device consists of a control gate and a floating gate on top of a conventional MOS transistor. A thin tunnel dielectric separates the floating gate from the channel. It must be thick enough to allow up to 10^5 writing and erasing cycles without breakdown—common thicknesses are 6–8 nm. Applying a high positive voltage (about 8–12 V) on the control gate raises the potential of the floating gate by capacitive coupling. The high electric field in the tunnel dielectric ($\approx 10^9$ V/m) leads to Fowler–Nordheim tunneling of electrons from the substrate to the floating gate. The charge on the floating gate changes the threshold voltage of the underlying MOS transistor and is retained even if the control gate voltage is removed. A retention time of 10 years is required for consumer applications such as memory cards. Though EEPROM cells offer random access for writing and erasing of individual bits, Flash cells can be programmed selectively but erased only at once. This has the advantage of lower cell size. Due to the high electric field in the dielectric, degradation or even breakdown of the dielectric is a major concern. A comprehensive survey of NVM technology is given in Refs. [169] and [170].

3.2.1.1. Static SILC in EEPROMs The speed of the programming and erasing process is one of the main figures of merit of an EEPROM cell. Therefore, strong electric fields are applied at the control gate to allow Fowler–Nordheim tunneling of carriers during programming and erasing cycles. However, due to this repeated high-field stress, trap centers in the dielectric are formed, which allow trap-assisted tunneling at low fields and thus reduce the retention time of the devices. This additional current at low bias is known as stress-induced leakage current (SILC) and represents one of the major reliability concerns in contemporary EEPROM devices [112, 135]. In the left part of Fig. 43, measured SILC after different stress times for a MOS capacitor with a dielectric thickness of 5.5 nm is shown [105]. The trap-assisted tunneling model outlined in Section 2.8.2 yields excellent agreement with the measured data if the trap concentration is used as a fitting parameter dependent on the stressing time (the model parameters are stated in the figure caption). The transition from the region of mainly trap-assisted tunneling for $V_{GS} < 5$ V to the region of Fowler–Nordheim tunneling for $V_{GS} > 5$ V is clearly visible. The right part of Fig. 43 shows the trap occupancy f_T across the gate dielectric of a MOS capacitor using the gate voltage as parameter. The regions near the gate (right) and near the substrate (left) are only sparsely occupied. Near the gate, the emission time is much smaller than the capture time, and near the substrate, the trap energy lies above the electron energy in the cathode. Some of the trapped electrons face a triangular barrier for the emission process, giving rise to an additional peak in the trap occupancy near the gate side (the anode) of the dielectric. This is due to the wave function interference in the Fowler–Nordheim region (the oscillations are also observed in the emission time of the traps shown in Fig. 17).

3.2.1.2. Transient SILC in EEPROMs It has been shown that the transient trap-assisted tunneling current can be described by a rate equation that gives rise to an exponential behavior of the tunneling current over time; see Section 2.8.2.4. The left part of Fig. 44 shows measurements of the gate current density of MOS capacitors as a function of time with dielectric thicknesses of 8.5 nm and 13.0 nm compared to simulations [104, 171]. Initially, the traps are empty, which can be achieved by applying flat band conditions. At $t = 0$ s, the gate voltage is turned on (−5.8 V and −8.3 V for the thinner and the thicker dielectric, respectively) and the traps are filled according to their specific capture and emission time constants. This charging current consists of an emission and a capture current, which may exceed the steady-state current by orders of magnitude. A good fit to the measured data can be achieved using the trap parameters indicated in the figure caption.

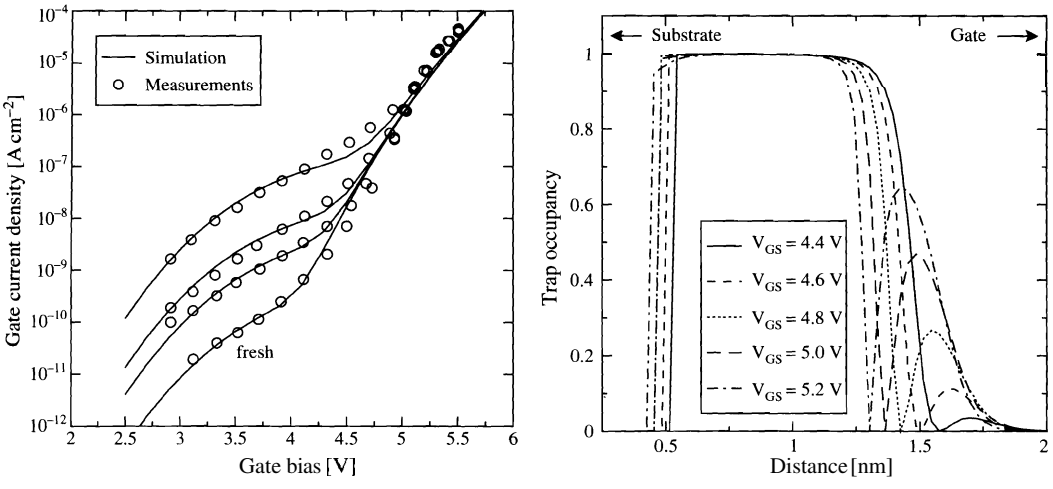


Figure 43. Comparison of simulations with measurements of a MOS capacitor with a dielectric thickness of 5.5 nm [105, 171] is shown on the left. The trap energy is 2.7 eV, the phonon energy 130 meV, and the Huang–Rhys factor is 10. The trap concentration was set to 9 x 10^{17} cm^{-3}, 10^{17} cm^{-3}, 3 x 10^{16} cm^{-3}, and 3 x 10^{15} cm^{-3} to fit the measurements (from top to bottom). The trap occupancy across the gate dielectric at different gate voltages is shown on the right.

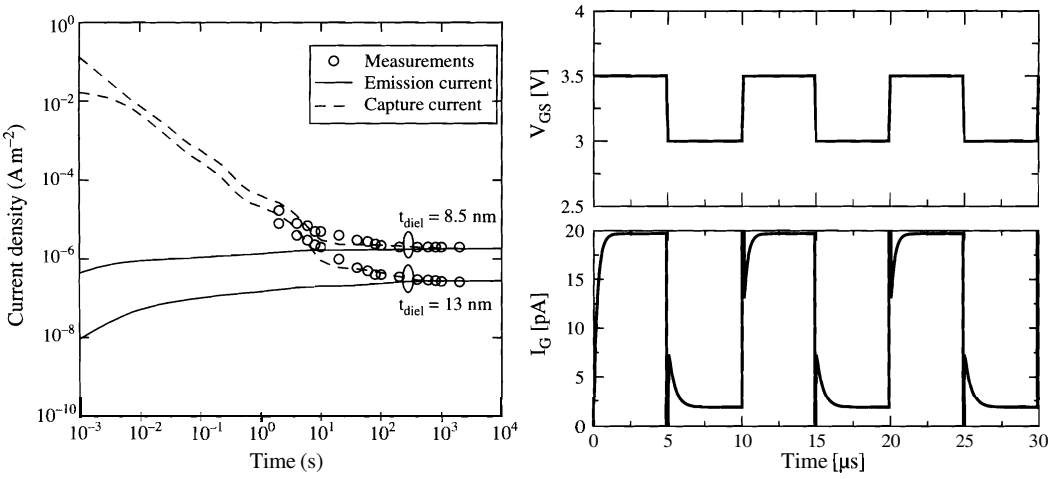


Figure 44. Transient capture and emission currents (left) of MOS capacitors at a gate bias of −5.8 V and −8.3 V [104, 171]. For the thinner dielectric, a trap energy of 2.5 eV and a trap concentration of 3 x 10^{18} cm^{-3} was used, whereas for the thicker dielectric, a trap concentration of 10^{18} cm^{-3} was found. The right figure shows transient simulation results of a MOS capacitor with a gate dielectric thickness of **3** nm and a trap energy level of 3 eV.

The right part of Fig. 44 shows the gate current of a MOS capacitor for an applied rectangular pulse with a frequency of 100 kHz assuming initial flat band conditions. It can be seen that the time constants of the trap filling and emptying processes are not equal but depend on the applied voltage, as different voltages lead to different capture and emission times. The spikes in this figure are due to the sudden voltage change, whereas the trap concentration remains constant: In the transition from 3.0 V to 3.5 V, the barrier shape changes suddenly, and traps are rapidly emptied. Traps near the cathode are filled, and it takes several microseconds until the new steady-state is reached. Thus, dielectric materials that have such a high trap concentration may lead to considerable problems for high-frequency applications.

For EEPROM devices, the charging and discharging characteristics are crucial: Programming and erasing should happen as fast as possible; therefore, high voltages are applied. The discharging current over time, on the other hand, determines the retention time and must be very low. Furthermore, the programming and erasing pulses must be carefully optimized to avoid over-erase, as the tunnel current density for positive and negative voltages on the floating gate is not equal. This is frequently addressed in the literature [172, 173].

3.2.2. Alternative Nonvolatile Memory Devices

Strong efforts are undertaken to improve the standard floating-gate EEPROM cell in terms of integration density, endurance, reliability, program time, erase time, and retention time. EEPROM devices with a tunnel window near the drain contact have been introduced to reduce the charge loss from the floating gate and thus reach higher retention time. However, due to the small area of the tunnel window, high voltages have to be used at the drain contact, which again reduces cell reliability.

Recently, Caywood et al. proposed a device structure where nonselected cells are isolated from the drain and source contacts by two additional side gates [174]. In this device, electrons tunnel from the inverted channel to the floating gate. The large area reduces programming and erasing time. Furthermore, the capacitive coupling between the control gate and the floating gate is higher than in the standard EEPROM cell, which allows use of lower programming and erasing voltages. No drain-source bias is applied for charging, thus the power consumption is low and the injected electrons are less likely to cause degradation of the dielectric. The control gate functions as a select transistor that isolates unselected cells from the high voltages at the shared source and drain contacts during read and write access of neighboring cells.

In contrast to the reduction of the cell footprint, integration density can also be increased by storing more than one bit on a standard EEPROM cell. This can be achieved by tailoring the programming and erasing pulses in such a way that the threshold voltage falls into one of 4, 8, or 16 voltage ranges. The different threshold voltages can be distinguished by the sensing circuits, resulting in two, three, or four bits that can be stored in the cell. However, charge loss must be extremely low over time, and the threshold voltages have to be detected very precisely.

Single-poly devices have been proposed to integrate NVM devices in standard CMOS logic processes, thus enabling an embedded memory. The control gate lies next to the floating gate, and capacitive coupling is achieved by a layer of highly doped silicon. Though such devices can readily be integrated into existing CMOS process flows, they come at the cost of a large footprint.

A different approach to store more than one bit in a single memory cell is to split the floating gate into two separate segments. If a nonuniform doping in the source and drain side of the channel is used, different amounts of charge can be stored in each floating gate. Such device structures are either achieved using separate metallic floating gates [97] or using a layer of trap-rich dielectric [175].

In the following sections, three of the most promising alternative EEPROM devices will be studied in detail. These are

- Quantum dot and trap-rich dielectric based devices: In these devices, charging and discharging is achieved by tunneling of electrons to and from localized trapping centers in the dielectric.
- Multibarrier tunneling devices consist of a floating gate—or memory node—which is separated from the control gate by several thin dielectric layers. By the use of a side gate, the tunneling current through these barriers can be controlled selectively. In contrast to EEPROMs, the tunneling current flows from the floating gate to the control gate and not to the channel. Extremely high I_{on}/I_{off} ratios can be achieved because the tunneling current is controlled by a separate side gate contact.
- Devices where the tunnel dielectric consists of stacked dielectrics that are engineered in such a way that they block tunneling in the off-state but allow strong tunneling in the on-state.

3.2.2.1. Nonvolatile Memory Devices Based on Trap-Rich Dielectrics A SONOS (silicon-oxide-nitride-oxide-silicon) device is a nonvolatile memory where the charge is stored in a layer of trap-rich dielectric material instead of a floating gate as in an EEPROM. Figure 45 shows an example where a layer of Si_3N_4 is sandwiched between two layers of SiO_2. Electrons tunneling from the substrate are trapped and redistribute themselves in separate trapping centers. This has the advantage that the charge is stored independently in the traps. A leaky path in the tunnel dielectric cannot lead to full charge loss, as it is the case in conventional EEPROM devices. Therefore, reliability and retention time is increased [75, 125, 176–185].

The band diagram along the dielectric of such a device is shown in Fig. 45 for the programming, storing, and erasing processes. By applying a positive voltage at the gate contact, electrons tunnel through the tunnel dielectric into the trap region. The traps are filled with

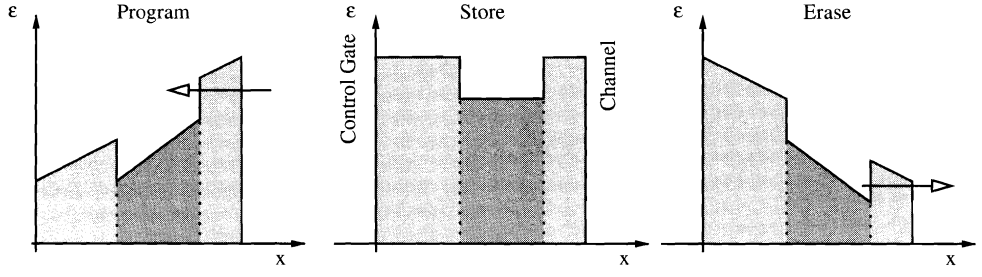


Figure 45. Conduction band edge in a SONOS device for the programming, storing, and erasing process.

electrons and become negatively charged. Because of the tunnel dielectric, this charge is stored even if the bias is removed. To erase the memory cell, a negative voltage is applied on the gate contact, leading to a reduced potential barrier and a high tunneling current of electrons out of the traps. Important device parameters are the charging and discharging current through the dielectric, the drain current in the on- and off-state, and the retention time.

The trap-assisted tunneling model can be applied to simulate device characteristics of this device, where three layers of SiO_2 have been used and the trap concentration and trap energy level in the middle layer was chosen to resemble a layer of silicon nitride. The transient trap occupancy for a discharging process starting from an initial condition of 2 V at the gate contact is shown in Fig. 46. Initially, the traps are filled. Over time, the electrons leak through the lower dielectric into the channel. After 10^9 s, almost no more charge is stored in the trap-rich dielectric.

3.2.2.2. Multibarrier Tunneling Devices One of the main shortcomings of conventional EEPROM devices is that the current in the on-state and off-state—the programming and leakage currents—flow through the same tunnel dielectric and face the same energy barrier. They cannot be optimized independently: Increasing the thickness of the tunnel dielectric reduces the leakage, but also reduces the on-state current and thus increases the programming time. Multibarrier tunneling devices offer a solution to this problem. Planar localized-electron device memory (PLEDM) cells have been presented by Nakazato et al. in Ref. [186], and promising results have been reported [187–190]. The principle of a PLEDM is to put a PLED transistor (PLEDTR) on top of the gate of a conventional MOSFET, as shown in Fig. 47. The charge on the memory node, which acts as a floating gate, is provided by tunneling of carriers through the PLED transistor, which consists of a stack of Si_3N_4 barriers sandwiched between layers of intrinsic silicon. Upper and lower barriers prevent diffusion from the polysilicon contacts, whereas the middle barrier—the central shutter barrier (CSB)—blocks the tunneling current in the off-state. The PLED transistor has two side

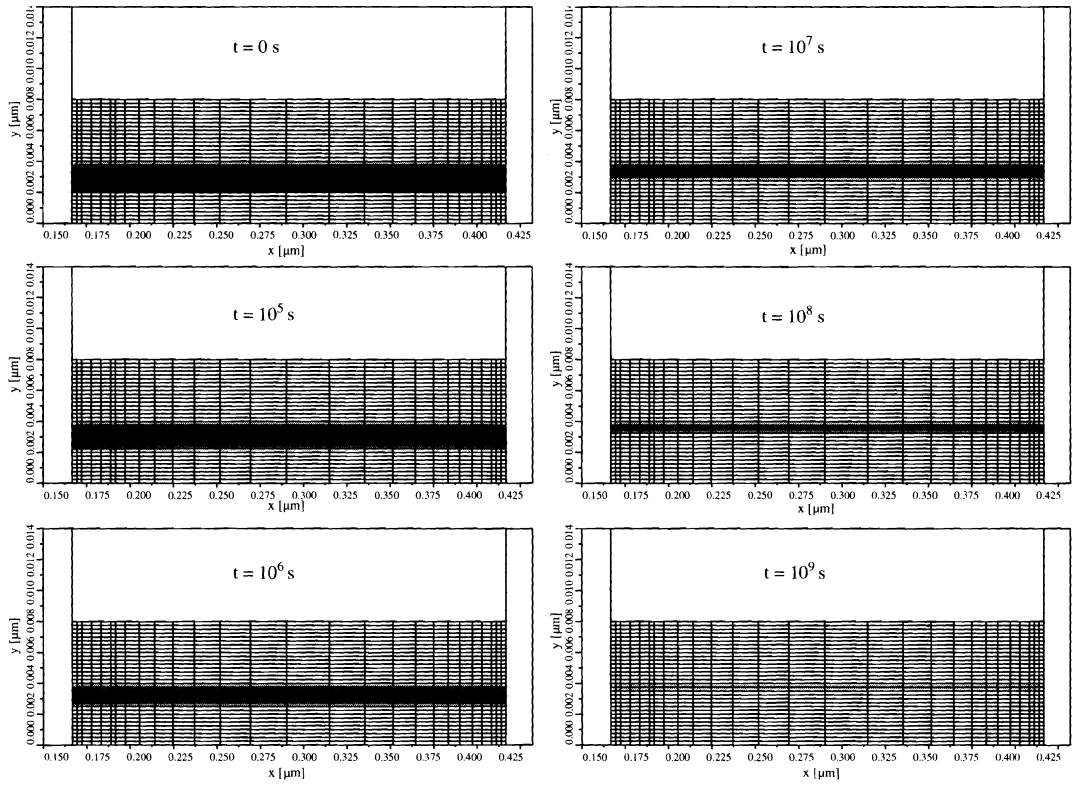


Figure 46. Transient trap occupancy in the trap-rich dielectric layer of a SONOS device that is discharged from $t = 0$ s to $t = 10^9$ **s.**

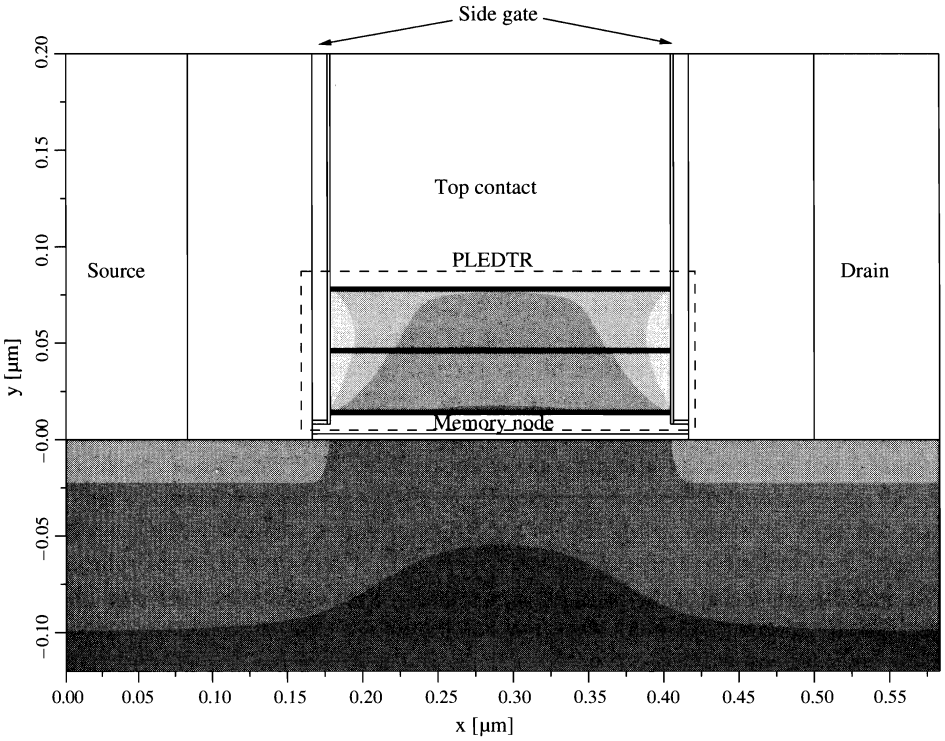


Figure 47. Conduction band edge energy in the PLEDM device.

gates that are separated by a thin dielectric layer. In the on-state, the energy barriers are heavily reduced by the voltage on the side gates, causing a strong tunneling current to flow at the interface to the side gate dielectric. In the off-state, however, the side gates are turned off, and the energy barrier blocks the leakage current. As in a conventional EEPROM, the charge on the memory node is used to control the underlying MOS transistor. Only a small amount of charge has to be added to or removed from the memory node to change the state of the memory cell.

For the simulation of such devices, measurement results for a single Si_3N_4 barrier diode [190] have been used to calibrate the model, as shown in Fig. 48 [191]. For calibration, the carrier mass in the dielectric was used as a fit parameter. Electron and hole masses of 0.5 m_0 and 0.8 m_0 were found to reproduce the data. The Si_3N_4 barrier was modeled with a barrier height of 5 eV and a conduction band offset of 2 eV to the silicon conduction band edge with the relative dielectric permittivity being 7.5.

The effect of the position and size of the central shutter barrier as well as the effect of shrinking the stack width have been investigated. Two cell states have been assumed: an on-state with 3 V applied on the top contact and the side gate, and an off-state with 0.8 V applied on the memory node and 0 V on the side gate. In both states, the charging and discharging current was extracted. The PLEDTR had a stack width of 180 nm and a stack height of 100 nm. The thickness of the upper and lower barriers was set to 2 nm. The left part of Fig. 49 shows the effect of different CSB thicknesses on the on- and off-current of the device. Though the on-current is hardly influenced by the different thicknesses, the off-current is very sensitive to it. Also, the position of the CSB is crucial, because for a CSB located near the memory node, the energy barrier will be reduced in the off-state by the charge on the memory node. If, on the other hand, the CSB is placed near the top contact, the energy barrier is not suppressed in the off-state, and the off-current is much lower. The on-current is also reduced by this effect, but the amount of reduction is much lower as compared to the off-current, due to the fact that the on-current mainly depends on the voltage of the side gate. Thus, the I_{on}/I_{off} ratio increases with the thickness of the central shutter barrier and is highest for a CSB located near the top contact. Such an asymmetry in

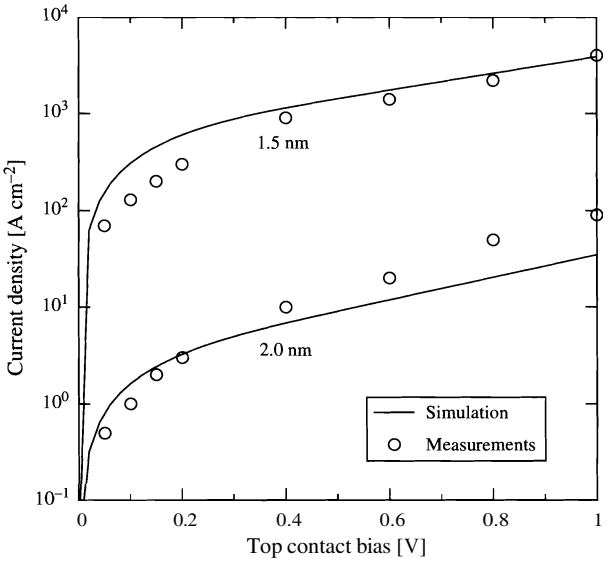


Figure 48. PLEDM calibration of the tunneling current density for a single Si_3N_4 layer with 1.5-nm and 2-nm thickness [191]. The measured values are taken from Ref. [190].

the IV characteristics depending on the position of the central shutter barrier has already been observed experimentally [190].

In Ref. [192], the feasibility of very narrow silicon-insulator stacks is shown. This encourages the assumption that a reduction of the stack width is possible. Figure 49 shows the on- and off-currents of the device with a CSB thickness of 10 nm for a stack width of 140 nm down to 20 nm. It can be seen that a reduction of the stack width leads to increasing on-currents and decreasing off-currents. The reason is that the current in the on-state, which mainly flows as a surface current near the side gate, is not reduced by the decreased width of the stack. It even increases for very low stack widths, which may be due to the fact that the energy barriers at the side of the stack merge for very low stack widths. The off-current, on the other hand, is directly proportional to the stack area and can thus be directly downscaled by shrinking the stack width. For a stack width of 20 nm, I_{on}/I_{off} ratios of more than 10^{32} can be reached.

3.2.2.3. Nonvolatile Memory Devices Based on Crested Barriers One of the most important figures of merit of a nonvolatile memory cell is its I_{on}/I_{off} ratio: A high on-current leads to low programming and erasing times, and a low off-current increases the retention

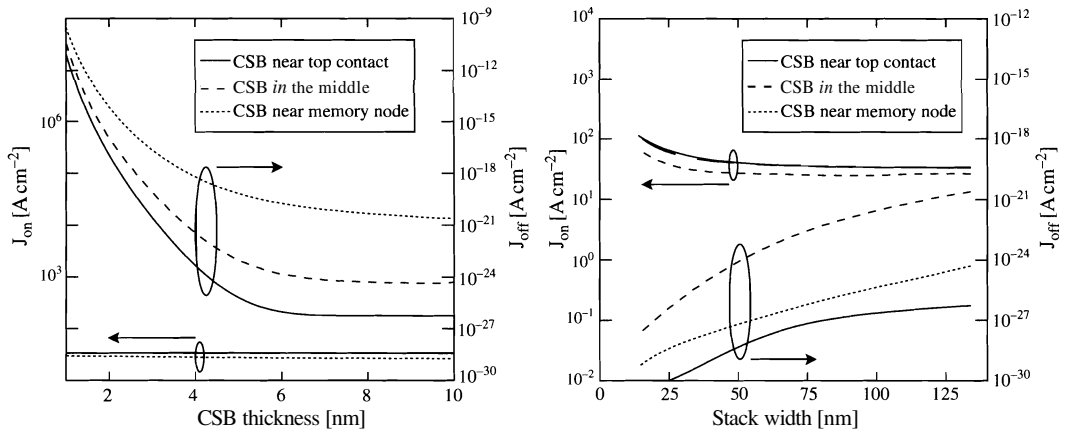


Figure 49. On-current density and off-current density as a function of the thickness of the central shutter barrier (left) and the stack width (right) [191].

time of the device. This ratio can be increased if, for a given device, the tunneling current in the on-state (the charging/discharging current) is increased or, in the off-state (during the retention time), decreased. With a single-layer dielectric it is not possible to tune on- and off-current independently. However, if the tunnel dielectric is replaced by a dielectric stack of varying barrier height as shown in Fig. 50, it becomes possible. In this figure, the device structure and the conduction band edge in the on- and off-state are shown. The device consists of a standard EEPROM structure, where the tunnel dielectric is composed of three layers. The middle layer has a higher energy barrier than the inner and outer layers. The flat-band case is indicated by the dotted lines.

In the on-state, a high voltage is applied on the top contact. The middle energy barrier is strongly reduced and gives rise to a high tunneling current. If the dielectric would consist of a single layer, the peak of the energy barrier would not be reduced. Thus, the on-current is much higher for the layered dielectric. In the off-state, a low negative voltage—due to charge stored on the memory node—is applied. The middle barrier is only slightly suppressed and blocks tunneling. The off-current is only slightly lower than for a single-layer dielectric. This behavior results in a high I_{on}/I_{off} ratio. A high suppression of the middle barrier in the on-state requires a low permittivity of the outer layers so that the potential drop in the outer layers is high [193]. This device design was first proposed by Capasso et al. in 1988 [194] based on AlGaAs–GaAs devices and later used by several authors [195, 196], where it became popular as *crested-barrier* memory or VARIOT (varying oxide thickness device).

The gate current density of the device depicted in Fig. 50 is shown as a function of the gate bias in the left part of Fig. 51. A stack thickness of 5 nm was chosen. Because the middle layers must have a high band gap, only few material combinations are possible. For the simulations, middle layers of Al_2O_3 and SiO_2 have been chosen, with outer layers of Y_2O_3, Si_3N_4, and ZrO_2. For comparison, full SiO_2 and Si_3N_4 stacks have also been simulated (the dotted and dash-dotted lines). Though Y_2O_3 shows a very high off-current, stacks with outer layers of Si_3N_4 or ZrO_2 and Al_2O_3 as middle layer show good ratios between the on-state (positive gate bias) current density and the off-state (negative gate bias) current density.

The important figure of merit, however, is the I_{on}/I_{off} ratio. In the right part of Fig. 51, the I_{on}/I_{off} ratio is shown for Si_3N_4 and ZrO_2 stacks with SiO_2 and Al_2O_3 middle layers as a function of the thickness of the middle layer. Also shown is the ratio for a layer of SiO_2 and Si_3N_4 alone. It is obvious that the ratio strongly depends on the thickness of the middle layer, and both minima and maxima can be observed. Only outer layers of Si_3N_4 lead to a

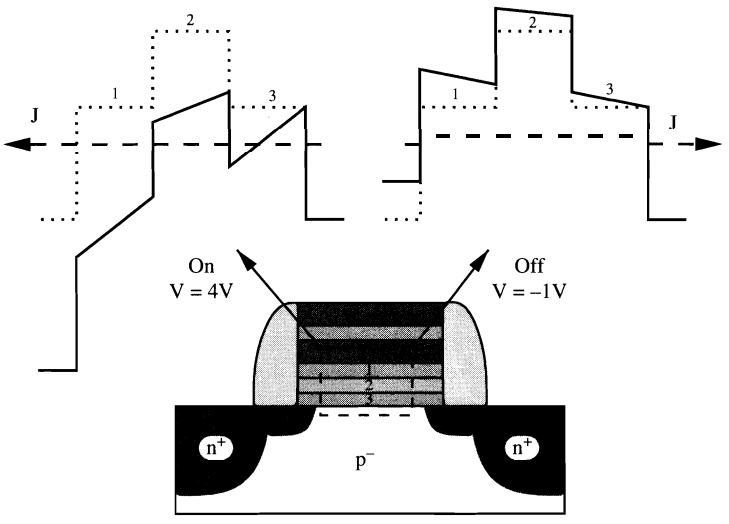


Figure 50. Device structure and operating principle of a nonvolatile memory based on crested barriers.

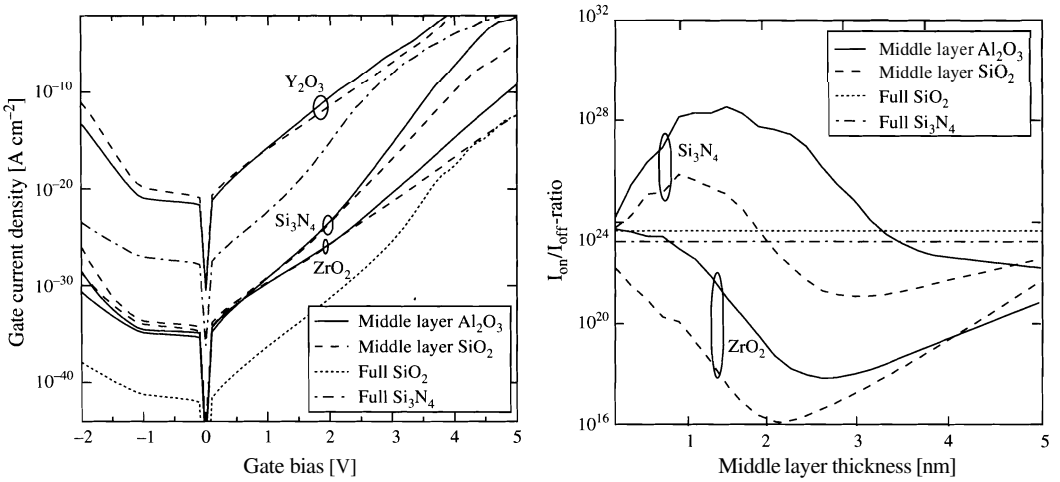


Figure 51. Gate current density as a function of the gate bias for different materials of the middle layer compared to full SiO_2 and Si_3N_4 layers (left). Ratio between the on-current and the off-current as a function of the middle layer thickness for different materials of the outer layers (Si_3N_4 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 (right).

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.

4. CONCLUSIONS

Tunneling effects in semiconductor devices were investigated. A hierarchy of tunneling models was outlined. Three main properties were identified to influence the tunneling process: The carrier energy distribution function, the transmission coefficient, and the presence of traps in the dielectric layer.

The energetic distribution of carriers was investigated using different approximations, such as the frequently applied Fermi–Dirac or Maxwell–Boltzmann statistics. However, these approximations are only valid near equilibrium. Comparisons with the results from Monte Carlo simulations showed that in turned-on devices, the distribution function strongly deviates from the ideal shape. Some non-Maxwellian models were reviewed, and it was found that a model that is based on the solution variables of a six-moments transport model accurately reproduces the Monte Carlo results.

The quantum-mechanical transmission coefficient can be computed from the solution of the stationary Schrodinger equation. Several approximations and analytical formulae were outlined. For a single-layer dielectric, the analytical WKB approximation or Gundlach's formula can be used. For arbitrary-shaped energy barriers, the numerical WKB, the transfer-matrix, or the quantum transmitting boundary method can be applied. It was found that the transfer-matrix method is prone to numerical problems due to the repeated matrix multiplications. The quantum transmitting boundary method turned out to be more robust.

Defects in the dielectric layer give rise to trap-assisted tunneling, which leads to an additional tunneling current at low bias. After reviewing several models from the literature, a recently presented inelastic trap-assisted tunneling model was adapted to avoid the numerical calculation of the overlap integral in the dielectric layer. This yielded a fully analytical model that was further developed to include transient trap charging and discharging effects.

Several examples were studied where a general distinction between tunneling in MOS transistors, where it is a parasitic effect, and tunneling in nonvolatile memory devices, where it is crucial for the device functionality, was made. Tunneling in MOS transistors was investigated, where special attention was paid to the investigation of the different tunneling paths from the gate to the channel and from the gate to the source and drain extension regions.

Furthermore, the importance of the carrier distribution functions for modeling of gate leakage in turned-on devices was shown. If a heated Maxwellian approximation was used for

the description of hot-carrier tunneling, the gate current density was heavily overestimated. This effect was found to be especially pronounced for devices with short gate lengths.

In future CMOS devices, the use of alternative dielectric materials instead of SiO_2 will make the reduction of the effective oxide thickness possible. Several candidate materials were studied, and it was found that they show a pronounced correlation between the barrier height and the permittivity. This makes optimization necessary to find the optimum layer composition. Furthermore, the investigation of a MOS capacitor with a ZrO_2 dielectric showed that the strong defect density makes the use of trap-assisted tunneling models a *sine qua non* for these materials.

In addition to MOS transistors, nonvolatile memory devices were studied. A general overview of nonvolatile memory technology was followed by an investigation of three selected device structures: devices where the floating gate contact is replaced by a layer of trap-rich dielectric, multibarrier tunneling devices, and devices that are based on crested barriers. Especially the multibarrier tunneling devices allow an extremely high I_{on}/I_{off} ratio. The trap-rich dielectric devices, on the other hand, are easier to fabricate and have a smaller footprint. Devices that are based on crested barriers allow tuning of the on- and off-current density independently. However, the I_{on}/I_{off} ratio heavily depends on the thicknesses of the dielectric layers, and simulation is necessary to find the optimum values. The investigated nonvolatile memory applications are expected to show high performance; however, the bad quality of the interface between the dielectric layers may offset the advantage in the I_{on}/I_{off} ratio.

ACKNOWLEDGMENTS

This work would not have been possible without the support of Hans Kosina and Tibor Grasser. Furthermore, the good cooperation with Byoung-Ho Cheong, Stefan Harasek, Francisco Jiménez-Molinos, and Helmut Puchner is gratefully acknowledged.

REFERENCES

1. J. P. Shiely, Ph.D. thesis, Duke University, 1999.
2. R. Clerc, Ph.D. thesis, Institut National Polytechnique de Grenoble, 2001.
3. C. B. Duke, "Tunneling in Solids." Academic Press, New York, 1969.
4. R. Tsu and L. Esaki, *Appl. Phys. Lett.* 22, 562 (1973).
5. N. Ashcroft and N. Mermin, "Solid State Physics." Harcourt College Publishers, Fort Worth, TX, 1976.
6. D. Cassi and B. Riccó, *IEEE Trans. Electron. Devices* 37, 1514 (1990).
7. A. Abramo and C. Fiegna, *J. Appl. Phys.* 80, 889 (1996).
8. K.-I. Sonoda, M. Yamaji, K. Taniguchi, C. Hamaguchi, and S. T. Dunham, *J. Appl. Phys.* 80, 5444 (1996).
9. C. Fiegna, F. Venturi, M. Melanotte, E. Sangiorgi, and B. Riccó, *IEEE Trans. Electron. Devices* 38, 603 (1991).
10. K. Hasnat, C.-F. Yeap, S. Jallepalli, S. A. Hareland, W.-K. Shih, V. M. Agostinelli, A. F. Tasch, and C. M. Maziar, *IEEE Trans. Electron. Devices* 44, 129 (1997).
11. T. Grasser, H. Kosina, C. Heitzinger, and S. Selberherr, *J. Appl. Phys.* 91, 3869 (2002).
12. T. Grasser, H. Kosina, and S. Selberherr, *J. Appl. Phys.* 90, 6165 (2001).
13. E. Nicollian and J. Brews, "MOS (Metal Oxide Semiconductor) Physics and Technology." Wiley, New York, 1982.
14. Y. Tsividis, "Operation and Modeling of the MOS Transistor." McGraw-Hill, New York, 1987.
15. S. M. Sze, "Physics of Semiconductor Devices," 2nd ed. Wiley, New York, 1981.
16. M. Levinshtein, S. Rumyantsev, and M. Shur, "Handbook Series on Semiconductor Parameters," Vol. 1. World Scientific, Singapore, 1996.
17. Y.-C. Yeo, T.-J. King, and C. Hu, *J. Appl. Phys.* 92, 7266 (2002).
18. W. Harrison, "Electronic Structure and the Properties of Solids." Dover Publications, New York, 1989.
19. E. H. Rhoderick and R. H. Williams, "Metal-Semiconductor Contacts." Oxford University Press, Oxford, UK, 1988.
20. W. Franz, in "Handbuch der Physik" (S. Flűgger, Ed.), Vol. XVII, p. 155. Springer, Berlin, 1956.
21. M. V. Fischetti, S. E. Laux, and E. Crabbé, *J. Appl. Phys.* 78, 1058 (1995).
22. M. Kleefstra and G. C. Herman, *J. Appl. Phys.* 51, 4923 (1980).
23. F. Jiménez-Molinos, F. Gámiz, A. Palma, P. Cartujo, and J. A. Lopez-Villanueva, *J. Appl. Phys.* 91, 5116 (2002).
24. G. Yang, K. Chin, and R. Marcus, *IEEE Trans. Electron. Devices* 38, 2373 (1991).
25. A. Schenk and G. Heiser, *J. Appl. Phys.* 81, 7900 (1997).
26. A. Schenk, "Advanced Physical Models for Silicon Device Simulation." Springer, Wien, 1998.
27. C. Fiegna, E. Sangiorgi, and L. Selmi, *IEEE Trans. Electron. Devices* 40, 2018 (1993).
28. Z. A. Weinberg, *J. Appl. Phys.* 53, 5052 (1982).

29. W.-Y. Quan, D. M. Kim, and M. K. Cho, *J. Appl. Phys.* 92, 3724 (2002).
30. L. Larcher, A. Paccagnella, and G. Ghidini, *IEEE Trans. Electron. Devices* 48, 271 (2001).
31. B. Majkusiak, *IEEE Trans. Electron. Devices* 37, 1087 (1990).
32. E. Schrödinger, *Annalen der Physik* 79, 361 (1926).
33. A. Messiah, "Quantum Mechanics," Dover, New York, 2000.
34. S. Gasiorowicz, "Quantum Physics." John Wiley & Sons, New York, 1995.
35. A. Hadjadj, G. Salace, and C. Petit, *J. Appl. Phys.* 89, 7994 (2001).
36. S. Nagano, M. Tsukiji, E. Hasegawa, and A. Ishitani, *J. Appl. Phys.* 75, 3530 (1994).
37. L. F. Register, E. Rosenbaum, and K. Yang, *Appl. Phys. Lett.* 74, 457 (1999).
38. H. Y. Yang, H. Niimi, and G. Lucovsky, *J. Appl. Phys.* 83, 2327 (1998).
39. N. Yang, W. K. Henson, J. R. Hauser, and J. J. Wortman, *IEEE Trans. Electron. Devices* 46, 1464 (1999).
40. M. I. Vexler, N. Asli, A. F. Shulekin, B. Meinerzhagen, and P. Seegebrecht, *Microelectronic Engineering* 59, 161 (2001).
41. J. Zhang, J. S. Yuan, Y. Ma, and A. S. Oates, *Solid-State Electron.* 44, 2165 (2000).
42. K. H. Gundlach, *Solid-State Electron.* 9, 949 (1966).
43. M. Abramowitz and I. A. Stegun, "Handbook of Mathematical Functions." Dover, New York, 1972.
44. A. Shanware, J. P. Shiely, and H. Z. Massoud, in "Proceeding of the International Electron Devices Meeting," pp. 815–818, IEEE Press, Piscataway, NJ, 1999.
45. M. O. Vassell, J. Lee, and H. F. Lockwood, *J. Appl. Phys.* 54, 5208 (1983).
46. Y. Ando and T. Itoh, *J. Appl. Phys.* 61, 1497 (1987).
47. B. Zimmermann, E. Marclay, M. Ilegems, and P. Gueret, *J. Appl. Phys.* 64, 3581 (1988).
48. G. Yong, *Phys. Rev. B* 50, 17249 (1994).
49. R. Clerc, A. Spinelli, G. Ghibaudo, and G. Pananakakis, *J. Appl. Phys.* 91, 1400 (2002).
50. D. K. Ferry and S. M. Goodnick, "Transport in Nanostructures." Cambridge University Press, Cambridge, UK, 1997.
51. W. W. Lui and M. Fukuma, *J. Appl. Phys.* 60, 1555 (1986).
52. K. F. Brennan, *J. Appl. Phys.* 62, 2392 (1987).
53. D. C. Hutchings, *Appl. Phys. Lett.* 55, 1082 (1989).
54. J.-G. S. Demers and R. Maciejko, *J. Appl. Phys.* 90, 6120 (2001).
55. B. A. Biegel, Ph.D. thesis, Stanford University, 1997.
56. J. N. Schulman and Y.-C. Chang, *Phys. Rev. B* 27, 2346 (1983).
57. D. Y. K. Ko and J. C. Inkson, *Phys. Rev. B* 38, 9945 (1988).
58. T. Usuki, M. Saito, M. Takatsu, R. A. Kiehl, and N. Yokoyama, *Phys. Rev. B* 52, 8244 (1995).
59. D. Z. Y. Ting, E. T. Yu, and T. C. McGill, *Phys. Rev. B* 45, 3583 (1992).
60. W. R. Frensley and N. G. Einspruch, Eds., "Heterostructures and Quantum Devices, VLSI Electronics: Microstructure Science." Academic Press, New York, 1994.
61. C. S. Lent and D. J. Kirkner, *J. Appl. Phys.* 67, 6353 (1990).
62. A. P. Gnädinger and H. E. Talley, *Solid-State Electron.* 13, 1301 (1970).
63. F. Stern, *Phys. Rev. B* 5, 4891 (1972).
64. W. Magnus and W. Schoenmaker, *Microelectronics Reliability* 41, 31 (2001).
65. E. Anemogiannis, E. N. Glytsis, and T. K. Gaylord, *IEEE J. Quant. Electron.* 29, 2731 (1993).
66. E. Cassan, *J. Appl. Phys.* 87, 7931 (2000).
67. A. Schenk, in "Proceedings of the European Solid-State Device Research Conference" (H. Ryssel, G. Wachutka, and H. Grünbacher, Eds.), pp. 9–16. Frontier Group, 2001.
68. A. T. M. Fairus and V. K. Arora, *Microelectronics Journal* 32, 679 (2000).
69. N. Matsuo, Y. Takami, and Y. Kitagawa, *Solid-State Electron.* 46, 577 (2002).
70. S. Padmanabhan and A. Rothwarf, *IEEE Trans. Electron. Devices* 36, 2557 (1989).
71. M. J. van Dort, P. H. Woerlee, and A. J. Walker, *Solid-State Electron.* 37, 411 (1994).
72. G. Gildenblatt, B. Gelmont, and S. Vatannia, *J. Appl. Phys.* 77, 6327 (1995).
73. P. J. Price, *Phys. Rev. B* 45, 9042 (1992).
74. P. J. Price, *Appl. Phys. Lett.* 82, 2080 (2003).
75. A. Thean and J. P. Leburton, *IEEE Electron. Device Lett.* 22, 148 (2001).
76. A. Ghetti, A. Hamad, P. J. Silverman, H. Vaidya, and N. Zhao, in "Proceedings of the Simulation of Semiconductor Processes and Devices," pp. 239–242. IEEE Press, Piscataway, NJ, 1999.
77. S. H. Lo, D. A. Buchanan, Y. Taur, and W. Wang, *IEEE Trans. Electron. Devices* 18, 209 (1997).
78. S. Mudanai, Y. Fan, Q. Ouyang, A. F. Tasch, and S. K. Banerjee, *IEEE Trans. Electron. Devices* 47, 1851 (2000).
79. S. Mudanai, L. F. Register, A. F. Tasch, and S. K. Banerjee, *IEEE Electron. Device Lett.* 22, 145 (2001).
80. F. Rana, S. Tiwari, and D. A. Buchanan, *Appl. Phys. Lett.* 69, 1104 (1996).
81. W.-K. Shih, E. X. Wang, S. Jallepalli, F. Leon, C. M. Maziar, and A. F. Tasch, Jr., *Solid-State Electron.* 42, 997 (1998).
82. E. Cassan, P. Dollfus, S. Galdin, and P. Hesto, *IEEE Trans. Electron. Devices* 48, 715 (2001).
83. A. Dalla Serra, A. Abramo, P. Palestri, L. Selmi, and F. Widdershoven, *IEEE Trans. Electron. Devices* 48, 1811 (2001).
84. Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, Eds., "Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide." SIAM, Philadelphia, 2000.
85. R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, *J. Appl. Phys.* 81, 7845 (1997).

86. R. C. Bowen, W. R. Frensley, G. Klimeck, and R. K. Lake, *Phys. Rev. B* 52, 2754 (1995).
87. C. Bowen, C. L. Fernando, G. Klimeck, A. Chatterjee, D. Blanks, R. Lake, J. Hu, J. Davis, M. Kulkarni, S. Hattangady, and I.-C. Chen, in "Proceedings of the International Electron Devices Meeting," pp. 35.1.1–35.1.4. IEEE Press, Piscataway, NJ, 1997.
88. G. Klimeck, R. Lake, C. Bowen, W. R. Frensley, and T. S. Moise, *Appl. Phys. Lett.* 67, 2539 (1995).
89. C. L. Fernando and W. R. Frensley, *J. Appl. Phys.* 76, 2881 (1994).
90. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Numerical Recipes in C." Cambridge University Press, Cambridge, UK, 1997.
91. W. R. Frensley, *Superlattices & Microstructures* 11, 347 (1992).
92. "Nanotechnology Engineering Modeling Program (NEMO) Version 3.0." Raytheon TI Systems, 1996.
93. N. Arora, "MOSFET Models for VLSI Circuit Simulation." Springer, Berlin, 1993.
94. C.-H. Choi, K.-H. Oh, J.-S. Goo, Z. Yu, and R. W. Dutton, in "Proceedings of the International Electron Devices Meeting," pp. 30.6.1–30.6.4. IEEE Press, Piscataway, NJ, 1999.
95. C.-H. Choi, K.-Y. Nam, Z. Yu, and R. W. Dutton, *IEEE Trans. Electron. Devices* 48, 2823 (2001).
96. Y.-S. Lin, H.-T. Huang, C.-C. Wu, Y.-K. Leung, H.-Y. Pan, T.-E. Chang, W.-M. Chen, J.-J. Liaw, and C. H. Diaz, *IEEE Trans. Electron. Devices* 49, 442 (2002).
97. H. Lin, J. T.-Y. Chen, and J.-H. Chang, *Solid-State Electron.* 46, 1145 (2002).
98. S. Schwantes and W. Krautschneider, in "Proceedings of the European Solid-State Device Research Conference" (H. Ryssel, G. Wachutka, and H. Grünbacher, Eds.), pp. 471–474. Frontier Group, 2001.
99. R. H. Fowler and L. Nordheim, *Proc. Roy. Soc. A* 119, 173 (1928).
100. M. Lenzlinger and E. H. Snow, *J. Appl. Phys.* 40, 278 (1969).
101. K. F. Schuegraf and C. Hu, *IEEE Trans. Electron. Devices* 41, 761 (1994).
102. K. F. Schuegraf, C. C. King, and C. Hu, in "Proceedings of the Symposium on VLSI Technology," 1992, pp. 18–19.
103. S. Aritome, R. Shirota, G. Hemink, T. Endoh, and F. Masuoka, *Proc. IEEE* 81, 776 (1993).
104. R. Moazzami and C. Hu, in "Proceedings of the International Electron Devices Meeting," pp. 139–142. IEEE Press, Piscataway, NJ, 1992.
105. E. Rosenbaum and L. F. Register, *IEEE Trans. Electron. Devices* 44, 317 (1997).
106. S.-I. Takagi, N. Yasuda, and A. Toriumi, *IEEE J. Solid-State Circuits* 46, 348 (1999).
107. R. Rofan and C. Hu, *IEEE Electron. Device Lett.* 12, 632 (1991).
108. J. Wu, L. F. Register, and E. Rosenbaum, in "Proceedings of the International Reliability Physics Symposium," 1999, pp. 389–395.
109. B. Riccó, G. Gozzi, and M. Lanzoni, *IEEE Trans. Electron Devices* 45, 1554 (1998).
110. K. Sakakibara, N. Ajika, K. Eikyu, K. Ishikawa, and H. Miyoshi, *IEEE Trans. Electron. Devices* 44, 1002 (1997).
111. A. Ghetti, E. Sangiorgi, J. Bude, T. W. Sorsch, and G. Weber, *IEEE Trans. Electron. Devices* 47, 2358 (2000).
112. C.-M. Yih, Z.-H. Ho, M.-S. Liang, and S. S. Chung, *IEEE Trans. Electron. Devices* 48, 300 (2001).
113. A. I. Chou, K. Lai, K. Kumar, P. Chowdhury, and J. C. Lee, *Appl. Phys. Lett.* 70, 3407 (1997).
114. K. Komiya and Y. Omura, *Microelectronic Engineering* 59, 61 (2001).
115. T.-K. Kang, M.-J. Chen, C.-H. Liu, Y. J. Chang, and S.-K. Fan, *IEEE Trans. Electron. Devices* 48, 2317 (2001).
116. M. Lenski, T. Endoh, and F. Masuoka, *J. Appl. Phys.* 88, 5238 (2000).
117. S.-I. Takagi, N. Yasuda, and A. Toriumi, *IEEE Trans. Electron. Devices* 46, 335 (1999).
118. W. J. Chang, M. P. Houng, and Y. H. Wang, *J. Appl. Phys.* 89, 6285 (2001).
119. W. J. Chang, M. P. Houng, and Y. H. Wang, *J. Appl. Phys.* 90, 5171 (2001).
120. L. Larcher, A. Paccagnella, and G. Ghidini, *IEEE Trans. Electron. Devices* 48, 285 (2001).
121. D. Ielmini, A. S. Spinelli, M. A. Rigamonti, and A. L. Lacaita, *IEEE Trans. Electron. Devices* 47, 1258 (2000).
122. D. Ielmini, A. S. Spinelli, M. A. Rigamonti, and A. L. Lacaita, *IEEE Trans. Electron. Devices* 47, 1266 (2000).
123. D. Ielmini, A. S. Spinelli, A. L. Lacaita, A. Martinelli, and G. Ghidini, *Solid-State Electron.* 45, 1361 (2001).
124. D. Ielmini, A. S. Spinelli, A. L. Lacaita, and G. Ghidini, *Solid-State Electron.* 46, 417 (2002).
125. D. Ielmini, A. S. Spinelli, A. L. Lacaita, and A. Modelli, *Microelectronic Engineering* 59, 189 (2001).
126. D. Ielmini, A. S. Spinelli, A. L. Lacaita, and A. Modelli, *Solid-State Electron.* 46, 1749 (2002).
127. A. Ghetti, *Microelectronic Engineering* 59, 127 (2001).
128. C. Chaneliere, J. L. Autran, and R. A. B. Devine, *J. Appl. Phys.* 86, 480 (1999).
129. M. Houssa, R. Degraeve, P. W. Mertens, M. M. Heyns, J. S. Leon, A. Halliyal, and B. Ogle, *J. Appl. Phys.* 86, 6462 (1999).
130. M. Houssa, M. Tuominen, M. Naili, V. Afanas'ev, A. Stesmans, S. Haukka, and M. M. Heyns, *J. Appl. Phys.* 87, 8615 (2000).
131. D. Caputo, F. Irrera, S. Salerno, S. Spiga, and M. Fanciulli, in "Proceedings of the 4th European Workshop on Ultimate Integration of Silicon" (E. Sangiorgi and L. Selmi, Eds.), pp. 89–92. University of Udine, Udine, Italy, 2003.
132. B. DeSalvo, G. Ghibaudo, G. Pananakakis, B. Guillaumot, and G. Reimbold, *Solid-State Electron.* 44, 895 (2000).
133. E. Kameda, T. Matsuda, Y. Emura, and T. Ohzone, *Solid-State Electron.* 42, 2105 (1998).
134. J. A. López-Villanueva, J. A. Jiménez-Tejada, P. Cartujo, J. Bausells, and J. E. Carceller, *J. Appl. Phys.* 70, 3712 (1991).
135. F. Jiménez-Molinos, A. Palma, F. Gámiz, J. Banqueri, and J. A. Lopez-Villanueva, *J. Appl. Phys.* 90, 3396 (2001).

136. A. Palma, A. Godoy, J. A. Jimenez-Tejada, J. E. Carceller, and J. A. Lopez-Villanueva, *Phys. Rev. B* 56, 9565 (1997).
137. J. H. Zheng, H. S. Tan, and S. C. Ng, *J. Phys.: Condensed Matter* 6, 1695 (1994).
138. W. B. Fowler, J. K. Rudra, M. E. Zvanut, and F. J. Feigl, *Phys. Rev. B* 41, 8313 (1990).
139. M. Herrmann and A. Schenk, *J. Appl. Phys.* 77, 4522 (1995).
140. A. Gehring, H. Kosina, T. Grasser, and S. Selberherr, in "Proceedings of the 4th European Workshop on Ultimate Integration of Silicon" (E. Sangiorgi and L. Selmi, Eds.), pp. 131–134. University of Udine, Udine, Italy, 2003.
141. Minimos-NT 2.1 User's Guide, Institute for Microelectronics, Wien, 2004.
142. J. Cai and C.-T. Sah, *J. Appl. Phys.* 89, 2272 (2001).
143. H. Z. Massoud and J. P. Shiely, *Microelectronic Engineering* 36, 263 (1997).
144. Y. Shi, T. P. Ma, S. Prasad, and S. Dhanda, *IEEE Trans. Electron. Devices* 45, 2355 (1998).
145. M. Städele, B. Tuttle, B. Fischer, and K. Hess, *J. Comput. Electron.* 1, 153 (2002).
146. M. Städele, F. Sacconi, A. D. Carlo, and P. Lugli, *J. Appl. Phys.* 93, 2681 (2003).
147. F. Sacconi, M. Povolotskyi, A. D. Carlo, P. Lugli, and M. Städele, in "Proceeding of the 4th European Workshop on Ultimate Integration of Silicon" (E. Sangiorgi and L. Selmi, Eds.), pp. 125–128. University of Udine, Udine, Italy, 2003.
148. S. H. Lo, D. A. Buchanan, and Y. Taur, *IBM J. Res. Dev.* 43, 327 (1999).
149. S. Jallepalli, J. Bude, W.-K. Shih, M. R. Pinto, C. M. Maziar, and A. F. Tasch, Jr., *IEEE Trans. Electron. Devices* 44, 297 (1997).
150. *MEDICI User's Manual*, Synopsys, Mountain View, CA, 2003.
151. A. Gehring, T. Grasser, H. Kosina, and S. Selberherr, *J. Appl. Phys.* 92, 6019 (2002).
152. A. Gehring, T. Grasser, H. Kosina, and S. Selberherr, *Electron. Lett.* 39, 691 (2003).
153. L. Selmi, A. Ghetti, R. Bez, and E. Sangiorgi, *Microelectronic Engineering* 36, 293 (1997).
154. E. M. Vogel, K. Z. Ahmed, B. Hornung, K. Henson, P. K. McLarty, G. Lucovsky, J. R. Hauser, and J. J. Wortman, *IEEE Trans. Electron. Devices* 45, 1350 (1998).
155. A. Gehring, H. Kosina, and S. Selberherr, *J. Comput. Electron.* 2, 219 (2003).
156. Y.-Y. Fan, R. E. Nieh, J. C. Lee, G. Lucovsky, G. A. Brown, L. F. Register, and S. K. Banerjee, *IEEE Trans. Electron. Devices* 49, 1969 (2002).
157. S. Harasek, Ph.D. thesis, Technische Universität Wien, 2003.
158. A. Gehring, S. Harasek, E. Bertagnolli, and S. Selberherr, in "Proceedings of European Solid-State Device Research Conference" (J. Franca and R. Freitas, Eds.), pp. 473–476. Frontier Group, 2003.
159. P. Tanner, S. Dimitrijev, and H. B. Harrison, *Electron. Lett.* 31, 1880 (1995).
160. D. A. Antoniadis, I. J. Djomehri, K. M. Jackson, and S. Miller, "Well-Tempered" Bulk-Si NMOSFET Device Home Page, available at http://www-mtl.mit.edu/Well/.
161. W. D. Brown and J. Brewer, "Nonvolatile Semiconductor Memory Technology," IEEE Press, Piscataway, NJ, 1998.
162. A. Concannon, S. Keeney, A. Mathewson, R. Bez, and C. Lombardi, *IEEE Trans. Electron. Devices* 40, 1258 (1993).
163. S. Keeney, R. Bez, D. Cantarelli, F. Piccinin, A. Mathewson, L. Ravazzi, and C. Lombardi, *IEEE Trans. Electron. Devices* 39, 2750 (1992).
164. A. Kolodny, S. T. K. Nieh, B. Eitan, and J. Shappir, *IEEE Trans. Electron. Devices* 33, 835 (1986).
165. K. T. San, C. Kaya, D. K. Y. Liu, T.-P. Ma, and P. Shah, *IEEE Electron. Device Lett.* 13, 328 (1992).
166. R. Bouchakour, N. Harabech, P. Canet, P. Boivin, and J. M. Mirabel, in "Proceedings of the International Symposium on Circuits & Systems," 2001, pp. 822–825.
167. R. Duane, A. Concannon, P. O'Sullivan, M. O'Shea, and A. Mathewson, *Solid-State Electron.* 45, 235 (2001).
168. D. Kahng and S. M. Sze, *Bell Syst. Tech. J.* 46, 1288 (1967).
169. P. Pavan, R. Bez, P. Olivo, and E. Zanoni, *Proc. IEEE* 86, 1248 (1997).
170. P. Cappelletti, C. Golla, P. Olivo, and E. Zanoni, "Flash Memories." Kluwer Academic Publishers, Boston, 2000.
171. A. Gehring, F. Jiménez-Molinos, H. Kosina, A. Palma, F. Gámiz, and S. Selberherr, *Microelectron. Reliab.* 43, 1495 (2003).
172. P. Canet, R. Bouchakour, N. Harabech, P. Boivin, J. M. Mirabel, and C. Plossu, in "Proceedings of the 43rd IEEE Midwest Symposium on Circuits and Systems," 2000, pp. 1144–1147.
173. M. K. Cho and D. M. Kim, *IEEE Electron. Device Lett.* 21, 399 (2000).
174. J. M. Caywood, C. J. Huang, and Y. J. Chang, *IEEE Trans. Electron. Devices* 49, 802 (2002).
175. B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Frommer, and D. Finzi, *IEEE Electron. Device Lett.* 21, 543 (2000).
176. M. H. White, D. A. Adams, and J. Bu, *IEEE Circuits Devices* 22 (2000).
177. K.-T. Chang, W.-M. Chen, C. Swift, J. M. Higman, W. M. Paulson, and K.-M. Chang, *IEEE Electron. Device Lett.* 19, 253 (1998).
178. G. Iannaccone and P. Coli, *Appl. Phys. Lett.* 78, 2046 (2001).
179. A. Thean and J. P. Leburton, *IEEE Electron. Device Lett.* 20, 286 (1999).
180. J. J. Welser, S. Tiwari, S. Rishton, K. Y. Lee, and Y. Lee, *IEEE Electron. Device Lett.* 18, 278 (1997).
181. B. DeSalvo, G. Ghibaudo, G. Pananakakis, P. Masson, T. Baron, N. Buffet, A. Fernandes, and B. Guillaumot, *IEEE Trans. Electron. Devices* 48, 1789 (2001).
182. K. Han, I. Kim, and H. Shin, *IEEE Trans. Electron. Devices* 48, 874 (2001).
183. H. I. Hanafi, S. Tiwari, and I. Khan, *IEEE Trans. Electron. Devices* 43, 1553 (1996).

184. Y.-C. King, T.-J. King, and C. Hu, *IEEE Trans. Electron. Devices* 20, 409 (1999).
185. X. Tang, X. Baie, J.-P. Colinge, C. Gusting, and V. Bayot, *IEEE Trans. Electron. Devices* 49, 1420 (2002).
186. K. Nakazato, P. J. A. Piotrowicz, D. G. Hasko, H. Ahmed, and K. Itoh, in "Proceedings of the International Electronic Devices Meeting," pp. 179–182. IEEE Press, Piscataway, NJ, 1997.
187. H. Mizuta, K. Nakazato, P. J. A. Piotrowicz, K. Itoh, T. Teshima, K. Yamaguchi, and T. Shimada, in "Proceedings of the Symposium on VLSI Technology," 1998, pp. 128–129.
188. N. Nakazato, K. Itoh, H. Mizuta, and H. Ahmed, *Electron. Lett.* 35, 848 (1999).
189. K. Nakazato, K. Itoh, H. Ahmed, H. Mizuta, T. Kisu, M. Kato, and T. Sakata, in "Proceedings of the International Solid-State Circuits Conference," 2000, p. TA 7.4.
190. H. Mizuta, M. Wagner, and K. Nakazato, *IEEE Trans. Electron. Devices* 48, 1103 (2001).
191. A. Gehring, T. Grasser, B.-H. Cheong, and S. Selberherr, *Solid-State Electron.* 46, 1545 (2002).
192. H. Fukuda, J. L. Hoyt, M. A. McCord, and R. F. W. Pease, *Appl. Phys. Lett.* 70, 333 (1997).
193. J. D. Casperson, L. D. Bell, and H. A. Atwater, *J. Appl. Phys.* 92, 261 (2002).
194. F. Capasso, F. Beltram, R. J. Malik, and J. F. Walker, *IEEE Electron. Device Lett.* 9, 377 (1988).
195. K. K. Likharev, *Appl. Phys. Lett.* 73, 2137 (1998).
196. B. Govoreanu, P. Blomme, M. Rosmeulen, J. V. Houdt, and K. D. Meyer, *IEEE Electron. Device Lett.* 24, 99 (2003).
197. W. Harrison, "Solid State Theory." Dover, New York, 1979.
198. M. LeRoy, E. Lheurette, O. Vanbesien, and D. Lippens, *J. Appl. Phys.* 93, 2966 (2003).
199. C. M. Osburn, I. Kim, S. K. Han, I. De, K. F. Yee, S. Gannavaram, S. J. Lee, C.-H. Lee, Z. J. Luo, W. Zhu, J. R. Hauser, D.-L. Kwong, G. Lucovsky, T. P. Ma, and M. C. Öztürk, *IBM J. Res. Dev.* 46, 299 (2002).
200. G. D. Wilk, R. M. Wallace, and J. M. Anthony, *J. Appl. Phys.* 89, 5243 (2001).
201. J. Robertson, *J. Vac. Sci. Technol.* 18, 1785 (2000).
202. H.-S. P. Wong, *IBM J. Res. Dev.* 46, 133 (2002).