ECS Transactions, 1 (5) 693-703 (2006)
10.1149/1.2209316, copyright The Electrochemical Society

EFFICIENT CALCULATION OF LIFETIME BASED DIRECT TUNNELING THROUGH STACKED DIELECTRICS

M. Karner[1], A. Gehring[2], S. Holzer[3], H. Kosina[1], and S. Selberherr[1]

[1]Institute for Microelectronics
Technische Universität Wien
Gußhausstraße 27–29, A-1040 Wien, Austria
[2]AMD Saxony, Wilschdorfer Landstrasse 101,
D-01109 Dresden, Germany
[3]Christian Doppler Laboratory for TCAD in Microelectronics
at the Institute for Microelectronics

Abstract. We present an efficient simulation method for lifetime based tunneling in CMOS devices through layers of high-κ dielectrics, which relies on the precise determination of quasi-bound states (QBS). The QBS are calculated with the perfectly matched layer (PML) method. Introducing a complex coordinate stretching allows artifical absorbing layers to be applied at the boundaries. The QBS appear as the eigenvalues of a linear, non-Hermitian Hamiltonian where the QBS lifetimes are directly related to the imaginary part of the eigenvalues. The PML method turns out to be a numerically stable and efficient method to calculate QBS lifetimes for the investigation of direct tunneling through stacked gate dielectrics.

INTRODUCTION

The continuous progress in the development of MOS field-effect transistors within the last decades goes hand in hand with down-scaling the device feature size. To enable further device down-scaling to the deca nanometer channel length regime, it is necessary to reduce the effective oxide thicknesses (EOT) below 2 nm, which will result in high gate leakage currents. The use of high-κ gate dielectrics provides an option to reduce the gate leakage current of future CMOS devices while retaining a good control over the inversion charge (1).

Gate dielectric stacks consisting of high-κ dielectric layers such as Si_3N_4, Al_2O_3, Ta_2O_5, HfO_2, or ZrO_2 have been suggested as alternative dielectrics. Parameter values for these materials taken from (2)-(8) are summarized in Tab. I.

Apart from interface quality and reliability, the dielectric permittivity and the conduction band offset to silicon are of utmost importance as they determine the gate current density through the layer. Furthermore, at the interface to the underlying silicon substrate, an interface layer exists which is either created unintentionally

Table I: Dielectric permittivity, band gap, and conduction band offset of dielectric materials.

	Permittivity κ/κ_0 [1]	Band gap $\mathcal{E}_\mathrm{g}$ [eV]	Band offset $\Delta\mathcal{E}_\mathrm{C}$ [eV]
SiO_2	3.9	8.9 – 9.0	3.0 – 3.5
Si_3N_4	7.0 – 7.9	5.0 – 5.3	2.0 – 2.4
Ta_2O_5	23.0 – 26.0	4.4 – 4.5	0.3 – 1.5
TiO_2	39.0 – 170.0	3.0 – 3.5	0.0 – 1.2
Al_2O_3	7.9 – 12.0	5.6 – 9.0	2.78 – 3.5
ZrO_2	12.0 – 25.0	5.0 – 7.8	1.4 – 2.5
HfO_2	16.0 – 40.0	4.5 – 6.0	1.5

during processing or intentionally deposited to improve the interface quality. Unfortunately, materials with high permittivity have a low band offset and vice versa, so that a trade-off between these parameters has to be found. However, for investigation of tunneling phenomena and especially for optimization purposes, accurate, and yet efficient simulation models are necessary.

CALCULATION OF DIRECT TUNNELING USING A LIFETIME BASED APPROACH

Calculation of tunneling currents is frequently based on the assumption of a three-dimensional continuum of states at both sides of the gate dielectric and the conservation of parallel momentum. Then, the tunneling current can be described by the Tsu-Esaki formula, (9)

$$J_\mathrm{3D} = \mathrm{q}\int_{\mathcal{E}_\mathrm{min}}^{\mathcal{E}_\mathrm{max}} TC(\mathcal{E}_x, m_\mathrm{diel})N(\mathcal{E}_x, m_\mathrm{D})\,\mathrm{d}\mathcal{E}_x\ , \qquad [1]$$

where $TC(\mathcal{E}_x, m_\mathrm{diel})$ is the transmission coefficient and $N(\mathcal{E}_x, m_\mathrm{D})$ the supply function. Two electron masses enter this equation: The density-of-states mass in the plane parallel to the interface, $m_\mathrm{D} = 2m_\mathrm{t}^* + 4\sqrt{m_\mathrm{t}^* m_\mathrm{l}^*}$, which, equals $2.052m_0$ for (100) silicon with $m_\mathrm{l}^* = 0.92m_0$ and $m_\mathrm{t}^* = 0.19m_0$, and the electron mass in the dielectric m_diel, which is commonly used as a fit parameter (10).

However, in the inversion layer of a MOS-structure, the strong electric field leads to quantum confinement. Whenever electrons are confined or partially confined in movement, this gives rise to bound or quasi bound states (QBS), and the assumption of continuum tunneling is no longer valid. In the inversion layers of MOS-FETs, a major, if not the dominant, source of tunneling electrons is represented by quasi bound states (11). The QBS tunneling current is proportional to $\sum n_i/\tau_i$ where n_i

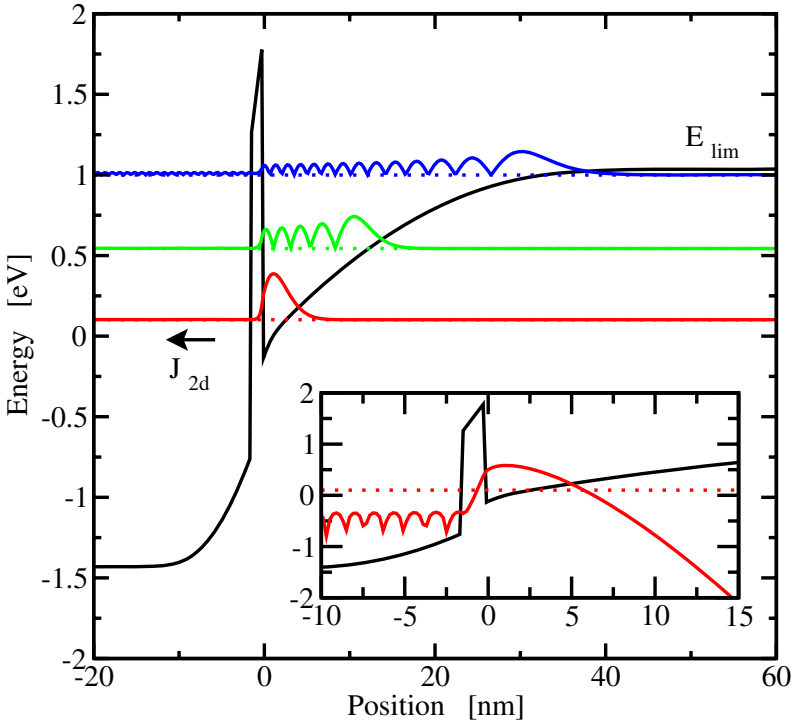

Fig. 1: The potential well of an nMOS inversion layer and its eigenstates assuming closed boundary conditions. The inset displays the wave function of the first QBS on a logarithmic scale.

and τ_i denote the carrier concentration and the lifetime of the QBS with index i, respectively. To take into account the tunneling current from both, continuum and quasi-bound states, [1] has to be replaced by

$$J = J_{2D} + J_{3D} = \frac{k_B T q}{\pi \hbar^2} \sum_{i,\nu} \frac{g_\nu m_\parallel}{\tau_\nu(\mathcal{E}_{\nu,i}(m_q))} \ln\left(1 + \exp\left(\frac{\mathcal{E}_F - \mathcal{E}_{\nu,i}}{k_B T}\right)\right) + q \int_{\mathcal{E}_{min,1}}^{\mathcal{E}_{max}} TC(\mathcal{E}_x, m_{diel}) N(\mathcal{E}_x, m_D)\, d\mathcal{E}_x \ . \quad [2]$$

Here, the symbols g_ν, $m_\parallel$, and m_q denote the valley degeneracy, parallel, and quantization masses respectively ($g = 2$: $m_\parallel = m_t$, $m_q = m_l$ and $g = 4$: $m_\parallel = \sqrt{m_l m_t}$, $m_q = m_t$), $\tau_\nu(\mathcal{E}_{\nu,i})$ is the lifetime of the quasi-bound state $\mathcal{E}_{\nu,i}$, and the integration in the Tsu-Esaki formula starts from $\mathcal{E}_{min,1} = E_{lim}$ as indicated in Fig 1. The following considerations are focused on the tunneling current J_{2D} originating from the QBS.

Within our simulation framework the QBS are obtained from the single particle, time-independent, effective mass SCHRÖDINGER equation:

$$-\frac{\hbar^2}{2} \nabla \cdot \left(\tilde{m}^{-1} \nabla \Psi(\mathbf{x})\right) + V(\mathbf{x})\Psi(\mathbf{x}) = \mathcal{E}\Psi(\mathbf{x}). \quad [3]$$

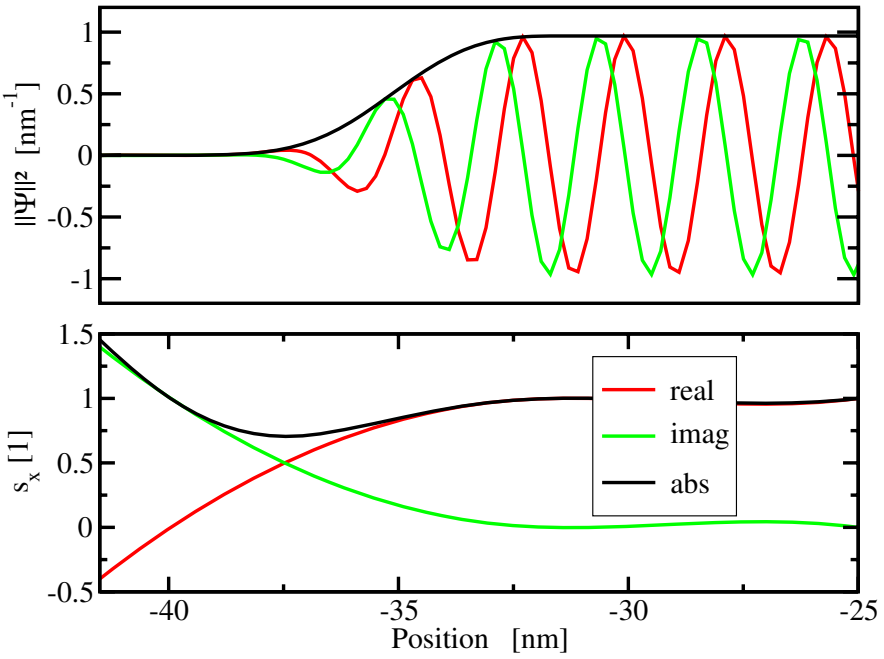

Fig. 2: The wave function of the first QBS and the complex stretching function are displayed in the perfectly matched layer region as well as its transition to the physical region.

Several methods have been proposed to calculate the quasi-bound states and their respective lifetimes (12). In a first approximation the energy levels of the QBS can be estimated by the eigenvalues of the Hamiltonian of the closed system as displayed in Fig. 1. Since closed boundaries are assumed, no information about the broadening and the associated QBS lifetimes is available. It is to note that bound states cannot carry any current, since their wavefunctions Ψ fulfill the relation: $\Psi\nabla\Psi^* - \Psi^*\nabla\Psi = 0$.

A semi-classical approximation based on corrected closed-boundary eigenvalues, which uses a classical formulation of the lifetime (escape time) is pointed out in (13). However, using the closed-boundary eigenvalues for the calculation of open-boundary QBS lifetimes seems to be questionable.

A more rigorous way to apply open boundary conditions to (3) is the quantum transmitting boundary method (QTBM) (14) where a computationally intensive scanning of the derivative of the phase of the reflection coefficient (12) or the reflection coefficient itself (15) yields the desired QBS lifetimes. These methods are especially demanding in the presence of strong confinement (high lifetimes).

PERFECTLY MATCHED LAYER METHOD

Recently, a method based on absorbing boundary conditions (known as the Perfectly Matched Layer (PML) method) for SCHRÖDINGER's equation has been applied for band structure calculations in III-V heterostructure devices (16). In the present work the PML formalism which is often used in electromagnetics, has been applied to determine the energy levels and the lifetime broadening of QBS in MOS inversion layers. In contrast to the QTBM, the Hamiltonian of the system is still linear. Thus, all QBS are calculated in one step and no iteration or scanning procedures are needed.

The basic principle is to add non-physical absorbing layers at the boundary of the simulation region (physical region). This procedure prevents reflections at the boundary of the physical region. The artificial absorbing layers allow the application of Dirichlet boundary conditions, and the QBS are determined by the eigenvalues of the non-Hermitian Hamiltonian of the system. This yields the desired QBS which are the eigenstates of the open system, although Dirichlet boundary conditions are applied. The absorbing property of the PML region is achieved by introducing stretched coordinates

$$\tilde{x} = \int_0^x s_x(\tau)\,\mathrm{d}\tau \qquad [4]$$

in (3). The evaluation of the gradient operator ∇ in one dimension yields:

$$\frac{\partial}{\partial \tilde{x}} = \frac{1}{s_x(x)}\frac{\partial}{\partial x}. \qquad [5]$$

In the artificial layers the stretching function $s_x(x)$ is given as $s_x(x) = 1 + (\alpha + \imath\beta)x^n$, with $\alpha = 1$, $\beta = 1.4$, and $n = 2$, while it is unity in the physical region as displayed in Fig. 2. Adding absorbing layers at the boundary of the physical simulation region, the Hamiltonian becomes non-Hermitian and admits complex eigenvalues $\mathcal{E} = \mathcal{E}_r + \imath\mathcal{E}_i$. The QBS lifetimes are related to the imaginary parts of the eigenvalues as $\tau_i = \hbar/2\mathcal{E}_i$.

To better clarify the PML method, let us assume a constant potential $V(z)$ in the PML region. Then, within this region, the wave function can be written as a plane wave $\Psi(x) = \Psi_0 \exp(\imath\tilde{k_x}x)$ with the wave vector $\tilde{k_x} = k_x/s_x$. Considering two points in the PML region x_1, $x_2 = x_1 + \mathrm{d}x$ the wave vector at the point x_2 can be approximated as

$$k_x(x_2) \approx \frac{s_x(x_2)}{s_x(x_1)} k_x(x_1) = (1 + (\alpha + \imath\beta)\,\mathrm{d}x)\ . \qquad [6]$$

Therefore, the parameter α scales the phase velocity of the plane wave, while β acts as a damping parameter. Since this damping coefficient is greater than zero in the absorbing region, the envelope of the wave functions decay to zero, as can be seen in Fig. 2. These parameters, as well as the thickness of the absorbing layer can

be varied over a wide range with virtually no influence on the results, as long as there are no reflections at the boundaries. However, to achieve this goal, the complex stretching function and its first derivative have to be continuous.

In the gate region, using QTBM or assuming closed boundary conditions results in a superposition of two plane waves in opposite directions, which can bee seen in the inset of Fig. 1. In contrast, when using PML, there are no reflected waves. The wave function is a traveling wave with a constant envelope function. In the absorbing layer, the wave functions are gradually decaying to zero (Fig. 2). The QBS, however, are reproduced correctly.

For an arbitrary potential well a comparison between the PML method and the established methods has been carried out in (17). Very good agreement between the established QTBM and the PML formalism has been obtained.

Furthermore, the computational effort of the PML and QTBM approaches was compared. Fig. 3 shows the CPU time necessary to calculate 1, 3, and 30 quasi-bound states with the QTB and PML methods as a function of the spatial resolution. For the QTBM, an equidistant grid in energy space was used to determine the lifetime broadening of the QBS. Although the dimension of the system increases due to the additional points in the PML region, the computational effort of the PML method has shown to be in almost all cases lower than that of the QTBM.

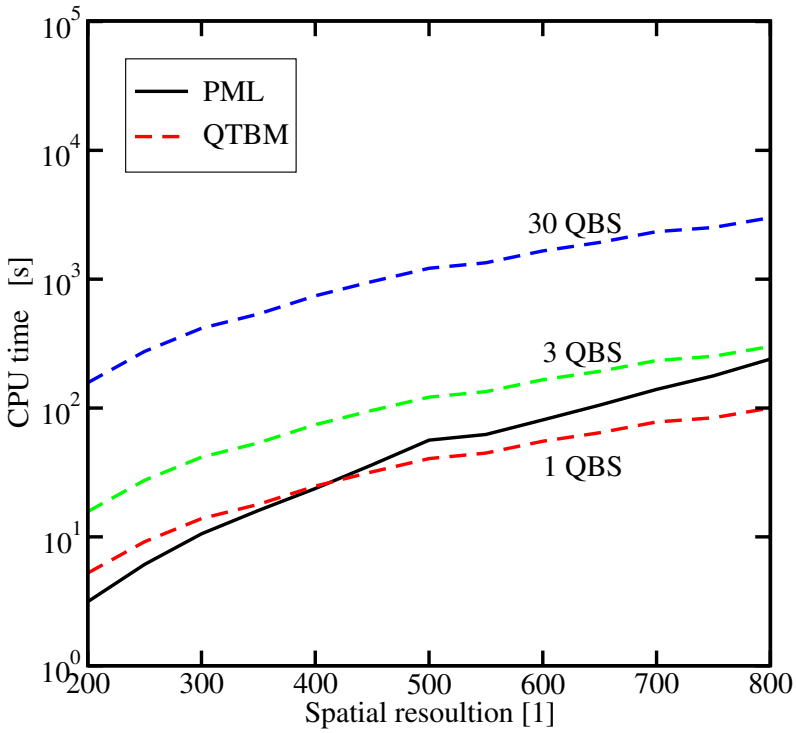

Fig. 3: Comparison of the CPU time demand for the PML, and the QTB methods.

APPLICATION TO DEVICE SIMULATION

With the described method, the gate leakage currents of nMOS transistors with a gate length of 50nm have been evaluated. The gate current density has been evaluated for a stacked SiO_2-Si_3N_4 and a single SiO_2 layer gate dielectric having nearly the same EOT. A doping of N_A=3×10^{17}cm^{-3} in the bulk and N_D=1×10^{19}cm^{-3} in the poly-silicon gate was assumed.

For the investigation of gate leakage currents in the whole device, the conduction band edge has been acquired from the device simulator MINIMOS-NT (18). It is displayed for strong inversion at a gate bias of 1.2 V and V_{DS}=0.0V in Fig. 4, and at drain bias of 0.6 V in Fig. 5. Several one-dimensional cuts through the simulation region are shown in Fig.6.

As a post-processing step on these cuts the QBS energy levels and the related lifetimes have been evaluated using the PML formalism. Based on an accurate computation of the QBS lifetimes, the tunneling current has been estimated according to (2). For the stacked gate dielectric some of the extracted quasi-bound states are shown in Fig. 7 considering the transversal mass as the quantization mass at V_{GB}=1.2V. The energy levels, the QBS lifetimes, and their contribution to the total current density are listed in Tab. II.

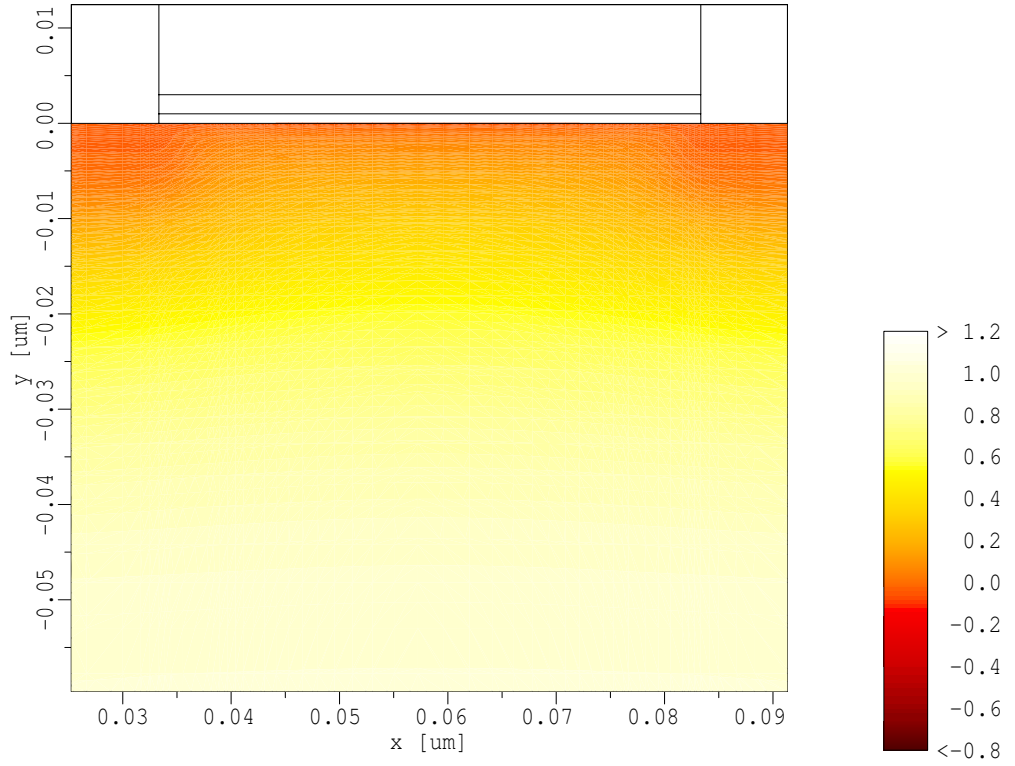

Fig. 4: The band edge energy [eV] of the nMOS device with a stacked gate dielectric evaluated at a gate bias of 1.5 V and a drain voltage of 0 V.

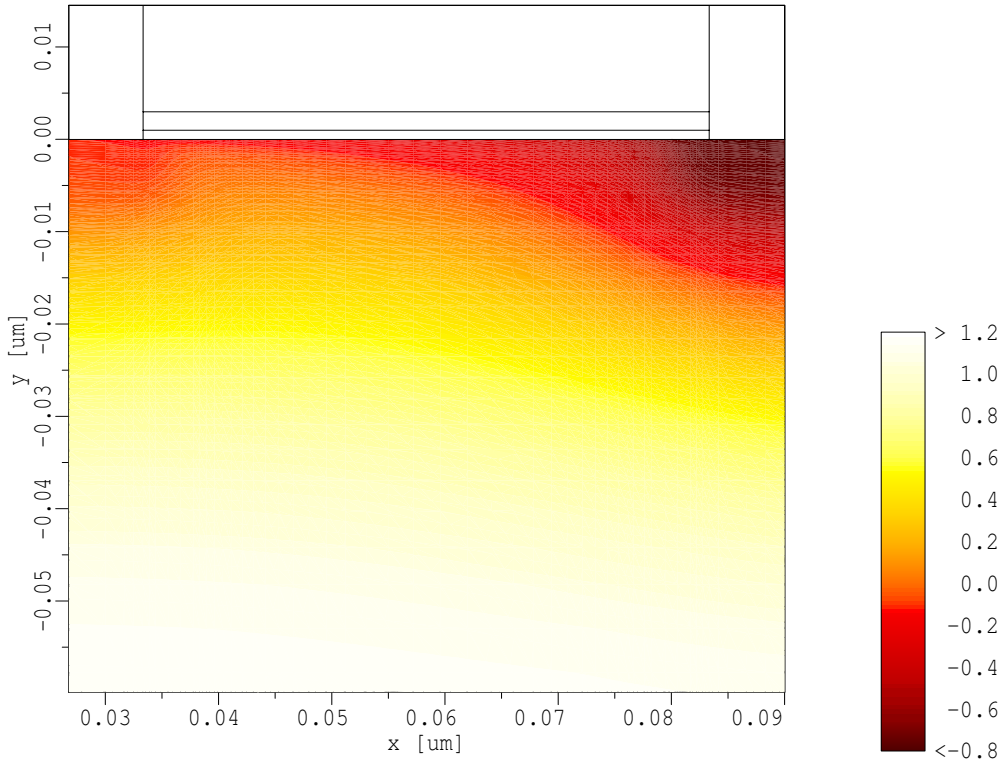

Fig. 5: The band edge energy [eV] of the nMOS device with a stacked gate dielectric evaluated at a gate bias of 1.5 V and a drain voltage of 0.6 V.

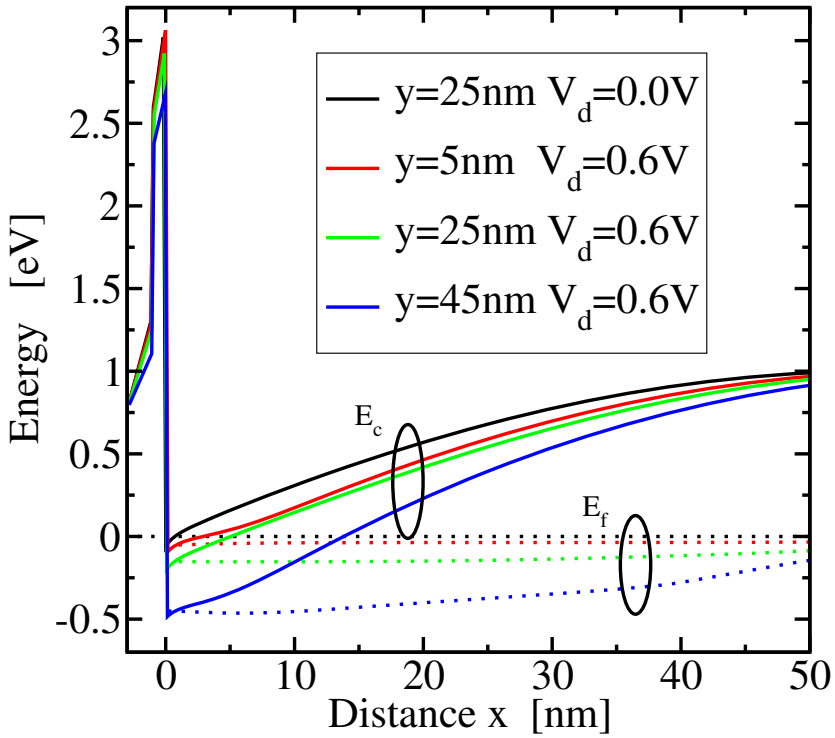

Fig. 6: Cuts of conduction band edge energy of the nMOS transistor. The y-coordinate is relative to the beginning of the gate contact.

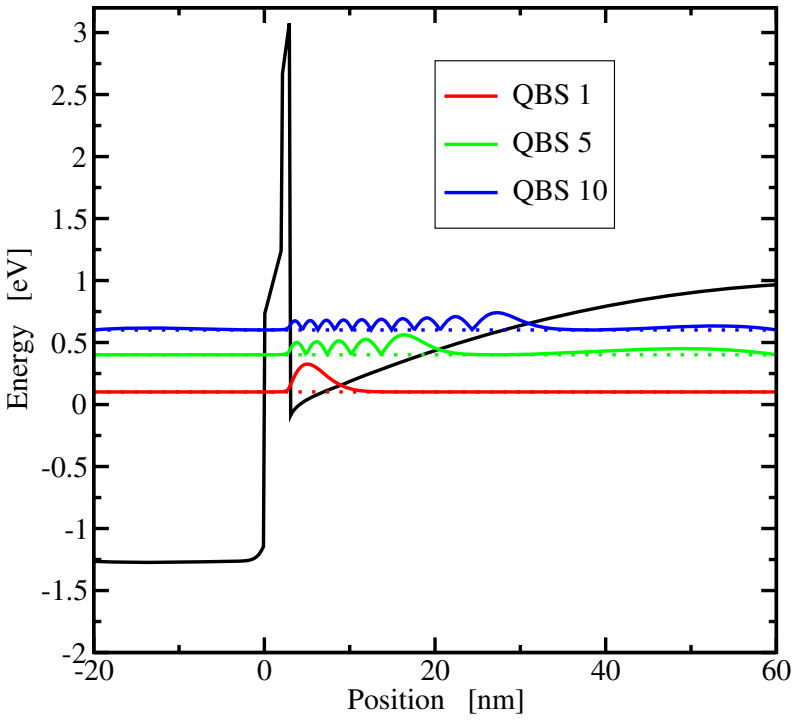

Fig. 7: Potential barrier and eigenstates assuming open boundary conditions using the PML technique.

Table II: The QBS of the MOS capacitor for a gate bias of 1.2V, the corresponding lifetimes, and their contribution to the total gate current density.

QBS	$\mathcal{E}_\mathrm{r}$ [eV]	τ_l [s]	J_G [A cm^{-2}]
1	0.054	2.1×10^{-4}	3.2×10^{-3}
2	0.210	8.5×10^{-5}	2.0×10^{-5}
3	0.326	3.7×10^{-5}	5.1×10^{-8}
5	0.507	8.5×10^{-6}	1.9×10^{-10}

The resulting IV-characteristics as a function of the gate voltage for zero drain bias of the two structures are compared in Fig. 8. It can be seen that the gate current leakage of the stacked dielectric is considerably smaller. Furthermore, we have to point out that the Tsu-Esaki approach overestimates the gate current leakage under inversion conditions. Thus, the use of the more sophisticated lifetime based approach is mandatory for accurate modeling of direct tunneling through stacked gate dielectrics under inversion conditions.

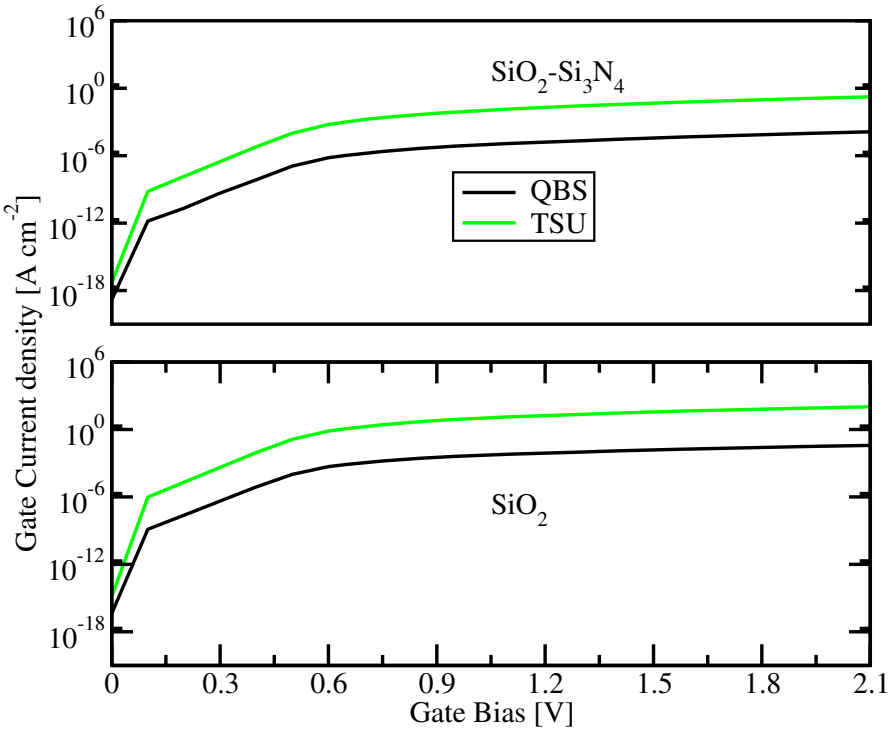

Fig. 8: The gate current density for a single SiO_2 layer as well as for a stacked SiO_2-Si_3N_4 dielectric calculated from the Tsu-Esaki formula and the lifetime based approach.

SUMMARY AND CONCLUSION

We presented an efficient approach for the estimation of lifetime based tunneling currents through stacked gate dielectrics. The lifetimes of quasi bound states (QBS) have been evaluated with the perfectly matched layer (PML) formalism. The traditional approach requires a computationally very demanding scanning procedure. The QBS lifetimes appear as the complex eigenvalues of a non-Hermitian Hamiltonian. Since the equation to be solved is linear, highly efficient algorithms are available. Moreover, the PML approach was used to evaluate QBS in the conduction band on several cuts of the MOS inversion layer and its contribution to the total gate leakage current was determined. For typical device parameters, the QBS tunneling is the dominant tunneling component. The PML formalism represents an efficient and numerically stable method to determine QBS. Therefore, it is appropriate for integration in a device simulator for the investigation of direct tunneling phenomena.

ACKNOWLEDGMENTS

This work has been partly supported by the European Commission, project SINANO, IST 506844 and the Austrian Science Fund, special research program IR-ON (F25).

REFERENCES

1. E. M. Vogel *et al.*, *IEEE Trans.Electron Devices*, **45**, 1350 (1998).
2. J. Zhang *et al.*, *Solid-State Electron.*, **44**, 2165 (2000).
3. J. D. Casperson, L. D. Bell, and H. A. Atwater, *J.Appl.Phys.*, **92**, 261 (2002).
4. 4. M. LeRoy *et al.*, *J.Appl.Phys.*, **93**, 2966 (2003).
5. C. M. Osburn *et al.*, *IBM J.Res.Dev.*, **46**, 299 (2002).
6. G. D. Wilk, R. M. Wallace, and J. M. Anthony, *J.Appl.Phys.*, **89**, 5243 (2001).
7. J. Robertson, *J.Vac.Sci.Technol.*, **18**, 1785 (2000).
8. H.-S. P. Wong, *IBM J.Res.Dev.*, **46**, 133 (2002).
9. R. Tsu and L. Esaki, *Appl.Phys.Lett.*, **22**, 562 (1973).
10. Khairurrijal *et al.*, *J.Appl.Phys.*, **87**, 3000 (2000).
11. A. Gehring and S. Selberherr, *Proc. SISPAD*, 25 (2004).
12. E. Cassan, *J.Appl.Phys.*, **87**, 7931 (2000).
13. A. Dalla Serra *et al.*, *IEEE Trans.Electron Devices*, **48**, 1811 (2001).
14. C. L. Fernando and W. R. Frensley, *J.Appl.Phys.*, **76**, 2881 (1994).
15. R. Clerc *et al.*, *J.Appl.Phys.*, **91**, 1400 (2002).
16. S. Odermatt, M. Luisier, and B. Witzigmann, *J.Appl.Phys.*, **97**, 046104 (2005).
17. M. Karner *et al.*, *Proc. SISPAD*, 35 (2005).
18. Institute for Microelectronics, TU-Wien, *MINIMOS-NT 2.1 User's Guide* (2004).