Efficient calculation of lifetime based direct tunneling through stacked dielectrics

M. Karner · A. Gehring · H. Kosina

Abstract We present the efficient simulation of 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 using 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 an elegant, numerically stable, and efficient method to calculate QBS lifetimes for the investigation of direct tunneling through stacked gate dielectrics.

1. Introduction

The continuous progress in the development of MOS field-effect transistors within the last years goes hand in hand with down-scaling the device feature size. To enable further device down-scaling to the sub-100 nm channel length regime it is necessary to reduce the effective oxide thicknesses (EOT) below 2 nm for MOSFETs resulting 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].

M. Karner (✉) · H. Kosina
Institut für Mikroelektronik, Technische Universität Wien,
Gußhausstraße 27–29, A-1040 Wien, Austria
e-mail: karner@iue.tuwien.ac.at

A. Gehring
AMD Saxony, Wilschdorfer Landstrasse 101, D-01109 Dresden,
Germany

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. The properties [2–8] are summarized in Table 1.

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

2. Calculation of direct tunneling using a lifetime based approach

Calculation of tunneling currents is traditionally 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]

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

where $TC(\mathcal{E}_x, m_{diel})$ is the transmission coefficient and $N(\mathcal{E}_x, m_D)$ the supply function. Two electron masses enter this equation: The density-of-states mass in the plane parallel to the interface $m_D = 2m_t^* + 4\sqrt{m_t^* m_l^*}$, which, for

Table 1 Dielectric permittivity, band gap, and conduction band offset of dielectric materials

	Permittivity κ/κ_0 (1)	Band gap $\mathcal{E}_g$ (eV)	Band offset $\Delta\mathcal{E}_C$ (eV)
SiO$_2$	3.9	8.9–9.0	3.0–3.5
Si$_3$N$_4$	7.0–7.9	5.0–5.3	2.0–2.4
Ta$_2$O$_5$	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$_2$O$_3$	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

(100) silicon with $m_1^* = 0.92m_0$ and $m_t^* = 0.19m_0$ equals $2.052m_0$, and the electron mass in the dielectric m_{diel}, which is commonly used as fit parameter [10].

However, in the inversion layer of MOS devices, 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 more valid.

In the inversion layers of MOS-structures, 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 and τ_i denote the carrier concentration and the lifetime of the QBS with index i, respectively.

To take account for the tunneling current from both, continuum and quasi-bound states, (1) has to be replaced by

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

where the symbols g_ν, $m_\parallel$, and m_q denote the valley degeneracy, parallel, and quantization masses ($g = 2$: $m_\parallel = m_t$, $m_q = m_1$ and $g = 4$: $m_\parallel = \sqrt{m_1 m_t}$, $m_q = m_t$), $\tau_\nu(\mathcal{E}_{\nu,i})$ 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} which follows from the QBS.

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

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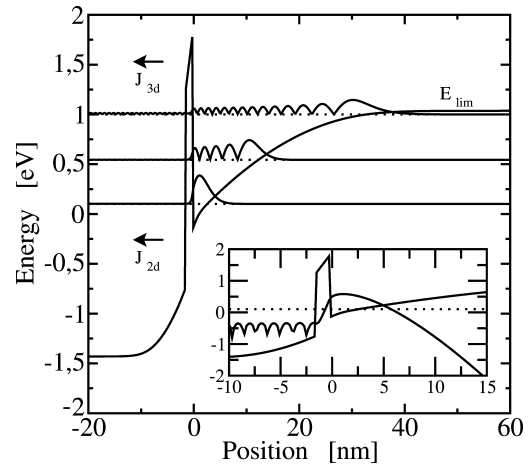$$

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

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. Also, bound states cannot carry any current, since their wavefunctions Ψ fulfill the relation:

$$
\Psi \nabla \Psi^* - \Psi^* \nabla \Psi = 0. \quad (4)
$$

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 cumbersome in the presence of strong confinement (high lifetimes).

3. Perfectly matched layer method

Recently, a method based on absorbing boundary conditions (called 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.

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, linear 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

$$\tilde{x} = \int_0^x s_x(\tau)\, d\tau \qquad (5)$$

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

$$\frac{\partial}{\partial \tilde{x}} = \frac{1}{s_x(x)} \frac{\partial}{\partial x}. \qquad (6)$$

Within the PML region, the stretching function $s_x(x)$ is given as $s_x(x) = 1 + (\alpha + \iota\beta)x^n$, 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 $\mathcal{E} = \mathcal{E}_r + \iota\mathcal{E}_i$, where the QBS lifetimes are related to the imaginary parts of the eigenvalues as

$$\tau_i = \hbar/2\mathcal{E}_i. \qquad (7)$$

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.

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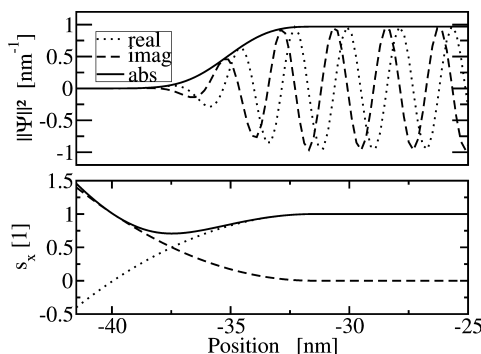
$\Psi(x) = \Psi_0 \exp(\iota \tilde{k}_x x)$ with the wave vector $\tilde{k}_x = k_x/s_x$. Considering two points in the PML region x_1, $x_2 = x_1 + dx$ the wave vector at the point x_2 can be approximated as

$$k_x(x_2) \approx \frac{s_x(x_2)}{s_x(x_1)} k_x(x_1) = (1 + (\alpha + \iota\beta)\, dx). \qquad (8)$$

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 within the absorbing region the envelope of the wave functions decay to zero. This 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 yields 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 (see 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]. A 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. Figure 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
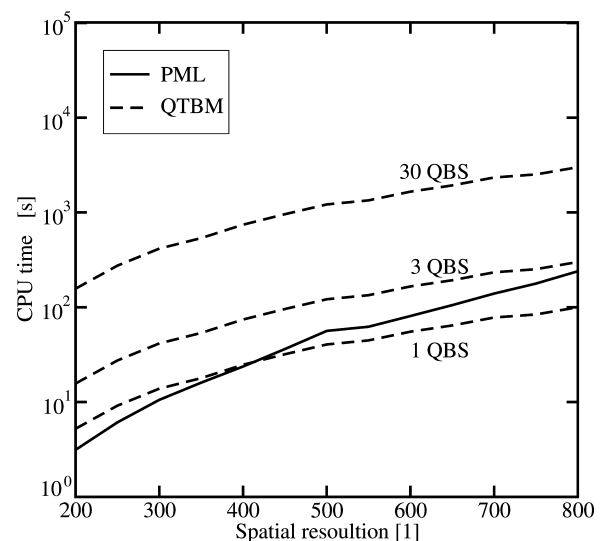

Fig. 2 The wave function of the first QBS and the complex stretching function are displayed in the perfectly matched layer region as well as its transition to the physical region

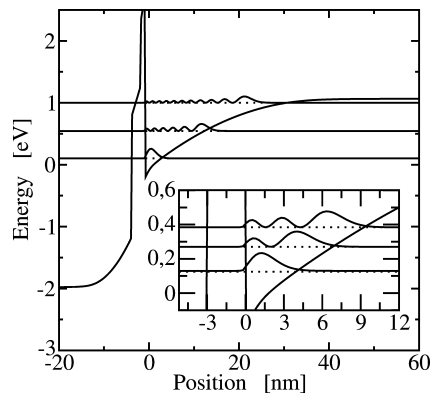**Fig. 3** Comparison of the CPU time demand for the PML, and the QTB methods

Fig. 4 The potential barrier and its eigenstates assuming open boundary conditions using the PML technique. The inset zooms the wavefunction of the first, the second and the third QBS

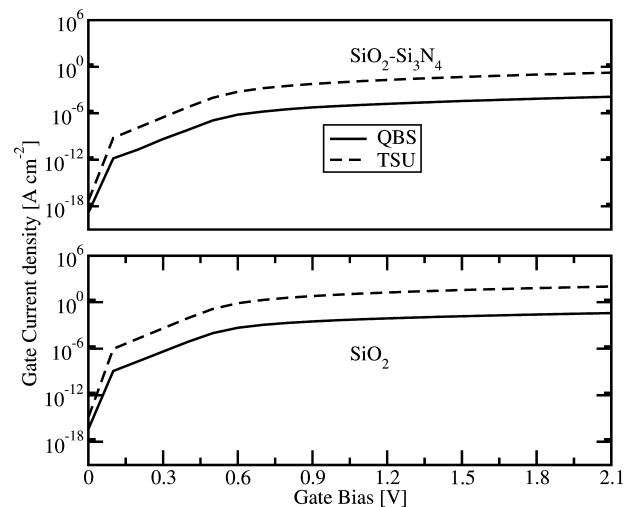**Fig. 5** The gate current density for a single SiO_2 layer as well as for a stacked SiO_2-Si_3N_4 dielectric calculated from the Tsu-Esaki formula and the lifetime based approach

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.

4. Application

With the described methodology, the gate current density has been evaluated for a stacked SiO_2-Si_3N_4 and a single SiO_2 layer gate dielectric. A doping of $N_A = 1 \times 10^{17} cm^{-3}$ in the bulk and $N_D = 1 \times 10^{19} cm^{-3}$ in the poly gate was assumed.

For the investigation of gate leakage currents in MOS transistors the conduction band edge has been acquired from a self-consistent SCHRÖDINGER-POISSON solver. As a post-processing step, the QBS 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. 4 considering the transversal mass as the quantization mass at a bias of 1.5 V. The energy levels, the QBS lifetimes, and their contribution to the total current density are listed in Table 2. To make a comparison, the resulting IV-characteristics of the structures are shown in Fig. 5.

It can be seen that the gate leakage current 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.

5. Summary and conclusion

We presented an efficient lifetime based approach for the simulation of tunneling currents through high-κ materials. The calculation of the QBS lifetimes has been performed using the perfectly matched layer formalism.

In contrast to the traditional QTBM approach which requires a computationally very demanding scanning procedure, the QBS lifetimes appear as the complex eigenvalues of a non-Hermitian, linear Hamiltonian. Since the equations to be solved are linear, highly efficient algorithms are available.

The PML approach was used to evaluate the QBS in MOS inversion layers and to investigate the impact on direct tunneling through the dielectric layer which is, for typical device parameters, the dominant tunneling component. The PML formalism represents an efficient and numerically stable method to determine the QBS which is the key factor for the integration of QBS tunneling in a device simulator for rigorous simulation of gate leakage currents through stacked gate dielectrics.

Table 2 The QBS of the MOS capacitor for a gate bias of 1.5 V, the corresponding lifetimes, the values of the supply function, and their contribution to the gate current density

QBS	$\mathcal{E}_r$ (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}

References

1. Vogel, E.M. et al.: IEEE Trans. Electron Devices **45**, 1350 (1998)
2. Zhang, J. et al.: Solid-State Electron. **44**, 2165 (2000)
3. Casperson, J.D., Bell, L.D., Atwater, H.A.: J. Appl. Phys. **92**, 261 (2002)
4. LeRoy, M. et al.: J. Appl. Phys. **93**, 2966 (2003)
5. Osburn, C.M. et al.: IBM J. Res. Dev. **46**, 299 (2002)
6. Wilk, G.D., Wallace, R.M., Anthony, J.M.: J. Appl. Phys. **89**, 5243 (2001)
7. Robertson, J.: J.Vac.Sci.Technol. **18**, 1785 (2000)
8. Wong, H.-S.P.: IBM J. Res. Dev. **46**, 133 (2002)
9. Tsu R., Esaki, L.: Appl. Phys. Lett. **22**, 562 (1973)
10. Khairurrijal et al. J. Appl. Phys. **87**, 3000 (2000)
11. Gehring A., Selberherr, S.: in Proc. SISPAD (Munchen, 2004), pp. 25–28
12. Cassan, E.: J. Appl. Phys. **87**, 7931 (2000)
13. Dalla Serra, A. et al.: IEEE Trans. Electron Devices **48**, 1811 (2001)
14. Fernando, C.L. Frensley, W.R.: J. Appl. Phys. **76**, 2881 (1994)
15. Clerc, R. et al.: J. Appl. Phys. **91**, 1400 (2002)
16. Odermatt, S., Luisier, M., Witzigmann, J.: Appl. Phys. **97**, 046104 (2005)
17. Karner, M. et al.: in Proc. SISPAD (2005)