The Influence of Localized States on Gate Tunnel Currents – Modeling and Simulation

A. Wettstein, A. Schenk, A. Scholze, and W. Fichtner
Integrated Systems Laboratory, ETH Zürich, Switzerland
Phone: +41 1 632 66 89, FAX: +41 1 632 11 94
E-mail: schenk@iis.ee.ethz.ch

I. Introduction

In the numerical simulation of ultra-small MOSFETs with oxide thicknesses in the range (2...4) nm gate leakage currents have to be modeled on a sound physical base. The main mechanisms apart from oxide non-idealities are direct and resonant tunneling (turning into Fowler-Nordheim tunneling at large biases). The self-consistent simulation of direct tunneling using a fully analytical model was presented in Ref. [1] assuming plane waves both in the gate electrode and the silicon substrate. Here we study the impact of the confinement of carriers in the inversion channel (quasi 2D states) on the size of the direct tunnel current. This will be done based on a Poisson-Schrödinger solver integrated with the device simulator DESSIS-*ISE*, and by applying Bardeen's perturbational method [2].

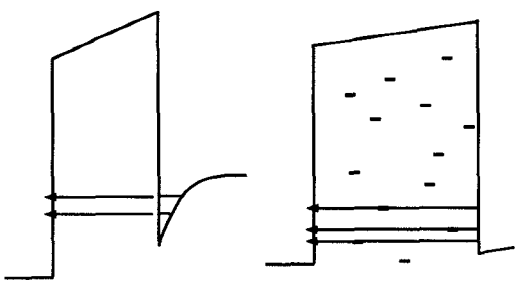

Fig. 1. Direct tunneling out of confined states in the inversion channel (left) and resonant tunneling via oxide trap levels (right).

As experimentally evidenced, direct tunneling cannot account for the strong gate leakage when the oxide thickness becomes larger than 3 nm. A straightforward explanation for these currents is in terms of resonant tunneling via quasi 0D states induced by oxide traps (see Fig. 1) which starts to dominate over direct tunneling as soon as the tunnel length exceeds 3 nm. An analytical model of zero-phonon resonant tunneling via oxide traps [3] was implemented into DESSIS-*ISE*, and simulations for various configurations of resonance levels were performed.

II. Direct Tunneling

To obtain the tunnel current through the gate oxide, we compute the eigenfunctions ψ_i^ν and eigenvalues E_i^ν of the 1D Schrödinger equation

$$\left(-\frac{\partial}{\partial z}\frac{\hbar^2}{2m_z^\nu(z)}\frac{\partial}{\partial z}+V(z)\right)\psi_i^\nu = E_i^\nu\psi_i^\nu \qquad (1)$$

for electrons localized perpendicular to the interface in a MOSFET channel. Here, i labels the different eigenvalues, ν labels the conduction band valleys of silicon, and $m_z^\nu(z)$ is the effective mass component in quantization direction z. As little is known about the dispersion relation in the SiO_2 gap, using (1) with some value $m_z^\nu(z) = m_{ox}^\nu$ in the oxide must be considered as a fit.

The Schrödinger equation is solved within a "quantum box" that extends from the gate/oxide interface to some point sufficiently deep in the silicon bulk (e. g. in 50 nm distance from the Si-SiO_2 interface). At the end points of the quantum box we assume boundary conditions of the form $|\psi_i^{\nu\prime}/\psi_i^\nu| = |k_z(E_i^\nu)|$. Here, $k_z(E_i^\nu)$ is the local wave number computed from the energy, the effective mass, and the potential at the boundaries. For end points at which the wave function attenuates, these boundary conditions correspond to an infinitely extending constant potential outside the box. The boundary conditions are arbitrary at end points where the wave function oscillates, however, they produce less artefacts in the total electron density than e. g. zero boundary conditions.

Eq. (1) is solved by guessing an eigenvalue E_i^ν and inserting it into the Schrödinger equation. The resulting ordinary differential equation is solved for the left and the right part of the quantum box using the CPM(1) method [4]. If the partial solutions can be matched at a properly chosen intermediate point, the guessed value was really an eigenvalue; otherwise, the mismatch allows to compute a better estimate [5].

The decay rates into unbound states in the gate can be determined by perturbation theory. The perturbation is given by the difference of the real potential – which allows free motion on the gate side of the oxide – and the potential assumed to compute the localized eigenstates. The calculation then goes straightforward as demonstrated by Bardeen [2]. By integrating over degrees of freedom perpendicular to the quantization direction and summing over all eigenstates, one obtains for the direct tunnel cur-

0-7803-3775-1/97/$10.00

rent from/into quasi 2D states in the MOS channel

$$j_{dt} = \frac{q\sqrt{2m_g}L}{8\pi m_{ox}^2} \sum_{i,\nu} m_{xy}^{\nu} \int_0^{\infty} dE \, \frac{\Theta(\tilde{E} - E_{c,g})}{\sqrt{\tilde{E} - E_{c,g}}} \times$$
$$\times \, |\psi_{\tilde{E}} \partial_z \psi_i^{\nu} - \psi_i^{\nu} \partial_z \psi_{\tilde{E}}|^2_{z=z_0} \, \Delta f(E_i^{\nu} + E) \, , \quad (2)$$

with $\tilde{E} = E_i^{\nu} + (1 - m_{xy}^{\nu}/m_g)E$. Δf is the difference of the Fermi functions in the channel and the gate. L is the thickness of the gate, which is assumed large enough to neglect quantization. L cancels with the normalization constant of the wave functions $\psi_{\tilde{E}}$ in the gate. $E_{c,g}$ labels the conduction band edge in the gate. z is the direction perpendicular to the Si-SiO_2 interface and z_0 is the location of the gate-SiO_2 interface. m_{ox} and m_g are effective masses in the oxide and the gate; m_z and $m_{xy}^{\nu} = m_t\sqrt{m_l/m_z^{\nu}}$ are the silicon effective masses in z- and xy-direction, respectively. m_l and m_t are the longitudinal and transverse mass components of the electrons in silicon. E_i^{ν} and ψ_i^{ν} are determined by the numerical solution of the 1D Schrödinger equation. In order to calculate the correct tunnel current from (1) and (2), a sufficient number of eigensolutions has to be considered. For strong negative gate biases this number can be quite large (up to about 300 for the curves presented in Fig. 2), because most of the electrons injected from the gate have a high energy.

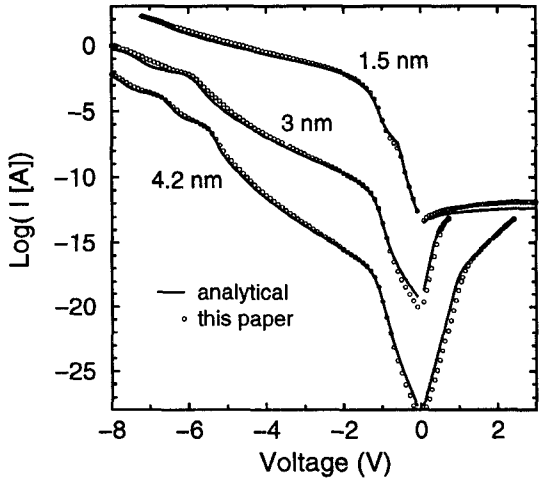

Fig. 2. Simulated direct tunnel currents for various MOS capacitors (p-Si, ⟨100⟩, $m_{ox} = 0.42\,m_0$).

Fig. 2 shows IV-characteristics of MOS capacitors with different oxide thicknesses obtained with the analytical transmission coefficient of Ref. [1] and the full quantum-mechanical treatment, respectively. Note that the two sets of curves result from completely independent implementations, but using equal effective masses. The close agreement for negative biases $< -1V$ is not surprising because of the absence of confinement.

But even in strong inversion (p-Si, $N_A = 10^{18}\,cm^{-3}$) the effect of quantization is rather small. In the case of ultra-thin oxides the reverse current is limited by thermal generation of electrons in the Si depletion region,

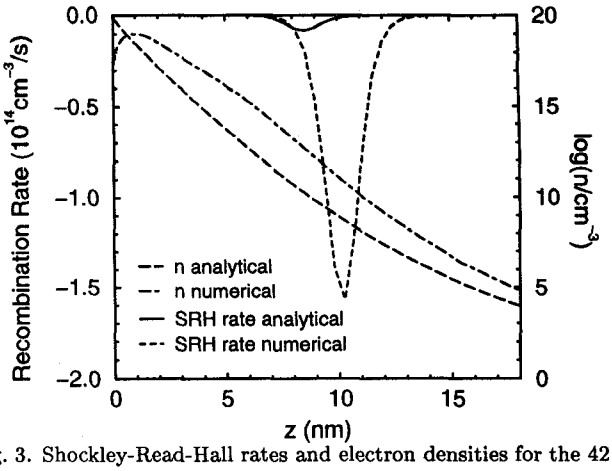

Fig. 3. Shockley-Read-Hall rates and electron densities for the 42 Å device at $V_G = 1.5\,V$.

i.e. the current is proportional to n_i/τ (τ - minority carrier lifetime). In the region of quantum confinement the intrinsic density n_i is roughly reduced by a factor $\Delta n = \exp[-(E_1 - E_c)/k_BT]$, where E_1 is the energy level of the bottom of the lowest subband. However, since not only the total charge, but also its distribution is changed, n_i in the region of maximum Shockley-Read-Hall generation is larger than in the classical case, and therefore, the full quantum-mechanical treatment yields a slightly larger current as compared to the analytical case. Adjusting the lifetime can easily absorb this difference.

For the 42 Å oxide, the tunnel resistance of the barrier dominates the current, which is reduced now as result of the confinement by about one order of magnitude at the most. This is caused by the interplay of two opposite effects: the increase of the tunnel probability due to the split-off of the lowest subband at the Si-SiO_2 interface by about $E_1 - E_c = \hbar\Theta_z(9\pi/8)^{2/3}$ with $\hbar\Theta_z = (q^2F^2\hbar^2/2m_z)^{1/3}$ (exact for triangular potential), and a corresponding decrease of the occupation probability by about a factor Δn. The knee is caused by the onset of strong inversion. Here both curves approach as the Fermi level approaches E_1. In the bias range where the difference is most pronounced, resonant tunneling will outnumber direct tunneling. Hence, we conclude that the influence of the quasi 2D states on direct tunneling is negligible in all cases.

III. Resonant Tunneling

The resonant tunnel current via a single trap level is evaluated by [3]

$$j_{res}(\Phi_t) = \frac{8qk_BT}{\pi\hbar L_t} \int_{ox} dz \, k_g k_{Si} \sqrt{\left|1 - \frac{E_g}{\Phi_g}\right| \left|1 - \frac{E_{Si}}{\Phi_{Si}}\right|} \times$$
$$\times \, \frac{\Phi_t}{\sqrt{\Phi_g \Phi_{Si}}} \, \frac{\mathcal{T}_l(z)\mathcal{T}_r(z)}{\sqrt{\mathcal{T}_l^2(z) + \mathcal{T}_r^2(z)}} \ln\left(\frac{1 + e^{\frac{E_{F,g} - E_t(z)}{k_BT}}}{1 + e^{-\frac{E_{F,Si} - E_t(z)}{k_BT}}}\right) \quad (3)$$

where $E_t(z) = E_{c,g} + \Phi_g - qFz - \Phi_t$ is the oxide trap level. The WKB transmission coefficients $\mathcal{T}_{l,r}(z)$ of the partial barriers (separated by the trap at position z_0) are given by

$$\mathcal{T}_l(z) = \exp\left\{\frac{4}{3}\left(\frac{\Phi_t}{\hbar\Theta_{ox}}\right)^{3/2} - \frac{4}{3}\left(\frac{\Phi_t + qFz}{\hbar\Theta_{ox}}\right)_+^{3/2}\right\},$$
$$\mathcal{T}_r(z) = \exp\left\{\frac{4}{3}\left[\frac{\Phi_t + qF(z-d)}{\hbar\Theta_{ox}}\right]_+^{3/2} - \frac{4}{3}\left(\frac{\Phi_t}{\hbar\Theta_{ox}}\right)^{3/2}\right\}, \quad (4)$$

with $x_+ \equiv x\,\Theta(x)$. $k_{g,Si}$ are the momenta and $E_{g,Si}$ the kinetic energies in gate and silicon for the tunneling electrons, respectively. $\Phi_{g,Si}$ are barrier heights measured from the corresponding conduction band edges and Φ_t is the energy of the resonance level measured from the oxide conduction band edge. A homogeneous trap density N_t in z-direction has been assumed parameterized by the length $L_t = (\pi N_t r_t^2)^{-1}$. r_t is the localization radius of the trap. L_t can be interpreted as the thickness where the total cross section of all traps would equal the oxide area. Finally, $E_{F,g}$ and $E_{F,Si}$ denote the Fermi levels in gate and silicon, respectively.

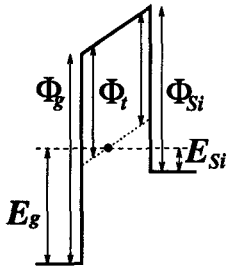

Fig. 4. Symbols used in equations (3) and (4)

Fig. 5 shows the total resonant tunneling current for three discrete oxide trap levels, which yield three local maxima in each branch. These maxima occur at voltages where $E_t(z^*) = E_{c,\alpha}$ ($\alpha = g$ for $V < 0$, $\alpha = Si$ for $V > 0$) with z^* given by the condition of maximum resonant tunnel current: $\mathcal{T}_l(z^*) = \mathcal{T}_r(z^*)$. The oscillatory behavior (known from the resonant tunnel diode) is less distinct for aluminium gates than for poly gates because of the large Fermi energy in the metal. In Fig. 6 a continuous ladder of trap levels was assumed ($10\,meV$ spacing). A reasonable fit to experimental data of Ref. [6] can be achieved using $L_t = 3\,m$ which corresponds to a total of 10 defects per resonance level in the whole oxide (volume $= 6.3 \times 10^{-11}\,\mathrm{cm}^3$).

IV. Summary

The direct tunnel current through gate oxides has been modeled by an analytical model and by the numerical solution of the Poisson-Schrödinger quantum-mechanical problem. Despite the well-know reduction of the channel charge density due to the quantization in the latter model, the tunnel currents differ only insignificantly between the two models. This has been attributed to the compensating effects of decreased occupation probability and increased transmission coefficient for higher energy states.

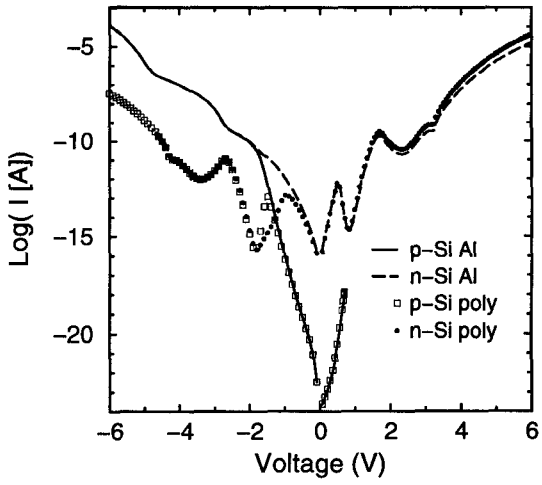

Fig. 5. Simulated resonant tunnel currents for an MOS capacitor assuming three oxide trap levels with $\Phi_t = 1.9\,\mathrm{eV}$, $2.4\,\mathrm{eV}$, and $2.9\,\mathrm{eV}$, relative weights 1, 10, and 1, and $L_t = 10\,\mathrm{cm}$ for weight 1 (oxide thickness 4.2 nm, p-Si, ⟨100⟩, $m_{ox} = 0.42\,m_0$).

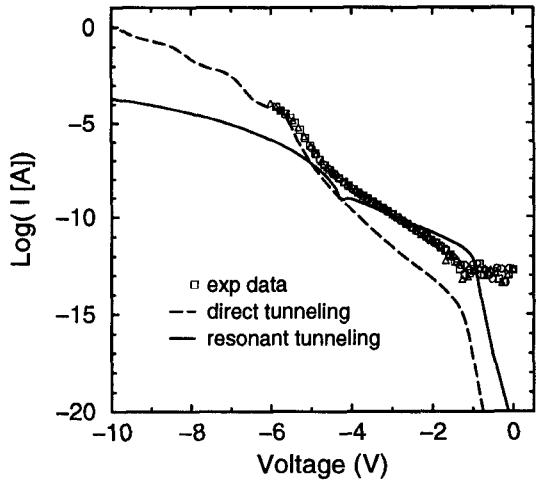

Fig. 6. Simulated resonant (solid) and direct (dashed) tunnel currents for an MOS capacitor with an oxide thickness of 4.2 nm assuming equidistant trap levels separated by 0.01 eV and equal weight $L_t = 3\,\mathrm{m}$.

Resonant tunneling can explain the measured gate leakage currents for oxide thicknesses larger than 3 nm and low biases. Assuming a dense distribution of resonance levels induced by oxide defects, the reasonable value of $1.6 \times 10^{11}\,\mathrm{cm}^{-3}$ for the spatial defect density turned out to yield good agreement between simulated and measured tunnel currents.

References

[1] A. Schenk. Modeling Tunneling through Ultra-Thin Gate Oxides. In *SISPAD'96*, pages 7–8, Tokyo, 1996.

[2] J. Bardeen. Tunneling from a Many-Particle Point of View. *Phys. Rev. Lett.*, 6(2):57–62, 1961.

[3] A. Schenk and M. Herrmann. A New Model for the Long Term Charge Loss in EPROMs. In *Ext. Abstracts Solid State Devices and Materials SSDM*, pages 494–96, Yokohama, Japan, 1994.

[4] L. GR. Ixaru. *Numerical Methods for Differential Equations and Applications*. Reidel, Dordrecht/Boston/Lancaster, 1984.

[5] John M. Blatt. Practical points concerning the solution of the Schrödinger equation. *J. Comp. Phys.*, 1:382–396, 1967.

[6] H. S. Momose, M. Ono, T. Yoshitomi, T. Ohguro, S. Nakamura, M. Saito, and H. Iwai. 1.5 nm Direct-Tunneling Gate Oxide Si MOSFET's. *IEEE Trans. Electron Devices*, 43(8):1233–41, 1996.