# Modeling of Tunneling Currents for Highly Degraded CMOS Devices

R. Entner*, A. Gehring†, H. Kosina†, T. Grasser*, and S. Selberherr†
*Christian Doppler Laboratory for TCAD in Microelectronics at the Institute for Microelectronics
Phone: +43-1-58801/36050, Fax: +43-1-58801/36099, E-mail: entner@iue.tuwien.ac.at
†Institute for Microelectronics, TU Wien, Gußhausstraße 27–29/E360, 1040 Wien, Austria

***Abstract*— We present a model for tunneling currents in highly degraded CMOS devices. In this field not only well established tunneling mechanisms like Fowler-Nordheim and direct tunneling are important to consider, but also defect-assisted tunneling mechanisms such as elastic and inelastic trap-assisted tunneling and hopping processes between defects. In our work the interaction of several defects in the tunneling process is taken into account. The multi-trap assisted tunneling current is dominant for heavily degraded devices with dielectric thicknesses above approximately 3-4 nm. The filling of traps with carriers leads to space charge and is thus changing device parameters such as threshold-voltage or saturation currents.**

## I. INTRODUCTION

State-of-the-art CMOS devices are subject to steadily growing stress conditions. As the reduction of the applied voltages does not keep up with the miniaturization of actual devices the electric fields across dielectric layers are constantly increasing. Especially for non-volatile memory cells high electric fields are necessary in order to achieve quick write and erase cycles. Due to the repeated high-field stress, defects can arise in the dielectric leading to tunneling currents even at low fields. This stress-induced leakage current (SILC) [1], [2] plays a major role in the determination of the retention times of non-volatile memory cells.

There are many approaches to model trap-assisted tunneling (TAT). One approach is to model the defect assisted tunneling process including a single trap. For each possible trap position tunneling from the cathode to the trap and further to the anode is considered, yielding a trap occupancy function which is used to calculate the tunneling current [3]. This single-TAT approach works very well for slightly degraded devices or devices with thin gate dielectrics. For thicker dielectrics with a high defect density it is reasonable to assume that also the interaction of two or more traps in the tunneling process takes place [4].

For the modeling of multi-trap assisted tunneling (multi-TAT) approaches like a two-trap process [5], [6] or a multi-trap process considering hopping of carriers between distinct defects [7] have been presented. Recently, anomalous charge loss in floating-gate memory cells has been reported [4], where a two-trap model was used to reproduce the measured data.

For correct modeling of such highly degraded devices an approach is presented which rigorously computes TAT current assisted by multiple traps [8]. In this model hopping processes between all oxide defects are taken into account. Also the filling of oxide traps with carriers, leading to space charge in the oxide and therefore to a shift of the threshold voltage, is accounted for.

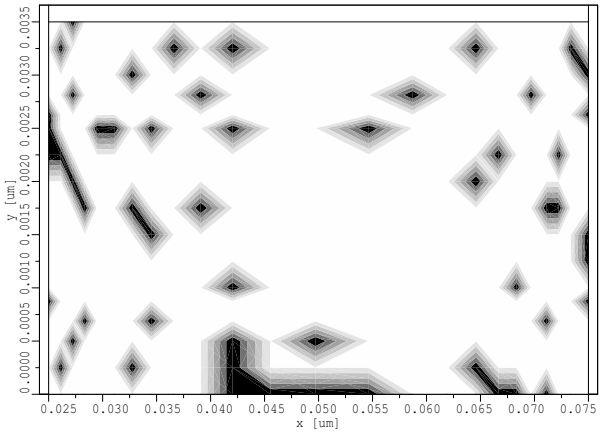


Figure 1: The formation of traps randomly distributed across the oxide of a MOS structure.

## II. MULTI-TRAP ASSISTED TUNNELING MODEL

The model for the simulation of SILC in highly degraded devices is based on inelastic, phonon-assisted tunneling [9] with the tunneling-rate proposed by Herrmann and Schenk [3]. These approaches are extended for modeling the interaction of multiple traps in the tunneling process.

### A. Inelastic Phonon-Assisted Tunneling

The defect-assisted tunneling process of an electron from the cathode to the anode via a trap is considered as two-step process. Electrons are captured from the cathode, relax to the energy level of the trap by emitting one or more phonons with the energy $\hbar\omega$, and are then emitted to the anode. This process is inelastic as the electron energy is not conserved during the tunneling process. Fig. 2 depicts this process including the phonon emission.

### B. Single-Trap Assisted Tunneling

When only one trap is considered in the tunneling process the tunneling current density can be modeled as the sum of capture and emission rates $R_i$ in each trap multiplied by the

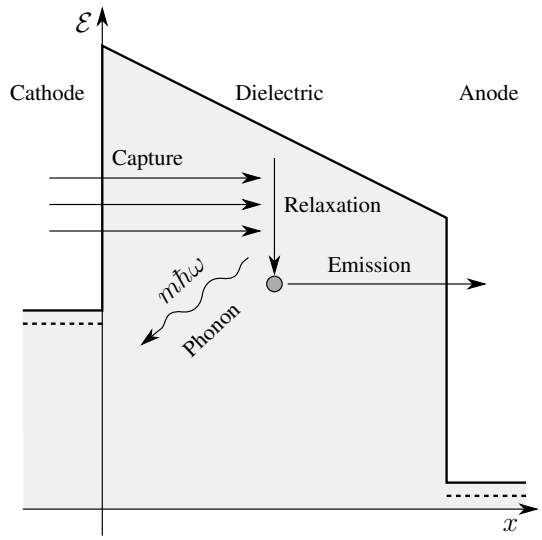


Figure 2: Inelastic tunneling process including a sole trap. The excess energy of the tunneling electron is released by means of phonon emission.

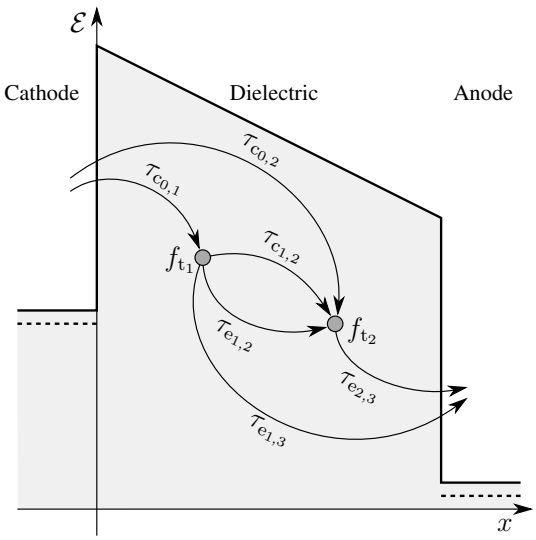


Figure 3: Multi-trap assisted tunneling process. The tunneling rate $R_i$ of a specific trap is determined by all capture and emission times to and from the trap.

trap cross section $\Delta x_i$,

$$J = q \sum_i R_i \Delta x_i \ . \tag{1}$$

The energetic position of the trap $\mathcal{E}_\mathrm{T}$ with respect to the conduction band edge determines the trap cross section [10]

$$\Delta x_i = \frac{\hbar}{\sqrt{2 m_\mathrm{diel} \mathcal{E}_\mathrm{T}}} \left(\frac{4\pi}{3}\right)^{1/3} \ , \tag{2}$$

where $m_\mathrm{diel}$ denotes the electron mass in the dielectric, which is used as a fitting parameter.

The single-TAT and the multi-TAT models differ in the way $R_i$ is calculated. When only single-trap processes are considered (see Fig. 2) the rates are determined by [3]

$$R_{\mathrm{c}_i} = \tau_{\mathrm{c}_i}^{-1} N_{\mathrm{t}_i} (1 - f_{\mathrm{t}_i}) \ , \ R_{\mathrm{e}_i} = \tau_{\mathrm{e}_i}^{-1} N_{\mathrm{t}_i} f_{\mathrm{t}_i} \, . \tag{3}$$

Here, $R_{\mathrm{c}_i}$ and $R_{\mathrm{e}_i}$ are the capture and emission rates of the considered trap, respectively, and $N_{\mathrm{t}_i}$ denotes the trap concentration. In the stationary case the capture and emission rates must be equal $R_{\mathrm{c}_i} = R_{\mathrm{e}_i} = R_i$. The trap occupancy $f_{\mathrm{t}_i}$ can be directly calculated as $f_{\mathrm{t}_i} = \tau_{\mathrm{c}_i}^{-1} / (\tau_{\mathrm{c}_i}^{-1} + \tau_{\mathrm{e}_i}^{-1})$ where the inverse capture and emission times can be evaluated as [3], [11]

$$\tau_{\mathrm{c}_i}^{-1} = \int_{\mathcal{E}_0}^{\infty} g_\mathrm{C}(\mathcal{E}) c_n(\mathcal{E}) \mathrm{T}_\mathrm{C}(\mathcal{E}) f_\mathrm{C}(\mathcal{E}) \, \mathrm{d}\mathcal{E} \ , \tag{4}$$

$$\tau_{\mathrm{e}_i}^{-1} = \int_{\mathcal{E}_0}^{\infty} g_\mathrm{A}(\mathcal{E}) e_n(\mathcal{E}) \mathrm{T}_\mathrm{A}(\mathcal{E}) (1 - f_\mathrm{A}(\mathcal{E})) \, \mathrm{d}\mathcal{E} \ . \tag{5}$$

In these expressions, $g_\mathrm{C}(\mathcal{E})$ and $g_\mathrm{A}(\mathcal{E})$ denote the density of states in the cathode and anode, respectively, and the symbols $c_n$ and $e_n$ are computed as

$$c_n(\mathcal{E}) = c_0 \sum_m L_m \delta(\mathcal{E} - \mathcal{E}_m) \, , \tag{6}$$

$$e_n(\mathcal{E}) = c_0 \exp\left(-\frac{\mathcal{E} - \mathcal{E}_\mathrm{T}}{\mathrm{k_B} T_\mathrm{L}}\right) \sum_m L_m \delta(\mathcal{E} - \mathcal{E}_m) \, , \tag{7}$$

with

$$c_0 = (4\pi)^2 \Delta x_i^2 (\hbar\Theta_0)^3 / (\hbar \mathcal{E}_{\mathrm{g,SiO2}}) \ , \tag{8}$$

$$(\hbar\Theta_0) \ = (\mathrm{q}^2 \hbar^2 F^2 / (2 \ m_\mathrm{diel}))^{1/3} \ . \tag{9}$$

The summation index $m$ denotes the number of discrete phonon emissions, $\mathcal{E}_m$ is the phonon energy, and $L_m$ is the multiphonon transition probability [3]. The symbols $f_\mathrm{C}$ and $f_\mathrm{A}$ are the Fermi distributions, $\mathrm{T_C}$ and $\mathrm{T_A}$ the transmission coefficients from the cathode and the anode, $F$ the electric field in the dielectric, and $\mathcal{E}_{\mathrm{g,SiO2}}$ the band gap of $SiO_2$. The transmission coefficients were evaluated by a numerical WKB method which yields reasonable accuracy for single-layer dielectrics. This model has been used in a more or less similar form by various authors [12], [9], [11].

### C. Considering Multiple Traps

Recently, however, anomalous charge loss in memory cells has been observed and was explained by conduction through a second trap [5]. The single-trap model can be extended for this case, and the rate equations become (see Fig. 3)

$$\begin{aligned} &\overbrace{\tau_{\mathrm{c}_{0,1}}^{-1} N_{\mathrm{t}_1} (1 - f_{\mathrm{t}_1})}^{R_{\mathrm{c}_1}} - \overbrace{(\tau_{\mathrm{e}_{1,2}}^{-1} N_{\mathrm{t}_1} f_{\mathrm{t}_1} (1 - f_{\mathrm{t}_2}) + \tau_{\mathrm{e}_{1,3}}^{-1} N_{\mathrm{t}_1} f_{\mathrm{t}_1})}^{R_{\mathrm{e}_1}} = 0 \ , \\ &\underbrace{\tau_{\mathrm{c}_{0,2}}^{-1} N_{\mathrm{t}_2} (1 - f_{\mathrm{t}_2}) + \tau_{\mathrm{c}_{1,2}}^{-1} N_{\mathrm{t}_2} f_{\mathrm{t}_1} (1 - f_{\mathrm{t}_2})}_{R_{\mathrm{c}_2}} - \underbrace{\tau_{\mathrm{e}_{2,3}}^{-1} N_{\mathrm{t}_2} f_{\mathrm{t}_2}}_{R_{\mathrm{e}_2}} = 0 \ , \end{aligned}$$

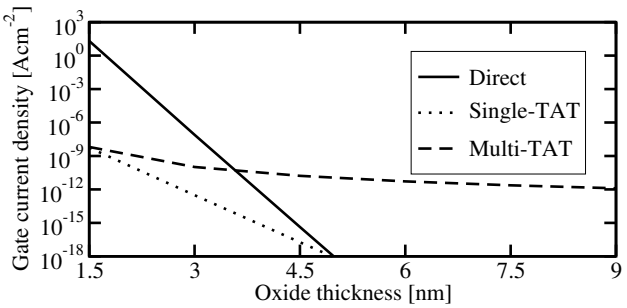


Figure 4: SILC simulations for a set of MOS devices at 1 V gate bias. The oxide has a constant trap concentration.

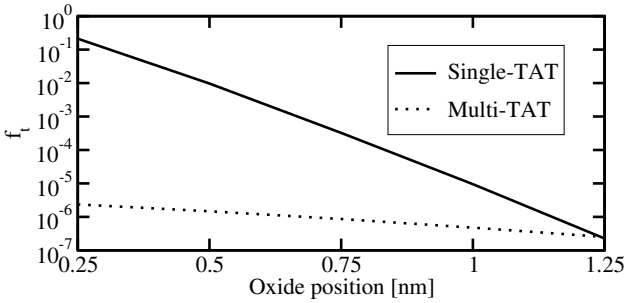


Figure 5: The trap occupancy $f_{\mathrm{t}}$ in the oxide of a 1.5 nm MOS transistor at 1 V gate bias.

where instantaneous transitions between occupied and free traps are assumed. For thicker dielectrics it is quite reasonable to assume that an arbitrary number of traps assist in the conduction process. We therefore extend the model to $n$ traps where the capture and emission rates are evaluated as

$$R_{\mathrm{c}_k} = \sum_{i=0}^{k-1} \tau_{\mathrm{c}_{i,k}}^{-1} N_{\mathrm{t}_k} f_{\mathrm{t}_i} (1 - f_{\mathrm{t}_k}) \,, \tag{10}$$

$$R_{\mathrm{e}_k} = \sum_{i=k+1}^{n+1} \tau_{\mathrm{e}_{k,i}}^{-1} N_{\mathrm{t}_k} f_{\mathrm{t}_k} (1 - f_{\mathrm{t}_i}) \,. \tag{11}$$

The values for $f_{\mathrm{t}_0}$ and $f_{\mathrm{t}_{n+1}}$, which are the trap occupation probabilities at the cathode and the anode, are set to 1 and 0 respectively. This way the cathode acts as electron source and the anode as electron sink.

From the capture and emission rates the following equation system can be set up

$$\sum_{i=0}^{k-1} \tau_{\mathrm{c}_{i,k}}^{-1} N_{\mathrm{t}_k} f_{\mathrm{t}_i} (1 - f_{\mathrm{t}_k}) = \sum_{i=k+1}^{n+1} \tau_{\mathrm{e}_{k,i}}^{-1} N_{\mathrm{t}_k} f_{\mathrm{t}_k} (1 - f_{\mathrm{t}_i}) \qquad k = 1, 2, \ldots n$$

from which the values of all trap occupation probabilities have to be calculated. This is performed within MINIMOS-NT using a Newton method with the cost function $F_k$ for a trap at position $k$

$$F_k(\mathbf{f}) = \sum_{i=0}^{k-1} \tau_{\mathrm{c}_{i,k}}^{-1} f_{\mathrm{t}_i} (1 - f_{\mathrm{t}_k}) - \sum_{j=k+1}^{n} \tau_{\mathrm{e}_{k,j}}^{-1} f_{\mathrm{t}_k} (1 - f_{\mathrm{t}_j}) = 0 \qquad 1 \le k \le n \,,$$

and the values of the derivatives in the Jacobian matrix

$$\frac{\partial F_i}{\partial f_{\mathrm{t}_j}} = \begin{cases} \tau_{\mathrm{e}_{i,j}}^{-1} f_{\mathrm{t}_i} & \text{for } i < j, \\ -\sum_{k=0}^{i-1} \tau_{\mathrm{c}_{k,i}}^{-1} f_{\mathrm{t}_k} - \sum_{k=i+1}^{n} \tau_{\mathrm{e}_{i,k}}^{-1} (1 - f_{\mathrm{t}_k}) & \text{for } i = j, \\ \tau_{\mathrm{c}_{j,i}}^{-1} (1 - f_{\mathrm{t}_i}) & \text{for } i > j. \end{cases}$$

A typical number of unknowns of the equation system is 15, depending on the dielectric thickness and trap energy. The computational effort remains negligible compared to the total device simulation time. The multi-trap assisted tunneling current density can then be obtained from the capture or emission rate

$$J = q \sum_i \left[ \sum_{j=0}^{i-1} \tau_{\mathrm{c}_{j,i}}^{-1} N_{\mathrm{t}_i} f_{\mathrm{t}_j} (1 - f_{\mathrm{t}_i}) \right] \Delta x_i \,. \tag{12}$$

## III. APPLICATION

The implementation of these models into the device- and circuit-simulator MINIMOS-NT [13] allows the two- and three-dimensional study of single- and multi-trap assisted tunneling as well as direct tunneling mechanisms. Fig. 1 shows the gate oxide with randomly distributed traps generated by a transient stress simulation. Such results can be used for the investigation of trap-assisted tunneling models.

### *A. Tunneling Current*

Fig. 4 shows a comparison of SILC simulation results assuming three different tunneling mechanisms, namely direct tunneling, single-TAT, and multi-TAT. The models have been applied to a set of MOS transistors with gate dielectric thicknesses ranging from 1.5 nm up to 9 nm. The gate is biased at 1 V, source and drain are kept at 0 V. For both, the single-TAT and the multi-TAT simulations, the trap energy is set to 2.8 eV below the dielectric conduction band with a constant trap density of $9 \times 10^{17}\,\mathrm{cm}^{-3}$ across the oxide. In the multi-trap simulation the tunneling current is several orders of magnitude higher than in the single-trap simulation. This is due to the fact that the multi-TAT current includes the single-TAT component as a limiting case. The multi-TAT model considers the capture and emission processes from the cathode and to the anode, respectively, and also the capture and emission processes involving all other trap centers. This leads to the comparably high multi-TAT component in devices with thicker oxides. It has to be considered, though, that this high current is mainly due to the assumption of uniformly distributed trap concentrations across the oxide. The direct tunneling component loses importance for thicker dielectrics

but dominates for thin dielectrics as found in logic CMOS devices. For miniaturized devices with thicker oxides and higher trap densities multi-TAT processes become increasingly important.

### *B. Trap Occupancy*

Fig. 5 depicts the resulting trap occupancy within the oxide. A MOS transistor with 1 V gate bias was simulated. It can be seen that the trap occupancy $f_\mathrm{t}$ is remarkably lower in the multi-TAT case. The reason is the higher probability for electrons to tunnel to one of the neighbor traps compared to tunneling to the anode as it is the only possibility in the single-TAT model.

### *C. Threshold Voltage Shift*

The implementation of this model into the device simulator MINIMOS-NT allows the simulation of the effect of charged defects on the threshold voltage of memory devices. The space charge density in the dielectric is calculated as

$$\rho(x) = \mathrm{q} f_\mathrm{t}(x) N_\mathrm{t}(x) \tag{13}$$

and added to the Poisson's equation

$$\nabla\varepsilon\nabla\psi = -(\rho_0 + \rho_\mathrm{TAT})\,. \tag{14}$$

Fig. 6 outlines the threshold voltage $V_\mathrm{t}$ for different oxide thicknesses. The direct tunneling model, applying the commonly used Tsu-Esaki approach, does not account for the filling of traps in the oxide. Therefore the threshold voltage is not shifted compared to the simulation without a tunneling model. The new multi-TAT model predicts an increase in $V_\mathrm{t}$. This higher threshold voltage is due to the filled and therefore negatively charged traps.

## IV. CONCLUSION

We presented a new trap-assisted tunneling model for the simulation of CMOS devices with highly degraded dielectrics. The model considers inelastic tunneling processes between several defects in the gate dielectric. The model has been applied to a variety of devices with different oxide thicknesses. It has been compared to single-trap and direct tunneling processes. We show that for thicker dielectrics with high defect density the inclusion of multiple traps is crucial for the simulation of both, quantum mechanical gate currents and the shift of threshold voltages. Thin dielectrics, on the other hand side may be as well treated using the single-TAT model.


## ACKNOWLEDGEMENT

This work has been partly supported by the European Commission, project SINANO, IST 506844.


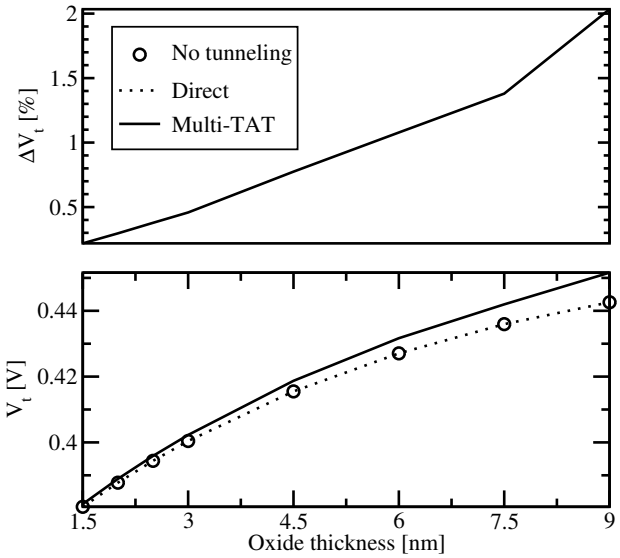


Figure 6: Comparison of the threshold voltage $V_\mathrm{t}$ of MOSFET structures with different oxide thicknesses.

## REFERENCES


[1] R. Moazzami and C. Hu, in *Proc. Intl. Electron Devices Meeting* (1992), pp. 139–142.

[2] B. Riccó, G. Gozzi, and M. Lanzoni, IEEE Trans.Electron Devices **45**, p.1554 (1998).

[3] M. Herrmann and A. Schenk, J.Appl.Phys. **77**, p.4522 (1995).

[4] F. Schuler, R. Degraeve, P. Hendrickx, and D. Wellekens, in *Intl. Reliability Physics Symposium* (2002), pp. 26–33.

[5] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and A. Modelli, Solid-State Electron. **46**, p.1749 (2002).

[6] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and M. J. van Duuren, IEEE Trans.Electron Devices **51**, p.1288 (2004).

[7] L. Larcher, IEEE Trans.Electron Devices **50**, p.1246 (2003).

[8] R. Entner *et al.*, in *Proc. Nanotech 2005 Vol. 3* (2005), pp. 45–48.

[9] W. J. Chang, M. P. Houng, and Y. H. Wang, J.Appl.Phys. **89**, p.6285 (2001).

[10] A. Palma *et al.*, Phys.Rev.B **56**, p.9565 (1997).

[11] F. Jiménez-Molinos *et al.*, J.Appl.Phys. **90**, p.3396 (2001).

[12] A. Gehring *et al.*, Microelectron.Reliab. **43**, p.1495 (2003).

[13] *MINIMOS-NT 2.1 User's Guide*, Institut für Mikroelektronik, Technische Universität Wien, Austria, 2004, `www.iue.tuwien.ac.at/software/`.