Impact of Multi-Trap Assisted Tunneling on Gate Leakage of CMOS Memory Devices

R. Entner*, A. Gehring**, H. Kosina**, T. Grasser*, and S. Selberherr**

* Christian Doppler Laboratory for TCAD in Microelectronics at the Institute for Microelectronics
Gußhausstraße 27–29, A–1040 Vienna, Austria, entner@iue.tuwien.ac.at
** Institute for Microelectronics, Vienna University of Technology

ABSTRACT

Dielectrics of state-of-the-art memory cells subject to repeated high field stress can have a high defect density. Thus, not only direct tunneling but also trap-assisted tunneling plays an important role. In this work a new approach for modeling gate leakage currents through highly degraded dielectrics is proposed. By rigorous simulation we show that multi-trap assisted tunneling becomes important for highly degrad dielectrics with thicknesses above approximately 4 nm, there it exceeds the single-trap assisted and direct tunneling components.

Keywords: dielectric layers, reliability, trap assisted tunneling, device modeling

1 INTRODUCTION

While logic CMOS devices feature dielectric thicknesses below 1.2 nm, non-volatile memory cells rely on tunneling oxides as thick as 7 nm. In order to speed up the programming and erasing process strong electric fields are applied across the dielectric. Due to the repeated high-field stress, trap centers in the insulator are created, which lead to trap-assisted tunneling at low bias, forming stress-induced leakage current (SILC).

Modeling this gate leakage current for such devices is of paramount interest, because it determines the retention time. Thicker dielectrics subject to high field stress may have a high defect density. Thus not only direct tunneling but also trap-assisted tunneling (TAT) currents play an important role [1]. The trap-assisted current component has been found to stem from inelastic tunneling assisted by phonon emission [2].

For device simulation, trap-assisted tunneling is commonly modeled as a single-trap [3] or two-trap [4] process. Recently, multi-trap models considering hopping processes have been presented [5]. The single-trap model was found to accurately reproduce experimental data of slightly stressed dielectrics [6]. Recently, however, anomalous charge loss in floating-gate memory cells after program/erase stress cycles has been observed [7]. Due to the high defect density in those cells it is reasonable to assume that more than one trap is involved in the tunneling process. For correct modeling of such highly degraded devices a new approach is presented which rigorously computes TAT current assisted by multiple traps. In contrast to the model presented in [5], where conduction across discrete paths is assumed, hopping processes between all traps are taken into account. The space charge of occupied traps is accounted for in the Poisson equation to estimate the resulting V_t shift.

2 THE MODEL

For the simulation of trap-assisted tunneling currents the current density across the insulator is modeled as the sum of the capture and emission rates R_i in each trap times the trap cross section Δx_i,

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

The energetic position of the trap with respect to the conduction band edge $\mathcal{E}_\mathrm{T}$ determines the trap cross section [8]

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

where m_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. 1) the rates are determined by [9]

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

Here, R_{c_i} and R_{e_i} are the capture and emission rates of the considered trap, respectively, and N_{t_i} denotes the trap concentration. In the stationary case the capture and emission rates must be equal, hence $R_{\mathrm{c}_i} = R_{\mathrm{e}_i} = R_i$. The trap occupancy f_{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], [9]

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

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

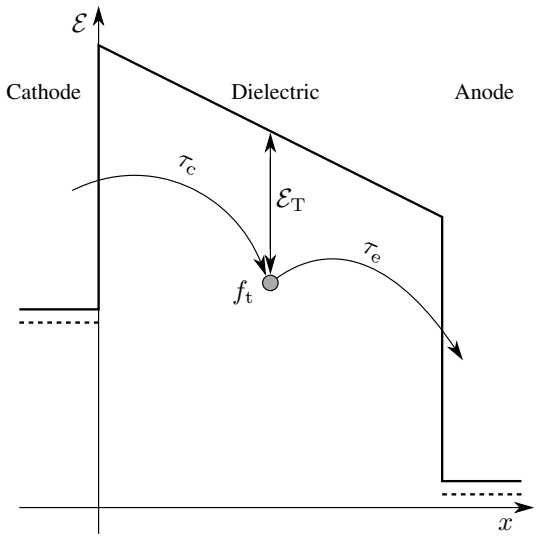

Figure 1: Single-trap assisted tunneling process. The tunneling rate R_i of a specific trap is determined by the capture time from the cathode and the emission time to the anode.

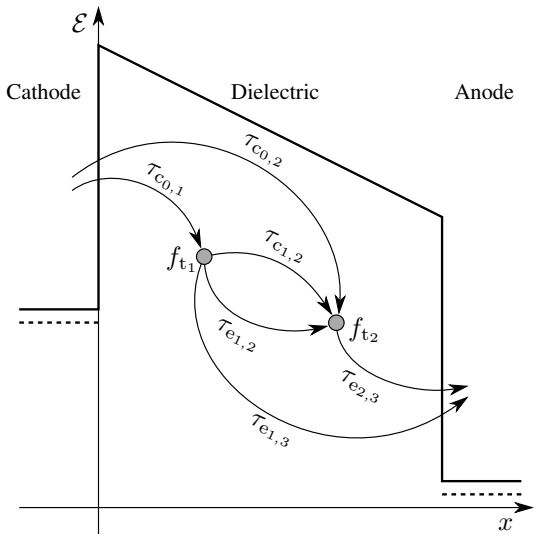

Figure 2: Multi-trap assisted tunneling process. The tunneling rate R_i of a specific trap is determined by all capture and emission times to and from the trap.

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

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

$$e_n(\mathcal{E}) = c_0 \exp\left(-\frac{\mathcal{E} - \mathcal{E}_T}{k_B T_L}\right) \sum_m L_m \delta(\mathcal{E} - \mathcal{E}_m) \tag{7}$$

with

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

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

The summation index m denotes the discrete phonon emissions, $\mathcal{E}_m$ is the phonon energy, and L_m is the multiphonon transition probability [9]. The symbols f_c and f_a are the Fermi distributions, T_c and T_a the transmission coefficients from the cathode and the anode, F the electric field in the dielectric, and $\mathcal{E}_{g,SiO2}$ 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 [1]–[3].

Recently, however, anomalous charge loss in memory cells has been observed and was explained by conduction through a second trap [4]. The single-trap model can be extended for this case, and the rate equations become (see Fig. 2)

$$\overbrace{\tau_{c_{0,1}}^{-1} N_{t_1}(1 - f_{t_1})}^{R_{c_1}} - \overbrace{(\tau_{e_{1,2}}^{-1} N_{t_1} f_{t_1}(1 - f_{t_2}) + \tau_{e_{1,3}}^{-1} N_{t_1} f_{t_1})}^{R_{e_1}} = 0 \ ,$$

$$\underbrace{\tau_{c_{0,2}}^{-1} N_{t_2}(1 - f_{t_2}) + \tau_{c_{1,2}}^{-1} N_{t_2} f_{t_1}(1 - f_{t_2})}_{R_{c_2}} - \underbrace{\tau_{e_{2,3}}^{-1} N_{t_2} f_{t_2}}_{R_{e_2}} = 0 \ ,$$

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 assists in the conduction process. We therefore extend the model to n traps where the capture and emission rates are evaluated as

$$R_{c_k} = \sum_{i=0}^{k-1} \tau_{c_{i,k}}^{-1} N_{t_k} f_{t_i}(1 - f_{t_k}) \ , \tag{10}$$

$$R_{e_k} = \sum_{i=k+1}^{n+1} \tau_{e_{k,i}}^{-1} N_{t_k} f_{t_k}(1 - f_{t_i}) \ . \tag{11}$$

The values for f_{t_0} and f_{t_n}, 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. The values for the other trap occupation probabilities have to be evaluated from the equation system. This is performed within MINIMOS-NT using the Newton method. Typical dimensions of the equation system to be solved are, depending on the dielectric thickness and trap energy, up to 15×15. The computational effort remains negligible compared to the total device simulation time. From either the capture or emission rates the multi-trap assisted tunneling current density J is obtained.

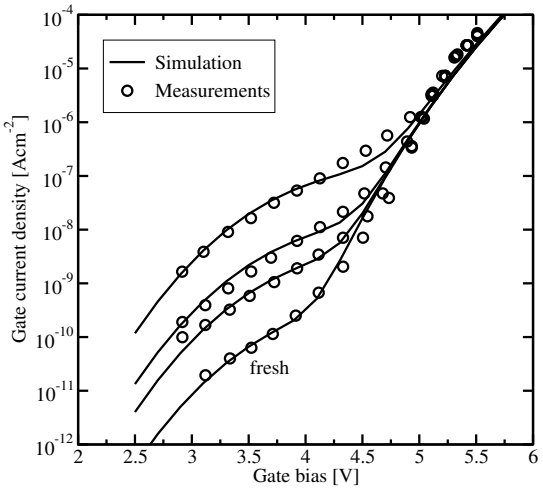

Figure 3: Single-TAT simulations of a MOS capacitor with a 5.5 nm dielectric and N_t=$9 \times 10^{17}\,\mathrm{cm}^{-3}$ ... $3 \times 10^{15}\,\mathrm{cm}^{-3}$ (top to bottom).

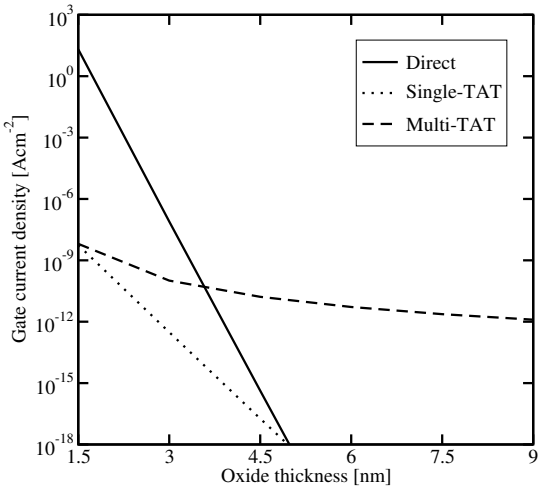

Figure 4: SILC simulations for a set of MOS devices at 1 V gate bias. The oxide has a constant trap concentration of $9 \times 10^{17}\,\mathrm{cm}^{-3}$.

3 APPLICATION

The implementation of these models into the device- and circuit-simulator MINIMOS-NT [10] allows the two- and three-dimensional study of single- and multi-trap assisted tunneling. Fig. 3 shows measured SILC [6] after different stress times for a MOS capacitor and the representative simulation results using a single-TAT simulation. It can be clearly seen that for slightly degraded dielectrics the single-TAT model yields excellent agreement with the measured data. Here the trap concentration is used as fitting parameter. For highly degraded devices it has been shown [4], however, that the SILC cannot be explained by conduction over solitary traps. It must be assumed that the large SILC is due to defect interaction of nearby traps.

Fig. 4 shows a comparison of the 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. It can be clearly seen that 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, but also the capture and emission processes involving all other trap centers. This fact 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.

Fig. 5 depicts the trap occupancy within the oxide. A MOS transistor with 1 V gate bias was simulated. It can be seen that the trap occupancy f_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.

The implementation of this model into the device simulator MINIMOS-NT allows to simulate the effect on the threshold voltage of memory devices. The space charge density in the dielectric is calculated as $\rho(x) = \mathrm{q} f_t(x) N_t(x)$. Fig. 6 outlines the threshold voltage V_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_t. This higher threshold voltage is due to the tunneling current in the oxide and the filled and therefore negatively charged traps.

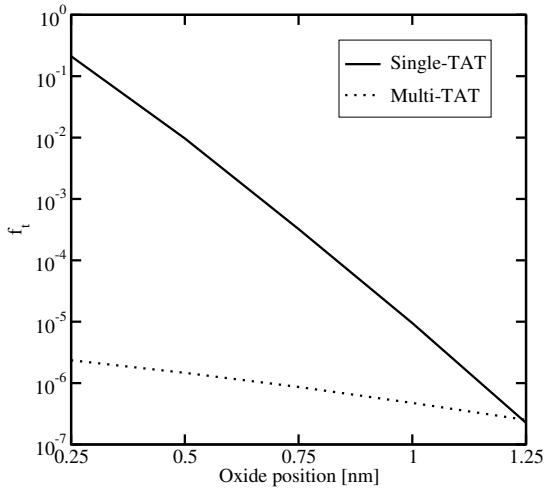

Figure 5: The trap occupancy f_t in the oxide of a 1.5 nm MOS transistor at 1 V gate bias.

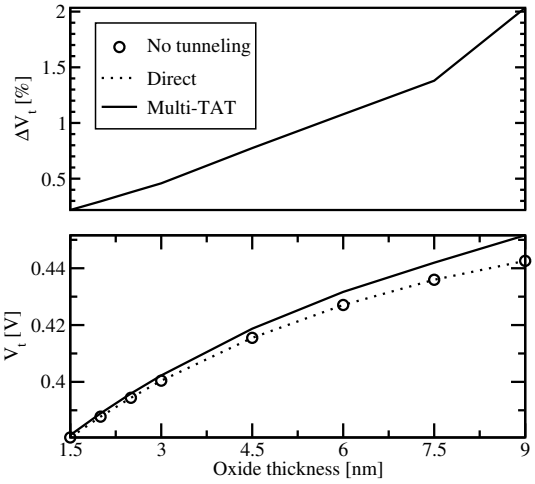

Figure 6: Comparison of the threshold voltage V_t of MOSFET structures with different oxide thicknesses.

4 CONCLUSION

We presented a new trap-assisted tunneling model which takes the interaction of several traps for the creation of conducting paths into account. The model is applied to devices with varying oxide thicknesses. Comparing single-TAT and direct tunneling reveals that for highly degraded devices with oxide thicknesses between 3 nm and 8 nm the inclusion of a multi-TAT model is crucial. With the implementation of the multi-TAT model in the multi-purpose device- and circuit simulator MINIMOS-NT arbitrary device geometries can be evaluated.

5 ACKNOWLEDGMENT

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

REFERENCES

[1] A. Gehring and S. Selberherr, "Modeling of Tunneling Current and Gate Dielectric Reliability for Nonvolatile Memory Devices," *IEEE Trans.Dev.Mat.Rel.*, vol. 4, no. 3, pp. 306–319, 2004.

[2] W. J. Chang, M. P. Houng, and Y. H. Wang, "Simulation of Stress-Induced Leakage Current in Silicon Dioxides: A Modified Trap-Assisted Tunneling Model considering Gaussian-Distributed Traps and Electron Energy Loss," *J.Appl.Phys.*, vol. 89, no. 11, pp. 6285–6293, 2001.

[3] F. Jiménez-Molinos, A. Palma, F. Gámiz, J. Banqueri, and J. A. Lopez-Villanueva, "Physical Model for Trap-Assisted Inelastic Tunneling in Metal-Oxide-Semiconductor Structures," *J.Appl.Phys.*, vol. 90, no. 7, pp. 3396–3404, 2001.

[4] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and A. Modelli, "Modeling of Anomalous SILC in Flash Memories Based on Tunneling at Multiple Defects," *Solid-State Electron.*, vol. 46, no. 11, pp. 1749–1756, 2002.

[5] L. Larcher, "Statistical Simulation of Leakage Currents in MOS and Flash Memory Devices with a New Multiphonon Trap-Assisted Tunneling Model," *IEEE Trans.Electron Devices*, vol. 50, no. 5, pp. 1246–1253, 2003.

[6] E. Rosenbaum and L. F. Register, "Mechanism of Stress-Induced Leakage Current in MOS Capacitors," *IEEE Trans.Electron Devices*, vol. 44, no. 2, pp. 317–323, 1997.

[7] F. Schuler, R. Degraeve, P. Hendrickx, and D. Wellekens, "Physical Description of Anomalous Charge Loss in Floating Gate Based NVM's and Identification of its Dominant Parameter," in *Intl. Reliability Physics Symposium*, pp. 26–33, 2002.

[8] A. Palma, A. Godoy, J. A. Jimenez-Tejada, J. E. Carceller, and J. A. Lopez-Villanueva, "Quantum Two-Dimensional Calculation of Time Constants of Random Telegraph Signals in Metal-Oxide-Semiconductor Structures," *Phys.Rev.B*, vol. 56, no. 15, pp. 9565–9574, 1997.

[9] M. Herrmann and A. Schenk, "Field and High-Temperature Dependence of the Long Term Charge Loss in Erasable Programmable Read Only Memories: Measurements and Modeling," *J.Appl.Phys.*, vol. 77, no. 9, pp. 4522–4540, 1995.

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