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]

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

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

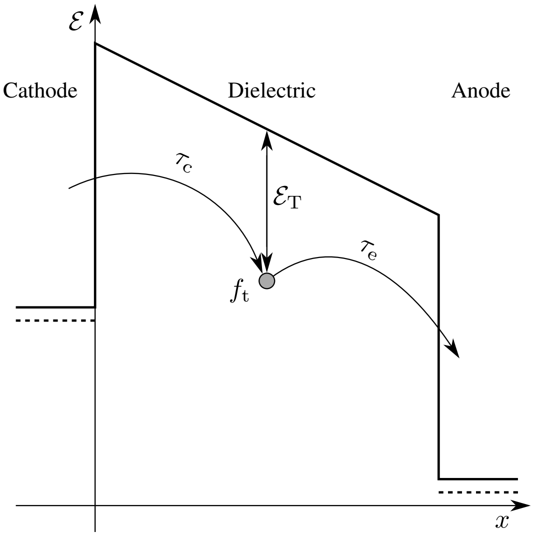

Figure 1: Single-trap assisted tunneling process. The tunneling rate R_i of a specific trap is determined by the capture time from the cathode and the emission time to the anode.

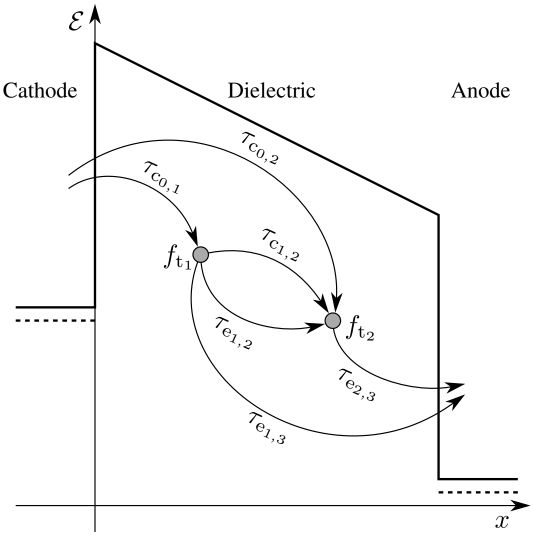

Figure 2: Multi-trap assisted tunneling process. The tunneling rate R_i of a specific trap is determined by all capture and emission times to and from the trap.

In these expressions, $g_{\mathrm{C}}(\mathcal{E})$ 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 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, $\mathrm{T_c}$ 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 [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)

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

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

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

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

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

The values for f_{t_0} and $f_{\mathrm{t}_\mathrm{n}}$, 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.