PERGAMON

Microelectronics Reliability 43 (2003) 1495–1500

MICROELECTRONICS RELIABILITY

www.elsevier.com/locate/microrel

Modeling of retention time degradation due to inelastic trap-assisted tunneling in EEPROM devices

A. Gehring*, F. Jiménez-Molinos†, H. Kosina, A. Palma†, F. Gámiz†, and S. Selberherr

Institute for Microelectronics, Vienna University of Technology Gusshausstrasse 27–29, A-1040 Vienna, Austria

†*Departamento de Electrónica, Facultad de Ciencias, Universidad de Granada, Avenida Fuentenueva s/n 18071-Granada, Spain*

Abstract

We present a model to describe inelastic trap-assisted tunneling through dielectric layers. The numerically expensive calculation of the transition matrix element inside the barrier is avoided by using approximative barrier shapes which lead to closed-form expressions for the wave function. The model is used for the simulation of discharging characteristics in stressed electrically erasable programmable read-only-memories (EEPROM). The trap occupancy in the gate oxide shows peaks in the middle of the layer and near the anode due to wave function interference. The model is further developed to study transient stress-induced leakage current where good agreement with measured data is achieved.

1. Introduction

The excess low field current through gate oxides observed after writing-erasing cycles in EEPROMs is mainly responsible for the degradation of the retention time. It is now generally accepted that this stress-induced leakage current (SILC) is caused by inelastic trap-assisted tunnel transitions and that these traps are created by the electric high-field stress during the writing and erasing processes [1–3]. SILC has been widely studied and modeled in metal-oxide-semiconductor (MOS) capacitors [2–5] and EEPROM devices. Recently, a new model for trap-assisted inelastic tunneling has been presented [6]. This model is based on the theory of non-radiative capture and emission of electrons by multiphonon processes [7]. Therefore, it does not require capture cross sections or emission rates as fitting parameters, but calculates them for each trap, position, and bias. The free parameters of the model are the trap energy level, the trap concentration, and the Huang-Rhys factor [8]. In this paper, an adaptation of the model facilitating its implementation in device simulators is shown. The main step is the transformation of the barriers in order to get analytical expression for the electron wave functions and the capture and emission probabilities. A fully analytical model is obtained using this approximation and it has been implemented in the general-purpose device simulator MINIMOS-NT [9]. Furthermore, an iterative formulation for transient SILC where capture and emission times are updated at each time step has been developed.

The paper is organized as follows. In Section 2, we derive the physical model and describe the approximations for triangular and trapezoidal barriers, followed by the derivation of expressions for the steady-state and transient current. In Section 3, we present steady-state and transient simulation results and compare them with numerical solutions and measurements. Finally, conclusions are drawn in Section 4.

*Corresponding author. Tel.: +43-1-58801/36016; fax:+43-1-58801/36099.

E-mail address: Gehring@iue.tuwien.ac.at (A. Gehring).

0026-2714/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0026-2714(03)00265-8

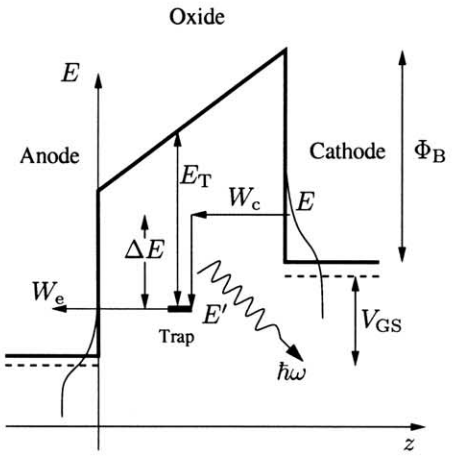

Figure 1. The trap-assisted tunneling model.

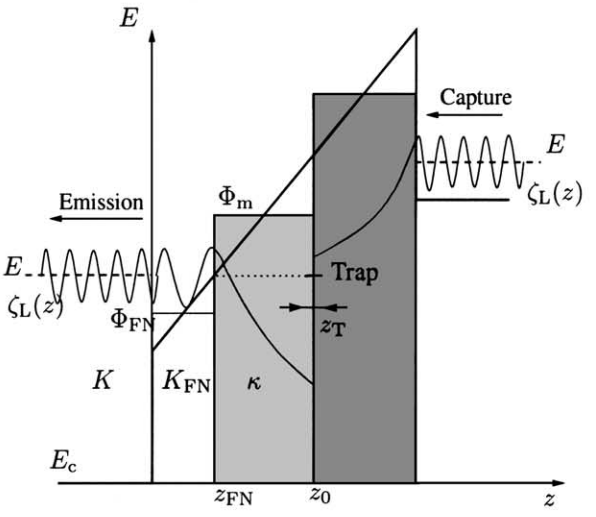

Figure 2. Approximation of the energy barrier.

2. The Trap-Assisted Tunneling Model

The tunneling model is based on [6] where Jimenez *et al.* describe a two-step inelastic tunneling process via traps in the oxide, incorporating energy loss by phonon relaxation.

2.1. Capture and Emission Probabilities

Fig. 1 shows the basic two-step process of an electron tunneling from a region with higher Fermi energy (the cathode) to a region with lower Fermi energy (the anode). The initial electron energy E is assumed to be $3/2k_BT$ above the cathode conduction band edge. During the capture process, the difference in total energy between the initial and final state is released by means of phonon emission. The probability that an electron with energy E is captured by a trap located at position z and energy E' is given by [10]

$$W_c(z,E',E) = \frac{\pi}{\hbar^2\omega}|V_e|^2 S\left(1-\frac{p}{S}\right)^2 \cdot I_p(\xi) \cdot \exp\left[-(2n+1)S + \frac{\Delta E}{2k_BT}\right]. \quad (1)$$

Here, S is the Huang-Rhys factor, $\hbar\omega$ is the energy of the phonons involved in the transitions, $\Delta E = E - E'$, and $p = \Delta E/\hbar\omega$ is the number of phonons emitted due to this energy difference. The population of phonons n is given by the Bose-Einstein statistics

$$n = \left[\exp\left(\frac{\hbar\omega}{k_BT}\right) - 1\right]^{-1}, \quad (2)$$

and the function $I_p(\xi)$ denotes the modified Bessel function of order p, with $\xi = 2S[n(n+1)]^{1/2}$.

The term $|V_e|^2$ in (1) denotes the transition matrix element which is given by [7]

$$|V_e|^2 = 5\pi S(\hbar\omega)^2 \frac{a_T^2}{V} \int_{z_0-z_T/2}^{z_0+z_T/2} |\zeta_L(z)|^2 \, dz. \quad (3)$$

In this expression, z_T is the side length of the trap cube, estimated as $z_T = a_T(4\pi/3)^{1/3}$ with the trap radius $a_T = \hbar/\sqrt{2m^*_{ox}E_T}$ and E_T being the energy difference between the trap energy and the conduction band edge as shown in Fig. 1. The symbol $\zeta_L(z)$ denotes the envelope function in the direction normal to the interface. If the traps are positive when empty, the transition element $|V_e|^2$ is multiplied by the Sommerfeld factor C_0 [10]. For the emission of electrons from the trap to the anode, elastic tunneling is assumed. Hence, the probability of emission to the anode W_e^a is equal to the probability of capture from the anode W_c^a, which is calculated from (1).

The numerical evaluation of the wave functions in the oxide degrades the computational efficiency of a multipurpose device simulator where simulation speed is a critical topic. Therefore, we used an analytic approximation for the wave function in the dielectric. We assumed plane waves and replaced the triangular or trapezoidal barriers by a series of rectangular barriers as shown in Fig. 2. With this approximation, the integral in (3) yields two analytical expressions for the wave function valid for the direct (trapezoidal barrier) and Fowler-Nordheim (triangular barrier) region.

2.2. Capture and Emission Times

Once the capture and emission probabilities have been obtained, the corresponding times can be calculated. The inverse of the capture time is given by [6]

$$\tau_c^{-1}(z) = \int_{E'}^{\infty} W_c(z, E', E) \cdot N_c(E) f_c(E)\, dE \quad (4)$$

where $N_c(E)$ is the density of states and $f_c(E)$ the electron energy distribution function in the cathode. For the above stated assumption that all electrons are captured from the same energy level $E_c + 3/2 k_B T$ in the cathode, this expression can be simplified to

$$\tau_c^{-1}(z) \approx W_c(z, E', E_c + \frac{3}{2} k_B T) \cdot n_c, \quad (5)$$

where n_c is the carrier concentration in the cathode. The inverse of the emission time is [6]

$$\tau_e^{-1}(z) = \int_{E'}^{\infty} W_e(z, E', E) \cdot N_a(E)(1 - f_a(E))\, dE.$$

Assuming $f_a(E) \approx 0$ and elastic tunneling for the emission process, the emission time becomes

$$\tau_e^{-1}(z) \approx W_e(z, E', E') \cdot N_a(E'). \quad (6)$$

2.3. Steady-State Current

From the capture and emission times, the steady-state trap-assisted tunnel current is calculated by an integration across the oxide

$$J_{TAT} = q \int_0^{t_{ox}} \frac{N_t(z)}{\tau_c(z) + \tau_e(z)}\, dz, \quad (7)$$

where N_t denotes the trap concentration [11]. The total tunneling current is found as the sum of the trap-assisted tunneling current and the direct tunneling current, for which we used the Tsu-Esaki expression [12]

$$J_t = J_0 \int_0^{\infty} T(\mathcal{E}_t) \ln \left[\frac{1 + \exp\left(\frac{\mathcal{E}_f - \mathcal{E}_t}{k_B T} \right)}{1 + \exp\left(\frac{\mathcal{E}_f' - \mathcal{E}_t}{k_B T} \right)} \right] d\mathcal{E}_t$$

where $J_0 = 4\pi m_{ox} q k_B T / h^3$, $\mathcal{E}_f$ and $\mathcal{E}_f'$ denote the Fermi energies at the semiconductor-oxide interfaces. The transmission coefficient is calculated according to the WKB approximation. The total tunneling current is the sum of J_{TAT} and J_t.

2.4. Transient Current

We consider a spatial trap distribution $N_t(z)$ across the dielectric layer, where z points in direction of tunneling current flow. The rate equation for the concentration of occupied traps in position z is

$$N_t(z) \frac{df_T(z,t)}{dt} = N_t(z)\,[1 - f_T(z,t)]\,\tau_c^{-1}(z,t) - N_t(z) f_T(z,t) \tau_e^{-1}(z,t), \quad (8)$$

where $f_T(z,t)$ is the occupancy function of the traps and $\tau_c(z,t)$ and $\tau_e(z,t)$ are the time constants for capture and emission of electrons by a trap placed at position z. In the static case, capture and emission processes are in equilibrium and $df_T(z,t)/dt = 0$. In the transient case, however, the trap occupancy is not constant. Capture and emission times include transitions from the cathode and the anode

$$\tau_c^{-1}(z,t) = \tau_{ca}^{-1}(z,t) + \tau_{cc}^{-1}(z,t) \quad (9)$$

$$\tau_e^{-1}(z,t) = \tau_{ea}^{-1}(z,t) + \tau_{ec}^{-1}(z,t) \quad (10)$$

where τ_{ca} and τ_{cc} are the capture times to the anode and to the cathode, and τ_{ea} and τ_{ec} the corresponding emission times. To get the local trap occupancy, the differential equation (8) must be solved. If the capture and emission times τ_c^{-1} and τ_e^{-1} are constant over time, like in a decharging process with constant potential distribution, we can solve (8) analytically

$$f_T(z,t) = f_T(z,0) \exp\left(-\frac{t}{\tau_m(z,t)} \right) + \frac{\tau_m(z,t)}{\tau_c(z,t)} \cdot \left[1 - \exp\left(-\frac{t}{\tau_m(z,t)} \right) \right]$$

with $\tau_m^{-1} = \tau_c^{-1} + \tau_e^{-1}$. A more general approach is found if we look at the change of the trap distribution at discrete times. Integration of (8) in the time variable between t_i and t_{i+1} and changing to discrete time steps yields

$$f_T(z,t_i) - f_T(z,t_{i-1}) \approx \tau_c^{-1}(z,t_{i-1}) \cdot \Delta t_i - \tau_m^{-1}(z,t_{i-1}) \cdot \overline{f_i} \Delta t_i$$

where the abbreviations $\Delta t_i = t_i - t_{i-1}$ and $\overline{f_i} = [f_T(z,t_i) + f_T(z,t_{i-1})]/2$ have been used. Thus it is possible to write the trap distribution over time in the following recursive manner:

$$f_T(z,t_i) = A_i + B_i \cdot f_T(z,t_{i-1}) \quad (11)$$

$$A_i = \frac{\tau_c^{-1}(z,t_i) \cdot \Delta t_i}{1 + C_i} \quad (12)$$

$$B_i = \frac{1 - C_i}{1 + C_i} \quad (13)$$

$$C_i = \frac{\tau_m^{-1}(z,t_i) \cdot \Delta t_i}{2}. \quad (14)$$

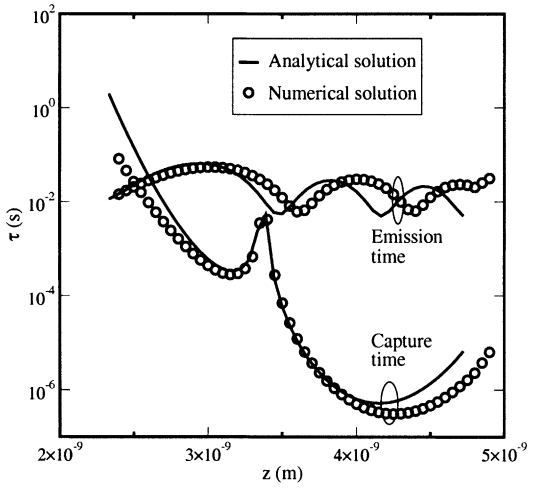

Figure 3. Capture and emission times using analytical and numerical solutions at $V_{GS} = -7\,V$.

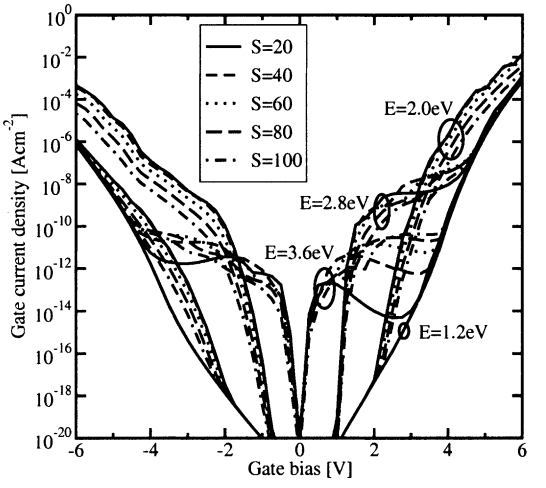

Figure 4. Influence of the model parameters E (trap energy level) and $S\hbar\omega$ for a fixed value of $\hbar\omega$=20 meV.

Once the time-dependent occupancy function in the dielectric is known, the tunnel current through one of the interfaces is

$$J_{\mathrm{TAT,i}}(t) = q \int_0^{t_{\mathrm{ox}}} N_{\mathrm{t}}(z) \cdot \tau_{\mathrm{i}}^{-1}(z,t)\,\mathrm{d}z, \qquad (15)$$

where i denotes the considered interface (either anode or cathode) and τ_{i} is calculated from

$$\tau_{\mathrm{i}}^{-1}(z,t) = \tau_{\mathrm{c,i}}^{-1}(z,t) - f_{\mathrm{T}}(z,t) \cdot \left[\tau_{\mathrm{c,i}}^{-1}(z,t) + \tau_{\mathrm{e,i}}^{-1}(z,t)\right]$$

3. Results

To check the validity of the approximations for the wave functions, the resulting capture and emission times have been compared to results using a Schrödinger-Poisson solver for a MOS capacitor with parameters of E_{T}=2.8 eV, $S\hbar\omega$=1.6 eV, and a trap concentration of $N_{\mathrm{t}} = 10^{19}$ cm^{-3}. As can be seen in Fig. 3, the analytical and the numerical results are very close. The typical emission time oscillations in the Fowler-Nordheim regime can be observed. Fig. 4 shows the dependence of the gate current density on the model parameters E (trap energy level) and $S\hbar\omega$ for a fixed phonon energy of $\hbar\omega$=20 meV. For a low trap energy level, traps are located near the conduction band edge in the oxide, and direct tunneling prevails. With increasing trap energy level, the trap-assisted component becomes stronger and exceeds the direct tunneling current for low bias.

Fig. 5 shows a comparison to measured SILC [2] after different stress times for a MOS capacitor with an oxide thickness of 5 nm. The model parameters are given in the figure caption. A good agreement with the measurements can be achieved if the trap concentration is used as a fitting parameter dependent on the stressing time. Note the transition from the region of mainly trap-assisted tunneling for $V_{\mathrm{GS}} < 5\,V$ to the region of Fowler-Nordheim tunneling for $V_{\mathrm{GS}} > 5\,V$.

The incorporation of the presented model into the de-

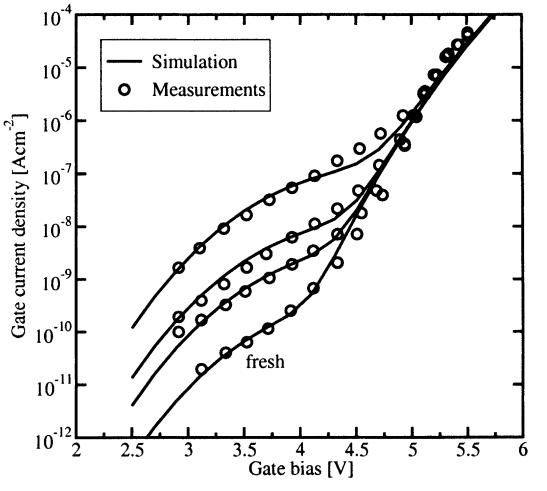

Figure 5. Comparison to measurement data of an MOS capacitor with an oxide thickness of 5.5 nm [2]. The model parameters are E_{T}=2.7 eV, $S\hbar\omega$=1.3 eV, and N_{t}= 9.0 × 10^{17} cm^{-3}, 1.0 × 10^{17} cm^{-3}, 3.0 × 10^{16} cm^{-3}, and 3.0 × 10^{15} cm^{-3} (from top to bottom).

vice simulator MINIMOS-NT allows to study the behavior of the trap occupancy in the oxide of MOS capacitors. Fig. 6 shows the trap occupancy f_T across the gate oxide of a MOSFET with an oxide thickness of 3 nm, using the gate voltage as parameter. The position of occupied traps is sketched in the inset. The regions near the gate (right) and near the substrate (left) are only sparsely occupied. Near the gate, the emission time is much smaller than the capture time, and near the substrate, the trap energy lies above the electron energy in the cathode. This characteristic shape of the trap occupancy may explain why a Gaussian or similar trap-distribution in space is commonly used to fit measurement data [4, 13]. Furthermore, the fact that some trapped electrons face a triangular barrier for the emission process gives rise to an additional peak in the trap occupancy near the gate side (the anode) of the oxide, which has not been observed before.

The resulting gate current density of a transient simulation is compared with measurements in Fig. 7 for MOS capacitors with oxide thicknesses of 8.5 and 13.0 nm. Initially, the traps are empty, which can be achieved by applying flat band conditions. At $t = 0$, the gate voltage is turned on and the traps are filled according to their specific capture and emission time constants. This charging current consists of an emission and a capture current which may exceed the steady-state current by orders of magnitude. The simulation results are compared to measurements [1] and a good fit could be achieved using the trap parameters indicated in the figure caption.

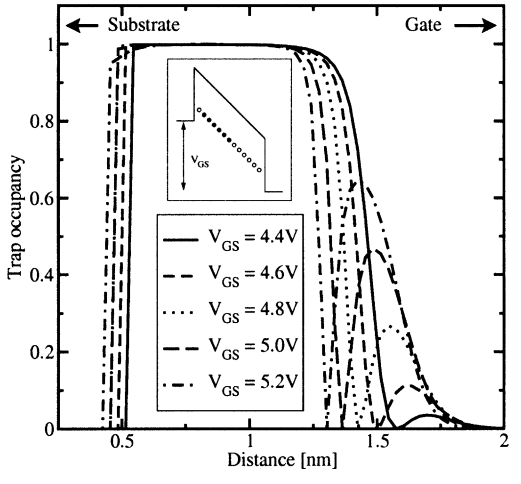

Figure 6. The trap occupancy across the gate oxide for a MOS capacitor with 3 nm oxide thickness.

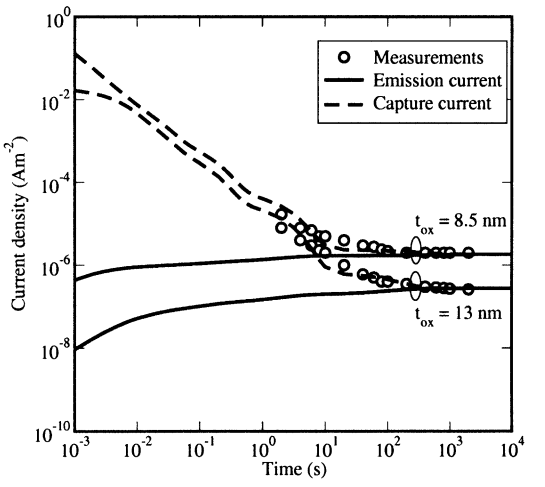

Figure 7. Transient capture and emission currents of MOS capacitors [1] at $V_{GS} = -5.8$ V and $V_{GS} = -8.3$ V with $S\hbar\omega$=1 eV, E_T=2.5 eV and $N_t = 3 \cdot 10^{18}$ cm^{-3} for the thinner oxide and $S\hbar\omega$=1.26 eV, E_T=2.5 eV and $N_t = 10^{18}$ cm^{-3} for the thicker oxide.

Fig. 8 shows the gate current of an MOS capacitor for an applied rectangular pulse assuming initial flat band conditions. It can be seen that the time constants of the trap filling and emptying processes depend on the applied voltage. The spikes are due to the sudden voltage change while the trap concentration remains constant: In the transition from 3.0 V to 3.5 V, the barrier shape changes suddenly, and traps are rapidly emptied. Traps near the cathode are filled and after several μs until the new steady state is reached.

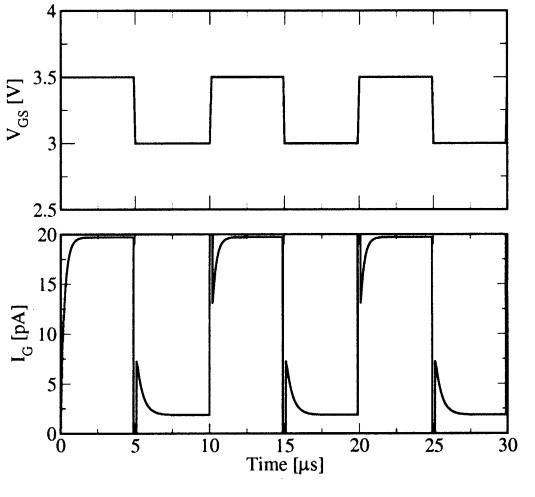

Figure 8. Transient simulation of a MOS capacitor with a gate oxide thickness of 3 nm and E_T=3 eV.

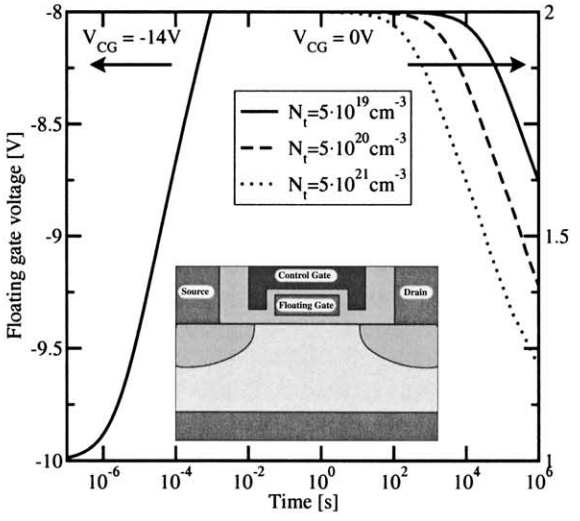

Figure 9. Decharging curve of an EEPROM. The floating gate is charged at a control gate voltage of -14V (left axis) and is then left floating at a control gate voltage of 0 V (right axis).

Using MINIMOS-NT, the model can be applied to the simulation of the decharging characteristics of EEPROM devices. Fig. 9 shows the floating gate voltage for the charging (left) and decharging (right) processes with different trap concentrations in the oxide. The device structure is shown in the inset. A tunneling oxide thickness of 8.5 nm was assumed. It can be seen that an increase of the trap concentration by an order of magnitude reduces the retention time by about the same factor.

4. Conclusion

A model based on inelastic trap-assisted tunneling has been adapted and implemented into the device simulator MINIMOS-NT to study stress-induced leakage current. In order to get a fully analytical model, the linear energy barrier in the dielectric was transformed into a staircase-shaped barrier. The results obtained with this analytical model have been compared to the results using a Schrödinger-Poisson solver and found to be sufficiently accurate. A formulation for the simulation of the transient component of SILC has also been presented. Therefore, this model allows the simulation of steady and transient SILC in MOS devices. Comparisons with measurements of transient and steady-state gate leakage in MOS capacitors show good agreement. The model allows to predict the charging and decharging characteristics of EEPROM devices based only on physical parameters like trap concentration, trap energy, and electron mass in the dielectric.

References

[1] Moazzami R, Hu C. Stress-induced current in thin silicon dioxide films. Proceedings of the Intl. Electron Devices Meeting 1992 p. 139–142.

[2] Rosenbaum E, Register LF. Mechanism of stress-induced leakage current in MOS capacitors. IEEE Transactions Electron Devices 1997;44(2):317–323.

[3] Takagi SI, Yasuda N, Toriumi A. A new I-V model for stress-induced leakage current including inelastic tunneling. IEEE J Solid-State Circuits 1999;46(2):348–354.

[4] Riccó B, Gozzi G, Lanzoni M. Modeling and simulation of stress-induced leakage current in ultrathin SiO_2 films. IEEE Transactions Electron Devices 1998;45(7):1554–1560.

[5] Ielmini D, Spinelli AS, Rigamonti MA, Lacaita AL. Modeling of SILC based on electron and hole tunneling - part I: transient effects. IEEE Transactions Electron Devices 2000;47(6):1258–1265.

[6] Jimenez-Molinos F, Palma A, Gamiz F, Banqueri J, Lopez-Villanueva JA. Physical model for trap-assisted inelastic tunneling in metal-oxide-semiconductor structures. J Appl Phys 2001;90(7):3396–3404.

[7] Palma A, Godoy A, Jimenez-Tejada JA, Carceller JE, Lopez-Villanueva JA. Quantum two-dimensional calculation of time constants of random telegraph signals in metal-oxide-semiconductor structures. Physical Review B 1997;56(15):9565–9574.

[8] Fowler WB, Rudra JK, Zvanut ME, Feigl FJ. Hysteresis and Franck-Condon relaxation in insulator-semiconductor tunneling. Physical Review B 1990;41(12):8313–8317.

[9] Institut für Mikroelektronik, MINIMOS-NT user's guide 2002.

[10] Zheng JH, Tan HS, Ng SC. Theory of non-radiative capture of carriers by multiphonon processes for deep centres in semiconductors. J Phys: Condensed Matter 1994;6(9):1695.

[11] Herrmann M, Schenk A. Field and high-temperature dependence of the long term charge loss in erasable programmable read only memories: measurements and modeling. J Appl Phys 1995;77(9):4522–4540.

[12] Tsu R, Esaki L. Tunneling in a finite superlattice. Appl Phys Lett 1973;22(11):562–564.

[13] Larcher L, Paccagnella A, Ghidini G. A model of the stress induced leakage current in gate oxides. IEEE Transactions Electron Devices 2001;48(2):285–288.