Impact of the Charge Transport in the Conduction Band on the Retention of Si–Nitride Based Memories

E. Vianello, F. Driussi, P. Palestri, A. Arreghini, D. Esseni, L. Selmi, N. Akil, M. van Duuren and D.S. Golubovic

DIEGM, Univ. of Udine, Via delle Scienze 208, 33100 Udine, Italy, FAX: +39-0432558251
*NXP–TSMC Research Center, Kapeldreef 75, 3001 Leuven, Belgium
email: elisa.vianello@uniud.it

Abstract—An improved model for charge injection through ONO gate stacks, that comprises carrier transport in the conduction band of the silicon nitride (Si₃N₄), is used to investigate the program/retention sequence of Si₃N₄ based (SONOS/TANOS) non-volatile memories without making assumptions on the initial distribution of the trapped charge at the beginning of retention.

We show that carrier transport in the Si₃N₄ layer impacts the spatial charge distribution and consequently several other aspects of the retention transient. The interpretation of the Arrhenius plots of the high temperature retention data, typically used to infer the spatial charge distribution and consequently several other aspects of the retention transient. The interpretation of the Arrehnius plots to cross the tunnel barrier (Jₔ) and the current density (Jₛi) impinging the Si–SiO₂ interface (calculated under the free electron gas approximation). The current density from the substrate to the Si₃N₄ bandgap (Band–to–Trap Tunneling J₂, (2) in Fig. 1) is calculated assuming an inelastic process, that can occur only if the trap energy level lies below the injection energy level. We compute this contribution as J₂=qRBT(N_T−n_T)Δx, where N_T is the total trap density, n_T is the occupied trap density and Δx the mesh spacing. RBT=σJₔT_p/q is the Band–to–Trap tunneling rate, where σ is the capture cross section of traps.

In the Si₃N₄ layer the CB and the traps exchange carriers through capture and emission processes ((3) in Fig. 1). The emission rate (Rₑ) is modeled according to the Poole-Frenkel (PF) equation [9], and the capture rate (R_C) is proportional to the concentration of free traps. We have also considered the tunneling of carriers from the Si₃N₄ traps to the gate J₀=qRₑBn₁Δx ((6) in Fig. 1), where Rₑ is the Trap–to–Band tunneling rate, given by the product of T_p times a constant attempt–to–escape frequency νₑ [4].

The electron transport in the CB of the Si₃N₄ is described with a Drift-Diffusion (DD) relation: J_DD=qµ(−n∂V/∂x + kT∂n/∂x), where V is the electrostatic potential.
mobilities in the order of 0.1–1 cm²/Vs [8], a field of 10–100 MV/cm is needed to reach the expected saturation velocity of 10⁷ cm/s [11].

The boundary conditions for the DD equation at the Si₃N₄/tunnel–oxide interface are set by the direct and Fowler–Nordheim tunneling components of J₁, and by J₅=qnVT (Eq. (5) in Fig. 1). Here vT is an average tunnel–out velocity:

\[ v_T = \frac{\int_0^\infty f(E_\perp)v_\perp(E_\perp)T_P(E_\perp)dE_\perp}{\int_0^\infty f(E_\perp)dE_\perp} \]

where \( E_\perp \) is the kinetic energy in the x direction, \( f(E_\perp) \) is the energy distribution of electrons (product of occupation times density of states) and \( v_\perp(E_\perp) \) the normal velocity (considering a single parabolic band minimum).

Fig. 1.(b) shows the carrier fluxes during retention: this case is simpler than the programming phase, as the injection currents \( J_1 \) and \( J_2 \) are negligible.

Given the relatively narrow trap energy distribution found in [7], for the sake of simplicity a unique discrete trap energy level (\( E_T \), referred to the CB minimum) is considered in this work. The model ignores the direct interaction among the Si₃N₄ traps: assuming a uniform distribution, a trap density in the order of 10¹⁹–10²⁰ cm⁻³ [4] corresponds to an average distance between traps of 2–4.5 nm, so that the probability of direct tunneling appears to be very small. Since we are analyzing the program state, the hole contribution is neglected.

### B. Numerical Implementation

Fig. 2 schematically represents the balance of fluxes for each spatial bin of the silicon nitride. These fluxes are linked together by two continuity equations: one for the free electrons in the CB (\( n_1 \)) and one for the trapped charge (\( n^T_1 \)). Considering also the Poisson equation, this approach leads to a system of three non linear partial differential equations in space (\( x \)) and time (\( t \)). To numerically solve these equations, we adopted a simple discretization in space and time (index \( i \) and \( k \), respectively):

\[
\begin{align*}
\frac{n^k_{i} - n^{k-1}_{i}}{\Delta t} & = -\frac{J_{DD,i+1}^k + J_{DD,i}^k - J_{m,FN,i}^k}{q \Delta x} - R_{E_{n,0}}^k n^k_1 + R_{B_{T,0}}^k n^T_1, \\
\frac{n^T_{i+1} - n^{T,k-1}_{i}}{\Delta t} & = R_{D_{T,0}}^k n^k_1 - R_{B_{T,0}}^k n^{T,k}_1 + R_{D_{B,0}}^k (N_T - n^{T,k}_1) - R_{D_{B,0}}^k n_1^{T,k}, \\
\frac{V_{i+1}^{k+1} - 2V_i^{k+1} + V_{i-1}^{k+1}}{\Delta x^2} & = q \left( n^k_1 + n^{T,k}_1 \right) \frac{1}{\varepsilon_0 \varepsilon_N}
\end{align*}
\]

where \( J_{m,FN} \) is the modified Fowler–Nordheim component of \( J_1 \). The time discretization is tackled by means of the Backward–Euler scheme, which is stable irrespective of the time and space discretization granularity. At each time step the resulting system of non linear equations is solved with the Full Newton scheme.

In the gate and substrate we solved the nonlinear Poisson equation (which also provides boundary conditions for the Poisson equation in Si₃N₄), leading to the instantaneous update of the potential profile consistently with the gate voltage (\( V_G \)).

### C. Simulation Procedure

The proposed model allows us to perform simulations of retention transients immediately following the program phase, thus obtaining curves as those in Fig. 3. Consequently, a priori assumptions on the spatial distribution of the trapped charge at the beginning of the retention phase (\( \tau_r=0 \)) are not necessary.

Fig. 4 (left plot) shows the simulated band profile during the program transient of a SONOS device with a gate stack of 2/6/9 nm for \( V_G=12.5 \) V. The right plot shows typical free electron density \( n_1 \) (filled symbols) and trapped charge density \( n^T_1 \) profiles (open symbols) along the vertical position in the Si₃N₄ layer. It can be seen that \( n^T_1 \) is several orders of magnitude larger than \( n_1 \). Furthermore \( n^T_1 \) peaks at the interface with the top oxide because free electrons injected directly in the nitride CB drift toward the top oxide interface, pile up there and are eventually trapped. The increase of \( n^T_1 \) near the tunnel oxide interface is instead related to the inelastic Band-to-Trap Tunneling (Fig. 1.(a), (2)).
III. Model Calibration

As a first step, it is important to estimate the value of the trap energy ($E_T$, Fig. 1(b)), the effective mass of electrons in the nitride ($m^*_N$) and the attempt–to–escape frequency ($\nu_T$), which are the model parameters with the strongest influence on the retention transient. To this purpose, the analysis of retention at high temperature of SONOS with thick oxides can be useful. In this regime, the thermal emission of carriers, which is linked to the trap energy, is the dominant discharge mechanism [6], [7] and the activation energy ($E_A$) extracted from the Arrhenius plot of the retention time at high temperature can be usefully related to $E_T$.

However, the relationship between $E_A$ and $E_T$ is not straightforward; indeed $E_A$ is the overall result of three different phenomena: the emission of charge from the traps to the CB by the PF effect ((3) in Fig. 1), the transport of emitted electrons in the CB toward the interfaces ((4) in Fig. 1), and the tunneling of carriers through the barriers ((5) in Fig. 1). In the literature, most of the authors only consider the first mechanism, i.e. the thermal emission in the CB [4]; therefore according to these models, $E_A$ coincides with $E_T$. In [7] the role of the oxide barriers was included, but only the emission above their top edge contributed to the charge loss. In the present model, instead, we have considered all the three mechanisms and the resulting $E_A$ value can be linked to $E_T$ in a more appropriate way.

Fig. 5 reports the activation energy values of the retention time (for a given threshold voltage decrease from the initial programmed state), as a function of the assumed $E_T$. $E_A$ is extracted from simulations of high temperature retention on three gate stacks with relatively thick tunnel oxide. As we see, $E_T$ values of about 1.0–1.2 eV [12] have to be chosen in order to reproduce $E_A$ values compatible with those extracted from experiments in [6], [13] (≈1.75–1.9 eV). As sketched in the inset of Fig. 5, $E_A$ is related to the average energy of escaping electrons, that is the maximum of the product $f(E_L)v_L(E_L)T_P(E_L)$. Since the CB offset between SiO$_2$ and Si$_3$N$_4$ is approximately 1 eV, the difference between the $E_A$ and $E_T$ values of Fig. 5 indicates that in thick tunnel oxide devices the emission occurs fairly close to the top of the barrier, thus confirming the assumption made in [7] and casting doubts on the simplifications in [4].

In order to calibrate the Trap–to–Band tunneling ((6) in Fig. 1(b)), which is the dominant charge loss mechanism in devices with thin tunnel oxide at room temperature [7], we have performed measurements and simulations on a 2/6/9 nm SONOS device. Fig. 6 shows the impact on the simulated retention of the two parameters that govern the Band–to–Trap tunneling component, i.e. $m^*_N$ (upper graph) and $\nu_T$ (lower graph). The simulator well reproduces the measurements for $m^*_N=0.25 m_0$ and $\nu_T=5 \times 10^8$ s$^{-1}$. The difference between these values and those in [7] is discussed in Sec. IV.

IV. Results

Having calibrated the main model parameters we now try to reproduce the retention characteristics of SONOS cells featuring both thin and thick oxides. For all the devices the nitride was a standard LPCVD and the top oxide was a HTO. The 2/6/9 nm device has a n+ poly gate while the other cells have a metal gate (TiN). Fig. 7 shows the measured and simulated retention of the 2/6/9 nm device programmed with pulses of 1 ms and different $V_G$. The charge distribution and the $\Delta V_T$ at the beginning of the retention are the ones calculated by the simulator at the end of the program pulse. Note the good agreement between measurements and simulations, in terms of both initial state (that is the outcome of the corresponding program phase) and retention dynamics.

In order to evaluate the role of charge redistribution during retention, we have performed measurements also on devices with relatively thick oxides at room temperature. In these conditions we suppress most of the discharge mechanisms and we expect to put in evidence the possible effects of carrier transport in Si$_3$N$_4$. We have programmed two SONOS devices in the neutral state with $V_G=12$ V for 10 ms. An increase of the threshold voltage during retention (Fig. 8, symbols) is observed in both samples as in Fig. 10 of [3].

The retention simulations well reproduce the $\Delta V_T$ increase (see lines in Fig. 8). This increase is explained by the slow shift of the trapped charge centroid toward the tunnel oxide illustrated in Fig. 9, which is due to the complex balance of drift, diffusion, PF emission and capture fluxes. In these thick oxides the spatial profiles of the trapped charge are still

Fig. 5. Simulated activation energy versus $E_T$ for devices featuring thick tunnel oxide and for $T$ between 500 K and 600 K. The dashed lines indicate the range of $E_A$ values extracted from the experiments in [6], [13]. The inset shows how $E_A$ can be related to the escape energy of electrons.
consistent with the measurements in [14] but different from those assumed in [7], consequently $m^*_N$ and $\nu_T$ values are also different. Fig. 8 also reports the impact of the attempt–to–escape frequency for the PF emission, $\nu_{PF}$, on $\Delta V_T$: an increase of $\nu_{PF}$ accelerates the redistribution mechanisms. Instead the effect of $\mu$ (see Fig. 10) is more complex, because it influences both the program and the retention phases. At the end of the program phase the charge is stored closer to the tunnel oxide for lower mobilities, resulting in a smaller effect of the redistribution, hence smaller $\Delta V_T$ variations. Simulations reproduce the experiments at best with $\mu \simeq 1$ cm$^2$/V$\cdot$s.

V. CONCLUSIONS

In summary, we presented an improved model for charge transport and trapping in ONO stacks that allows us to simulate a complete program/retention experiment, thus avoiding a priori assumptions on the initial distribution of the trapped charge. The model sheds new light on the relations between the activation energy of retention (at high $T$) and the trap energy. Moreover it provides a simple explanation for the $\Delta V_T$ increase during retention of thick tunnel oxide samples. Our analysis points out the need for an understanding of the spatial profile of the trapped charge in order to eliminate the ambiguity between some of the model parameters.

Acknowledgments: This work was partially funded by the Italian MIUR (FIRB RBIP06Y5JJ project).

REFERENCES