This is the peer reviewd version of the followng article:

Impact of the Charge Transport in the Conduction Band on the Retention of Si-Nitride Based Memories / E., V., Driussi, F., Palestri, P., A., A., Esseni, D., Selmi, L., N., A., M., V.D., D. S., G.. - STAMPA. - (2008), pp. 107-110. (ESSDERC 2008 - 38th European Solid-State Device Research Conference Edimburgo (GB) 15-19 Settembre 2008) [10.1109/ESSDERC.2008.4681710].

IEEE Computer Society
Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

12/11/2026 18:28

(Article begins on next page)

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. Golubović*
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 (Si_3N_4), is used to investigate the program/retention sequence of Si_3N_4 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 Si_3N_4 layer impacts the spatial charge distribution and consequently several other aspects of the retention transient. The interpretation of the Arrehnius plots of the high temperature retention data, typically used to infer the trap depth from the retention activation energy is discussed. The model provides a simple explanation of the small threshold voltage increase observed during retention experiments of thick tunnel oxide ONO stacks.

I. Introduction

Due to difficult scaling of the industry standard floating gate (FG) memory cells, alternative non volatile memory (NVM) technologies are gaining importance [1], [2]; among them, localized–trapping silicon–nitride or high–k based cells (SONOS/TANOS) allow for a significant reduction of the tunnel oxide thickness [2], with beneficial effects on the scaling of the programming voltage. Unfortunately, aggressive scaling of the tunnel oxide causes poor retention; this latter, then, needs to be understood in detail [3]–[6].

Simple retention models identified electron tunneling through the tunnel oxide and thermal emission from the nitride as the two dominant charge loss processes in SONOS cells [4], [7]. However, the threshold voltage dynamics during retention could be affected also by the charge distribution at the beginning of the retention experiment [7], and by the spatial redistribution of the stored charge due to the transport in the nitride. At this regard we note that, although the effective electron mobility in Si_3N_4 reported by a few authors is low ($\sim$0.1 cm^{-2}/(Vs) [8]), it is still large enough to induce significant changes on the charge spatial distribution even at low fields.

In order to shed new light on these aspects, in the following we present an improved numerical model for program and retention transients in SONOS cells, which accounts also for the electron transport in the Conduction Band (CB) of the Si_3N_4 layer. This improved model allows us to simulate the programming and retention sequence; therefore, differently from [3]–[5], [7] no assumption is needed on the spatial distribution of the trapped charge at the beginning of the retention phase.

Simulations are compared with experiments performed on SONOS cells with ONO stacks of 2/6/9 nm (tunnel–oxide/Si_3N_4/top–oxide thicknesses), 4/6/6 nm and 6/6/6 nm.

II. The Model

A. Physics

The charge fluxes accounted for by our model in program and retention operations are schematically depicted in Figs. 1.(a) and 1.(b) respectively. The current density injected from the substrate toward the Si_3N_4 CB (J_1, (1) in Fig. 1) is calculated as the integral over energy of the product of the WKB probability to cross the tunnel barrier (T_P) and the current density (J_{Si}) impinging the Si–SiO_2 interface (calculated under the free electron gas approximation). The current density from the substrate to the Si_3N_4 bandgap (Band–to–Trap Tunneling J_2, (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 $J_2{=}qR_{BT}(N_T - n^T)\Delta x$, where N_T is the total trap density, n^T is the occupied trap density and Δx the mesh spacing. $R_{BT}{=}\sigma J_{si}T_P/q$ is the Band–to–Trap tunneling rate, where σ is the capture cross section of traps.

In the Si_3N_4 layer the CB and the traps exchange carriers through capture and emission processes ((3) in Fig. 1). The emission rate (R_E) 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 Si_3N_4 traps to the gate $J_6{=}qR_{TB}n^T\Delta x$ ((6) in Fig. 1), where R_{TB} is the Trap–to–Band tunneling rate, given by the product of T_P times a constant attempt–to–escape frequency ν_T [4].

The electron transport in the CB of the Si_3N_4 ((4) in Fig. 1) is described with a Drift-Diffusion (DD) relation: $J_{DD}{=}q\mu(-n\frac{\partial}{\partial x}V + kT\frac{\partial}{\partial x}n)$, where V is the electrostatic

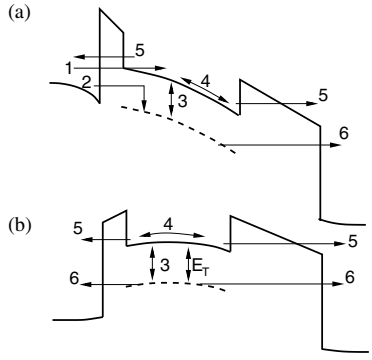


Fig. 1. Mechanisms involved during programming (a) and retention (b) of SONOS cells: 1) Tunnel–In, 2) Band–to–Trap Tunneling, 3) Emission–Capture events, 4) Electron Transport, 5) Tunnel–Out and, 6) Trap–to–Band Tunneling.

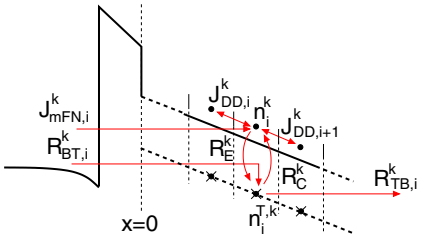


Fig. 2. Flux balance for each bin of the Si_3N_4 layer.

potential, n is the free electron density and μ the (constant) electron mobility. In order to avoid stability issues related to the Poisson/Drift Diffusion coupling we have however rewritten J_{DD} in the form proposed by [10]. Note that, with expected mobilities in the order of 0.1–1 $cm^2/(Vs)$ [8], a field of 10–100 MV/cm is needed to reach the expected saturation velocity of 10^7 cm/s [11].

The boundary conditions for the DD equation at the Si_3N_4/tunnel–oxide interface are set by the direct and Fowler–Nordheim tunneling components of J_1 and by $J_5=qnv_T$ ((5) in Fig. 1). Here v_T is an average tunnel–out velocity:

$$v_T = \frac{\int_0^\infty f(E_\perp) v_\perp(E_\perp) T_P(E_\perp) dE_\perp}{\int_0^\infty f(E_\perp) dE_\perp} \qquad (1)$$

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 Si_3N_4 traps: assuming a uniform distribution, a trap density in the order of 10^{19}–10^{20} 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) and one for the trapped charge (n^T). 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):

$$\begin{aligned}
&\frac{n_i^k - n_i^{k-1}}{\triangle t} = -\frac{J^k_{DD,i+1} - (J^k_{DD,i} + J^k_{mFN,i})}{q\triangle x} - R^k_C n^k_i + R^k_E n^{T,k}_i \\
&\frac{n_i^{T,k} - n_i^{T,k-1}}{\triangle t} = R^k_C n^k_i - R^k_E n^{T,k}_i + R^k_{BT,i}(N_T - n^{T,k}_i) - R^k_{TB,i} n^{T,k}_i \\
&\frac{V^k_{i+1} + V^k_{i-1} - 2V^k_i}{\triangle x^2} = q\frac{n^k_i + n^{T,k}_i}{\varepsilon_0 \varepsilon_N}
\end{aligned} \qquad (2)$$

where J_{mFN} 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 Si_3N_4), 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 (τ_r=0) are not necessary.

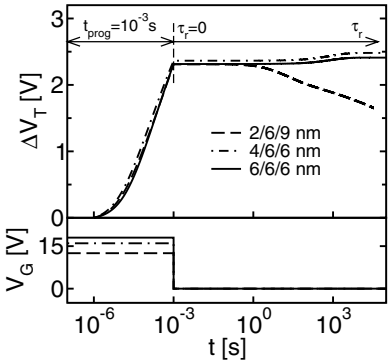


Fig. 3. Simulated program and retention transients for different gate stacks.

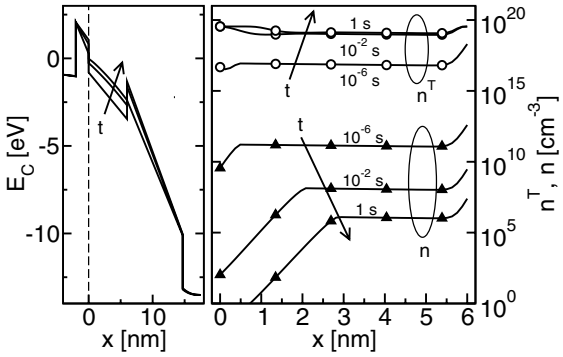


Fig. 4. Left plot: simulated time evolution of the band profile for a 2/6/9 nm device at V_G=12.5 V. Right plot: time evolution of the trapped charge (open symbols) and of the free electrons (filled symbols) in the nitride.

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 (filled symbols) and trapped charge density n^T profiles (open symbols) along the vertical position in the Si_3N_4 layer. It can be seen that n^T is several orders of magnitude larger than n. Furthermore n^T 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 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 (ν_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] ($\sim$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 $f(E_\perp)v_\perp(E\perp)T_P(E\perp)$. Since the CB offset between SiO_2 and Si_3N_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].

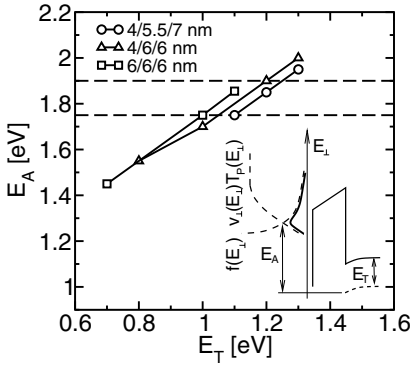


Fig. 5. Simulated activation energy versus E_T for devices featuring thick tunnel oxide and for T between 500 K and 600 K. The dashed lines indicate the range of E_A values extracted from the experiments in [6], [13]. The inset shows how E_A can be related to the escape energy of electrons.

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 ν_T (lower graph). The simulator well reproduces the measurements for m_N^*=0.25 m_0 and ν_T=5·10^8 s^{-1}. The difference between these values and those in [7] is discussed in Sec. IV.

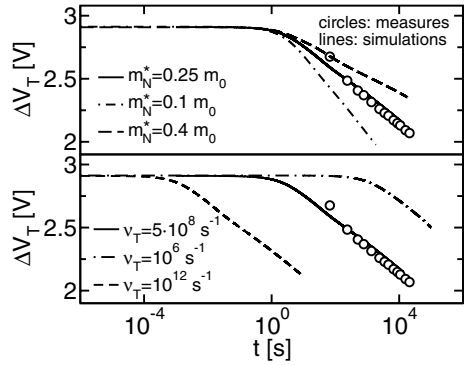


Fig. 6. Upper graph: simulated retention transients with various nitride electron effective mass m_N^*; lower graph: retention transients with different attempt–to–escape frequencies ν_T. The device (2/6/9 nm) is initially programmed at V_G=13 V for 10 ms. Symbols are reference measurements.

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 ΔV_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_3N_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 ΔV_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

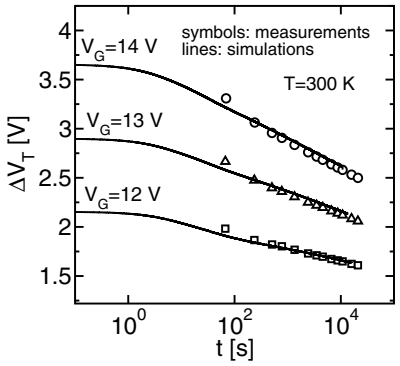


Fig. 7. Measured (symbols) and simulated (lines) retention characteristics for a 2/6/9 nm device at V_G=0 V. The initial ΔV_T for the retention experiments are obtained applying pulses of 1 ms and V_G as indicated in the figure.

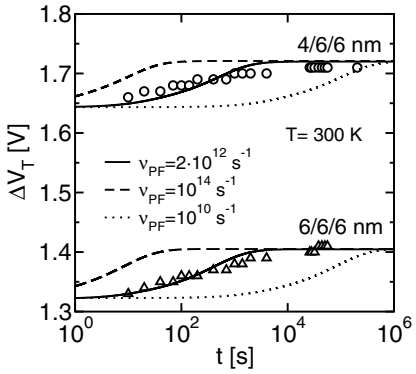


Fig. 8. Measured retention characteristics of a 6/6/6 nm cell (triangles) and of a 4/6/6 nm device (circles). Note the ΔV_T increase due to the charge redistribution, that can be reproduced by the model (lines) with a calibrated value of ν_{PF}.

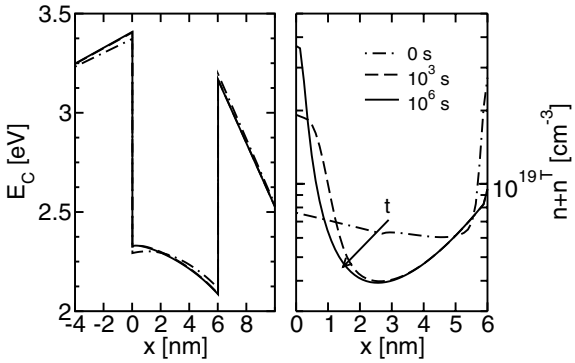


Fig. 9. Simulated evolution of the CB profile (left plot) and of the total charge distribution in the nitride of the 6/6/6 nm device of Fig. 8 at 300 K during the retention transient between the initial state and 10^6 s.

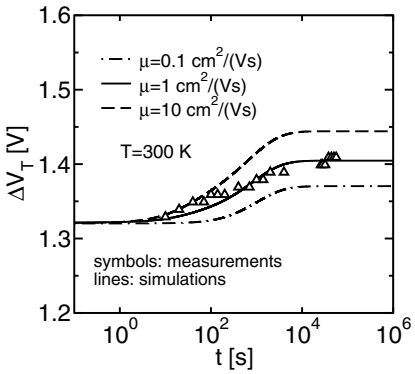


Fig. 10. Effect of μ in the nitride CB on the simulated retention transient (lines). Measurements (symbols) on the 6/6/6 nm device are the same of Fig. 8.

consistent with the measurements in [14] but different from those assumed in [7], consequently m_N^* and ν_T values are also different. Fig. 8 also reports the impact of the attempt–to–escape frequency for the PF emission, ν_{PF}, on ΔV_T: an increase of ν_{PF} accelerates the redistribution mechanisms. Instead the effect of μ (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 ΔV_T variations. Simulations reproduce the experiments at best with $\mu \simeq 1$ cm^2/(sV).

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 hight T) and the trap energy. Moreover it provides a simple explanation for the ΔV_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

[1] K. Kim and S. Y. Lee, "Memory Technology in the Future", *Microelectronic Engineering*, vol. 84, pp. 1976–1981, 2007.

[2] K. Prall, "Scaling Non–Volatile Memory Below 30 nm", *Proc. of NVSMW*, pp. 5–10, 2007.

[3] A. Furnémont *et al.*, "Physical Understanding and Modeling of SANOS Retention in Programmed State", *SSE*, vol. 52, pp. 577-583, 2008.

[4] Y. Wang and M. H. White, "An Analytical Retention Model for SONOS Nonvolatile Memory Devices in the Excess Electron State", *SSE*, vol. 49, pp. 97-107, 2005.

[5] S.-H. Gu *et al.*, "Numerical Simulation of Bottom Oxide Thickness Effect on Charge Retention in SONOS Flash Memory Cells", *IEEE TED*, vol. 54, n. 1, pp. 90-97, 2007.

[6] C. M. Compagnoni *et al.*, "Experimental Study of Data Retention in Nitride Memories by Temperature and Field Acceleration", *IEEE EDL*, vol. 28, n. 7, pp. 628–630, 2007.

[7] A. Arreghini *et al.*, "Characterization and Modeling of long term retention in SONOS Non Volatile Memories", *Proc. of the ESSDERC*, p. 406, 2007.

[8] I. Ay *et al.*, "Steady-State and Transient Photoconductivity in Hydrogeneted Amorphous Silicon Nitride Films", *Solar Energy Mat. & Solar Cells*, vol. 80, pp. 209-216, 2003.

[9] J. Frenkel, "On Pre-Breakdown Phenomena in Insulators and Electronic Semi–Conductors", *Phys. Rev.*, vol. 54, n. 8, pp. 647-648, 1938.

[10] D. L. Scharfetter and H. K. Gummel,, "Large–Signal Analysis of a Silicon Read Diode Oscillator", *IEEE TED*, vol. 16, n. 1, pp. 64-77, 1969.

[11] K. A. Nasyrov *et al.*, "Charge Transport Mechanism in Metal–Nitride–Oxide–Silicon Structure", *IEEE EDL*, vol. 23, n. 6, pp. 336-338, 2002.

[12] M. Naich *et al.*, "Exoelectron Emission Studies on Trap Spectrum in Ultrathin Amorphous Si_3N_4 Films", *SSE*, vol. 48, pp. 477-482, 2004.

[13] Y. Rozin *et al.*, "Novel Techniques for Data Retention and Leff Mesurements in Two Bit *micro*FLASH Memory Cells", *AIP Proceedings*, vol. 550, n. 8, pp. 181-185, 2001.

[14] A. Arreghini *et al.*, "Experimental Extraction of the Charge Centroid and of the Charge Type in the P/E Operations of the SONOS Memory Cells", *IEDM 2006 Tech. Digest*, p. 499-502, 2006.