Physical Modeling for Programming of TANOS Memories in the Fowler–Nordheim Regime

Christian Monzio Compagnoni, *Member, IEEE*, Aurelio Mauri, Salvatore Maria Amoroso, Alessandro Maconi, and Alessandro S. Spinelli, *Senior Member, IEEE*

***Abstract*—This paper presents a physics-based model that is able to describe the TANOS memory programming transients in the Fowler–Nordheim uniform tunneling regime across the bottom-oxide layer. The model carefully takes into consideration the trapping/detrapping processes in the nitride, the limited number of traps available for charge storage, and their spatial and energetic distribution. Results are in good agreement with experimental data on TANOS devices with different gate-stack compositions, considering a quite extended range of gate biases and times. The reduced gate-bias sensitivity of the programming transients with respect to the floating-gate cell is explained in terms of a finite number of nitride traps and a thinner extension of the nitride trapping region as the gate bias is increased. The model represents a valid contribution for the investigation of the achievable performances of the TANOS technology.**

***Index Terms*—Flash memories, semiconductor device modeling, TANOS memories, tunneling program/erase (P/E).**

I. Introduction

THE TANOS memory cell is considered today the most practical evolution of the floating-gate Flash cell for NAND architectures, allowing improved reliability and scaling perspectives [1]–[4]. Stress-induced leakage current immunity, strongly reduced cell-to-cell parasitic interference, and the possibility to decrease the thickness of the gate dielectric stack and, therefore, the program/erase (P/E) biases appear as the main promises of the TANOS technology [5], [6]. However, a clear quantitative assessment of these benefits and of the possible drawbacks is still lacking, and no detailed perspective of the achievable TANOS technology performances has been reported so far.

In this paper, we investigate the programming dynamics of TANOS memory devices, presenting a new physics-based model that is able to reproduce the threshold voltage (V_T) transients over an extended range of programming biases and times.

Manuscript received February 18, 2009; revised May 13, 2009. First published July 28, 2009; current version published August 21, 2009. This work was supported in part by the European Commission under FP7 Research Contract 214431 “GOSSAMER” and in part by MIUR under FIRB Project RBIP06YSJJ. The review of this paper was arranged by Editor C.-Y. Lu.

C. Monzio Compagnoni, S. M. Amoroso, and A. Maconi are with the Dipartimento di Elettronica e Informazione, Politecnico di Milano, 20133 Milano, Italy, and also with the Italian University Nano-Electronics Team, 20133 Milano, Italy (e-mail: monzio@elet.polimi.it).

A. Mauri is with the R&D—Technology Development, Numonyx, 20041 Agrate Brianza, Italy.

A. S. Spinelli is with the Dipartimento di Elettronica e Informazione, Politecnico di Milano, 20133 Milano, Italy, with the Italian University Nano-Electronics Team, 20133 Milano, Italy, and also with the Istituto di Fotonica e Nanotecnologie, Consiglio Nazionale delle Ricerche, 20133 Milano, Italy.

Digital Object Identifier 10.1109/TED.2009.2026315

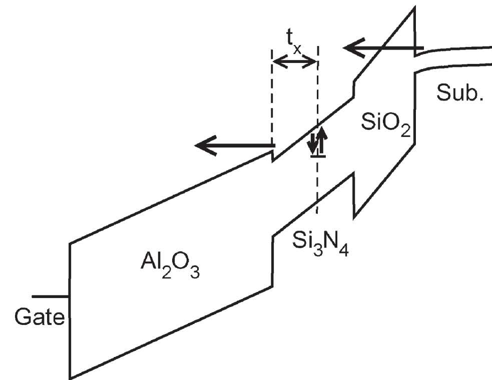

Fig. 1. Schematics for the band profile in the TANOS structure during a program operation, highlighting the electron fluxes taking place in the device.

The analysis that we present in this paper is more extended and complete with respect to that we reported in [7], starting from a simplified analytical model allowing the understanding of some basic properties of V_T transients in nitride memories. A more refined numerical model is then presented to achieve a good quantitative agreement between experimental and modeling results. The model includes an accurate description of the trapping/detrapping processes in the nitride layer, taking into account the finite number of traps available for charge storage, their energetic and spatial distribution, and the nonzero energy relaxation length of injected carriers [8], [9]. With all these features, the model explains the reduced gate control over V_T transients with respect to floating-gate cells, which affects the achievable programming performances. The good agreement between experimental and modeling results that is found on samples with different gate-stack compositions makes the model a valuable contribution for the investigation of the TANOS technology performances from the P/E standpoint.

II. First-Order Analytical Model for the Program Operation

Fig. 1 shows the band diagram along a TANOS memory device during programming at a high gate voltage V_G, giving rise to Fowler–Nordheim (FN) electron tunneling through the bottom-oxide layer. The tunneling current is strictly related to the electric field F in the bottom oxide, which can be straightforwardly calculated as

$$F = \frac{V_G - \Delta V_T}{EOT} \tag{1}$$

where we have assumed that the gate work function equals the silicon electron affinity and that a planar charge density Q

0018-9383/$26.00 © 2009 IEEE

is trapped inside the nitride at a distance t_x from the Al_2O_3 interface, giving rise to the threshold-voltage shift ΔV_T

$$\Delta V_T = -Q\left(\frac{t_{\rm HK}}{\epsilon_{\rm HK}} + \frac{t_x}{\epsilon_N}\right) = -Q\cdot\theta. \tag{2}$$

In the previous equations, ϵ_N and $\epsilon_{\rm HK}$ are the nitride and the Al_2O_3 dielectric constants, $t_{\rm HK}$ is the Al_2O_3 thickness, $\theta = t_{\rm HK}/\epsilon_{\rm HK} + t_x/\epsilon_N$ is the reciprocal of the capacitance from the trapping point to the gate, and EOT is the equivalent oxide thickness of the gate stack, given by

$$EOT = t_{\rm OX} + t_{\rm HK}\frac{\epsilon_{\rm ox}}{\epsilon_{\rm HK}} + t_N\frac{\epsilon_{\rm ox}}{\epsilon_N} \tag{3}$$

where $t_{\rm ox}$ and $\epsilon_{\rm ox}$ are the bottom-oxide thickness and dielectric constant, respectively, and t_N is the nitride thickness.

The density of electron-filled traps in the nitride $n'_t = -Q/q$ (units: $\rm cm^{-2}$) increases as a consequence of the capture of a part of the FN electron flow coming from the substrate, according to the relation [10]

$$\frac{dn'_t}{dt} = \frac{J}{q}\sigma\left(N_t - n'_t\right) - n'_t\langle e\rangle \tag{4}$$

where σ is the capture cross section of the nitride traps (units: $\rm cm^2$), N_t is the total trap density (units: $\rm cm^{-2}$), and $\langle e\rangle$ is the electron detrapping rate (units: $\rm s^{-1}$), e.g., by thermal or tunneling emission. Note that the product $\sigma \times N_t$ should be lower than one for the previous equation to keep its validity. Taking the time derivative of (2) and combining it with (4), the following equation for the time evolution of ΔV_T can be obtained:

$$\begin{aligned}\frac{d\Delta V_T}{dt} &= \left[J\sigma\left(N_t - n'_t\right) - qn'_t\langle e\rangle\right]\cdot\theta \\ &= J\sigma\left(N_t\theta - \frac{\Delta V_T}{q}\right) - \Delta V_T\langle e\rangle\end{aligned} \tag{5}$$

where we have considered θ as a time-independent constant, i.e., any variation of the charge centroid in the nitride with time was neglected. In order to solve (5), the following FN formula for the tunneling current through the bottom oxide can be used:

$$J = AF^2 e^{-B/F} \tag{6}$$

where A and B depend on the physical parameters of the potential barrier [11], [12]. By means of (1) and (6), (5) can be expressed as a function of the electric field F

$$\frac{dF}{dt} = -AF^2 e^{-B/F}\left[\frac{\sigma N_t\theta}{EOT} - \frac{\sigma}{q}\left(\frac{V_G}{EOT} - F\right)\right] + \left(\frac{V_G}{EOT} - F\right)\langle e\rangle. \tag{7}$$

A. Large Number of Trapping Sites and No Detrapping

Neglecting electron detrapping (i.e., putting $\langle e\rangle = 0$) and assuming that $N_t \gg n'_t$, only the first term in the square brackets on the RHS of (7) can be considered, obtaining

$$\frac{dF}{dt} = -\frac{\sigma N_t\theta}{EOT}AF^2 e^{-B/F} \tag{8}$$

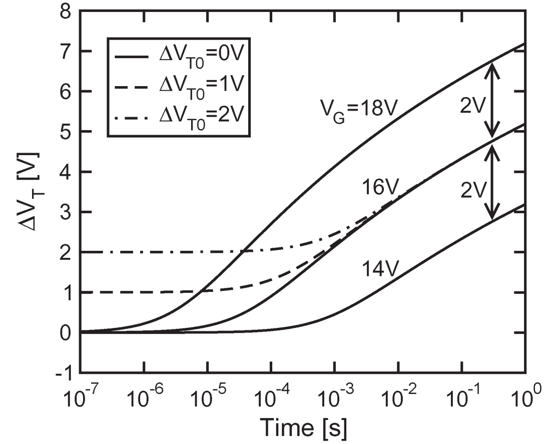

Fig. 2. Programming V_T transients calculated by (10) for $V_G = 14$, 16, and 18 V. Results for $V_G = 16$ V are also shown for $\Delta V_{T0} = 1$ and 2 V.

TABLE I
PARAMETERS USED TO CALCULATE THE ΔV_T TRANSIENTS IN FIG. 2

Parameter	Value
t_{OX}	4.5 nm
t_N	6 nm
t_x	3 nm
t_{HK}	15 nm
ϵ_{OX}	$3.9\epsilon_0$
ϵ_N	$6\epsilon_0$
ϵ_{HK}	$10.3\epsilon_0$
σ	10^{-16} cm^2
N_T	10^{14} cm^{-2}
A	10^{-5} A/V^2
B	2.4×10^8 V/cm

which can be straightforwardly integrated to obtain the time evolution of the electric field F during programming

$$F = \frac{B}{\ln\left(\frac{\sigma N_t\theta AB}{EOT}t + e^{B/F_i}\right)} \tag{9}$$

where F_i is the electric field at the beginning of the program operation ($t = 0$), e.g., $F_i = V_G/EOT$ assuming that no charge is initially stored in the nitride. Using (1), (9) gives the ΔV_T evolution with time

$$\Delta V_T = V_G - \frac{EOT\cdot B}{\ln\left(\frac{\sigma N_t\theta AB}{EOT}t + e^{B/F_i}\right)}. \tag{10}$$

Fig. 2 shows the calculated ΔV_T transients obtained at V_G equal to 14, 16, and 18 V, assuming the device parameters reported in Table I. For a time t that is sufficiently long to lose the ΔV_T dependence on the different initial electric field F_i, the curves are only vertically shifted by the difference in their programming V_G. This means that, for a fixed time t, the gate sensitivity factor $GSF = \partial\Delta V_T/\partial V_G$ for the transients is equal to one, as normally obtained on floating-gate memory cells. Note that this result does not derive from the trapping of all the electrons injected into the nitride layer but from the hypothesis of a very large number of trapping sites available for electron storage. From (1), in fact, when the programming voltage is increased from V_{G1} to V_{G2}, the same bottom-oxide electric field F is obtained when the difference between the threshold-voltage shift obtained at $V_{G,2}(\Delta V_{T,2})$

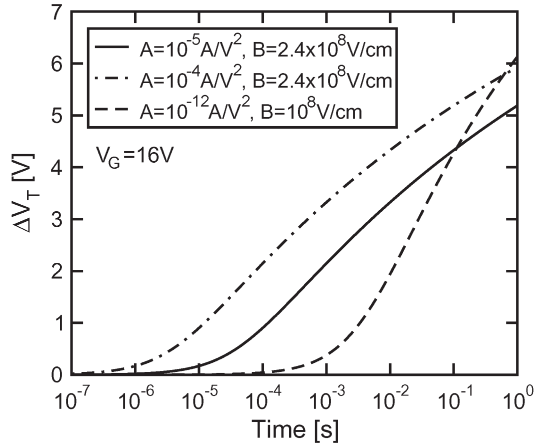

Fig. 3. Programming V_T transients calculated by (10) for different values of the tunneling parameters A and B.

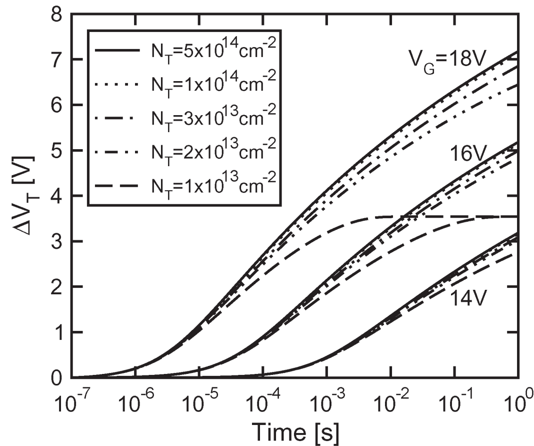

Fig. 4. Programming V_T transients calculated by the numerical integration of (11) for different N_T values.

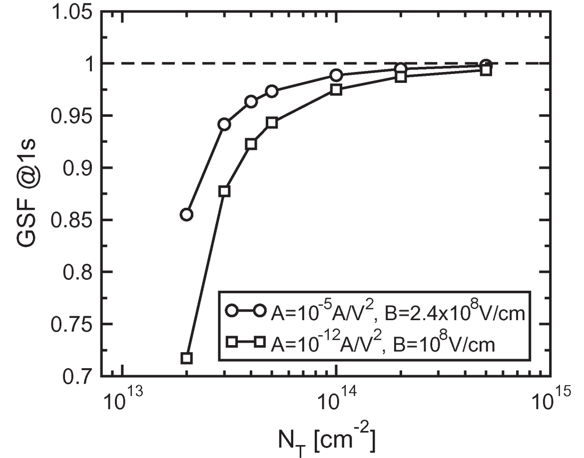

Fig. 5. GSF extracted from Fig. 4 at $t = 1$ s as a function of the trap density N_T.

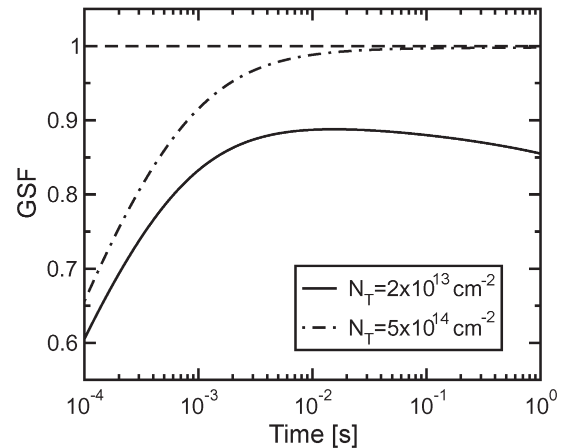

Fig. 6. GSF as a function of the programming time for $N_T = 5 \times 10^{14}$ cm^{-2} and $N_T = 2 \times 10^{13}$ cm^{-2}.

and $V_{G,1}(\Delta V_{T,1})$ is equal to $\Delta V_{T,2} - \Delta V_{T,1} = V_{G2} - V_{G1}$. When this condition is reached, the same ΔV_T evolution takes place in both cases, as no limitations in the trapping dynamics come from the charge that has already been stored, owing to the large number of available trapping sites. As a result, the ΔV_T transients are only vertically shifted by $V_{G2} - V_{G1}$.

Fig. 2 also shows that, for a fixed programming voltage V_G, the V_T transients display a converging behavior when different ΔV_T values are assumed at $t = 0$ (e.g., $\Delta V_{T0} = 1$ and 2 V in the figure), as usually obtained in floating-gate memory devices. Moreover, (10) states that the shape of these transients is affected only by a change in EOT or B. In fact, any change in σ, θ, A, or N_T determines only a horizontal shift of the curves in the logarithmic time axis, as shown in Fig. 3 when changing A from 10^{-5} to 10^{-4} A/V^2 for a fixed B. Instead, the figure shows that decreasing B from 2.4×10^8 to 10^8 V/cm increases the slope of the ΔV_T transient: This is due to the less field-dependent tunneling current characteristics that result from the reduction of B, allowing a smaller decrease of the programming current for a given charge stored in the nitride layer. Finally, note that the change in shape of the ΔV_T curve does not modify the GSF of the transients, which is still equal to one.

B. Effect of the Finite Trap Density

When the hypothesis of $N_t \gg n_t'$ is removed, in the case of no charge detrapping, (7) becomes

$$\frac{dF}{dt} = -\sigma A F^2 e^{-B/F} \left[\frac{N_t \theta}{EOT} - \frac{1}{q} \left(\frac{V_G}{EOT} - F \right) \right]. \quad (11)$$

Fig. 4 shows the ΔV_T transients calculated by the numerical integration of this equation for different N_T values, maintaining the $\sigma \cdot N_t$ product constant. The reduction of N_T from 5×10^{14} to 2×10^{13} cm^{-2} determines a lowering of the transients as the achieved ΔV_T gets higher, as a result of the smaller number of empty traps available [13], [14]. The corresponding GSF is shown in Fig. 5 for $t = 1$ s, where a value nearly equal to 0.85 is obtained for $N_T = 2 \times 10^{13}$ cm^{-2}. A slightly larger value of the GSF can be extracted for smaller programming times, as shown in Fig. 6, considering the time range between 1 ms and 1 s. Below 1 ms, instead, the GSF rapidly drops to zero, as the programming time is not long enough to make the transients lose their dependence on the initial F_i, as a ΔV_T that is nearly equal to 0 is maintained at $V_G = 14$ V. For comparison, the

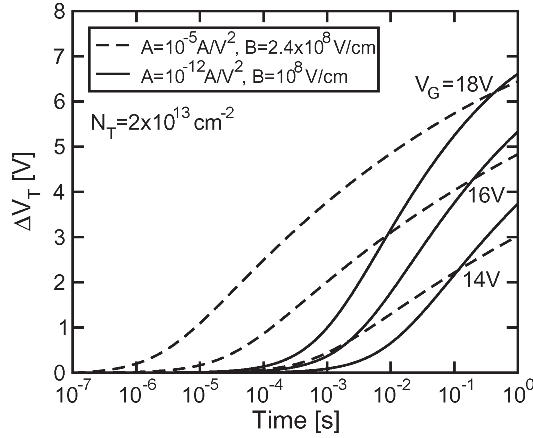

Fig. 7. Programming V_T transients calculated for $B = 2.4 \times 10^8$ V/cm and $B = 10^8$ V/cm in the case of $N_T = 2 \times 10^{13}$ cm^{-2}.

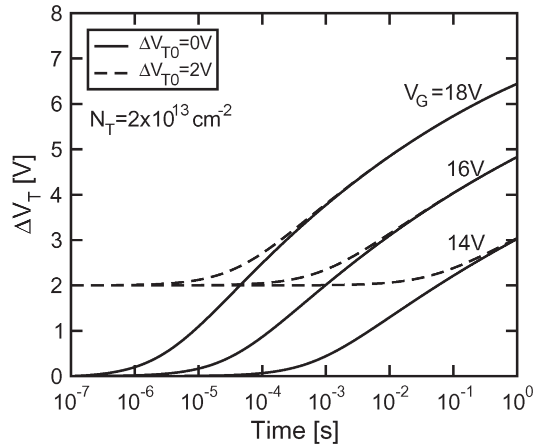

Fig. 8. Programming V_T transients calculated for $N_T = 2 \times 10^{13}$ cm^{-2} and a different initial ΔV_{T0}.

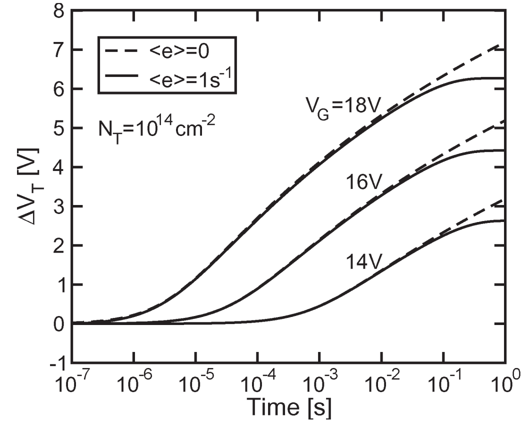

Fig. 9. Programming V_T transients neglecting electron detrapping and assuming a constant $\langle e \rangle = 1$ s^{-1}.

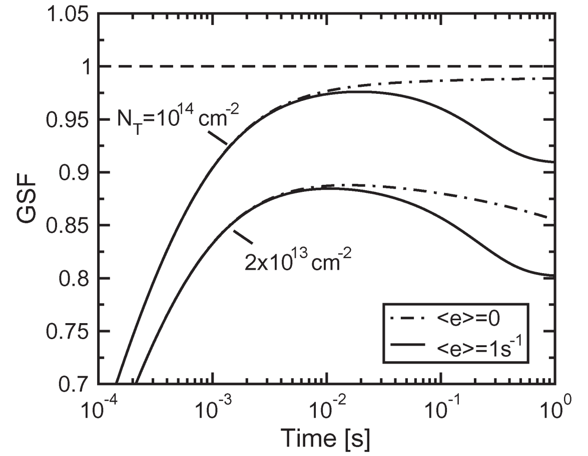

Fig. 10. Effect of electron detrapping on the GSF as a function of programming time.

figure also shows the GSF for $N_T = 5 \times 10^{14}$ cm^{-2}, displaying a convergent behavior toward one for increasing programming times. In the case of very low N_T values, a clear saturation of ΔV_T can be achieved, as shown in Fig. 4 for $N_T = 10^{13}$ cm^{-2}. The saturation level does not depend on the programming voltage and is determined by the filling of all the available traps in the nitride. As a consequence, when the ΔV_T curves reach the saturation level, the GSF rapidly falls to zero.

As in the case of a very large number of available traps, the gate sensitivity factor does not depend on A or σ, as these terms appear as multiplying factors in the second member of (11), therefore determining only a horizontal shift of the ΔV_T curves in the logarithmic time axis. Instead, Fig. 5 shows that the reduction of B gives rise to a slight decrease of the GSF, due to the steeper ΔV_T transients (see Fig. 7) and the flatter $J{-}F$ characteristics. Finally, Fig. 8 shows that the converging behavior of the ΔV_T transients for a different initial ΔV_{T0} is preserved, even taking into account the finite number of traps in the nitride. In fact, for a fixed programming voltage V_G, the different ΔV_{T0} modifies the initial field F_i and the short time behavior of the ΔV_T transients but does not modify the regime (F_i-independent) behavior of ΔV_T, as the number of available traps for a given $\Delta V_T > \Delta V_{T0}$ is not changed.

C. Effect of Nonzero $\langle e \rangle$

Fig. 9 shows the ΔV_T transients calculated by (7) when a constant electron detrapping rate $\langle e \rangle = 1$ s^{-1} is assumed, with $N_T = 10^{14}$ cm^{-2}. Electron detrapping makes the transients saturate at a maximum ΔV_T level which is not determined by the filling of all the available traps but by an equal rate of trapping and detrapping processes. As a consequence, the saturation level depends on the programming voltage, while the time at which saturation occurs is barely affected by V_G. The effect of electron detrapping on the GSF is shown in Fig. 10: For short programming times, the curves for $\langle e \rangle \neq 0$ coincide with those for $\langle e \rangle = 0$, while they slightly drop to a lower value, determined by the separation of the maximum ΔV_T levels for the different V_G, when saturation occurs.

III. Physics-Based Numerical Model for Programming

The analysis of the program operation presented in the previous section allowed a first-order evaluation of the dependence of the programming ΔV_T transients on the main physical and device parameters of TANOS memories. A reduced gate control on the threshold-voltage shift achieved at a fixed programming time was shown to appear as a result of the finite number

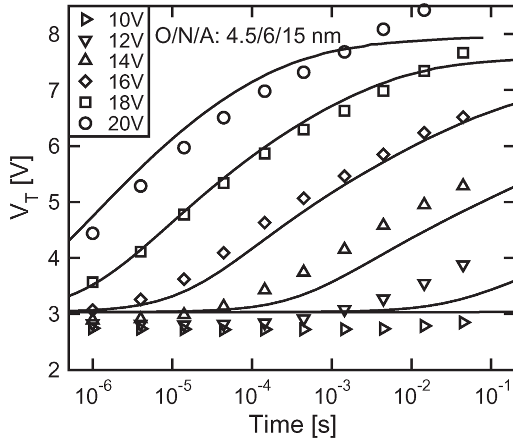

Fig. 11. Comparison between (symbols) experimental data and (lines) modeling results obtained by the numerical physics-based model.

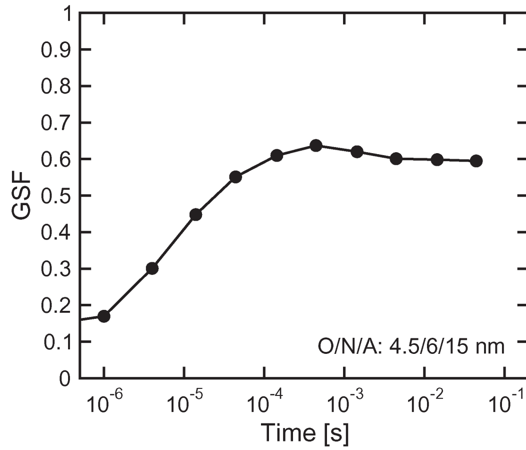

Fig. 12. GSF extracted from the experimental data in Fig. 11 as a function of the programming time. Only the transients at $V_G = 14$, 16, and 18 V were used for the calculation of the GSF.

of traps in the nitride layer. This reduced gate control was quantified by means of the GSF, representing a fundamental parameter for the determination of the achievable programming performances of the TANOS technology. A value of this parameter lower than one represents a drawback of nitride with respect to floating-gate storage, critically limiting the possibility to scale the programming voltages.

Fig. 11 shows the experimental programming transients at different V_G's ranging from 10 to 20 V on a TANOS device with 4.5 nm of bottom oxide, 6 nm of nitride, and 15 nm of Al_2O_3. Fig. 12 shows the GSF extracted as a function of time by fitting the $V_T - V_G$ relation of the curves for $V_G = 14$, 16, and 18 V: A maximum value of 0.65 is extracted, which is in the 0.5–0.7 range commonly reported by other authors [3], [13], [15], highlighting a quite low gate control over ΔV_T with respect to the floating-gate case. As V_T shifts as large as 5 V are observed for $V_G = 20$ V, N_T should be larger than $N_T = 2 \times 10^{13}$ cm^{-2} in order to avoid an early saturation of the transients. This means that the simplified model presented in the previous section cannot completely describe the programming transients in our devices, as GSFs larger than 0.85 are shown in Fig. 5 for the standard B parameter of the $J - F$ characteristics. The possibility to further reduce the sensitivity factor by reducing B would require an extreme modification of the $J - F$ relation, which can be obtained only by engineering the bottom dielectric tunneling barrier. In order to obtain a good agreement between experimental and modeling results, we developed a new physics-based model [7] that is able to overcome the main limitation of the analytical model previously presented, i.e., the inability to carefully deal with the distributed electron trapping inside the nitride.

A. Numerical Model for the Program Operation

In order to remove the approximations assumed in the analytical model of Section II, at a fixed time t during programming, the potential profile along the 1-D TANOS stack is calculated by self-consistently solving the Schrödinger–Poisson equations, in order to carefully evaluate the device electrostatics, and the electron tunneling current through the bottom oxide is calculated using the WKB approximation, considering quasi-bound states in the substrate. Different from simplified models for the trapping processes [13], [16], the drift-diffusion equation is solved to evaluate the electron concentration in the nitride conduction band [10], [17], including the possibility for carrier capture and release in the nitride traps. Electron release from previously filled traps by both thermal emission and tunneling was included, considering in the latter case both detrapping to the nitride conduction band and directly to the gate through the top dielectric material. When solving the drift-diffusion equations, at the nitride–alumina interface, the presence of the conduction band discontinuity was accounted for, considering the output current as the product of the electron drift term and the tunneling transparency through the Al_2O_3 potential barrier. As, during high-voltage programming, electrons enter the nitride with a large energy (see the band diagram in Fig. 1), we introduced an energy relaxation rate λ_E (in eV/nm) (not included in [17]) to describe their thermalization to the nitride conduction band [8], [9], neglecting the trapping of the high-energy electron flow in the nitride region between the bottom-oxide interface and the thermalization point of the carriers. This latter assumption corresponds to considering the trapping cross section a rapidly decreasing function of energy [18]. Further numerical details on the model can be found in the Appendix.

B. Results

Fig. 11 shows that a good agreement can be obtained between experimental data and the results from our new physics-based numerical model, using the parameters reported in Table II. Note that nitride and alumina dielectric constants and electron affinities were chosen in accordance to the results presented in [19] and were not treated as fitting parameters. Moreover, the bottom-oxide, nitride, and alumina thicknesses given by TEM measurements were used. Fig. 13 shows that, with the same set of parameters, a good agreement between experimental and modeling results can be obtained also on TANOS devices with a different stack composition. A Gaussian energy distribution of the nitride traps was assumed, extracting its average distance from the nitride conduction band and its standard deviation when fitting the available data with simulation results,

TABLE II
MAIN PARAMETERS USED FOR THE SIMULATIONS

Parameter	Value
Nitride relative dielectric constant	6
Al_2O_3 relative dielectric constant	10.3
Nitride electron affinity	2.05 eV
Al_2O_3 electron affinity	1.25 eV
TaN work function	4.76 eV
Average traps energy depth	1.9 eV
Traps energy spread	0.12 eV
Traps cross-section	8×10^{-15} cm^2
Escape frequency from the traps	2×10^{10} Hz
Electron energy-loss rate	3.1 eV/nm

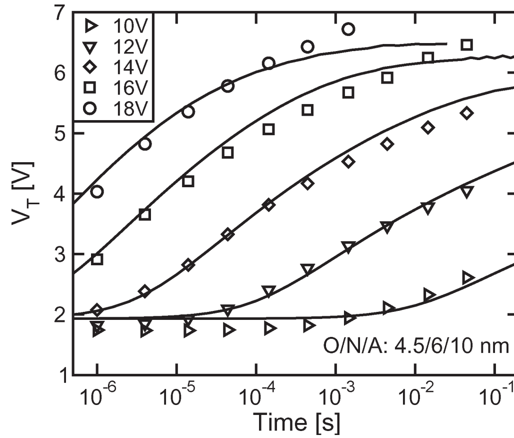

Fig. 13. Same as in Fig. 11 but for a TANOS device with a different stack composition.

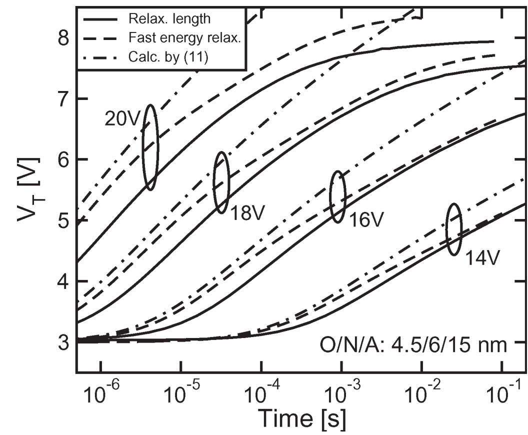

Fig. 14. Effect of the nonzero energy relaxation length on the V_T transients. The curves calculated by means of (11) are also shown.

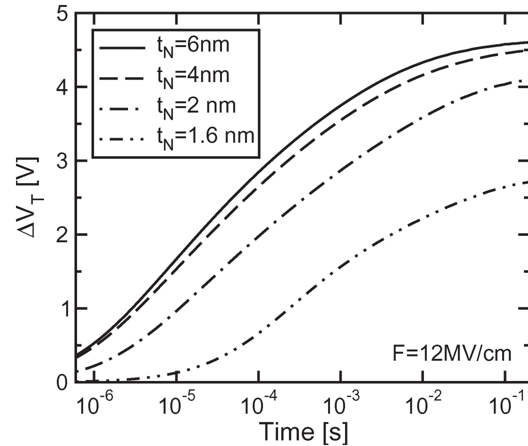

Fig. 15. ΔV_T transients for a TANOS device with 4.5 nm of bottom oxide, 15 nm of Al_2O_3, and different thicknesses of the nitride layer. The same electric field $F = 12$ MV/cm in the bottom oxide at the beginning of the program operation was assumed.

obtaining 1.9 and 0.12 eV, respectively. A hyperbolic cosine profile was instead extracted for the spatial distribution of the traps (reported in [7]), highlighting a larger trap density at the nitride–bottom-oxide and nitride–alumina interfaces [6], [17]. The effect of the nonzero energy relaxation length on the programming transients is shown in Fig. 14, highlighting a further reduction of the GSF with respect to the case of an instantaneous energy relaxation of electrons when they enter the nitride. In fact, the introduction of λ_E and the assumption that only thermalized electrons can be trapped make only a fraction of the nitride "active" in the trapping process. Moreover, the thickness of this layer decreases as the programming bias is increased, due to the larger energy of electrons entering the nitride, making the V_T transients closer. Fig. 14 also shows the programming transients calculated by means of the analytical model presented in Section II, neglecting electron detrapping and assuming the parameters in Table II [in order to correctly take into account the gate work function and the substrate potential drop at inversion, 1 V was subtracted from V_G using (11)]. The curves nearly match the results obtained when an energy relaxation length equal to zero is assumed, considering ΔV_T values sufficiently below the saturation level. This confirms that the analytical model in Section II can be used to have first-order results on the programming transients.

Finally, note that our model can explain the experimental evidences reported in [6], [20]–[22], stating that trapping into very thin nitride layers is negligible. In fact, Fig. 15 shows that, for a fixed electric field at the beginning of the program operation equal to 12 MV/cm, the reduction of the nitride thickness results into lower programming transients, with a rapid decrease of the achievable ΔV_T when the nitride layer is reduced below 2 nm. This behavior can be explained considering that, when the nitride thickness approaches the energy relaxation length for electrons entering the nitride, trapping is critically reduced. Moreover, as a limiting case, our model predicts that, for nitride thicknesses below 1.6 nm, no trapping takes place in the nitride, as electrons leak through the alumina layer before relaxing their energy.

As a final remark, note that all the previous modeling results were calculated assuming that no trapping takes place in the bottom oxide and alumina, considering them as ideal trap-free dielectrics. However, the possibility for the alumina to present a nonnegligible density of trapping sites, strictly dependent on its deposition and annealing process conditions, has been previously reported [19], [23]. Trapping in the Al_2O_3 layer may limit the possibility to observe the clear saturating behavior of the programming transients that is predicted at high V_G by our model in Figs. 11 and 13 and in the possibility for small V_T variations even for nitride thicknesses below 1.6 nm. More important, the role of alumina should be carefully investigated referring to the TANOS device working conditions, highlighting its impact on the technology reliability.

IV. Conclusion

This paper has presented a physics-based model that is able to describe high-voltage programming on TANOS memories, representing a valid contribution for the investigation of the achievable performances of this technology. The modeling results are in good agreement with experimental data on TANOS devices with different gate-stack compositions, considering an extended range of gate biases and times. In particular, the reduced gate-bias sensitivity of the programming transients with respect to the floating-gate cell was explained in terms of a finite number of nitride traps and a thinner extension of the nitride trapping region as the gate bias is increased.

Appendix

In order to evaluate the free electron charge in the nitride conduction band self-consistently with trapping and detrapping, the following system of equations should be solved:

$$\begin{cases} J(x,t) = qn_c(x,t)\mu_n F(x,t) + qD_n \frac{\partial n_c(x,t)}{\partial x} \\ \frac{\partial n_c(x,t)}{\partial t} = \frac{1}{q}\frac{\partial J(x,t)}{\partial x} - \frac{\partial n_t(x,t)}{\partial t} \\ \frac{\partial n_t(x,t)}{\partial t} = \frac{J(x,t)}{q}\sigma\left(N_t - n_t(x,t)\right) - n_t(x,t)\langle e\rangle \end{cases} \quad (12)$$

describing the electron drift-diffusion, continuity, and trapping/detrapping, respectively. In the previous equations, N_t and n_t are now the total trap density and the filled trap density per unit volume (units: cm^{-3}), n_c is the electron density in the nitride conduction band, and μ_n and D_n are the electron mobility and diffusion coefficient, respectively. For the solution of the system, the following boundary conditions are applied:

$$J(x_r,t) = qn_c(x_r,t)\mu_n F(x_r,t) + qD_n \frac{\partial n_c(x_r,t)}{\partial x} + J_{\text{in}}$$

$$J(x_2,t) = qn_c(x_2,t)\mu_n F(x_2,t)T$$

$$n_c(x,0) = 0$$

$$n_t(x,0) = 0$$

where x_2 is the coordinate of the nitride–alumina interface and x_r is that of the point where electrons entering the nitride reach the bottom of the conduction band according to the energy relaxation rate λ_E. In correspondence to this point, all the electron currents injected from the substrate J_{in} are summed to the electron drift-diffusion current.

The solution of the system (12) resulted in the calculated curves in Figs. 11, 13, and 15. For electron mobilities as low as $\mu_n = 10^{-3}$ $\mathrm{cm}^2/\mathrm{Vs}$ and nitride/alumina tunneling barriers as high as 1.2 eV, a negligible impact of the solution of the drift-diffusion equation [the first equation in (12)] is found. This means that, for the explored V_G and programming-time range, the solution of the second and the third equations of the system (with $n_c = 0$) is enough to obtain the programming transients, revealing a negligible impact of the free electron charge accumulation in the nitride conduction band.

Acknowledgment

The authors would like to thank A. L. Lacaita from Politecnico di Milano and P. Cappelletti, E. Camerlenghi, R. Bez, and L. Baldi from Numonyx for the discussions and support.

References

[1] C. H. Lee, K. I. Choi, M. K. Cho, Y. H. Song, K. C. Park, and K. Kim, "A novel SONOS structure of $SiO_2/SiN/Al_2O_3$ with TaN metal gate for multi-giga bit Flash memories," in *IEDM Tech. Dig.*, 2003, pp. 613–616.

[2] Y. Shin, J. Choi, C. Kang, C. Lee, K.-T. Park, J.-S. Lee, J. Sel, V. Kim, B. Choi, J. Sim, D. Kim, H.-J. Cho, and K. Kim, "A novel NAND-type MONOS memory using 63 nm process technology for multi-gigabit Flash EEPROM," in *IEDM Tech. Dig.*, 2005, pp. 337–340.

[3] Y. Park, J. Choi, C. Kang, C. Lee, Y. Shin, B. Choi, J. Kim, S. Jeon, J. Sel, J. Park, K. Choi, T. Yoo, J. Sim, and K. Kim, "Highly manufacturable 32 Gb multi-level NAND Flash memory with 0.0098 μm^2 cell size using TANOS (Si-Oxide–Al_2O_3–TaN) cell technology," in *IEDM Tech. Dig.*, 2006, pp. 29–32.

[4] J. S. Sim, J. Park, C. Kang, W. Jung, Y. Shin, J. Kim, J. Sel, C. Lee, S. Jeon, Y. Jeong, Y. Park, J. Choi, and W.-S. Lee, "Self aligned trap-shallow trench isolation scheme for the reliability of TANOS (TaN/AlO/SiN/Oxide/Si) NAND Flash memory," in *Proc. Non-Volatile Semicond. Memory Workshop*, 2007, pp. 110–111.

[5] C.-H. Lee, J. Choi, C. Kang, Y. Shin, J.-S. Lee, J. Sel, J. Sim, S. Jeon, B.-I. Choe, D. Bae, K. Park, and K. Kim, "Multi-level NAND Flash memory with 63 nm-node TANOS (Si-Oxide–SiN–Al_2O_3–TaN) cell structure," in *VLSI Symp. Tech. Dig.*, 2006, pp. 21–22.

[6] T. Ishida, T. Mine, D. Hisamoto, Y. Shimamoto, and R. Yamada, "Anomalous electron storage decrease in MONOS nitride layers thinner than 4 nm," *IEEE Electron Device Lett.*, vol. 29, no. 8, pp. 920–922, Aug. 2008.

[7] A. Mauri, C. Monzio Compagnoni, S. Amoroso, A. Maconi, F. Cattaneo, A. Benvenuti, A. S. Spinelli, and A. L. Lacaita, "A new physics-based model for TANOS memories program/erase," in *IEDM Tech. Dig.*, 2008, pp. 555–558.

[8] E. Suzuki, Y. Hayashi, and H. Yanai, "Transport processes of electrons in MNOS structures," *J. Appl. Phys.*, vol. 50, no. 11, pp. 7001–7006, Nov. 1979.

[9] T. Tomita, Y. Kamakura, and K. Taniguchi, "Energy relaxation length for ballistic electron transport in SiO_2," *Phys. Stat. Sol.*, vol. 204, no. 1, pp. 129–132, Nov. 1997.

[10] P. C. Arnett, "Transient conduction in insulators at high fields," *J. Appl. Phys.*, vol. 46, no. 12, pp. 5236–5243, Dec. 1975.

[11] M. Lenzlinger and E. H. Snow, "Fowler–Nordheim tunneling into thermally grown SiO_2," *J. Appl. Phys.*, vol. 40, no. 1, pp. 278–283, Jan. 1969.

[12] Y. L. Chiou, J. F. Gambino, and M. Mohammad, "Determination of the Fowler–Nordheim tunneling parameters from the Fowler–Nordheim plot," *Solid State Electron.*, vol. 45, no. 10, pp. 1787–1791, Oct. 2001.

[13] E.-S. Choi, H.-S. Yoo, K.-H. Park, S.-J. Kim, J.-R. Ahn, M.-S. Lee, Y.-O. Hong, S.-G. Kim, J.-C. Om, M.-S. Joo, S.-H. Pyi, S.-S. Lee, S.-K. Lee, and G.-H. Bae, "Modeling and characterization of program/erasure speed and retention of TiN-gate MANOS (Si-Oxide–SiN_x–Al_2O_3–Metal gate) cells for NAND Flash memory," in *Proc. Non-Volatile Semicond. Memory Workshop*, 2007, pp. 83–84.

[14] A. Arreghini, F. Driussi, D. Esseni, L. Selmi, M. J. van Duuren, and R. van Schaijk, "Experimental extraction of the charge centroid and of the charge type in the P/E operation of SONOS memory cells," in *IEDM Tech. Dig.*, 2006, pp. 499–502.

[15] A. Furnemont, M. Rosmeulen, A. Cacciato, L. Breuil, K. De Meyer, H. Maes, and J. Van Houdt, "Physical understanding of SANOS disturbs and VARIOT engineered barrier as a solution," in *Proc. Non-Volatile Semicond. Memory Workshop*, 2007, pp. 94–95.

[16] A. Furnemont, M. Rosmeulen, A. Cacciato, L. Breuil, K. De Meyer, H. Maes, and J. Van Houdt, "A consistent model for the SANOS programming operation," in *Proc. Non-Volatile Semicond. Memory Workshop*, 2007, pp. 96–97.

[17] A. Paul, C. Sridhar, and S. Mahapatra, "Comprehensive simulation of program, erase and retention in charge trapping Flash memories," in *IEDM Tech. Dig.*, 2006, pp. 393–396.

[18] M. Lax, "Cascade capture of electrons in solids," *Phys. Rev.*, vol. 119, no. 5, pp. 1502–1523, Sep. 1960.

[19] C. Scozzari, G. Albini, M. Alessandri, S. Amoroso, P. Bacciaglia, A. Del Vitto, G. Ghidini, A. Grossi, A. Mauri, A. Modelli, R. Piagge, A. Sebastiani, and P. Tessariol, "Al_2O_3 optimization for charge trap memory application," in *Proc. ULIS*, 2008, pp. 191–194.

[20] H.-T. Lue, P.-Y. Du, S.-Y. Wang, E.-K. Lai, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, "A novel gate-sensing and channel-sensing transient analysis method for real-time monitoring of charge vertical location in SONOS-type devices and its applications in reliability studies," in *Proc. IRPS*, 2007, pp. 177–183.

[21] S.-Y. Wang, H.-T. Lue, P.-Y. Du, C.-W. Liao, E.-K. Lai, S.-C. Lai, L.-W. Yang, T. Yang, K.-C. Chen, J. Gong, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, "Reliability and processing effects of bandgap-engineered SONOS (BE-SONOS) Flash memory and study of the gate-stack scaling capability," *IEEE Trans. Device Mater. Rel.*, vol. 8, no. 2, pp. 416–425, Jun. 2008.

[22] C.-H. An, M. Soo Lee, and H. Kim, "Effects of Si_3N_4 thickness on the electrical properties of oxide–nitride–oxide tunneling dielectrics," *J. Electrochem. Soc.*, vol. 155, no. 11, pp. G247–G252, 2008.

[23] M. Specht, H. Reisinger, F. Hofmann, T. Schulz, E. Landgraf, R. J. Luyken, W. Rosner, M. Grieb, and L. Risch, "Charge trapping memory structures with Al_2O_3 trapping dielectric for high-temperature applications," *Solid State Electron.*, vol. 49, no. 5, pp. 716–720, May 2005.

Christian Monzio Compagnoni (M'08) was born in Busto Arsizio, Italy, in 1976. He received the Laurea (*cum laude*) degree in electronics engineering and the Ph.D. degree in information technology from the Politecnico di Milano, Milano, Italy, in 2001 and 2005, respectively.

Since 2002, he has been with the Dipartimento di Elettronica e Informazione, Politecnico di Milano, where he became an Assistant Professor in 2006. He is also with the Italian University Nano-Electronics Team, Milano, Italy. His research activities include characterization and modeling of advanced nonvolatile memories and MOS devices.

Dr. Monzio Compagnoni received the Outstanding Paper Award at the IRPS in 2008 and was a member of the memory committee of the IRPS in 2009.

Aurelio Mauri was born in 1969. After classical studies, he received the M.S. degree in plasma physics (*cum laude*) from the University of Milano, Milano, Italy, in 1995.

Then, he spent a research period with Rutherford Appleton Laboratory, U.K., and with ENEA, Frascati, Italy, on the characterization and modeling of high-density plasmas produced by laser confinement. In 1996, he started to work for a semiconductor company focused on the chemistry treatment of silicon surfaces. From 1998 to 2000, he spent a "nonworking" time helping the young generation. After that period, he came back to the semiconductor industries (STMicroelectronics) in the power device development. In 2004, he joined the nonvolatile technology development of STMicroelectronics in the TCAD group working particularly on NOR/NAND memories and then with the same function in the R&D—Technology Development, Numonyx, Agrate Brianza, Italy. He is a coauthor of more than 20 scientific conference papers on different physics topics.

Salvatore Maria Amoroso was born in Catania, Italy, in 1983. He received the B.S. and M.S. degrees in physics engineering from the Politecnico di Milano, Milano, Italy, in 2005 and 2008, respectively, where he is currently working toward the Ph.D. degree in information technology in the Dipartimento di Elettronica e Informazione.

He is also with the Italian University Nano-Electronics Team, Milano, Italy. His research activities include modeling and numerical simulation of semiconductor devices, with particular interest on innovative nonvolatile memories.

Alessandro Maconi was born in Carate Brianza, Italy, in 1983. He received the Laurea degree in electronics engineering from the Politecnico di Milano, Milano, Italy, in 2008, where he is currently working toward the Ph.D. degree in the Dipartimento di Elettronica e Informazione.

He is also with the Italian University Nano-Electronics Team, Milano, Italy. His research activities mainly involve characterization and modeling of advanced nonvolatile memories, with particular interest to TANOS memories.

Alessandro S. Spinelli (M'99–SM'07) was born in Bergamo, Italy, in 1966. He received the Laurea (*cum laude*) and Ph.D. degrees in electronics engineering from the Politecnico di Milano, Milano, Italy.

In 1995, he was a Visiting Scholar at the University of Tennessee Space Institute, Tullahoma, TN, where he worked on single-molecule detection in solutions, and in 1996, he was a Consultant with the Central Department of Research and Development, STMicroelectronics, Agrate Brianza, Italy. In 1997, he became an Assistant Professor with the Politecnico di Milano, where he is currently a Full Professor of electronics. He was with the Università degli Studi dell'Insubria, Como, Italy, as an Associate Professor of electronics in 1998. In 2001, he was a Visiting Professor at the Institute National Polytechnique de Grenoble, Grenoble, France. He is also currently with the Italian University Nano-Electronics Team, Milano, Italy, and also with the Istituto di Fotonica e Nanotecnologie, Consiglio Nazionale delle Ricerche, Milano. He has conducted experimental and theoretical research in electronics instrumentation and microelectronics and coauthored more than 100 papers published in international journals or presented at international conferences. His current research interests include experimental characterization and modeling of nonvolatile memory cell reliability, development of innovative nonvolatile memories, modeling and simulation of nanoscale MOS devices, and circuit development for biological signal readout.

He has served in the technical committees of the IEDM and IRPS conferences.