Experimental and Simulation Analysis of Program/Retention Transients in Silicon Nitride-Based NVM Cells

Elisa Vianello, Francesco Driussi, Antonio Arreghini, Pierpaolo Palestri, David Esseni, Luca Selmi, Nader Akil, Michiel J. van Duuren, and Dusan S. Golubović

Abstract—A new characterization technique and an improved model for charge injection and transport through ONO gate stacks are used to investigate the program/retention sequence of silicon nitride-based (SONOS/TANOS) nonvolatile memories. The model accounts for drift-diffusion transport in the conduction band of silicon nitride (SiN). A priori assumptions on the spatial distribution of the charge at the beginning of the program/retention operations are not needed. We show that the carrier transport in the SiN layer impacts the spatial distribution of the trapped charge and, consequently, several aspects of program and retention transients. A few model improvements allow us to reconcile the apparent discrepancy between the values of silicon nitride trap energies extracted from program and retention experiments, thus reducing the number of model parameters.

Index Terms—Modeling, silicon nitride, TANOS/SONOS, transport and trapping properties.

I. INTRODUCTION

LOCIALIZED-TRAPPING silicon nitride- and high-k-based memory cells are good candidates to extend nonvolatile-memory (NVM) technology beyond the 45-nm node, thanks to low P/E voltages, SILC immunity, and small capacitive coupling between adjacent cells. The design of the ONO stack is constrained by a number of tradeoffs: A thin tunnel oxide (for SONOS) and a high-k top oxide (for TANOS) allow for reduced P/E voltages and for limited erase saturation [1], [2], but strongly enhance charge detrapping, eventually compromising retention.

Recently, many authors addressed the modeling of silicon nitride (SiN)-based NVM cells. Most of these models are focused on the description of the charge exchange between the substrate/control gate and the charge-trapping layer, often neglecting the charge transport in the SiN layer, either by assuming the SiN layer as a single point (i.e., null spatial extension) [2], [3] or by assuming a priori the shape of the spatial charge distribution [4]–[9].

Comprehensive understanding and modeling of program and retention characteristics are consequently still missing, and the mechanisms that are responsible for limited program efficiency and poor retention are not completely clear yet. Moreover, the values of the model parameters are often treated as mere fitting parameters with limited correlation to their physical meaning.

To shed new light on these aspects, we developed an improved numerical model for program and retention transients in SONOS cells, which accounts for a number of physical effects, among which the electron transport in the conduction band (CB) of the SiN layer. This improved model allows us to simulate the programming and retention sequence; therefore, unlike in [4]–[9], no a priori assumption is needed on the spatial distribution of the charge at the beginning of the retention phase. This is a critical point, particularly for thin-tunnel-oxide devices, where fast charge loss and detrapping at the beginning of the retention transient may lead to an uncertain determination of program/retention curves. In order to address this problem, we propose an improved characterization technique, and we exploit it to gather measured data for model calibration. This paper extends significantly beyond the results in [10]. We present a new characterization technique and several model improvements that allow us to gain new insights in the operation of nitride-based NVM cells.

This paper proceeds as follows. The model for the charge transport and trapping in ONO stacks and the simulation procedure are described in Section II. In order to infer the trap-energy depth from retention experiments, the interpretation of the Arrhenius plots of the high-temperature retention data is discussed in Section III-A. Simulations of retention transients are reported in Section III-B. In Section IV, we investigate, by means of measurements and simulations, the program efficiency reduction at large ΔVprog typically observed in SiN-based NVM cells. Conclusions are drawn in Section V.

II. MODEL

A. Physics

To describe the time evolution of the charge through the stack, a 1-D model for charge transport is used. Various conduction mechanisms through the ONO stack have been accounted for, as schematically shown in Fig. 1.
to-band tunneling, 5—electron transport, and 6—tunnel out.

1—Tunnel in, 2—band-to-trap tunneling, 3—emission-capture events, 4—trap-to-band tunneling, 5—electron transport, and 6—tunnel out.

The upper sketch refers to a positive gate voltage (\(V_G\)), hence to the programming phase. Let us now analyze the individual contribution of the current fluxes.

1) The current density \(J_1\) injected from the substrate toward the SiN CB at position \(x\) (flux 1 in Fig. 1) is given by

\[
\begin{align*}
J_1(x) = & \begin{cases} 
J_{1,DT-FN}(x), & x > 0 \\
J_{1,DFN}(x), & x = 0 
\end{cases}
\end{align*}
\]

where \(J_{1,DT-FN}\) is the current density through the tunnel oxide due to the direct and Fowler–Nordheim tunneling components, which is injected at the tunnel oxide/SiN interface, while \(J_{1,DFN}(x)\) is due to the electrons that tunnel through the tunnel oxide and part of the silicon nitride barrier (i.e., the modified Fowler–Nordheim component) [11] and are thus injected at any point \(x\) in the nitride. \(J_{1,DT-FN}\) (i.e., at the tunnel oxide interface, \(x = 0\)) is calculated as

\[
\begin{align*}
J_{1,DT-FN} &= \int_{E_{CO}}^{\infty} J_{Si}(E_\perp) T_{P}(E_\perp, x) dE_\perp \\
J_{Si}(E_\perp) &= q \sum_{j=1}^{n_v} f_j(E_\perp) v_{\perp,j}(E_\perp)
\end{align*}
\]

where \(E_\perp\) is the kinetic energy (referred to the Si CB at the Si/SiO\(_2\) interface) in the \(x\)-direction, \(f(E_\perp)\) is the electron distribution in normal energy (i.e., the occupation times the density of states), \(v_{\perp,j}(E_\perp)\) is the normal velocity, \(n_v = 6\) is the number of valleys, \(T_{P}(E_\perp, x)\) is the tunneling probability due to the direct and Fowler–Nordheim mechanisms, and \(E_{CO}\) is the bottom of the SiN CB at \(x = 0\) (with respect to the CB of Si at the Si/SiO\(_2\) interface). \(J_{1,DFN}(x)\) is generated at different positions inside the silicon nitride layer according to

\[
J_{1,DFN}(x) = J_{Si}(E_\perp) T_{P}(E_\perp, x) \Delta E_\perp
\]

where \(x\) corresponds to the exit point of electrons with silicon substrate energy \(E_\perp < E_C(x)\), with \(E_C(x)\) being the \(x\)-dependent bottom of the SiN CB. \(\Delta E_\perp\) is the range of kinetic energies corresponding to the spatial bin \(\Delta x\) centered in \(x\). The tunneling probability \(T_{P}(E_\perp, x)\) from the substrate to the \(x\)-position in the SiN CB is calculated by means of the WKB approximation using a parabolic dispersion relationship in the SiN band gap [12].

We assumed, for the sake of simplicity, that the absolute value of the effective mass for the nitride CB \(m^{*}_{N,CB}\) is equal to the absolute value of the effective mass for the nitride valence band and that the nitride electron tunneling mass \(m^{*}_{N}\) depends on the energy value inside the band gap as [12]

\[
m^{*}_{N}(x) = m^{*}_{N,CB} \left(1 - \frac{2E(x)}{E_{g,N}}\right)
\]

where \(E(x) = E_C(x) - E_\perp\) and \(E_{g,N}\) is the nitride energy gap. Since we are considering only the electron contribution, we use (4) only for \(E(x) < E_{g,N}/2\). A constant effective tunneling mass \(m^{*}_{ox} = 0.5m_0\) is assumed in SiO\(_2\) because of the much larger gap than in SiN.

2) The current density \(J_2\) from the substrate to the SiN trap sites (band-to-trap tunneling, flux 2 in Fig. 1) is calculated assuming an inelastic process that can occur only if the trap-energy level lies below the injection energy. This assumption is consistent with the evidence of an inelastic trap-assisted tunneling in SiO\(_2\) [13]. The band-to-trap tunneling contribution is

\[
J_2(x) = J_{Si}(E_\perp) \sigma T_{P}(E_\perp, x) \left(N_T - n^T(x)\right) \Delta x \Delta E_\perp
\]

where \(\sigma\) is the capture cross section of the traps, which we have assumed to be independent of the electric field [4, 14]; \(N_T\) is the total trap density; \(n^T(x)\) is the density of occupied traps; and \(\Delta x\) is the mesh spacing in nitride. As this material is usually uniformly grown in CVD chambers, it is reasonable to consider the spatial trap distribution that is uniform along the vertical direction, so \(N_T\) is assumed to be constant in the \(x\)-direction. However, evidence of nonuniformity distributions has been reported in [15] and [16]. This point will be further discussed in Section IV-B.

3) The traps in the SiN layer exchange carriers with the CB through capture and emission processes (flux 3 in Fig. 1). The emission rate \(R_E(x)\) is modeled according to the Poole–Frenkel (PF) equation [17]

\[
R_E(x) = \nu_{PF} \exp \left(\frac{\beta \sqrt{F(x) - E_T}}{kT}\right), \quad \beta = \frac{e^3}{\pi \varepsilon_0 \varepsilon_\infty}
\]

where \(\nu_{PF}\) is the attempt-to-escape frequency for emission, \(F(x)\) is the electric field at position \(x\) inside the silicon nitride layer, \(E_T\) is the trap-energy depth referred to the nitride CB, and \(\varepsilon_\infty\) is the high-frequency SiN dielectric constant.
dielectric constant. The free-electron capture rate coefficient is
\[
R_C(x) = C_C \left( N_T - n^T(x) \right) \tag{7}
\]
where \( C_C \) is the capture coefficient, and it has the unit of cubic centimeter per second. It is worth noting that our model is simpler than a complete SRH theory. We only assume the capture rate to be proportional to the concentration of free traps.

4) The tunneling of carriers from the SiN traps to the gate and to the nitride CB (flux 4 in Fig. 1) is expressed as
\[
J_4(x) = qT_P(x)\nu_T n^T(x) \Delta x \tag{8}
\]
where \( \nu_T \) is a constant attempt-to-escape frequency [8] and \( T_P(x) \) is the tunneling probability from the silicon nitride traps to the gate CB. The nitride effective mass (dependent on the tunneling’s electron energy level) is particularly important for the modeling of the trap-to-band tunneling rate. Indeed, for small electric fields in the nitride layer (a typical condition in the retention phase), the tunneling path from the SiN traps to the gate/substrate CB lies at energies near the middle of the band gap, resulting in a low effective mass. For example, assuming an electron effective tunneling mass for the nitride CB of \( m_{N,CB}^{eff} = 0.42m_0 \) [11], a value of about 0.25\( m_0 \) is obtained for the electrons tunneling out of the nitride in retention conditions, which is in agreement with our previous results using a constant effective mass [10]. It is worth mentioning that, in the model, we consider the barrier lowering due to the PF effect when modeling the emission rate \( (R_E(x)) \) but not when modeling the trap-to-band tunneling. This assumption, although not entirely justified, provides the best agreement between experiments and simulations. This point will be further discussed in Section IV.

5) The electron transport in the CB of SiN (flux 5 in Fig. 1) is described with the drift–diffusion (DD) equation
\[
J_{DD}(x) = q\mu \left( -n(x) \frac{\partial}{\partial x} V(x) + \frac{kT}{q} \frac{\partial}{\partial x} n(x) \right) \tag{9}
\]
where \( V(x) \) is the electrostatic potential, \( n(x) \) is the free-electron density, and \( \mu \) is the (constant) electron mobility. Note that, with expected mobilities on the order of 0.1–1 cm²/V · s [18], an electric field of 10–100 MV/cm is needed to reach the expected saturation velocity of 10⁷ cm/s [14]. Therefore, the assumption of a constant mobility appears adequate for program and retention conditions and more justified than that of a saturated velocity at all fields.

6) The direct and Fowler–Nordheim tunneling components from the SiN CB to the gate CB, \( J_6(t_N) \), and to the substrate CB, \( J_6(0) \), (flux 6 in Fig. 1) are expressed as
\[
J_6(0) = qn(0)\nu_T(0) \\
J_6(t_N) = qn(t_N)\nu_T(t_N). \tag{10}
\]
(PE), we have a system of three nonlinear partial differential
equations in space ($x$) and time ($t$). To numerically solve the
system, we adopted a simple discretization in space and time
(indices $i$ and $k$, respectively)

$$
\frac{n^k_i - n^{k-1}_i}{\Delta t} = - \frac{J^k_{DD,i+1} - (J^k_{DD,i} + J^k_{1,i})}{q \Delta x} + \frac{R^k_C n^k_i + R^k_{E} n^{T,k}_i}{q \Delta x}
$$

$$
\frac{n^{T,k}_i - n^{T,k-1}_i}{\Delta t} = \frac{R^k_C n^k_i - R^k_{E} n^{T,k}_i + J^{2,i}_k}{q \Delta x} - J^k_{4,i}/q \Delta x
$$

$$
\frac{V^{k}_{i+1} + V^{k}_i - 2V^{k}_i}{\Delta x^2} = \frac{n^k_i + n^{T,k}_i}{\varepsilon \Delta x}
$$

(12)

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 nonlinear equations is solved with the full Newton
scheme.

While the two continuity equations are solved only in SiN, the PE is calculated through the whole structure. Therefore, we have to solve three different PEs: for the substrate, SiN, and polysilicon gate, respectively. In addition, we have adopted different parameters to describe the layers of our
structure (substrate, tunnel oxide, SiN, top oxide, and gate).

Therefore, to properly connect the potentials and preserve
the currents at the interfaces between two different materials,
appropriate equations are needed. The equations connecting
the currents at the SiN/tunnel oxide interface ($x = 0$) and at the
SiN/top oxide interface ($x = t_N$) are set by

$$
\begin{cases}
J^k_{DD,0} = J_0(0) \\
J^k_{DD,n+1} = J_0(t_N).
\end{cases}
$$

(13)

In order to avoid stability issues related to Poisson/DD coupling,
we have written $J^k_{DD}$ inside SiN (nodes from $i = 0$ to
$i = n - 1$) in the form proposed by [25]

$$
J^k_{DD,i+1} = \frac{\mu kT}{q \Delta x} \left\{ - n^{k+1}_i B \left[ \frac{q}{kT} (V^{k+1}_i - V^k_i) \right] + n^k_i B \left[ \frac{q}{kT} (V^k_i - V^{k+1}_i) \right] \right\},
$$

(14)

where $B(x)$ is the Bernoulli function

$$
B(x) = \frac{x}{e^x - 1}.
$$

(15)

To set the potentials at the boundaries between the different
regions, we have applied the Gauss theorem at all the interfaces.
In the gate and the substrate, we solve the PE, assuming that the
charge in the inversion layer is at equilibrium

$$
\frac{d^2 V(x)}{dx^2} = - \frac{q}{\varepsilon_0 \varepsilon_{Si}} \left[ n_0 e^{-qV(x)/(kT)} - n_0 e^{qV(x)/(kT)} \right] + N_{SUB/GATE}
$$

(16)

where $n_0$ and $p_0$ are the equilibrium electron and hole concen-
trations and $N_{SUB/GATE}$ is the net concentration of ionized
dopant atoms in either the substrate or the polysilicon gate.

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 nonlinear equations is solved with the full Newton
scheme.

While the two continuity equations are solved only in SiN, the PE is calculated through the whole structure. Therefore, we have to solve three different PEs: for the substrate, SiN, and polysilicon gate, respectively. In addition, we have adopted different parameters to describe the layers of our
structure (substrate, tunnel oxide, SiN, top oxide, and gate).

Therefore, to properly connect the potentials and preserve
the currents at the interfaces between two different materials,
appropriate equations are needed. The equations connecting
the currents at the SiN/tunnel oxide interface ($x = 0$) and at the
SiN/top oxide interface ($x = t_N$) are set by

$$
\begin{cases}
J^k_{DD,0} = J_0(0) \\
J^k_{DD,n+1} = J_0(t_N).
\end{cases}
$$

(13)

In order to avoid stability issues related to Poisson/DD coupling,
we have written $J^k_{DD}$ inside SiN (nodes from $i = 0$ to
$i = n - 1$) in the form proposed by [25]

$$
J^k_{DD,i+1} = \frac{\mu kT}{q \Delta x} \left\{ - n^{k+1}_i B \left[ \frac{q}{kT} (V^{k+1}_i - V^k_i) \right] + n^k_i B \left[ \frac{q}{kT} (V^k_i - V^{k+1}_i) \right] \right\},
$$

(14)

where $B(x)$ is the Bernoulli function

$$
B(x) = \frac{x}{e^x - 1}.
$$

(15)

To set the potentials at the boundaries between the different
regions, we have applied the Gauss theorem at all the interfaces.
In the gate and the substrate, we solve the PE, assuming that the
charge in the inversion layer is at equilibrium

$$
\frac{d^2 V(x)}{dx^2} = - \frac{q}{\varepsilon_0 \varepsilon_{Si}} \left[ n_0 e^{-qV(x)/(kT)} - n_0 e^{qV(x)/(kT)} \right] + N_{SUB/GATE}
$$

(16)

where $n_0$ and $p_0$ are the equilibrium electron and hole concen-
trations and $N_{SUB/GATE}$ is the net concentration of ionized
dopant atoms in either the substrate or the polysilicon gate.

No explicit correction is included to account for quantum-
mechanical effects at the Si–SiO₂ interface.

The boundary conditions that are sufficient to solve the proposed model are the gate-voltage waveform $V_G(t)$ and the initial concentration profiles $n(x, t = 0)$ and $n^T(x, t = 0)$ in the SiN CB and the traps, respectively. Once the boundary conditions are fixed, all the equations (Poisson, continuity, DD model, and tunneling) are solved together in a self-consistent manner at each time step using the Newton scheme. The threshold-voltage shift of the device at each time step can be calculated with this simple equation

$$
\Delta V_T(t) = - \frac{q t_{top}}{\varepsilon_{ox}} \int_0^{t_N} \left[ n(x, t) + n^T(x, t) \right] dx
$$

$$
+ \frac{q}{\varepsilon_N} \int_0^{t_N} x \left[ n(x, t) + n^T(x, t) \right] dx.
$$

(17)

The proposed model allows one to perform simulations of arbitrarily shaped gate-voltage waveforms, hence retention transients immediately following the programming phase (see, for example, the curves in Fig. 3). Consequently, a priori
assumptions on the spatial distribution of the trapped charge at
the beginning of retention ($\tau_r = 0$) are not necessary; in fact, this distribution results from the simulated program transient. This feature eliminates a relevant source of uncertainty.

### III. Determination of the Model Parameters

The model relies on the parameters that govern the capture/emission mechanisms and the charge-transport properties in silicon nitride. The capture rate $R_C$ (which is a function of $C_C$ and $N_T$) and the band-to-trap tunneling component (which is a function of $\sigma$) are the parameters that mostly affect the programming phase at low electric fields and low $\Delta V_T$ values. The numerical values of $\sigma$ and $C_C$ are not correlated
because, in our model, they are used to describe two different
processes: The $C_C$ coefficient models the capture of the electrons of the nitride CB in the traps, while the $\sigma$ value is instead used to model the current density from the substrate to the silicon nitride traps. The comparison between the measured and simulated program curves at low $\Delta V_T$ and $V_G$ allows
A. Relationship Between the Activation Energy and the Trap-Energy Depth

The value of trap-energy depth strongly influences the emission process and, consequently, the whole charging/discharging dynamics; therefore, an accurate analysis of this parameter is required first. The trap-energy distribution in the SiN layer was evaluated in the literature using different measurement techniques [26]–[28], but the interpretation is still partly controversial. The thermally stimulated exoelectron emission (TSEE) measurements performed in [28] yield a value in the range of about 1.0 eV. Another technique to estimate $E_T$ is the analysis of retention characteristics at high temperature, considering devices in which the discharge is due to the thermal emission of carriers and the tunneling through the tunnel or top oxide is made negligible by using thick tunnel and top oxides [10], [19]. Fig. 4 (left plot) shows the simulated retention transients at high temperature of SONOS devices with thick tunnel oxides (6 nm). By defining the retention time ($\tau_r$) as the time required to observe a predefined threshold-voltage shift ($\Delta V_T$), we extract $\tau_r$ as a function of temperature ($T$). Fig. 4 (right plot) shows $\tau_r$ versus the reciprocal of thermal energy $1/(kT)$, and an Arrhenius relation is observed, i.e., a linear relation in a semilog scale. The slope of this curve gives the activation energy of the phenomenon, which is found to be fairly independent of the $\Delta V_T$ value.

Fig. 5 shows the relationship between $E_A$ and $E_T$ obtained by applying the procedure of Fig. 4 to simulations performed by using different values of $E_T$ in the model. We simulated two SONOS devices featuring different gate stacks, but all with thick tunnel and top oxides, in order to suppress the tunneling components. The interpretation of the result of Fig. 5 is not trivial. Indeed, for small $\Delta V_T$, $\tau_r$ is inversely proportional to the us to fix $C_C = 5 \cdot 10^{-8} \text{cm}^3 \cdot \text{s}^{-1}$, $N_T = 4 \cdot 10^{19} \text{cm}^{-3}$ ([7]), and $\sigma = 2 \cdot 10^{-16} \text{cm}^2$ ([5]). Furthermore, we have adopted $\mu = 1 \text{cm}^2/\text{V} \cdot \text{s}$, as obtained in [10] from the analysis of the impact of the charge transport in the nitride CB on the retention of SONOS cells with thick tunnel oxides. The value of the other model parameters will be estimated in the following section.

B. Device Description and Retention Simulations

To validate the model, we have performed measurements on standard SONOS memory cells, featuring a gate stack of

Fig. 4. (Left graph) Simulated retention at different temperatures for the 6/6/6-nm device. $\Delta V_T$ is referred to the neutral state. A single trap-energy level ($E_T = 1.2$ eV) has been assumed. (Right graph) Arrhenius plot of the retention time $\tau_r$ calculated for two different threshold-voltage shift criteria ($\Delta V_T = 0.25$ V and $\Delta V_T = 0.5$ V).

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 [26], [27]. The inset shows how $E_A$ can be related to the escape energy of electrons. $f_N(E_\perp)$ is the electron distribution in normal energy (product of the occupation probability times the density of states) computed accordingly with a Maxwell–Boltzmann statistics.
reproduces nicely the experiments for $\nu_T$ respectively). The devices have a standard LP-CVD SiN and 2/6/9 nm (for tunnel oxide, silicon nitride, and HTO top oxide, respectively). The devices have a standard LP-CVD SiN and n+ poly gate.

Once the trap-energy depth (Section III-A) and the value for the SiN CB effective mass ($m_{N, CB} = 0.42m_0$ [8]) are fixed, the only parameter affecting the retention simulations is the attempt-to-escape frequency ($\nu_T$). This term governs the trap-to-band tunneling (flux 4 in Fig. 1(b)), which is the dominant charge-loss mechanism in devices with thin tunnel oxide at room temperature [19]. Fig. 6(a) shows the impact of $\nu_T$ on the simulated transient demonstrating that the simulator produces nicely the experiments for $\nu_T = 5 \cdot 10^7$ s$^{-1}$. The same numerical value has been adopted for the attempt-to-escape frequency of the PF emission, i.e., $\nu_{PF}$ ([6]). Obviously, there would be no need to assign the same value to these two variables. However, a clear distinction between the two is not straightforward based on the fitting of experiments. The chosen numerical value for $\nu_{PF}$ is on the same order of magnitude with that proposed in [29].

Fig. 6(b) shows that the same parameters provide a good agreement between measurements and simulations also for different $V_G$ and $\Delta V_T$. Note the good agreement between measurements and simulations in terms of both initial state and $\Delta V_T$'s. The device (2/6/9 nm) is initially programmed at $V_G = 13$ V for 1 ms. Symbols are reference measurements. (b) Measured (symbols) and simulated (lines) retention characteristics for the 2/6/9 nm. The initial $\Delta V_T$ for the retention experiments are obtained applying pulses of 1 ms and different gate voltages, as indicated in the figure.

2/6/9 nm (for tunnel oxide, silicon nitride, and HTO top oxide, respectively). The devices have a standard LP-CVD SiN and n+ poly gate.

Once the trap-energy depth (Section III-A) and the value for the SiN CB effective mass ($m_{N, CB} = 0.42m_0$ [8]) are fixed, the only parameter affecting the retention simulations is the attempt-to-escape frequency ($\nu_T$). This term governs the trap-to-band tunneling (flux 4 in Fig. 1(b)), which is the dominant charge-loss mechanism in devices with thin tunnel oxide at room temperature [19]. Fig. 6(a) shows the impact of $\nu_T$ on the simulated transient demonstrating that the simulator reproduces nicely the experiments for $\nu_T = 5 \cdot 10^7$ s$^{-1}$. The same numerical value has been adopted for the attempt-to-escape frequency of the PF emission, i.e., $\nu_{PF}$ ([6]). Obviously, there would be no need to assign the same value to these two variables. However, a clear distinction between the two is not straightforward based on the fitting of experiments. The chosen numerical value for $\nu_{PF}$ is on the same order of magnitude with that proposed in [29].

Fig. 6(b) shows that the same parameters provide a good agreement between measurements and simulations also for different $V_G$ and $\Delta V_T$. Note the good agreement between measurements and simulations in terms of both initial state and $\Delta V_T$'s. The device (2/6/9 nm) is initially programmed at $V_G = 13$ V for 1 ms. Symbols are reference measurements. (b) Measured (symbols) and simulated (lines) retention characteristics for the 2/6/9 nm. The initial $\Delta V_T$ for the retention experiments are obtained applying pulses of 1 ms and different gate voltages, as indicated in the figure.

IV. ANALYSIS OF THE PROGRAM TRANSIENT

A. Experiments

Fig. 6(b) shows a lack of experimental data for very short retention times. In fact, in conventional retention experiments, $\Delta V_T$ is measured by means of a gate-voltage sweep to detect the target drain current and which requires a few seconds after the end of the program pulse. The variable gate voltage and the relatively long time elapsed could affect the charge distribution in the nitride layer, hence the apparent $\Delta V_T$, particularly in thin-tunnel-oxide devices. To address these issues and better compare experiments and simulations, we have developed a new characterization technique for a very fast detection of $\Delta V_T$ after programming. This technique is based on the gate-voltage ($V_G$) waveform shown in Fig. 7(a). After the program pulse ($V_{prog}$), we lower $V_G$ to a value ($V_{read}$) that is slightly above the expected $V_T$ (estimated in advance with standard swept measurements), thus stopping the charge injection in SiN but still having the possibility to rapidly measure the drain current ($I_{DS}$) induced by a small $V_{DS}$. We then set $V_G = 0$ and put the device in retention mode. By applying short pulses ($V_{G} = V_{read}$ for $t_{read} = 10$ ms), we periodically stop the retention phase and measure again $I_{DS}$ without perturbing significantly the charge state in SiN [Fig. 7(a)]. $V_{read}$ is selected to achieve an optimum tradeoff between a short measurement time and good measurement accuracy. We verified that programming yields rigid shifts of the $I_{DS}$--$V_G$ characteristic, and thus, $I_{DS}$ can be converted to $\Delta V_T$ through a reference $I_{DS}$--$V_G$ curve. Fig. 7(b) shows typical $I_{DS}$ data for a SONOS device featuring a tunnel oxide thickness of $t_{ox} \approx 2$ nm (1.8 nm in TEM images). We clearly see that $I_{DS}$ increases with time at each pulse because detrapping decreases $V_T$.

Fig. 8 shows the program curves up to $t_{prog} = 0.1$ s and the $V_T$ relaxation after the last program pulse. We see that...
the detrapping at short times highlighted by our technique can have a big impact on the detected program curve; the effect is particularly evident at large program $V_T$ values, when $V_T$ is measured with a standard $V_G$ sweep, because the time needed to detect $V_T$ can be significant. Indeed, Fig. 9(a) compares the program curves measured with sequences of program/read pulses, such as those in Fig. 7(a), with the ones obtained with the conventional method for $V_T$ detection; in the latter case, detrapping at large $\Delta V_T$ alters the apparent program speed and makes the curve saturate at $\Delta V_T \approx 6$ V. This behavior is apparently very similar to that reported in [30]. As a result [Fig. 9(b)], the conventional $V_T$ detection underestimates the increase of $\Delta V_T$ for increasing $V_{prog}$ ($d\Delta V_T/dV_{prog}$). The implications of this behavior from the viewpoint of the calibration of SONOS program–retention models will be analyzed in the following section.

### B. Simulations

In Section III-A, a trap-energy depth of 1.2 eV was obtained from the analysis of retention experiments. If we now use the same value to model the programming phase, a dramatic saturation of the simulated program curve is observed (Fig. 10, squares), which is not seen in the experiments with fast $V_T$ detection. This discrepancy is due to a large tunneling-out current (fluxes 4 + 6 in Fig. 1) and can be reduced, without any change in the other model parameters with respect to those used for retention, only by assuming $E_T \approx 1.6$ eV for programming, consistently with previous models of the program curves ($E_T \approx 1.4−1.8$ eV [4], [5]). To solve this ambiguity on the $E_T$ value ($E_T \approx 1.2$ eV in retention and $E_T \approx 1.6$ eV in programming), we improved our model by considering the unfilled SiN traps as coulombic centers [Fig. 11(a)], whose fundamental eigenstate is adopted as the energy of the trapped electron. This level varies with the applied electric field in SiN ($F$) according to [31]

$$E_T(F) = \left| \phi(F) - R_{YN} \left[ 1 + \frac{36}{8} \left( \frac{F}{P} \right)^2 \right] \right|$$  \hspace{1cm} (18)

where $\phi(F)$ is the local maximum of the potential energy [see Fig. 11(a)] and

$$R_{YN} = \frac{m_e}{2} \left( \frac{q^2}{4\pi\varepsilon_{\infty}} \right)^2 P = \frac{q^5 m_e^2}{(4\pi\varepsilon_{\infty})^3 \hbar^4}.$$  \hspace{1cm} (19)

Here, $m_e$ is the free-electron mass, which is used as an approximation of the electron effective mass at high energy;
Note that, in order to experience both the PF effect and the trap-energy modulation with the applied electric field, the traps must be neutral when filled and must be positive when empty. A trap that is neutral when empty and that is charged when filled will not experience these effects because of the absence of the coulomb potential [32]. In this regard, we note that the real nature of the trapping center in SiN is not yet completely understood. Most of the authors refer to the amphoteric trap model [22], [23], where each trap can assume three different states, which are denoted as $D_0$, $D_0$, and $D_0^+$, hence a negatively, a neutral, and a positively charged state, respectively. Nevertheless, electron-spin-resonance measurements and density-functional-theory studies [33], [34] have demonstrated that it is energetically favorable for two adjacent $D_0$ traps to transform to one $D_0$ and one $D_0^+$ state, leading to an equal density of $D_0$ and $D_0^+$ traps in the equilibrium state of a neutral SiN layer. During programming, it is therefore reasonable to consider that the injected electrons are trapped by the $D_0$ states, which become $D_0$ traps, hence traps that are neutral when filled by an electron, which is consistent with our assumptions. Moreover, if we assume that the time scale involved in the experiments is small compared to the decay of $D_0$ states in couples of $D_0$ and $D_0^+$ traps, then, during retention, the charge loss is given by the emission of electrons from the $D_0$ state, which is consistent with the dependence of the trap-energy level on the applied field shown in Fig. 11.

The overall model [i.e., the one described in Section II plus the $E_T$ model of (18)] reproduces well the cell program characteristics for different $V_{\text{prog}}$’s (Fig. 12). The program speed slowdown at large $t_{\text{prog}}$ is due to both the reduction of the input current (fluxes 1 + 2 in Fig. 1) and the increase of the relative fraction of the electron tunneling through the top oxide [fluxes 4 + 6 in Fig. 1]. This latter is responsible also for $d\Delta V_T/dV_{\text{prog}} < 1$ shown by the fast-$V_T$-detection experiments in Fig. 9(b) [5]. Moreover, the simulations grasp with the same set of parameters also the long-term retention behavior shown in Fig. 13, where the retention transients start from different $\Delta V_T$ values obtained for fixed $V_{\text{prog}}$. Residual discrepancies at the beginning of retention are possibly due to a spatially nonuniform distribution of SiN traps [15], [16]. Indeed, a higher trap density that is close to the oxide interface would accelerate the initial discharging of the nitride layer, which is consistent with the experiments. Fig. 14 shows the simulated $n_T(x)$ during programming and retention. While the charge built up at the tunnel oxide interface is responsible for the detrapping in the first seconds, the flat part in the center of SiN governs the long-term retention ($t_{\text{ret}} > 100$ s) showing distribution dynamics similar to those proposed in [7] and [19]. Fig. 15 shows that the model can nicely reproduce the retention curves obtained with both fast $V_T$-detection and conventional gate-voltage sweep by appropriately taking into account for the delay in $V_T$-detection corresponding to the case of the constant $I_{DS}$ criterion.

V. Conclusion

In summary, we have presented a model for charge transport and trapping in ONO stacks that allows us to simulate a complete program/retention experiment, avoiding a priori assumptions on the distribution of the trapped charge during retention. The model identifies the relations between the
activation energy of the high-temperature retention experiments and the trap energy. An improved numerical model that includes a physically based treatment of the trap-energy levels allows us to reduce the number of model parameters that must be adjusted to fit the experiments and to reconcile the reported discrepancy between the trap-energy values in program and retention.

An accurate characterization of the program and retention transients of SiN-based memories has been devised to accurately measure the threshold-voltage transients due to charge detrapping after programming during the first seconds of retention. We show that, in thin-tunnel-oxide devices (such as those needed for low-voltage operations), this fast charge loss after programming can lead to substantial distortions in the measured transients. We devised a method to accurately measure the threshold-voltage transients due to charge trapping storage memory, in *Proc. NVSMW*, 2008, pp. 109–110.


Francesco Driussi received the Laurea and Ph.D. degrees in electronic engineering from the University of Udine, Udine, Italy, in 2000 and 2004, respectively, working on the reliability of Flash memory devices.

From October 2002 to September 2003, he held a student position at Philips Research Leuven, Leuven, Belgium. In 2005, he became a Research Associate with the University of Udine, where he is currently with DIEGM. His research activities are mainly in the field of nonvolatile-memory-cell reliability, and his interests include the characterization of device degradation and gate dielectric reliability. In particular, he has worked on substrate-enhanced hot-electron phenomena, and he has investigated on their implications for Flash EEPROM devices. More recently, he worked on the development of physical and statistical models for the study of SILC in large Flash memory arrays and of the oxide trap generation and distribution. He also is working on the characterization and modeling of SONOS memory cells and of the trapping properties of silicon nitride. Part of his current activities also includes the experimental characterization and modeling of carrier mobility in MOSFET devices featuring strained-silicon and high-$k$ gate stacks.

Elisa Vianello received the Laurea degree in electronic engineering from the Università degli Studi di Udine, Udine, Italy. She is currently working toward the Ph.D. degree, particularly on the simulation and characterization of Flash memories, between DIEGM, Università degli Studi di Udine, and the Instituto Nazionale Politecnico di Grenoble, Grenoble, France.

She was last year with the Laboratory of Electronics, Technology, and Instrumentation, CEA Grenoble, working on the electrical characterization of charge-trap high-$k$-based memories. Her research interests include modeling, characterization, and reliability of nonvolatile memory, focusing on reliability issues related to the statistical distribution of stress-induced leakage current and on the modeling and characterization of SONOS/TANOS memory cells.

Antonio Arreghini was born in Udine, Italy, in 1979. He received the M.S. (Laurea) degree in electronic engineering and the Ph.D. degree in industrial and information engineering from the University of Udine, Udine, Italy, in 2004 and 2008, respectively.

From July 2006 to July 2007, he was a Visiting Scientist with NXP Semiconductors, Leuven, Belgium, where he worked on the modeling and simulations of long-term retention, program, and erase transients of SONOS nonvolatile memories (NVMs). Since 2008, he has been with IMEC, Leuven. His research interests include the modeling and experimental characterization of charge-trapping NVM cells.

Pierpaolo Palestri received the Laurea degree (summa cum laude) in electronic engineering from the University of Bologna, Bologna, Italy, in 1998 and the Ph.D. degree in electronic engineering from the University of Udine, Udine, Italy, in 2003.

In 1998, he joined the Department of Electrical, Mechanical and Management Engineering (DIEGM), University of Udine, as a Research Assistant in the field of device simulation, where he became an Associate Professor in October 2001 and has been an Associate Professor since November 2005.

From July 2000 to September 2001, he held a postdoctoral position at Bell Laboratories, Lucent Technologies (currently Agere Systems), Murray Hill, NJ, where he worked on high-speed silicon–germanium bipolar technologies. His research interests include the modeling of carrier transport in nanoscale devices and the simulation of hot-carrier and tunneling phenomena in scaled MOSFETs and nonvolatile memory cells.

David Esseni received the Laurea and Ph.D. degrees in electronic engineering from the University of Bologna, Bologna, Italy, in 1994 and 1998, respectively.

In 2000, he was a Visiting Scientist with Bell Laboratories, Lucent Technologies (currently Agere Systems), Murray Hill, NJ. Since 2005, he has been an Associate Professor with the University of Udine, Udine, Italy, where he is currently with DIEGM. His research interests mainly include the characterization, modeling, and reliability of MOS transistors and nonvolatile memories (NVMs). In the field of NVM, he has worked on low-voltage and substrate-enhanced hot-electron phenomena and on several aspects of Flash EEPROM memories, including innovative programming techniques and reliability issues related to the statistical distribution of stress-induced leakage current. Since 2000, he has been involved in the field of advanced or innovative CMOS devices. In particular, he has experimentally investigated low-field mobility in ultrathin SOI MOS transistors and then started an activity of semiclassical transport modeling in advanced n- and p-MOS transistors. In this field, his research interests also include quantization models beyond the effective mass approximation, as well as the modeling and characterization of stress effects on CMOS technologies.
Luca Selmi was born in Rome, Italy, in 1961. He received the Ph.D. degree in electronic engineering from the University of Bologna, Bologna, Italy, in 1992. From April 1989 to June 1990, he was a Visiting Scientist with Hewlett-Packard Microwave Technology Division, Santa Rosa, CA. Since 2000, he has been a Full Professor of electronics with the University of Udine, Udine, Italy, where he is currently responsible of a research group on nanoelectronic devices with DIEGM. He held technical and/or coordination responsibility in several EC research projects (ULIS, NESTOR, SINANO, PULL-NANO, NANOSIL, GOSS-AMER, and GRAND) in the field of microelectronics and nanoelectronics. The research activities have been carried out in cooperation with many semiconductor research laboratories and companies worldwide, including Philips Research, NXP Semiconductors, Infineon Technologies, IMEC, LETI, IBM T. J. Watson Research Center, AT&T Bell Laboratories, and Hewlett-Packard Company. He has served as a Technical Program Committee Member of several electron device conferences (IEDM—twice, ESSDERC, INFOS, SISC, and ULIS) and a Co-organizer of the INFOS 2001 and ULIS 2003 and 2008 conferences. He coauthored approximately 150 papers in refereed journals and conference proceedings, including approximately 30 IEDM papers. His research interests include electron device modeling and characterization, with emphasis on CMOS scaling, Monte Carlo simulation techniques, hot-carrier effects, quasi-ballistic transport, nonvolatile memories (floating gate, CHISEL-based, localized trapping, SONOS, and TANOS), and semiconductor device reliability.

Michiel J. van Duuren was born in Groenlo, The Netherlands, in 1970. He received the M.S. degree in applied physics and the Ph.D. degree from the University of Twente, Enschede, The Netherlands, in 1993 and 1998, respectively, working on the subject of superconducting sensors.

In 1998, he joined Philips Research Laboratories, Eindhoven, The Netherlands, where he worked on reliability aspects, characterization methods, and physical modeling of floating-gate and SONOS memories. In 2001, he joined Philips Research Leuven (currently NXP Semiconductors), Leuven, Belgium, where he became the Project Leader of the embedded memories technology research project and currently with the NXP-TSMC Research Center, with activities on, for instance, charge-trapping, floating-gate, FinFET, nanocrystal, and phase-change memories.

Nader Akil received the M.S. degree in physics from the University of Reims Champagne-Ardenne, Reims, France, in 1995 and the Ph.D. degree from Paul Verlaine University-Metz, Metz, France, in 1998. He spent the last year of his Ph.D. studies with the Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN, working on light emission from silicon-based devices. From 1999 to 2000, he was a Postdoctoral Associate with the University of Twente, Enschede, The Netherlands, on nanoscale silicon-based devices for applications in micro total analysis systems. In 2000, he joined Philips Research Laboratories, Eindhoven, The Netherlands, and in 2001, he moved to Philips Research Leuven, Leuven, Belgium. In 2006, he joined NXP Semiconductors, Leuven, Belgium, where he worked on modeling and characterization of floating-gate and charge-trapping memories with the NXP-TSMC Research Center. He is currently with the R&D Department, Photovoltech NV, Tienen, Belgium.

Dusan S. Golubović received the M.S.E.E. degree from the University of Belgrade, Belgrade, Serbia, in 2000 and the Ph.D. degree in applied physics from the Katholieke Universiteit Leuven, Leuven, Belgium, in 2005. Since 2005, he has been a Senior Scientist with Philips Research Leuven (currently NXP Semiconductors), Leuven, where he is currently with the NXP-TSMC Research Center. His research interests include the process integration and characterization of advanced embedded nonvolatile memories.