Modeling the Operation of Charge Trap Flash Memory—Part II: Understanding the ISPP Curve With a Semianalytical Model

Devin Verreck, Franz Schanovsky, Antonio Arreghini, Geert Van den Bosch, Zlatan Stanojević, *Senior Member, IEEE*, Markus Karner, *Member, IEEE*, and Maarten Rosmeulen

***Abstract*— Flash memory with a charge trap layer (CTL), also known as silicon-oxide-nitride-oxide-silicon (SONOS), is the most common type in production, yet there is a lack of consensus on the physical modeling of its operation. In Part I, we therefore proposed a full TCAD model based on an energy relaxation approach and showed that it captures experimentally observed memory operation. This numerical model, however, comes with considerable complexity and computational cost. In Part II, we therefore construct a semianalytical model based on similar physical assumptions, called Pheido, to be as simple as possible. We first derive the model equations based on a balance of current densities, detailing the approximations made. We then use Pheido to analyze the various regimes of an experimental incremental step pulse programming (ISPP) curve and compare it to the full TCAD model derived in Part I. Finally, we investigate the impact of material and structural cell parameters on the ISPP curve, illustrating how the Pheido model offers wide utility at low computational cost.**

***Index Terms*— 3-D NAND, flash memory, incremental step pulse programming (ISPP), modeling, silicon-oxide-nitride-oxide-silicon (SONOS), VNAND.**

I. Introduction

IT IS quite remarkable for a technology so ubiquitous as NAND flash memory, that its basic operation is not yet fully understood. Commercial application of flash memories has grown explosively over the last decade, thanks to continuously decreasing cost per bit, enabled by vertical integration of flash strings, known as 3-D NAND or VNAND [1]. Three-dimensional NAND implementation was significantly simplified by a change of storage material from a metal floating gate (FG) to a charge trap layer (CTL). The operation of a CTL cell has particularities compared to an FG cell, however, notably the nonideal programming efficiency, which translates to a slope of the incremental step pulse programming (ISPP) curve significantly below one [2]. To study these aspects of the CTL flash cell, one approach is to explicitly model the various physical mechanisms numerically [3], [4], [5], [6]. In Part I of this article, we propose such a full TCAD model and show that it is able to capture the various experimental operating regimes of CTL cells by taking into account the energy relaxation of the injected charge carriers. The complexity of such models, however, comes with a significant computational cost. A different approach is to capture the main aspects of the memory operation in a simpler (semi)analytical or compact model [6], [7], [8]. Existing models, however, typically do not consider the energy relaxation of the carriers or are focused on transient rather than ISPP characteristics.

Here, we therefore expand on our work in [6] and propose a semianalytical memory operation model, called Pheido, which aims to be as simple as possible, while still capturing the main features of the physical model of Part I. Pheido allows for a deep understanding of the ISPP curve of CTL flash cells at little computational cost. We first derive the Pheido model and outline the approximations. Next, we compare results with experimental data and the full TCAD model from Part I. This comparison then allows us to gain deeper insight into the different regimes of the ISPP curve. Finally, we apply Pheido to investigate the impact of material and structural cell parameters on the ISPP curve, while checking its assumptions under a wide range of conditions.

II. Pheido, a Semianalytical Model

"Pheido" is Greek for parsimony and sparingness, and that is the goal of this model: to reproduce the essential features of the flash cell ISPP curve, while including the minimal necessary physical complexity or parameters. In this section, we derive the model and use it to analyze and explain the different regimes in an ISPP curve of a CTL flash cell.

Manuscript received 12 August 2023; revised 31 October 2023; accepted 27 November 2023. Date of publication 8 December 2023; date of current version 2 January 2024. This work was supported by imec's Industrial Affiliation Program for Storage Memory. The review of this article was arranged by Editor S. Alam. *(Corresponding author: Devin Verreck.)*

Devin Verreck, Antonio Arreghini, Geert Van den Bosch, and Maarten Rosmeulen are with imec, 3001 Leuven, Belgium (e-mail: devin.verreck@imec.be).

Franz Schanovsky, Zlatan Stanojević, and Markus Karner are with Global TCAD Solutions GmbH, 1010 Vienna, Austria.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2023.3339112.

Digital Object Identifier 10.1109/TED.2023.3339112

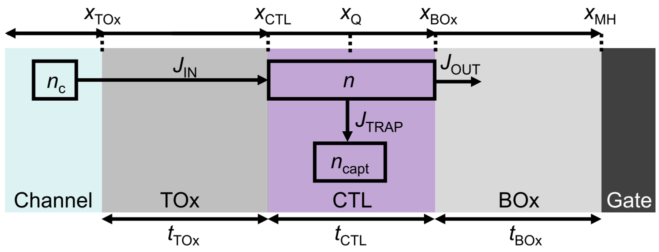

Fig. 1. Schematic of the SONOS cell structure and current density components during a programming pulse assumed in the Pheido model.

A. Model Derivation

The Pheido model assumes a conventional flash memory silicon-oxide-nitride-oxide-silicon (SONOS) stack as shown in Fig. 1, in which a number of current components flow when an electrostatic potential is applied. The stack consists of a channel, tunnel oxide (TOx), CTL, blocking oxide (BOx), and gate contact. When a potential is applied between the channel and the gate, such as during a programming pulse, currents arise that enter and leave the CTL, represented by the densities J_{IN} and J_{OUT}, respectively. The difference between them is J_{TRAP}, which is the current density that is captured by the CTL traps. The ISPP curve of such a cell consists of the threshold voltage shift (ΔV_{T}) after i pulses, which can be written as

$$\Delta V_{\mathrm{T},i} = \sum_{p=1}^{i} \Delta V_{\mathrm{T},p} = \sum_{p=1}^{i} \int_0^{t_p} \frac{dV_{\mathrm{T}}(t)}{dt} \tag{1}$$

with t_p the pulse duration and

$$\frac{dV_{\mathrm{T}}(t)}{dt} = \frac{J_{\mathrm{TRAP}}}{C_q} = \frac{1}{C_q}(J_{\mathrm{IN}} - J_{\mathrm{OUT}}) \tag{2}$$

where we call $((dV_{\mathrm{T}}(t))/dt)$ the voltage change rate (VCR) and with C_q the equivalent capacitance per unit area between the trapped charge and the gate, which depends on the geometry of the cell and the materials of the stack (see Appendix I). In the remainder of this section, we derive a simple physical description for J_{IN} and J_{OUT}, such that we can solve (2) as an ordinary differential equation.

First, we model J_{IN} with a Fowler–Nordheim formula, which depends exponentially on the field over the TOx (F_{TOx})

$$J_{\mathrm{IN}} = qn_c v_t e^{-\frac{B_{\mathrm{TOx}}}{F_{\mathrm{TOx}}}} \tag{3}$$

with q the elementary charge, n_c the carrier density in the inversion layer at the channel-TOx interface, v_t the carrier thermal velocity and

$$B_{\mathrm{TOx}} = \frac{4}{3}\frac{\sqrt{2qm^*}}{\hbar} E_b^{\frac{3}{2}} \tag{4}$$

with E_b the conduction band offset between the channel and the TOx, m^* the effective tunneling mass in the TOx, and $\hbar$ the reduced Planck's constant.

Next, we obtain an expression for J_{OUT} by assuming that it represents the current density that is not captured in the CTL and therefore reaches the CTL-BOx interface

$$J_{\mathrm{OUT}} = qn_{\mathrm{BOx}} v_{\mathrm{BOx}} \tag{5}$$

with n_{BOx} and v_{BOx} the carrier density and velocity at the CTL-BOx interface, respectively. This does not explicitly model the tunneling out through the BOx or any lateral escape but rather assumes that every carrier that reaches the BOx untrapped effectively disappears from the layer. We will evaluate this assumption in Section III with the full TCAD model developed in Part I. Continuing the derivation, we rewrite n_{BOx} in (5) using the expression we obtained in [9] for the carrier density at a location x in the CTL in the case where the carrier transport is governed by the drift-diffusion equations, with trapping according to a Shockley–Read–Hall (SRH) process

$$n(x) = \frac{J_{\mathrm{IN}}}{q\mu F_{\mathrm{CTL}}} \exp\left(-\int_0^x \frac{\sigma v_t N_t}{\mu F_{\mathrm{CTL}}} dx'\right) \tag{6}$$

with F_{CTL} the electric field in the CTL, μ the carrier mobility, N_t the density of available traps and σ the trap capture cross section. Equation (6) assumes that all tunneling current is injected into the CTL at the TOx-CTL interface. The carriers enter the CTL at an energy significantly above the conduction band edge, however, due to the high programming voltages applied. Some authors, including ourselves, have therefore posited that carriers first need to relax in energy before capture can occur (see also Part I) [7]. This means the injected carriers are effectively distributed over the CTL according to some shape function $S(x)$, with S normalized over the thickness of the CTL to preserve the conservation of current density

$$\int_{x_{\mathrm{CTL}}}^{x_{\mathrm{BOx}}} S(x)dx = 1. \tag{7}$$

Each point x_i in the CTL then receives a fraction of the injected current density equal to $S(x_i)J_{\mathrm{IN}}$. Based on (6), an expression can be derived for the carrier density that is injected at a point x_1 and which reaches x_2 without getting trapped

$$n(x_1, x_2) = \frac{S(x_1)J_{\mathrm{IN}}}{q\mu F_{\mathrm{CTL}}} \exp\left(-\int_{x_1}^{x_2} \frac{\sigma v_t N_t}{\mu F_{\mathrm{CTL}}} dx'\right). \tag{8}$$

The total carrier density that reaches the BOx without getting trapped can then be obtained as the integral of (8) over the thickness of the CTL

$$n_{\mathrm{BOx}} = \int_{x_{\mathrm{CTL}}}^{x_{\mathrm{BOx}}} n(x, x_{\mathrm{BOx}})dx \tag{9}$$

$$= \int_{x_{\mathrm{CTL}}}^{x_{\mathrm{BOx}}} \left[\frac{S(x)J_{\mathrm{IN}}}{q\mu F_{\mathrm{CTL}}} \exp\left(-\int_x^{x_{\mathrm{BOx}}} \frac{\sigma v_t N_t}{\mu F_{\mathrm{CTL}}} dx'\right)\right]dx. \tag{10}$$

In the assumption that $S(x)$ is uniform over the CTL and that F_{CTL} is the constant average field over the CTL, we obtain

$$\begin{aligned} n_{\mathrm{BOx}} &= \int_{x_{\mathrm{CTL}}}^{x_{\mathrm{BOx}}} \left[\frac{S(x)J_{\mathrm{IN}}}{q\mu F_{\mathrm{CTL}}} \exp\left(-\frac{\sigma v_t N_t}{\mu F_{\mathrm{CTL}}}(x_{\mathrm{BOx}} - x)\right)\right]dx \\ &= \frac{J_{\mathrm{IN}}}{qt_{\mathrm{CTL}}\mu F_{\mathrm{CTL}}} \frac{\mu F_{\mathrm{CTL}}}{\sigma v_t N_t}\left(1 - \exp\left(-\frac{\sigma v_t N_t}{\mu F_{\mathrm{CTL}}} t_{\mathrm{CTL}}\right)\right) \\ &= \frac{J_{\mathrm{IN}}}{qB_{\mathrm{CTL}}}\left(1 - e^{-\frac{B_{\mathrm{CTL}}}{\mu F_{\mathrm{CTL}}}}\right) \end{aligned} \tag{11}$$

with $B_{\mathrm{CTL}} = t_{\mathrm{CTL}}\sigma v_t N_t$. Inserting this expression into (5) and considering that $v_{\mathrm{BOx}} = \mu F_{\mathrm{CTL}}$ we obtain

$$J_{\mathrm{OUT}} = \frac{J_{\mathrm{IN}}\mu F_{\mathrm{CTL}}}{B_{\mathrm{CTL}}}\left(1 - e^{-\frac{B_{\mathrm{CTL}}}{\mu F_{\mathrm{CTL}}}}\right) \tag{12}$$

which shows that the outgoing current density is a fraction of the injected current density, determined by the CTL field, carrier mobility, and trapping properties of the CTL.

With expressions for J_{IN} and J_{OUT} derived in (3) and (12), respectively, we can now insert them into (2) for the VCR

$$\begin{aligned}\frac{dV_{\mathrm{T}}(t)}{dt} &= \frac{1}{C_q}\left(J_{\mathrm{IN}} - \frac{J_{\mathrm{IN}}\mu F_{\mathrm{CTL}}}{B_{\mathrm{CTL}}}\left(1 - e^{-\frac{B_{\mathrm{CTL}}}{\mu F_{\mathrm{CTL}}}}\right)\right) \\ &= \frac{1}{C_q}\underbrace{qn_c v_t e^{-\frac{B_{\mathrm{TOx}}}{F_{\mathrm{TOx}}}}}_{\text{Injection}}\underbrace{\left(1 - \frac{\mu F_{\mathrm{CTL}}}{B_{\mathrm{CTL}}}\left(1 - e^{-\frac{B_{\mathrm{CTL}}}{\mu F_{\mathrm{CTL}}}}\right)\right)}_{\text{Escape}}\end{aligned} \tag{13}$$

where we have grouped factors relating to the injection of carriers into and escape from the CTL. F_{TOx}, F_{CTL}, and B_{CTL} are time dependent variables. F_{TOx} is determined by the applied gate voltage, compensated by the charge that is captured in the CTL. At a given time step t_i, we have

$$F_{\mathrm{TOx}}(t_i) = \frac{(V_{\mathrm{G}} - \Delta V_{\mathrm{T}}(t_i))C_{\mathrm{tot}}}{C_{\mathrm{TOx}}t_{\mathrm{TOx}}} \tag{14}$$

where C_{TOx} is the capacitance over the TOx layer defined by the channel-TOx and TOx-CTL interfaces and C_{tot} is the total capacitance over the entire TOx-CTL-BOx stack (see Appendix I). For F_{CTL}, we assume that the average field over the CTL during the program operation is unaffected by the captured charge, even though the spatial profile changes

$$F_{\mathrm{CTL}}(t_i) = \frac{V_G C_{\mathrm{tot}}}{C_{\mathrm{CTL}}t_{\mathrm{CTL}}} \tag{15}$$

where C_{CTL} is the capacitance over the CTL (see Appendix I). Finally, B_{CTL} varies with time as the available trap density N_t changes during the program operation

$$N_t(t_i) = N_{t,0} - n_{\mathrm{capt}}(t_i) = N_{t,0} - \frac{\Delta V_{\mathrm{T}}(t_i)C_q}{qt_{\mathrm{CTL}}} \tag{16}$$

with $N_{t,0}$ the total density of traps. With these variables defined, (13) can be solved as an ordinary differential equation using numerical time integration. In our implementation, we use the ode45 solver of MATLAB [10]. Other assumptions for $S(x)$ lead to different expressions for (12) and (13) and are discussed in Appendix II.

Fig. 2 shows that the Pheido model can match experimental data for a 3-D NAND cell and illustrates the contribution of the different components of the VCR to the final ISPP curve. The data was measured on our in-house gate-all-around (GAA) test vehicle, which has three gates and is fabricated on a 300 mm platform with a BiCS-like flow [11]. If only the injection factor of (13) is included, combined with traps that do not fill up ($N_t = N_{t,0}$), the simulated ISPP slope quickly rises to 1 with increasing V_{PGM} and remains ideal for the remainder of the curve. With the inclusion of the escape factor of (13), the start of programming is delayed and the slope does not become ideal, but rather reaches a slowly rising plateau. Finally, if we include the filling of the traps as described in (16), the slope degrades at high V_{PGM}, with the peak slope remaining significantly below 1. With all three components activated, we are able to match the experimental data for the parameter values given in Table I. In the next section, we go into more detail on the different regimes of the ISPP curve, and how they can be understood in light of the Pheido model.

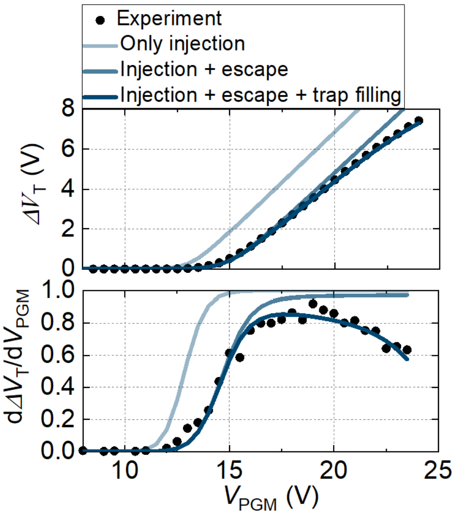

Fig. 2. ISPP curve (top) and slope (bottom) simulated with Pheido, illustrating the components of the model and comparing with an experimental GAA 3-D NAND cell with a memory hole diameter of 120 nm.

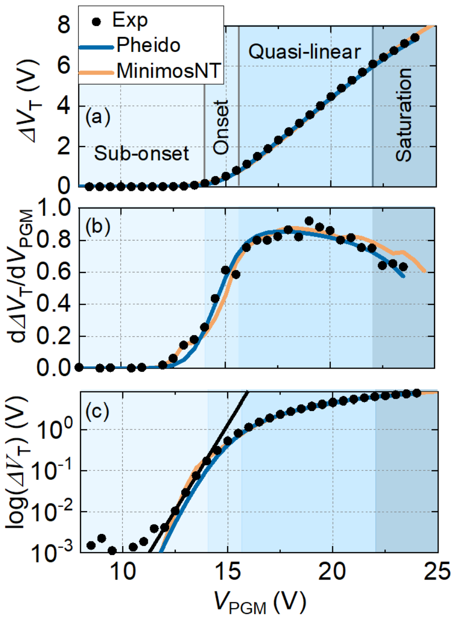

Fig. 3. (a) ISPP curve, (b) programming slope, and (c) logarithmic ISPP curve of a 3-D NAND flash cell with a memory hole diameter of 120 nm, comparing experimental data (Exp), a semianalytical model (Pheido) and the full TCAD approach from Part I (MinimosNT).

TABLE I
SIMULATION PARAMETERS: FIXED (TOP) AND CALIBRATED (BOTTOM)

Symbol	Meaning	Value
$t_{\mathrm{TOx/CTL/BOx}}$	thickness of TOx/CTL/BOx	6/6/6 nm
$\epsilon_{\mathrm{TOx/CTL/BOx}}$	permittivity in TOx/CTL/BOx	4.15/7.4/3.9 ϵ_0
t_{p}	pulse duration	100 μs
μ	electron mobility in CTL	0.07 cm^2/Vs
σ	capture cross section in CTL	1x10^{-16} cm^2
$m^*_{\mathrm{TOx/BOx}}$	effective mass in TOx/BOx	0.45 m_0
E_{b}	band offset channel-CTL	3.12 eV
n_{c}	inversion carrier density in channel	6e20 cm^{-3}
v_{t}	carrier thermal velocity in channel	1e7 cm/s
$N_{\mathrm{t,0}}$	trap density in CTL	5e19 cm^{-3}

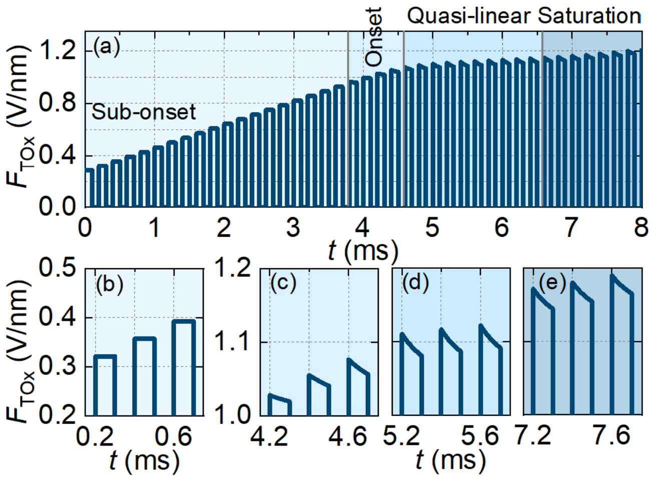

Fig. 4. Average electric field over the TOx during the ISPP operation. (a) Entire ISPP operation and representative pulses during (b) subonset, (c) onset, (d) quasi-linear, and (e) saturation regimes.

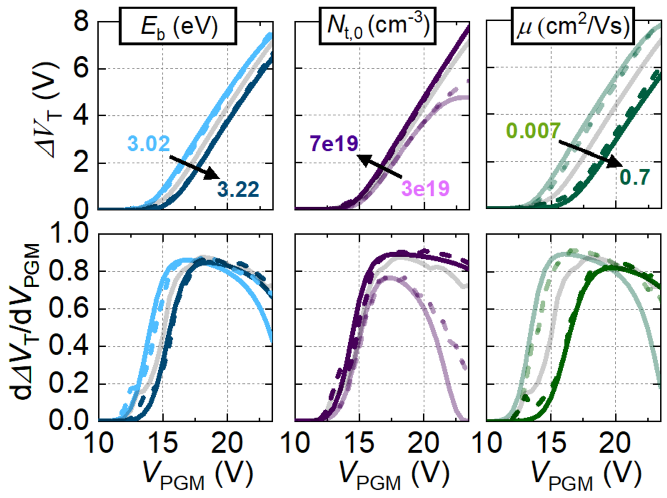

Fig. 5. ISPP curves (top) and slopes (bottom) for varying memory stack material parameters simulated with Pheido (solid) and MinimosNT (dashed), relative to the calibrated case of Table I (gray).

B. ISPP Curve Analysis

The ISPP curve and corresponding slope of a NAND flash cell can be subdivided into four regimes [see Fig. 3(a) and (b)]. The first is the subonset regime, in which the amount of trapped charge in the CTL is still small, such that it does not meaningfully affect the field over the TOx (F_{TOx}, see Fig. 4). With each pulse, F_{TOx} rises in constant steps with V_{PGM} and the injected tunneling current therefore increases exponentially. A significant fraction of these injected carriers are captured in the CTL as there are still many empty traps, causing the ΔV_{T} to rise exponentially as well. The ISPP curve is not usually plotted in a logarithmic scale, but we do so in Fig. 3(c) to confirm the presence of this exponential regime both in the simulated and experimental data. We then define the transition to the programming onset regime at the V_{PGM} for which the ISPP curve deviates from the exponential trend. At this point, the captured charge is sufficient to start partially compensating F_{TOx} [Fig. 4(c)]. This compensation rises both within a pulse and across different pulses as the charge is being captured. The field compensation rises according to the field dependence of the VCR, expressed in (13). If all injected carriers were captured in the CTL, the compensation would keep rising with each pulse until the step in V_{PGM} would be fully compensated and the ISPP slope would reach 1. In reality, however, some carriers escape: they reach the BOx without getting captured and either tunnel out toward the gate or migrate laterally. The probability of escape increases with V_{PGM} as the field in the CTL increases, and thereby the velocity of the carriers moving through it. This counteracts the increase in field compensation, and a gradually rising plateau is reached in the ISPP slope. At the same time, the traps are filling up, which tends to increase the escape of carriers even further. These combined effects result in a quasi-linear regime, in which the increasing injection is balanced out by the escape of carriers, resulting in an almost constant field compensation over a significant number of pulses. The field compensation at which this balance is obtained depends on the specifics of the memory stack, but is generally partial, corresponding to an ISPP slope below 1. When the CTL traps fill up to the point where their availability becomes the limiting factor, the ISPP curve enters the saturation regime. The field compensation drops, as the escape of carriers from the CTL wins out over the increase in injection. As a result, the ISPP slope decreases, with a rate determined by the trap density.

III. Parameter Impact on ISPP Curve

In this section, we use the Pheido model to study the impact of various material and structural parameters on the ISPP curve through their effect on the balance between carrier injection into and escape from the CTL.

A. Material Parameters

First, Fig. 5(a) shows that a change in band offset between the channel and the TOx results in a rigid shift of the ISPP curve, with the slope remaining the same. E_b only impacts the injection into the CTL and is independent of the captured charge. This corresponds to a constant change in the injection prefactor of (13). A similar effect would be seen for a change in the effective mass m^* of the TOx.

Next, Fig. 5(b) shows that a reduction in the trap density degrades the overall programming efficiency: due to the limited availability of traps, the peak ISPP slope in the quasi-linear regime is reduced and saturation sets in at lower V_{PGM} and is more pronounced. This can be understood from the role of N_t in (13) and (16). At the beginning of the ISPP curve, $N_t \approx N_{t,0}$, so B_{CTL} is large compared to μF_{CTL} and the escape factor is close to one. At the start of the quasi-linear regime, however, more and more traps are filled with each pulse, which means N_t begins to decrease as n_{capt} increases. At the V_{PGM} for which n_{capt} becomes significant compared to $N_{t,0}$, the decrease in N_t starts to affect the escape factor noticeably: the VCR field dependence goes down, corresponding to a degradation of the ISPP slope. As the traps are filled, N_t tends to zero and the programming saturates. Fig. 5(b) therefore shows that with lower $N_{t,0}$, the V_{PGM} at

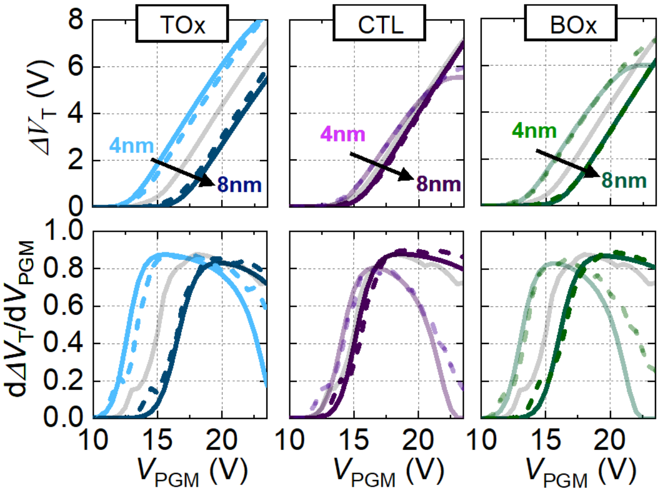

Fig. 6. ISPP curves (top) and slopes (bottom) for varying layer thicknesses at a fixed 120 nm memhole diameter, simulated with Pheido (solid) and MinimosNT (dashed), relative to the calibrated case of Table I (gray).

which the limited trap availability becomes apparent is also reduced. For sufficiently high $N_{t,0}$, on the other hand, the ISPP curve is not affected as sufficient traps are available for the entire V_{PGM} range.

Finally, Fig. 5(c) shows that the carrier mobility in the CTL has a strong impact on both the onset and the slope of the ISPP curve. A higher μ corresponds to a higher carrier velocity, such that more carriers escape without being captured. In (13), this corresponds to a reduction in the escape factor and hence a decrease in the VCR field dependence. In contrast to N_t, however, μ stays constant throughout the V_{PGM} range, so the effect on the saturation regime is less pronounced.

B. Structural Parameters

Fig. 6 shows that an increase in layer thickness of the TOx, CTL, and BOx affects both onset and slope of the ISPP curve: for all layers, the onset shifts to higher V_{PGM}, while the change in slope differs for each layer. The shift in onset in all three cases is a straightforward consequence of the reduction in F_{TOx} that results from the change in EOT of the memory stack. A larger EOT means a given F_{TOx} is reached at a larger V_{PGM}. This corresponds to a constant reduction in the injection prefactor in (13) and hence a rigid shift of the ISPP curve. As the CTL permittivity is larger than that of the other two layers, the EOT change for the same change in physical thickness is smaller, resulting in a smaller onset shift. The slight decrease in ISPP slope for an increase in TOx thickness results from a decrease in capacitive coupling between the captured charge and the channel. For the CTL, an increase in thickness results in an increase in B_{CTL} and a reduction in F_{CTL}, which both increase the escape factor and hence improve the ISPP slope. In the case of the BOx, the change in ISPP slope is opposite to the TOx trend: here, an increase in thickness increases the coupling of the captured charge to the channel relative to the gate. The good match between MinimosNT and Pheido curves for the BOx variation also shows that the approximation in Section II that carriers which reach the CTL-BOx interface disappear from the layer is justified.

IV. Conclusion

We derived a semianalytical model for the programming operation in charge trap NAND flash memory and showed that it can match the ISPP curve and slope of both experimental and full TCAD data. We used the model to define four regimes in the ISPP behavior: subonset, onset, quasi-linear, and saturation, and highlighted the role of the TOx electric field compensation in each. Finally, we illustrated the impact of material and structural cell parameters on the ISPP curve and demonstrated the wide validity of the Pheido model compared to a full TCAD approach.

Appendix I
Capacitance Expressions

This Appendix lists the expressions for the various capacitances in the memory cell. For a cylindrical cell, we have the expression for the capacitance between two concentric plates

$$C_{\mathrm{cyl}}(r_1, r_2) = \frac{2\pi\epsilon}{\ln(r_2/r_1)} \tag{17}$$

where r_1 and r_2 are the radii of the inner and outer cylindrical plate, respectively, and ϵ is the permittivity of the material in between. For the capacitance over the TOx, CTL, BOx, and the total stack, this amounts to

$$\begin{aligned} C_{\mathrm{TOx}} &= C_{\mathrm{cyl}}(x_{\mathrm{TOx}}, x_{\mathrm{CTL}}) = \frac{2\pi\epsilon_{\mathrm{TOx}}}{\ln(x_{\mathrm{CTL}}/x_{\mathrm{TOx}})} \\ C_{\mathrm{CTL}} &= C_{\mathrm{cyl}}(x_{\mathrm{CTL}}, x_{\mathrm{BOx}}) = \frac{2\pi\epsilon_{\mathrm{CTL}}}{\ln(x_{\mathrm{BOx}}/x_{\mathrm{CTL}})} \\ C_{\mathrm{BOx}} &= C_{\mathrm{cyl}}(x_{\mathrm{BOx}}, x_{\mathrm{MH}}) = \frac{2\pi\epsilon_{\mathrm{BOx}}}{\ln(x_{\mathrm{MH}}/x_{\mathrm{BOx}})} \\ C_{\mathrm{tot}}^{-1} &= C_{\mathrm{TOx}}^{-1} + C_{\mathrm{CTL}}^{-1} + C_{\mathrm{BOx}}^{-1} \end{aligned} \tag{18}$$

with x_{TOx}, x_{CTL}, x_{BOx}, and x_{MH} the radii of the various interfaces in the cell stack (see Fig. 1) and ϵ_{TOx}, ϵ_{CTL}, ϵ_{BOx}, and ϵ_{MH} the permittivities of the corresponding layers. C_q is defined as the capacitance between the stored charge centroid and the gate, which we write as a series connection

$$\begin{aligned} C_q^{-1} &= C_{\mathrm{cyl}}^{-1}(x_{\mathrm{Q}}, x_{\mathrm{BOx}}) + C_{\mathrm{cyl}}^{-1}(x_{\mathrm{BOx}}, x_{\mathrm{MH}}) \\ &= \left(\frac{2\pi\epsilon_{\mathrm{CTL}}}{\ln(x_{\mathrm{BOx}}/x_{\mathrm{Q}})}\right)^{-1} + \left(\frac{2\pi\epsilon_{\mathrm{BOx}}}{\ln(x_{\mathrm{MH}}/x_{\mathrm{BOx}})}\right)^{-1} \end{aligned} \tag{19}$$

with x_{Q} the radius of the charge centroid (see Fig. 1). For a planar cell, parallel plate capacitance formulas can be used.

Appendix II
Alternative Distribution Functions

Instead of a uniform distribution for $S(x)$ as in Section II, other assumptions are possible, which each lead to a different form of (11) and (13). Here, we consider three distributions:

1) all tunneling current injected at the TOx-CTL interface; 2) an exponential decay; and 3) a Gaussian.

If all tunneling current through the TOx is injected at the TOx-CTL interface, $S(x) = 1$ only for $x = x_{\mathrm{CTL}}$ and 0 elsewhere. Equation (11) can then be rewritten as

$$\begin{aligned} n_{\mathrm{BOx}} &= \int_{x_{\mathrm{CTL}}}^{x_{\mathrm{BOx}}} \left[\frac{S(x) J_{\mathrm{IN}}}{q\mu F_{\mathrm{CTL}}} \exp\left(-\frac{\sigma v_t N_t}{\mu F_{\mathrm{CTL}}} (x_{\mathrm{BOx}} - x) \right) \right] dx \\ &= \frac{J_{\mathrm{IN}}}{q\mu F_{\mathrm{CTL}}} \exp\left(-\frac{\sigma v_t N_t}{\mu F_{\mathrm{CTL}}} (x_{\mathrm{BOx}} - x_{\mathrm{CTL}}) \right) \\ &= \frac{J_{\mathrm{IN}}}{q\mu F_{\mathrm{CTL}}} \exp\left(-\frac{B_{\mathrm{CTL}}}{\mu F_{\mathrm{CTL}}} \right) \end{aligned} \tag{20}$$

such that (12) reduces to

$$J_{\mathrm{OUT}} = J_{\mathrm{IN}} \exp\left(-\frac{B_{\mathrm{CTL}}}{\mu F_{\mathrm{CTL}}} \right) \tag{21}$$

and (13) simplifies to

$$\begin{aligned} \frac{dV_{\mathrm{T}}(t)}{dt} &= \frac{1}{C_q} \left(J_{\mathrm{IN}} - J_{\mathrm{IN}} \exp\left(-\frac{B_{\mathrm{CTL}}}{\mu F_{\mathrm{CTL}}} \right) \right) \\ &= \frac{1}{C_q} \underbrace{q n_c v_t e^{-\frac{B_{\mathrm{TOx}}}{F_{\mathrm{TOx}}}}}_{\text{Injection}} \underbrace{\left(1 - e^{-\frac{B_{\mathrm{CTL}}}{\mu F_{\mathrm{CTL}}}} \right)}_{\text{Escape}} \end{aligned} \tag{22}$$

where the factors relating to injection and escape from the CTL can be identified. Note that the trapped charge centroid location x_{Q} in (19) becomes an additional fitting parameter as it can no longer be assumed to be in the middle of the CTL.

If the injected current decays exponentially across the CTL [7], $S(x) = Ae^{-\lambda x}$, and (11) can be rewritten as

$$\begin{aligned} n_{\mathrm{BOx}} &= \int_{x_{\mathrm{CTL}}}^{x_{\mathrm{BOx}}} \left[\frac{Ae^{-\lambda x} J_{\mathrm{IN}}}{q\mu F_{\mathrm{CTL}}} \exp\left(-\frac{\sigma v_t N_t}{\mu F_{\mathrm{CTL}}} (x_{\mathrm{BOx}} - x) \right) \right] dx \\ &= \frac{K J_{\mathrm{IN}}}{q\mu F_{\mathrm{CTL}}} \left[e^{-\lambda x_{\mathrm{BOx}}} - e^{-\frac{B_{\mathrm{CTL}}}{\mu F_{\mathrm{CTL}}} - \lambda x_{\mathrm{CTL}}} \right] \end{aligned} \tag{23}$$

where we define K for notational simplicity

$$K = \frac{A}{\frac{\sigma v_t N_t}{\mu F_{\mathrm{CTL}}} - \lambda} \tag{24}$$

such that (12) changes to

$$J_{\mathrm{OUT}} = K J_{\mathrm{IN}} \left[e^{-\lambda x_{\mathrm{BOx}}} - e^{-\frac{B_{\mathrm{CTL}}}{\mu F_{\mathrm{CTL}}} - \lambda x_{\mathrm{CTL}}} \right] \tag{25}$$

and (13) becomes

$$\begin{aligned} &\frac{dV_{\mathrm{T}}(t)}{dt} \\ &= \frac{1}{C_q} \underbrace{q n_c v_t e^{-\frac{B_{\mathrm{TOx}}}{F_{\mathrm{TOx}}}}}_{\text{Injection}} \underbrace{\left(1 - K \left[e^{-\lambda x_{\mathrm{BOx}}} - e^{-\frac{B_{\mathrm{CTL}}}{\mu F_{\mathrm{CTL}}} - \lambda x_{\mathrm{CTL}}} \right] \right)}_{\text{Escape}} \end{aligned} \tag{26}$$

where A can be obtained from the normalization in (7)

$$A = \frac{1}{\int_{x_{\mathrm{CTL}}}^{x_{\mathrm{BOx}}} e^{-\lambda x} dx} = \frac{-\lambda}{e^{-\lambda x_{\mathrm{BOx}}} - e^{-\lambda x_{\mathrm{CTL}}}}. \tag{27}$$

In addition, x_{Q} in (19) changes to $1/\lambda$.

If the injected current is distributed according to a Gaussian function, $S(x)$ takes the form

$$S(x) = \frac{1}{s\sqrt{2\pi}} \mathrm{e}^{\frac{-(x-M)^2}{2s^2}} \tag{28}$$

where M is the distribution mean, which is x_{Q} in (19), and s is the standard deviation. Equation (11) then changes to

$$n_{\mathrm{BOx}} = -\frac{J_{\mathrm{IN}}}{2q\mu F_{\mathrm{CTL}}} e^{Q(x_{\mathrm{BOx}})} (\mathrm{er}(P(x_{\mathrm{BOx}})) - \mathrm{er}(P(x_{\mathrm{CTL}}))) \tag{29}$$

where er is the error function, $A_{\mathrm{CTL}} = \sigma v_t N_t$ and

$$P(x) = \frac{A_{\mathrm{CTL}} s^2 + \mu F_{\mathrm{CTL}} M - \mu F_{\mathrm{CTL}} x}{\sqrt{2} \mu F_{\mathrm{CTL}} s} \tag{30}$$

$$Q(x) = \frac{A_{\mathrm{CTL}} \left(A_{\mathrm{CTL}} s^2 + 2M\mu F_{\mathrm{CTL}} - 2\mu F_{\mathrm{CTL}} x \right)}{2 F_{\mathrm{CTL}}{}^2 \mu^2}. \tag{31}$$

References

[1] J. Alsmeier, M. Higashitani, S. S. Paak, and S. Sivaram, "Past and future of 3D flash," in *IEDM Tech. Dig.*, Dec. 2020, pp. 6.1.1–6.1.4.

[2] K. Nam et al., "Origin of incremental step pulse programming (ISPP) slope degradation in charge trap nitride based multi-layer 3D NAND flash," *Solid-State Electron.*, vol. 175, Jan. 2021, Art. no. 107930.

[3] S. M. Amoroso, A. Maconi, A. Mauri, C. M. Compagnoni, A. S. Spinelli, and A. L. Lacaita, "Three-dimensional simulation of charge-trap memory programming—Part I: Average behavior," *IEEE Trans. Electron Devices*, vol. 58, no. 7, pp. 1864–1871, Jul. 2011.

[4] W.-C. Chen, H.-T. Lue, Y.-H. Hsiao, T.-H. Hsu, X.-W. Lin, and C.-Y. Lu, "Charge storage efficiency (CSE) effect in modeling the incremental step pulse programming (ISPP) in charge-trapping 3D NAND flash devices," in *IEDM Tech. Dig.*, Dec. 2015, pp. 5.5.1–5.5.4.

[5] N. Kariya et al., "8-1 A TCAD study on mechanism and countermeasure for program characteristics degradation of 3D semicircular charge trap flash memory," in *Proc. Int. Conf. Simul. Semiconductor Processes Devices (SISPAD)*, Sep. 2020, pp. 161–164.

[6] D. Verreck et al., "Understanding the ISPP slope in charge trap flash memory and its impact on 3-D NAND scaling," in *IEDM Tech. Dig.*, Dec. 2021, pp. 1–4.

[7] C. Monzio Compagnoni, A. Mauri, S. M. Amoroso, A. Maconi, and A. S. Spinelli, "Physical modeling for programming of TANOS memories in the Fowler–Nordheim regime," *IEEE Trans. Electron Devices*, vol. 56, no. 9, pp. 2008–2015, Sep. 2009.

[8] M. Kim, S. Kim, and H. Shin, "A compact model for ISPP of 3-D charge-trap NAND flash memories," *IEEE Trans. Electron Devices*, vol. 67, no. 8, pp. 3095–3101, Aug. 2020.

[9] F. Schanovsky et al., "A TCAD compatible SONOS trapping layer model for accurate programming dynamics," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2021, pp. 1–4.

[10] L. F. Shampine and M. W. Reichelt, "The MATLAB ODE suite," *SIAM J. Sci. Comput.*, vol. 18, no. 1, pp. 1–22, Jan. 1997.

[11] A. Arreghini et al., "Improvement of conduction in 3-D NAND memory devices by channel and junction optimization," in *Proc. IEEE 11th Int. Memory Workshop (IMW)*, May 2019, pp. 1–4.