Modeling the Operation of Charge Trap Flash Memory–Part I: The Importance of Carrier Energy Relaxation

F. Schanovsky, *Member, IEEE*, D. Verreck, Z. Stanojević, *Senior Member, IEEE*, S. Schallert, A. Arreghini, G. van den Bosch, M. Rosmeulen, and M. Karner, *Member, IEEE*

***Abstract*—We present a novel approach to the modeling of carrier energy relaxation during high-field phases in semiconductor-oxide-nitride-oxide-semiconductor (SONOS) flash memory gate stacks. We show that this method integrates well with TCAD simulators and that taking the energy relaxation of carriers into consideration solves two of the most prominent problems of trapping layer dynamics modeling: The missing slope degradation in incremental step-pulse programming (ISPP) simulations and the incompatibility of the resulting charge distributions with long-term room temperature charge retention measurements. This article consists of two parts where this part discusses the physical/TCAD level. The second part derives a semianalytical model specifically for programming that reduces the numerical complexity while still retaining the main physical assumptions and the applicability to experimental data.**

***Index Terms*—Charge trapping layer (CTL), energy relaxation, flash, incremental step pulse erase (ISPE), incremental step pulse programming (ISPP), semiconductor-oxide-nitride-oxide-semiconductor (SONOS), TCAD.**

I. Introduction

CHARGE trapping-based memories are a class of non-volatile memory cells where the information is stored as the charge state of point defects in a particularly defect rich layer of the gate stack. The most successful implementation of this concept today comes in the form of semiconductor-oxide-nitride-oxide-semiconductor (SONOS) gate statcks. SONOS memories are dominating the solid-state data storage market and will continue to do so for the foreseeable future due to various projected potentials for increase of the bit density ranging from further optimization of the cell geometry, to a better isolation of neighboring cells [1]. Despite this industrial importance and a multidecade long research history of charge-trapping-based memories in general and SONOS structures in particular, the physical understanding of the memory operation is still lacking in fundamental ways.

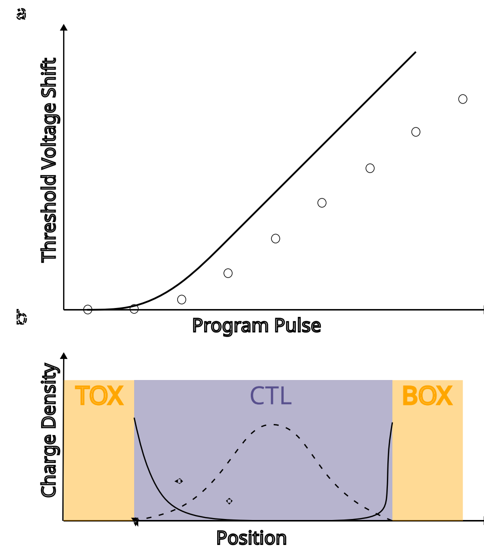

Fig. 1. Schematic illustration of the major challenges in the modeling of SONOS charging and discharging: (a) slopes of measured ISPP curves (circles) are usually grossly overestimated by the simulation (line) and (b) simulated charge distributions in the trapping layer (CTL) after programming usually have their minimum inside the trapping layer and rise sharply toward the surfaces (solid line) with the tunnel oxide (TOX) and BOX while the current understanding of room temperature retention requires the maximum inside the trapping layer and a drop toward the surfaces (solid line).

This fact is manifest in typical discrepancies between simulation and experiment, as well as contradictions between model assumptions as illustrated in Fig. 1. The most notorious shortcoming of established models show up in incremental step-pulse programming (ISPP) and incremental step pulse erase (ISPE) measurements [2], [3], [4], [5], [6]. The nonideality of experimental ISPP/ISPE slopes are hard to reproduce in simulations and only limited work is available in the literature.

Manuscript received 13 November 2023; accepted 27 November 2023. Date of publication 18 December 2023; date of current version 2 January 2024. The review of this article was arranged by Editor S. Alam. *(Corresponding author: F. Schanovsky.)*

F. Schanovsky, Z. Stanojević, S. Schallert, and M. Karner are with Global TCAD Solutions GmbH, 1010 Vienna, Austria (e-mail: f.schanovsky@globaltcad.com).

D. Verreck, A. Arreghini, G. van den Bosch, and M. Rosmeulen are with IMEC, 3001 Leuven, Belgium.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2023.3339076.

Digital Object Identifier 10.1109/TED.2023.3339076

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

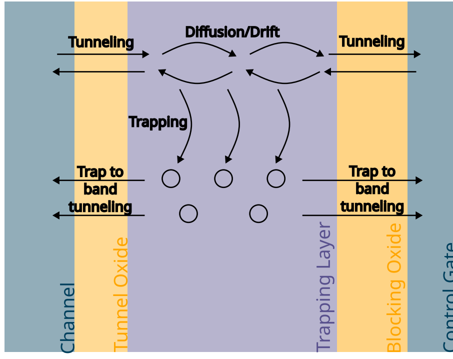

Fig. 2. Physical processes contributing to the charging and discharging kinetics of the SONOS trapping layer.

Another common issue is a discrepancy between the *resulting* trapped charge distributions predicted by the program model and the *initial* charge distributions used as basis for retention loss modelling. The program models usually result in a charge distribution that *exponentially decays* from the surface of the trapping layer toward the interior [7]. Retention loss around room temperature, on the other hand, is understood in terms of a charge distribution that strongly *increases* when going from the trapping layer surface to the inside [8].

In this work we propose the effect of energy relaxation of the carrier in the trapping layer as the missing part in the description of the trapping layer. As we will discuss in a bit more detail in Section III, the energy relaxation of carriers upon injection has received research interest for more than a decade [3], [9], [10], but as of yet no TCAD compatible model has been formulated. This work is the first part of a two-paper series. While this part concentrates on the physical aspects of carrier energy relaxation and the implementation in a TCAD simulator, the second part analyzes the high-level implications for the CTL dynamics using a semianalytical model of the charge trapping layer (CTL).

This article is structured as follows. Section II explains the conventional approach of continuity equation-based modeling of CTL dynamics. Section III highlights the necessity of considering energy relaxation. Section IV discusses our approach of modeling energy relaxation using shape functions. Section V presents several applications to real-world structures, showing the utility of our TCAD model for real-world applications. Finally, Section VII summarizes the findings.

II. Modeling of the Charge Trapping Layer

The physical processes that determine the charging and discharging behavior of the trapping layer are illustrated in Fig. 2:

1) tunneling of free carriers into the CTL conduction/valence band;
2) motion of the carriers within the trapping layer;
3) capture into and emission from localized states;
4) out-tunneling of free carriers in the CTL to gate and channel;
5) trap-to-band tunneling (relevant for long-term retention [8]).

The mathematical models that describe these processes are explained in the following.

A. Carrier Motion

The most central role in the modeling of the trapping layer dynamics is taken by the continuity equation as it not only describes the motion of the carriers but also connects all other physical processes together. The motion of carriers within the trapping layer is generally understood as occurring in the drift-diffusion limit due to the heavy scattering in the disordered nitride [4], [7], [8], [11], [12]. It is described using the continuity equation

$$\frac{\partial n}{\partial t} = -q_0^{-1}\vec{\nabla}\cdot\vec{J}_n - R_n \tag{1}$$

$$\frac{\partial p}{\partial t} = +q_0^{-1}\vec{\nabla}\cdot\vec{J}_p - R_p \tag{2}$$

with the current density $\vec{J}$ given as follows:

$$\vec{J}_n = -\mu_n\vec{E}n - D_n\vec{\nabla}n \tag{3}$$

$$\vec{J}_p = +\mu_p\vec{E}p + D_p\vec{\nabla}p \tag{4}$$

where n and p are the carrier concentrations, q_0 is the elementary charge, R_n and R_p are the recombination rates, μ_n and μ_p are the carrier mobilities, $\vec{E}$ is the electric field, and D_n and D_p are the diffusion coefficients. It is important to note here, that the transport modeled by the continuity equation assumes a *local equilibrium* distribution of carriers, i.e., an equilibrium energy distribution for each carrier type at every point in space.

B. Localized States

The capture and emission of carriers into and out of localized states (iii.) are described by the Shockley–Read–Hall (SRH) theory [4], [7], [8], [13] which describes the occupancy f of a trap at energy E_T as

$$\frac{\partial f}{\partial t} = (c_n + e_p)(1-f) - \left(e_n + c_p + \frac{1}{\tau_{\mathrm{TBT}}}\right)f \tag{5}$$

where c_n and c_p are the capture rates of electrons and holes, e_n and e_p are the corresponding emission rates, and τ_{TBT} is the trap-to-band tunneling emission time constant that is explained in more detail in the following. The capture and emission rates are defined as [13]

$$c_n = \sigma v_{\mathrm{th},n} n \tag{6}$$

$$c_p = \sigma v_{\mathrm{th},p} p \tag{7}$$

$$e_n = \sigma v_{\mathrm{th},n} N_c \exp\left(-\frac{E_c - E_T}{k_{\mathrm{B}}T}\right) \tag{8}$$

$$e_p = \sigma v_{\mathrm{th},p} N_v \exp\left(-\frac{E_T - E_v}{k_{\mathrm{B}}T}\right) \tag{9}$$

where σ is the capture cross section, v_{th} is the thermal velocity of the carriers, N_c and N_v are the conduction and valence effective density of states, and E_c and E_v are the corresponding band edges.

The SRH equations couple to the continuity equation (1) and (2) through the recombination rate R. As in the previous section, it is necessary to point out that the SRH model is derived under equilibrium considerations [13].

C. Trap-To-Band Tunneling

The trap-to-band tunneling is described by the lifetime τ_{TBT} in (5). A widely used model for this process is the simple model of Lundkvist, Lundstrom, and Svensson [14]

$$\tau_{\mathrm{TBT}} = \frac{\tau_0}{\mathrm{TC}_{\mathrm{WKB}}(E_T)} + \tau_1 \tag{10}$$

where $\mathrm{TC}_{\mathrm{WKB}}(E_T)$ is the quantum mechanical transmission coefficient at energy E_T for a particle with tunneling mass m_t through the barrier $V(x)$ as calculated from the Wentzel-Kramers-Brillouin approximation [15]

$$\mathrm{TC}_{\mathrm{WKB}}(E) = \exp\left(-\frac{\sqrt{2m_t}}{\hbar}\int\sqrt{V(x)-E}dx\right) \tag{11}$$

and τ_0 is the basic time constant for the process. As this model tends to overestimate the carrier emission during program and erase, we add a limiting time constant τ_1 [4].

D. Tunneling of Free Carriers

The tunneling of carriers through an insulator is conventionally modeled as an energy integral of the form [7], [8]

$$J_t = \frac{4\pi mq}{h^3}\int N(E)\mathrm{TC}(E)dE \tag{12}$$

where m is the tunnel mass of the carrier, q is the unit charge, h is Planck's constant, and $N(E)$ is the supply function at a given energy E. The transmission coefficient TC is again commonly calculated using the WKB approximation (11). The physical process described by this equation set is the *elastic transition of carriers* from one side of the insulator to the other. The tunnel current J_t is then added to the continuity equation (1) and (2) at the trapping layer surface.

III. Carrier Injection at High Fields

The set of equations laid out above are well established in the literature and usually readily available in commercial TCAD simulators. In the remainder of this article we will refer to this model set as **drift-diffusion only, or DD approach**, as the motion of carriers in the trapping layer is only described by the continuity equation with drift-diffusion currents.

What is often overlooked is the fact that there is a gap in the modeling with respect to the energy of the carriers during high-field (program or erase) phases. Fig. 3 illustrates this for a typical programming situation. As mentioned in Sections II-A and II-D, the insulator tunneling model of the DD approach assumes an *elastic* process, which means that the carriers retain most of their energy during the transition. The continuity equation on the other hand assumes a quasi-equilibrium at each point. By directly attaching the insulator tunneling current to the continuity equation, one implicitly assumes an instantaneous thermalization of the carriers, which is hard to justify considering the large energy gaps involved.

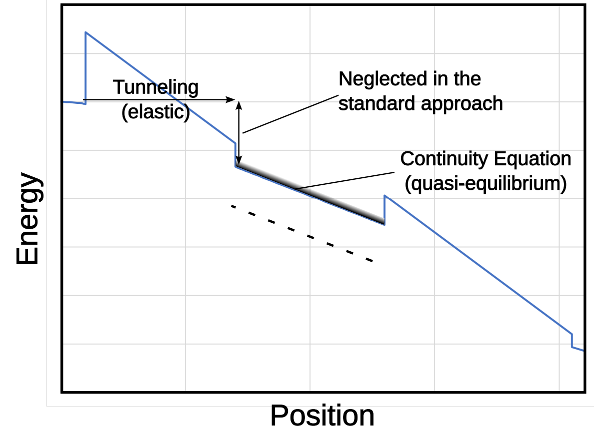

Fig. 3. Energy perspective on programming in the DD approach. The insulator tunneling is described as an elastic process, leading to a large kinetic energy at the point of injection. The continuity equation assumes an equilibrium distribution of carriers, as indicated by the shaded area. This leaves a large energy gap of up to 1.5 eV (depending on the gate voltage) between the carrier injection and the carrier motion that is completely neglected by the DD approach.

The effect of energy relaxation has been studied in the literature before [3], [9] mainly using the model of Tomita et al. [16], where the *average energy* of the carriers is described by a differential equation of the form

$$\frac{dE}{dx} = qF - \frac{E-E_0}{\lambda} \tag{13}$$

where q is the unit charge, F is the electric field, E_0 is the energy at which the thermalization is considered complete, and λ is the energy relaxation length.

This model captures the average behavior of an electron gas well, but has two major drawbacks.

1) It does not easily integrate with the continuity equation. In fact, the publications that describe energy relaxation this way do not have a continuity equation at all [3], [9].
2) It can not be easily generalized to three dimensions. Anything higher than one dimension actually requires the model to be expanded in phase space, which is computationally too expensive.

In the next section, we present our approach for modeling the energy relaxation of carriers, which does not have these limitations and integrates well with the continuity equation in arbitrary dimensions.

IV. Shape Function Approach to Energy Relaxation

When carriers are injected into the trapping layer at high energy, they will gradually lose their excess momentum through a cascade of scattering events that is random in nature. After a long enough relaxation process, the absorption of thermal energy from the lattice phonons starts to become relevant and the carrier energy assumes a thermal equilibrium with the lattice. At this point the carriers are called *thermalized*.

We propose a novel approach for the description of this energy relaxation process in a TCAD simulation. In our approach, instead of describing the energy relaxation explicitly, we focus on the *generation of thermalized carriers*,

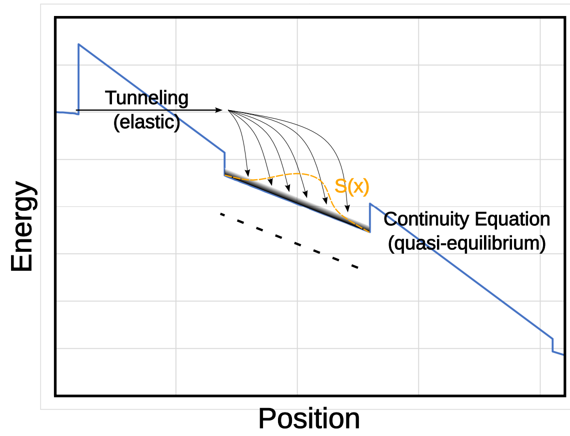

Fig. 4. Energy relaxation in the shape function picture. Carriers injected at a certain energy lose their kinetic momentum over a series of scattering events and eventually thermalize, i.e., assume an equilibrium energy distribution. The distribution of thermalized carriers produced by a given injection point and energy is described by the shape function $S(x)$.

as illustrated in Fig. 4. For this, we assume that carriers injected from a certain point $\vec{x}_0$ at a certain energy E_0 into the trapping layer will lead to a distribution of thermalized carriers $S(\vec{x}, \vec{x}_0, E_0)$. These thermalized carriers are then well described by the continuity equation (1) and (2) and can be easily added as an additional generation rate $G(\vec{x})$ that contains the complete information of the relaxation process

$$G(\vec{x}) = \frac{4\pi mq}{h^3} \int\int S(\vec{x}, \vec{x}_0, E) N'(E) dE d\vec{x}_0 \tag{14}$$

where the prefactor is the same as in (12) and $N'(E)$ is the one-sided supply function. Due to the prominent role played by the convolution integral in this method, we will refer to it as the **convolution-based injection or CI approach** in the remainder of the paper. The full determination of S involves the solution of the Boltzmann transport equation as well as a tunneling step, which is computationally forbidding. Thus, to make the method practical, approximations need to be made. We will show in the following that already a simple approximation gives very good results for describing the memory operation.

The approximation we choose is based on a succession of a simple tunneling step and an empirical description of the relaxation process. In accord with the illustration in Fig. 4, we determine the injection as an elastic tunneling process that starts at the channel and the gate. In order to remove the effort of the energy integral in (14), we apply the cold carrier approximation in the channel and gate, which is reasonable as long as the carrier energies are negligible compared to the tunnel barrier height. The tunnel currents J_{tn} and J_{tp} passing through the surface points $\vec{x}_s$ then simplify to

$$J_{tn}(\vec{x}_s) = v_n n_s \mathrm{TC}(E_c) \tag{15}$$

$$J_{tp}(\vec{x}_s) = v_p p_s \mathrm{TC}(E_v) \tag{16}$$

where n_s and p_s, and v_n and v_p are the concentration of electrons and holes at the injecting surface (e.g., the channel surface) and their thermal velocities, respectively. The generation rate (14) is then calculated by convolution of the tunnel current and the empirical shape functions S_n or S_p over the surface $\mathcal{S}$ as

$$G_n(\vec{x}) = \int_{\mathcal{S}} J_{tn}(\vec{x}_s) S_n(\vec{x} - \vec{x}_s) d\vec{x}_s \tag{17}$$

$$G_p(\vec{x}) = \int_{\mathcal{S}} J_{tp}(\vec{x}_s) S_p(\vec{x} - \vec{x}_s) d\vec{x}_s. \tag{18}$$

The chosen form of the empirical shape function is

$$S(\vec{d}, \vec{x}) = C \exp\left(\frac{(\vec{x}\cdot\vec{d} - x_0)^2}{2\sigma_{\mathrm{long}}^2}\right) \exp\left(\frac{\vec{x}\cdot\vec{x} - (\vec{x}\cdot\vec{d})^2}{2\sigma_{\mathrm{lat}}^2}\right) \tag{19}$$

where x_0 is the injection distance, $\vec{d}$ is the injection direction, and σ_{long} and σ_{lat} are the longitudinal and lateral spreads of the injection, respectively.

V. Application to Pulsed Programming and Erase

In the following, we apply the CI approach to several SONOS memory devices with different geometries. The convolution-based injection model (16)–(18) has been implemented into the *GTS Framework* commercial TCAD simulation framework [17]. All simulations have been performed using the semi-classical device simulator *MinimosNT* [18]. The CI approach is combined with a standard tunneling model so that the injection of carriers into the trapping layer is described exclusively using CI while the emission of carriers from the trapping layer conduction band is described by the standard Tsu-Esaki energy integral [15]. The selected geometries include planar capacitors that are modeled in 1-D, current state-of-the-art GAA VNAND structures that are simulated as 2-D structures with rotational symmetry, as well as next-generation trench memory test devices that require a full 3-D simulation.

A. Planar SONOS Stack

As a first step, we benchmark the CI method against measurement data from planar capacitor structures. The advantage of planar structures is their uniformity and absence of geometrical effects. The samples are $50 \times 50\,\mu$m square shaped on a p-type silicon substrate. The TOX is 7% SiON at 6 nm thickness, CTL is 6-nm Si_3N_4, the blocking oxide 7 nm of SiO_2. The TiN gate is separated from the blocking oxide through a 2-nm Al_2O_3 high-k layer.

ISPP protocols were executed on fresh devices with program pulse widths ranging from 1 μs to 10 ms and could be reproduced with remarkable accuracy. Fig. 5 shows the calibration result using the CI approach and the DD approach for comparison. The overestimation of the ISPP slope is obvious for the DD approach and requires an aggressive reduction of the trap concentration so that saturation reduces the program efficiency. No such limitation is present in the CI approach.

The origin of the slope reduction that comes from the CI approach can be understood from a simple semianalytical approximation that is discussed in detail in Part II of this article.

To evaluate the erase behavior, ISPE sequences were executed from a fresh device as well as several devices that

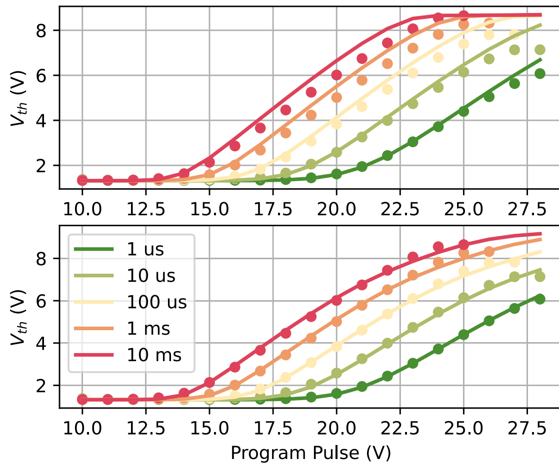

Fig. 5. Comparison of simulated **(lines)** and measured **(symbols)** ISPP curves for a range of pulse widths for the planar SONOS stack using the common DD **(top)** and the novel CI **(bottom)** approach. While the DD version clearly overestimates the slope of the ISPP curves, the experimental data are remarkably well reproduced by the CI approach. The CI type simulation uses acceptor traps with parameters $N_T = 2.88\ 10^{19}\ \text{cm}^{-3}$, $\sigma_n = 0.3\,\text{Å}^2$, and $\sigma_p = 0.05\,\text{Å}^2$. and electron injection parameters $x_0 = 3\,\text{nm}$, $\sigma_{\text{long}} = 1.2\,\text{nm}$. σ_{lat} has no effect in this simulation due to the lateral invariance of the 1-D simulation.

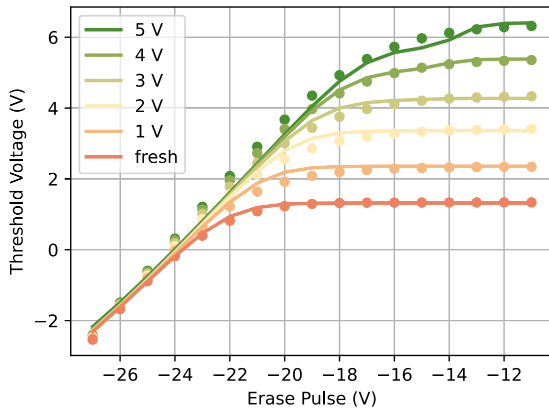

Fig. 6. Comparison of simulated (lines) and measured (symbols) ISPE curves starting from different program states of the planar SONOS stack. The voltages in the legend are the target ΔV_{th} values to which the devices were programmed before erase. In addition to the parameters indicated in Fig. 5, the erase modeling requires donor traps with parameters $N_T = 2.8\ 10^{19}\ \text{cm}^{-3}$, $\sigma_n = 0.05\,\text{Å}^2$, and $\sigma_p = 0.05\,\text{Å}^2$, as well as hole injection with parameters $x_0 = 3\,\text{nm}$, $\sigma_{\text{long}} = 1\,\text{nm}$.

were programmed to different ΔV_{th} targets using $100\,\mu\text{s}$ programming pulses. The erase pulse width was 1 ms. Very accurate fits could also be obtained in this scenario as shown in Fig. 6.

B. GAA VNAND

The CI approach also compares well against research-grade three-gate gate-all-around VNAND structures. The device samples presented here are similar to the ones used in the study reported in [19]. The basic layout of these devices is illustrated in Fig. 7. As is typical for these devices they consist of a SONOS gate stack that is deposited into an etched vertical hole from the outside inwards. The hole size for these structures is 120 nm, the TOX, CTL, and blocking oxide (BOX) are each 6 nm thick.

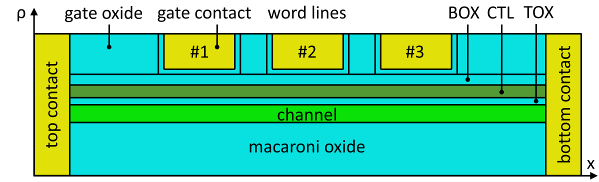

Fig. 7. Simulation structure for the three-word-line gate-all-around VNAND memory devices.

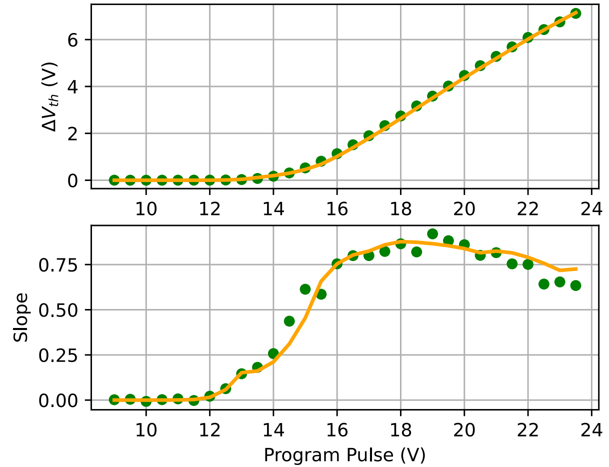

Fig. 8. Comparison of simulated **(lines)** and measured **(symbols)** ISPP curves **(top)** and slopes **(bottom)** for a gate-all-around VNAND structure. The simulation uses acceptor traps with the parameters $N_T = 5\ 10^{19}\ \text{cm}^{-3}$ and $\sigma_n = \sigma_p = 1\,\text{Å}^2$ and electron injection parameters $x_0 = 3\,\text{nm}$, $\sigma_{\text{long}} = 1.5\,\text{nm}$, and $\sigma_{\text{lat}} = 3\,\text{nm}$.

The simulation setup uses the same model set as for the planar structure with slightly different parameters as indicated in Fig. 8. The geometry is represented as a 2-D simulation simulation structure with rotational symmetry aound the x-axis as indicated in Fig. 7.

Fig. 8 shows the calibration results for the ISPP curve as well as the ISPP slope. Compared to their planar counterparts, the GAA devices show a slightly earlier onset of the programming and an increased slope of the linear part. This is due to the higher capacitance in the TOX of the cylindrical structure, see also Appendix I of the second part of this article. Both the ISPP curve and its slope are reproduced very well over the whole voltage range.

C. Trench

As a further testing ground, we evaluate the CI approach against a potential post-GAA-VNAND technology that has been introduced in 2017 [2]. The technology uses a process that is related to the production process of GAA VNAND devices, but features a further increase in the bit density. Due to its complex geometry, this device requires a fully three dimensional simulation to accurately capture its charging and discharging behavior, see Fig. 9.

The sample devices were produced for the study published in [20], which contains a more detailed description of the geometry and the processing. Due to the large statistical

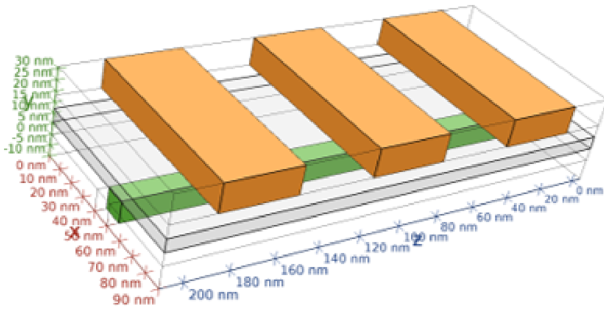

Fig. 9. Simulation structure used for the three-gate trench memory devices.

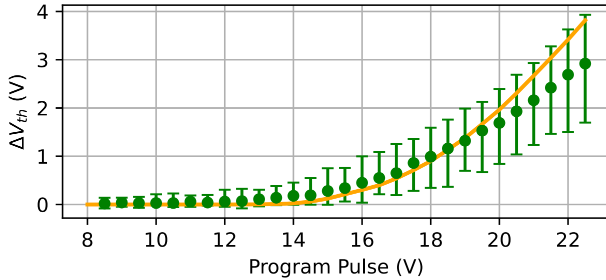

Fig. 10. Comparison of simulated and measured ISPP curves for the trench structure. The points and bars are indicating the average values and the confidence interval of the experimental data. The simulation lies within the expected range of the experimental data. The simulation uses acceptor traps with the parameters $N_T = 2.88\ 10^{19}\ \mathrm{cm}^{-3}$, $\sigma_n = \sigma_p = 1\ \text{Å}^2$ and electron injection parameters $x_0 = 3\ \mathrm{nm}$, $\sigma_{\mathrm{long}} = 1.5\ \mathrm{nm}$, $\sigma_{\mathrm{lat}} = 3\ \mathrm{nm}$.

variability of the measured data, error bars are added to the comparison in Fig. 10. While an accurate calibration to the average data could not be performed yet, the simulation results are within the error range of the measurements.

VI. Room Temperature Retention

Charge retention experiments at room temperature can provide valuable insight into the charge distribution inside the trapping layer. The temperature sensitivity of the long-time charge loss in charge trapping devices features two regions with a larger temperature sensitivity at high temperatures and a small one around room temperature [21]. This is generally explained as a combination of two different charge loss mechanisms: thermionic emission and trap-to-band tunneling. As the trap levels of the CTL traps are quite deep at around 1.5 eV from the conduction band edge [22], the thermionic emission of carriers from the traps is only relevant at high temperatures, but plays a negligible role at room temperature. The room temperature charge loss that is observed is generally understood to be governed by trap-to-band tunneling [23]. The transient emission of charge can then be interpreted in terms of a tunneling front that starts at the channel and moves into the CTL as the time proceeds, emptying the traps on its way. Therefore, the slope of the charge loss strongly depends on the distribution of charge in the trapping layer [8]. This poses another challenge to the simulation. While the DD approach results in a charge distribution that peaks at the trapping layer surfaces and falls toward the interior [7], the rather small

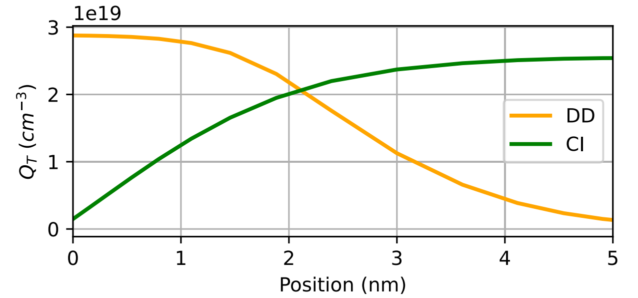

Fig. 11. Comparison of the charge distributions resulting from the DD and the CI approach after programming. The former shows the typical drop from the channel-facing surface (at position zero) toward the interior, while the latter shows the opposite trend that is more in line with the assumptions of published retention models [8].

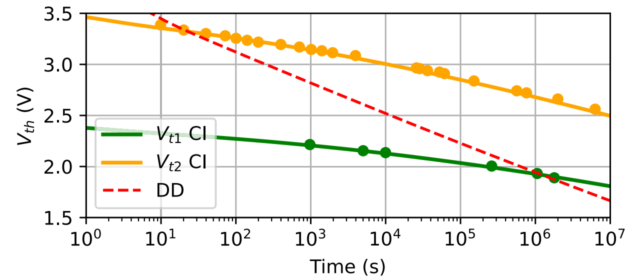

Fig. 12. Comparison of simulated (lines) and measured (symbols) room temperature charge loss for two different program states (data from [8]). The fits were made with charge distributions resulting from program simulations using the CI approach. The dashed line represents the typical charge loss starting from a charge distribution predicted by the DD approach, which is always too steep so that a meaningful calibration is not possible.

slopes in the charge loss require a charge distribution that rises from the surface toward the interior [8].

We have evaluated the impact of the CI-based carrier injection model on the charge distribution. Typical charge distributions resulting from the DD and the CI approach are shown in Fig. 11. While the DD approach shows a drop in the charge concentration from the surface of the trapping layer toward the interior, the CI-based model shows the opposite trend.

In order to compare against experimental data we use the measurement results of [8] that that were obtained from a structure with 2.2-nm TOX, 6-nm trapping layer, and 8-nm blocking oxide. The retention data are taken from two different programming states. As shown in Fig. 12, a very good fit could be obtained using the charge distributions resulting from the shape function-based injection. Using the charge distribution resulting from the simple insulator tunneling approach always results in a too high slope of the charge loss curve, which makes any attempts at a calibration futile.

VII. Summary and Conclusion

In this section, we have presented a new shape-function-based approach to modeling the carrier injection into the CTL of SONOS devices during programming and erase. We have demonstrated the feasibility and accuracy of the approach for different geometries ranging from simple planar memory stacks to post-GAA-VNAND trench structures. We found that

this approach resolves the two most prominent problems of SONOS simulation: The ISPP slope and the charge distribution in the trapping layer after programming.

References

[1] L. Heineck and J. Liu, "3D NAND flash status and trends," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2022, pp. 1–4.

[2] S.-C. Lai et al., "A bottom-source single-gate vertical channel (BS-SGVC) 3D NAND flash architecture and studies of bottom source engineering," in *Proc. IEEE 8th Int. Memory Workshop (IMW)*, May 2016, pp. 1–4.

[3] M. Hogyoku, Y. Yokota, and K. Nishitani, "TCAD simulation for capture/emission of carriers by traps in SiN: Trap-assisted tunneling model extended for capture of carriers injected via Fowler–Nordheim tunneling," *Jpn. J. Appl. Phys.*, vol. 61, Apr. 2022, Art. no. SC1087.

[4] 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.

[5] H.-T. Lue et al., "A novel double-density hemi-cylindrical (HC) structure to produce more than double memory density enhancement for 3D NAND flash," in *IEDM Tech. Dig.*, Dec. 2019, p. 28.

[6] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003811012030397X

[7] E. Vianello et al., "Experimental and simulation analysis of program/retention transients in silicon nitride-based NVM cells," *IEEE Trans. Electron Devices*, vol. 56, no. 9, pp. 1980–1990, Sep. 2009.

[8] A. Arreghini, N. Akil, F. Driussi, D. Esseni, L. Selmi, and M. J. van Duuren, "Long term charge retention dynamics of SONOS cells," *Solid-State Electron.*, vol. 52, no. 9, pp. 1460–1466, Sep. 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038110108001330

[9] 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.

[10] 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.

[11] A. Padovani, L. Larcher, D. Heh, and G. Bersuker, "Modeling TANOS memory program transients to investigate charge-trapping dynamics," *IEEE Electron Device Lett.*, vol. 30, no. 8, pp. 882–884, Aug. 2009.

[12] A. Padovani and L. Larcher, "A novel algorithm for the solution of charge transport equations in MANOS devices including charge trapping in alumina and temperature effects," in *Proc. Int. Conf. Simul. Semiconductor Processes Devices*, Sep. 2010, pp. 229–232.

[13] W. Shockley and W. T. Read Jr., "Statistics of the recombinations of holes and electrons," *Phys. Rev.*, vol. 87, pp. 835–842, Sep. 1952.

[14] L. Lundkvist, I. Lundström, and C. Svensson, "Discharge of MNOS structures," *Solid-State Electron.*, vol. 16, no. 7, pp. 811–823, Jan. 1973. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0038110173901780

[15] A. Gehring and S. Selberherr, "Tunneling models for semiconductor device simulation," in *Handbook of Theoretical and Computational Nanotechnology*. Los Angeles, CA, USA: American Scientific Publishers, 2006, pp. 469–543.

[16] T. Tomita, Y. Kamakura, and K. Taniguchi, "Energy relaxation length for ballistic electron transport in SiO_2," *Phys. Status Solidi (B)*, vol. 204, no. 1, pp. 129–132, Nov. 1997.

[17] *GTS Framework*. Accessed: Dec. 12, 2023. [Online]. Available: https://www.globaltcad.com/framework

[18] *GTS Minimos-NT*. Accessed: Dec. 12, 2023. [Online]. Available: https://www.globaltcad.com/mmnt

[19] 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.

[20] S. Rachidi et al., "Enabling 3D NAND trench cells for scaled flash memories," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2023, pp. 1–4.

[21] D. Oh et al., "TCAD simulation of data retention characteristics of charge trap device for 3-D NAND flash memory," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2015, pp. 1–4.

[22] A. Suhane et al., "Validation of retention modeling as a trap-profiling technique for SiN-based charge-trapping memories," *IEEE Electron Device Lett.*, vol. 31, no. 1, pp. 77–79, Jan. 2010.

[23] Y. Wang and M. H. White, "An analytical retention model for SONOS nonvolatile memory devices in the excess electron state," *Solid-State Electron.*, vol. 49, no. 1, pp. 97–107, Jan. 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038110104002400