Modeling the Operation of Charge Trap Flash Memory: A Monte Carlo Approach to Carrier Distribution and (De)trapping

Thomas Hellemans
Imec and KU Leuven
3001 Leuven, Belgium
thomas.hellemans@imec.be

Devin Verreck
Imec
3001 Leuven, Belgium
devin.verreck@imec.be

Antonio Arreghini
Imec
3001 Leuven, Belgium
antonio.arreghini@kuleuven.be

Geert Van den Bosch
Imec
3001 Leuven, Belgium
geert.vandenbosch@imec.be

Maarten Rosmeulen
Imec
3001 Leuven, Belgium
maarten.rosmeulen@imec.be

Michel Houssa
KU Leuven and Imec
3001 Leuven, Belgium
michel.houssa@kuleuven.be

Jan Van Houdt
Imec and KU Leuven
3001 Leuven, Belgium
jan.vanhoudt@imec.be

2024 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD) | 979-8-3315-1635-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/SISPAD62626.2024.10733042

***Abstract*—We propose a Monte Carlo framework including (de)trapping to describe the non-equilibrium operation of charge trap flash memories. We thereby base the empirical carrier distribution that was proposed in recent studies on physical parameters. After an outline of the simulation procedure, we show how carrier trapping and detrapping is included. Finally, we illustrate the simulators' capabilities via simulations of the programming and retention operations, highlighting the insight into the carrier dynamics that this approach enables.**

Keywords—Charge Trap Memory, Monte Carlo, Trapping, Detrapping, Carrier distribution, energy relaxation

I. Introduction

As charge trap memories have made the transition towards vertical 3D-NAND strings and memory cells have been scaled down over the last decade, amorphous silicon nitride ($a-Si_3N_4$) has been adopted as the storage material due to its easier integration process compared to a floating gate [1]. The resulting architecture is a semiconductor-oxide-nitride-oxide-semiconductor (SONOS) stack, in which neighboring memory cells share the same charge trap layer (CTL). Therefore, in order to further scale charge trap memory cells, it becomes increasingly more important to understand the carrier transport within the CTL itself. This understanding is crucial for accurately predicting phenomena such as programming dynamics or lateral charge migration towards neighboring cells during retention. Existing models for programming charge trap memory often overlook energy relaxation. Our group has recently proposed a TCAD model, co-developed with Global TCAD Solutions [3], and a semi-analytical model, called Pheido [2], that treat the energy relaxation empirically. Both rely on a shape function S, which characterizes the distribution of charge over the CTL and is usually assumed to take a Gaussian form. In this study, we develop a Monte Carlo framework to base this choice of S on a more physical basis and to calculate subsequently the charge distribution within the CTL. This is achieved by considering energy relaxation in combination with charge trapping and detrapping mechanisms.

II. Monte Carlo Procedure

We developed an ensemble Monte Carlo simulator for the Boltzmann Transport Equation that models the free-flight, scattering and (de)trapping of carriers in the CTL. Fig. 1 shows the procedure: carriers are injected and can end a free flight by scattering or by getting trapped. Both scattering and trapping are treated as instantaneous events localized in space.

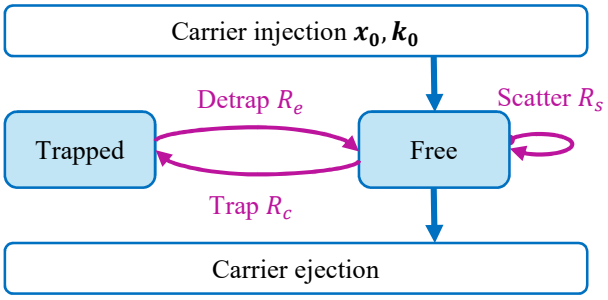

Fig. 1. Monte Carlo procedure, this flow is interrupted at fixed timesteps to solve the Poisson equation.

Trapping implies a transition to a localized trapped state from which particles can only escape by detrapping.

The **electron trajectory** in momentum-space is determined by the classical equations of motion considering a single isotropic parabolic energy band [4]. Both the duration of a free-flight (free state) and the duration of being trapped (trapped state) are determined by the self-scattering algorithm. Currently, only 1D electron transport is included.

For the first simulations, we lump the various **scattering** mechanisms which occur in amorphous trap rich insulators together in an effective inelastic scattering rate. We characterize this scattering rate by a fixed energy independent rate R_s, fixed energy loss ΔE and probability to scatter towards the front P_{front} (1D simulation).

Trapping, including trap filling, is described according to a simplified Shockley-Read-Hall theory as [5]:

$$R_c = \begin{cases} C_c\left(N_T - n_T(x,t)\right) & \text{for } E \le \Delta E \\ 0 & \text{for } E > \Delta E \end{cases} \quad (1)$$

with C_c the capture coefficient, N_T the total trap density and n_T the trapped charge density at position x and time t. Energy dependence for the trapping rate is included by stating that a particle cannot be trapped when its energy with respect to the conduction band exceeds the energy that can be emitted by a single scattering event.

We model **detrapping** with a Poole-Frenkel model in order to include field dependence for the emission rate [6]:

$$R_e = \nu_{PF} \exp\left(\frac{\beta\sqrt{|\mathcal{E}(x,t)|} - E_T}{k_b T}\right) \quad \text{with} \quad \beta = \sqrt{\frac{q^3}{\pi\epsilon_0\epsilon_\infty}} \quad (2)$$

with ν_{PF} the attempt to escape frequency for emission, $\mathcal{E}(x,t)$ the electric field at position x and time t, E_t the trap energy depth with respect to the conduction band minimum, k_b the Boltzmann constant, T the temperature, q the elementary charge and ϵ_∞ the high-frequency $a-Si_3N_4$ dielectric constant.

The **boundary conditions** imposed on the CTL interfaces are the same as those in the Pheido model [2], except for the injection energy. During programming, carrier injection at the tunnel oxide (TuOx) – CTL interface is described by a simplified Fowler-Nordheim equation:

$$J_{in} = q n_{ch} v_{th} \exp\left(\frac{B_{TuOx}}{\mathcal{E}_{TuOx}}\right) \tag{3}$$

with n_{ch} the carrier density in the inversion layer at the channel – TuOx interface, v_{th} the carrier thermal velocity and

$$B_{TuOx} = \frac{4}{3}\frac{\sqrt{2qm^*}}{\hbar} E_{b1}^{3/2} \tag{4}$$

given by the WKB approximation with E_{b1} the conduction band offset between the channel and the TuOx, m^* the effective tunnelling mass, and $\hbar$ the reduced Planck constant. Note that $\mathcal{E}_{TuOx}$ is a negative number (electric field pointing in negative x direction) when applying a positive gate bias, hence accelerating electrons in the positive x direction. The injection energy of the electrons with respect to the conduction band minimum of the CTL is given by (Fig. 2):

$$E_{in} = E_{b1} - E_{b1} - \mathcal{E}_{TuOx} t_{TuOx} \tag{5}$$

with E_{b2} the conduction band offset between the TuOx and CTL. All carriers that reach the CTL – BlOx interface disappear from the simulation, agnostic to the exact physical process by which this happens, e.g. lateral charge migration or tunnelling out through the BlOx. For retention, a similar worst-case scenario regarding the CTL interfaces is considered: all carriers that reach these interfaces disappear from the simulation.

The **Poisson equation** is discretized on a rectangular grid spanning the whole device (TuOx, CTL and BlOx). It is solved on fixed timesteps at which the ensemble of carriers is interrupted. During a single timestep, carrier trajectories are calculated using the potential calculated at the beginning of that timestep. Hence, timesteps must be sufficiently short to prevent significant potential deviations within them. During programming, the detrapping rate (field dependent) is much smaller than the change rate for the electric field. However, the Poole-Frenkel detrapping rate depends exponentially on the field, therefore, it's essential to re-evaluate the detrapping rate at least on the same the timescale as the variation in the electric field.

III. Memory Operation Simulations

We now illustrate the capabilities of the approach by simulating program and retention for the simple memory stack shown in Fig. 2. Other parameters can be found in Table 1. It's important to note that these have not been tailored to a specific measurement, so the depicted simulation results should be viewed solely as a demonstration of the simulators' capabilities.

The incremental step pulse **programming** (ISPP) curves produced by the MC simulator can be fitted with Pheido using a uniform shape function S(x) (Fig. 3). Analysis of the MC results indicates that at low gate voltages, the charge centroid resides near the BlOx interface and then gradually moves towards the CTL center at higher gate voltages. Thus, a uniform S(x) becomes a suitable approximation (Fig. 4a-4b).

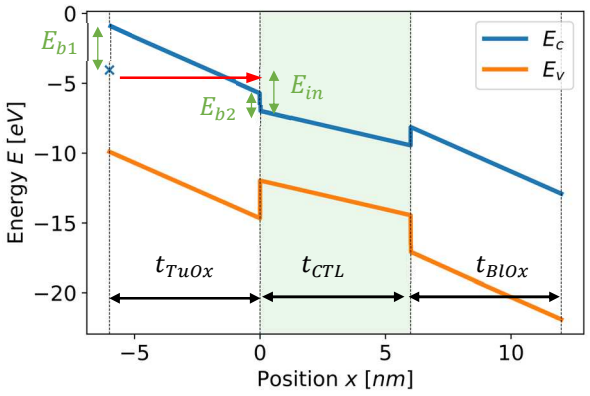

Fig. 2. Band diagram at the start of programming, i.e. without trapped charge. The shaded region corresponds to the charge trap layer (CTL). In this region the electron transport is described by the Monte Carlo framework. The Poisson equation is solved for the complete device (TuOx, CTL and BlOx).

TABLE I. Simulation Parameters

Parameter	Unit	Value	
		MC	Pheido
$\boldsymbol{A}$ (area planar)	nm^2	2.5e5	
$\boldsymbol{t_{TuOx/CTL/BlOx}}$	nm	6/6/6	
$\boldsymbol{\epsilon_{Tuox/CTL/BlOx}}$	ϵ_0	3.9/7.4/3.9	
$\boldsymbol{V_{step}}$	V	0.5	
$\boldsymbol{t_p}$	μs	100	
$\boldsymbol{n_{ch}}$	cm^{-3}	2e21	
$\boldsymbol{v_{th}}$	cm/s	1e7	
$\boldsymbol{E_{b1}}$	eV	3.15	
$\boldsymbol{E_{b2}}$	eV	1.3	–
$\boldsymbol{m^*}$	m_0	0.45	
$\boldsymbol{R_s}$	s^{-1}	4e15	–
$\boldsymbol{\Delta E}$	eV	0.2	–
$\boldsymbol{P_{front}}$	–	0.9	–
$\boldsymbol{C_c}$	cm^3/s	5e-8	–
$\boldsymbol{N_T}$	cm^{-3}	4e19	
$\boldsymbol{\nu_{PF}}$	s^{-1}	5e7	–
$\boldsymbol{E_t}$	eV	1.6	–
$\boldsymbol{T}$	K	300	–
$\boldsymbol{\mu}$	cm^2/Vs	–	0.07
$\boldsymbol{\sigma}$	cm^{-2}	–	2e-16

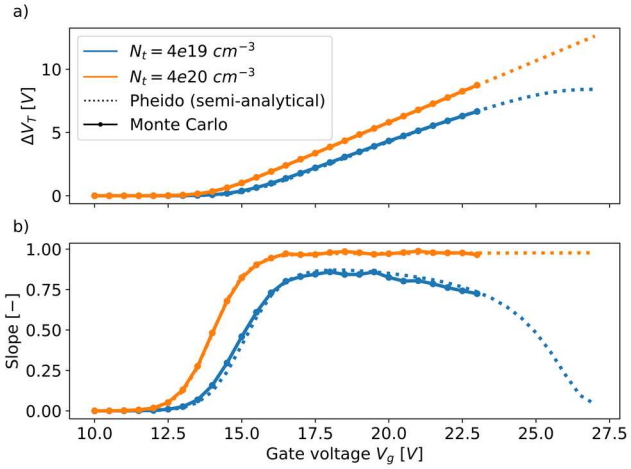

Fig. 3. ISPP curve from MC simulator and Pheido at different trap densities N_t. We use a voltage step $V_{step} = 0.5V$ and programming pulse $t_p = 100\mu s$.

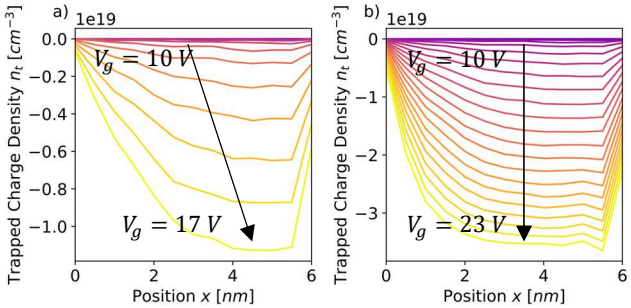

Fig. 4. Trapped charge density n_t inside the CTL in function of time during programming. Initially trapped charge builds up near the CTL – BlOx interface. At higher gate voltages a more uniform charge profile is formed.

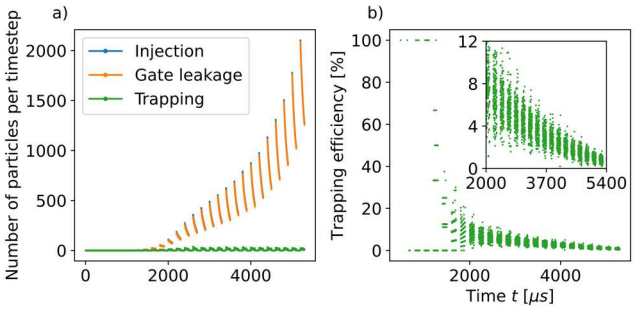

Fig. 5. (a) Particle current, expressed in number of particles per Poisson-timestep, in function of time during programming. We call the difference between the injection current and gate leakage current, the trapping current. Notice that this is not a physical current at one of the contacts of the device. (b) Trapping efficiency in function of time during programming. The trapping efficiency is the ratio of the trapping current and injection current.

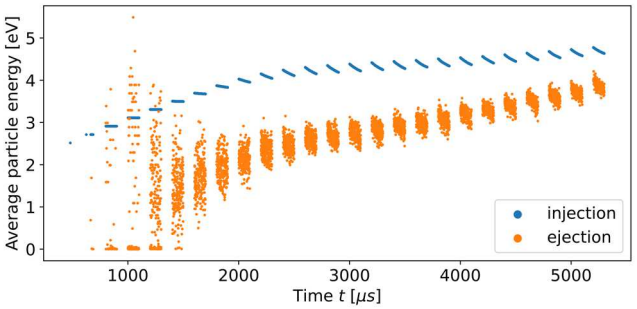

Fig. 6. Average particle injection and ejection energy for each Poisson-timestep as a function of time during programming. The graph shows that injection happens via Fowler-Nordheim tunneling through the TuOx. Additionally, it indicates that a significant amount of carriers reach the CTL – BlOx interface with energies exceeding the nitride – oxide conduction band offset.

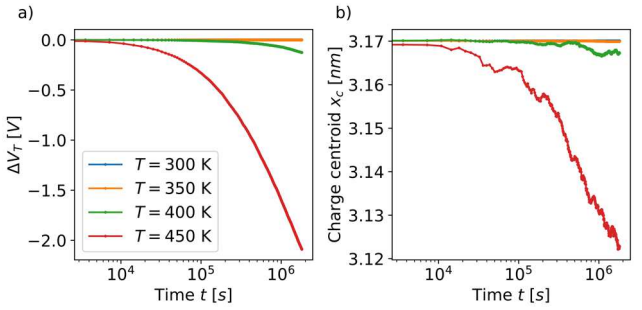

Fig. 7. (a) Shift in threshold voltage ΔV_T and (b) shift in charge centroid x_c during retention after ISPP up to $V_g = 23\ V$. The temperature dependence of the ΔV_T curve can be attributed to the temperature dependence of the emission rate (Poole-Frenkel).

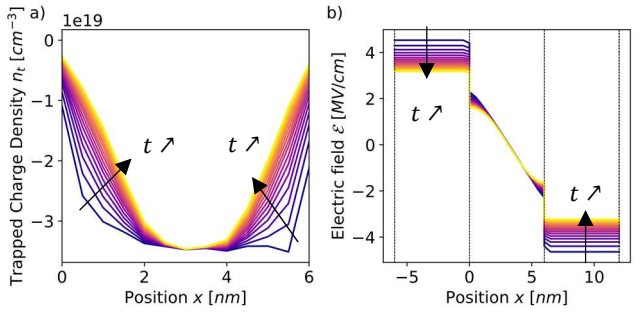

Fig. 8. (a) Charge density and (b) electric field during a 3-weeks retention simulation at temperature $T = 450\ K$. Charge loss initially occurs near the TuOx and BlOx interfaces.

The main advantage of the MC simulator over Pheido is the **detailed information on the carrier dynamics** which can be extracted. For example, Fig. 5a illustrates the injection current, gate leakage current and trapping current (injection minus leakage) as number of particles per Poisson timestep. The trapping efficiency in Fig. 5b is given by the ratio of the trapping current and injection current. Starting from the onset of programming, this value is mainly below 10% in the current simulation setup. The carrier dynamics can be studied further by examining the particles' injection and ejection energy, shown in figure Fig. 6. A specific tunneling-out model for the BlOx is omitted in the simulator, but the graph illustrates that a significant number of particles reach the CTL – BlOx interface with energies larger than the nitride – oxide conduction band offset. Again, this is only valid for the given set of scattering and (de)trapping mechanisms and parameters, which are not calibrated. Further information can be extracted in the form of energy distributions, the occurrence of detrappping events etc. (not shown).

Only **high temperature retention** simulations with the MC simulator are currently justified, due to the dominance of trap-to-band tunneling over Poole-Frenkel emission at room temperature [3]. Our simulation spans a retention time of three weeks with hourly electric field updates. The initial state is the final state after ISPP up to $V_g = 23\ V$. Poole-Frenkel detrapping is evident only at high temperatures (eq. 2, Fig. 7a) and throughout retention the charge centroid gradually shifts towards the CTL center (Fig. 7b). This is explained by Fig. 8, which demonstrates that for $T = 450K$ charge loss primarily occurs at the CTL interfaces due to higher electric fields compared to the CTL center (eq. 2).

IV. Conclusions

We propose a Monte Carlo framework which includes (de)trapping mechanisms and have demonstrated that this approach captures the programming and high-temperature retention behavior of charge trap flash memory. Additionally, the simulator offers valuable insights into carrier dynamics. Future work will focus on extending the framework with additional scattering and (de)trapping physics, and 2D/cylindrical coordinates.

Acknowledgment

This work was supported by imec's Industrial Affiliation Program for storage memories.

References

[1] A. Goda, "Recent Progress on 3D NAND Flash Technologies," *Electronics*, vol. 10, no. 24, Art. no. 24, Jan. 2021, doi: 10.3390/electronics10243156.

[2] D. Verreck, F. Schanovsky, A. Arreghini, G. Van den Bosch, Z. Stanojević, M. Karner, *et al.*, "Modeling the Operation of Charge Trap Flash Memory—Part II: Understanding the ISPP Curve With a Semianalytical Model," *IEEE Transactions on Electron Devices*, vol. 71, no. 1, pp. 554–559, Jan. 2024, doi: 10.1109/TED.2023.3339112.

[3] F. Schanovsky, D. Verreck, Z. Stanojević, S. Schallert, A. Arreghini, G. van den Bosch, *et al.*, "Modeling the Operation of Charge Trap Flash Memory–Part I: The Importance of Carrier Energy Relaxation," *IEEE Transactions on Electron Devices*, vol. 71, no. 1, pp. 547–553, Jan. 2024, doi: 10.1109/TED.2023.3339076.

[4] M. Lundstrom, *Fundamentals of carrier transport*, 2. ed., Digitally pr. version. Cambridge: Cambridge University Press, 2009.

[5] E. Vianello, F. Driussi, A. Arreghini, P. Palestri, D. Esseni, L. Selmi, *et al.*, "Experimental and Simulation Analysis of Program/Retention Transients in Silicon Nitride-Based NVM Cells," *IEEE Transactions on Electron Devices*, vol. 56, no. 9, pp. 1980–1990, Sep. 2009, doi: 10.1109/TED.2009.2026113.

[6] J. Frenkel, "On Pre-Breakdown Phenomena in Insulators and Electronic Semi-Conductors," *Phys. Rev.*, vol. 54, no. 8, pp. 647–648, Oct. 1938, doi: 10.1103/PhysRev.54.647.