# Modeling Carrier In- and Ejection for Charge Trap Flash Memory: Insights from Engineered SON(ON)OS Vehicles

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 and KU Leuven*
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

***Abstract* — Despite charge trap flash memories being commercialized, the detailed understanding of the underlying physics remains limited. Here, we therefore evaluate common assumptions for the in- and ejection of charge carriers in the trapping layer, by comparing to measurements of both standard SONOS and engineered SONONOS devices. We find a strong impact of these assumptions on the peak of the trapped charge profile as well as on the accumulation near the blocking oxide. Finally, we highlight the need for a non-local detrapping model.**

***Keywords — Charge Trap Memory, modeling, carrier injection, carrier ejection, charge distribution, SONOS***

## I. Introduction

Charge trap flash memories are programmed by electron tunneling from the channel through the tunnel oxide (TuOx) into the charge trap layer (CTL). These carriers dissipate energy within the CTL, get trapped, and hence alter the device's threshold voltage, enabling memory operation [1]. These processes can be modelled at different abstraction levels, ranging from semi-analytical compact models like Pheido [2] over numerical drift-diffusion calculations [3] to a Monte-Carlo framework to explicitly account for energy relaxation [4]. However, none of these abstraction levels offer a unified model that encompasses different operation regimes, i.e. programming, retention, and erase. To obtain new insights into the physical mechanisms, we propose to broaden our experimental parameter space by using specially designed SONOS and SONONOS vehicles. In this study, we specifically investigate how different in- and ejection models affect the trapped charge profile during programming. This programmed trapped charge profile will be crucial for understanding retention and erase behavior in future studies. The insights are broadly applicable since regardless of the simulation's abstraction level, a decision must be made regarding the in- and ejection model.

We use two specifically designed test vehicles. First, we study the impact on SONOS planar capacitors with varying CTL thickness (Fig. 1a). Next, we evaluate the validity of the SONOS calibration by applying it to SONONOS planar capacitors, in which a $4\,nm$ $SiO_2$ barrier is inserted in between two amorphous $Si_3N_4$ CTL layers. The total thickness of both CTLs is $24\,nm$ for each device such that the total EOT for all SONONOS devices is the same (Fig. 1b). Both SONOS and SONONOS devices feature a thin TuOx layer, enabling low-energy carrier injection into the CTL, hence minimizing the effects of energy dissipation. After processing of the devices, CET measurements indicated that the overall CET of the devices is approximately 5% smaller than targeted, hence in the simulations all thicknesses are reduced by 5%.

TABLE I. Device Architecture

|  | SONOS [nm] | SONONOS [nm] |
|---|---|---|
| $t_{TuOx}$ | 2 | 2 |
| $t_{CTL}$ | 24 – 18 – 12 – 6 | 24 – 18 – 12 – 6 |
| $t_{Bar}$ | - | 4 |
| $t_{CTL}$ | - | 0 – 6 – 12 – 18 |
| $t_{BlOx}$ | 8 | 4 |

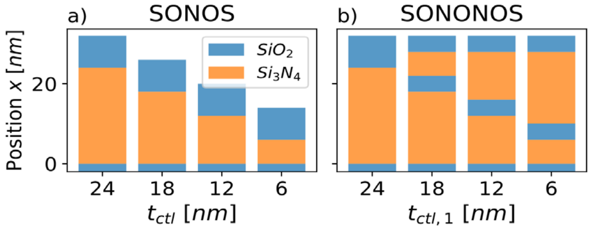

Fig. 1. Device architecture for SONOS (a) and SONONOS (b) devices. For SONONOS devices a $SiO_2$ barrier is inserted in between CTL1 and CTL2 such that the EOT for all SONONOS devices is the same.

## II. SONOS planar capacitors

### A. Reference calibration

To obtain a reference trapped charge density, we use a 1D numerical drift-diffusion model describing only electron transport in the CTL. The injection current $J_{inj}$ is calculated by a Fowler-Nordheim (FN) injection model for which the transmission coefficient $T_{inj}$ is evaluated at the conduction band (CB) minimum of the silicon channel ($E_C$):

$$J_{inj} = q n_{ch} v_{th} T_{inj}(E_c) \tag{1}$$

with $q$ the elementary charge, $n_{ch}$ the electron density in the channel and $v_{th}$ the carrier velocity. We treat $n_{ch}$ and $v_{th}$ as fitting parameters. The injection current is distributed over the CTL according to a uniform shape function [2,3]. All carriers that reach the CTL – BlOx interface are ejected from the CTL. Since drift current dominates within the CTL, the ejection current from the CTL CB to the gate CB can be expressed as:

$$J_{ej} = q\, n(x = t_{ctl})\, \mu\mathcal{E}\, (x = t_{ctl})\, T_{ej}(E_c) \tag{2}$$

This work was supported by imec's Industrial Affiliation Program for storage memories and Fonds Wetenschappelijk Onderzoek (FWO) Vlaanderen under grant number 1SS0925N.

with $n$ the free electron concentration in the CTL evaluated at the CTL-BlOx interface, $\mu$ the mobility within the CTL, $\mathcal{E}$ the electric field evaluated at the CTL-BlOx interface and $T_{ej}$ the transmission coefficient of the BlOx evaluated at the CTL conduction band energy $E_c$ . In the transmitting BlOx approximation, $T_{ej} = 1$.

Trapping is described by a capture cross section $\sigma$ and detrapping via a Poole-Frenkel (PF) mechanism. No direct trap-to-band tunneling to the channel, gate or CTL CB is included (Fig 2a). Fig. 4a shows the calibrated ISPP curves and Fig 3a-d the corresponding trapped charge profile for devices with $t_{ctl} = 6\,nm$ and $t_{ctl} = 24\,nm$. Note that all traps in the thinnest device are filled at high gate biases. Lowering the calibrated trap density $N_T$ further would prevent achieving the required $\Delta V_T$ values in the ISPP saturation regime. Thus, when assuming only bulk charging with a uniform trap density and no interface traps (as in the current model), any ISPP slope degradation in thick devices must be due to detrapping. Given that the electric field within the CTL is highest near the CTL-BlOx interface during program saturation, this region predominantly experiences PF detrapping. The inability to achieve a unified fit in the program saturation regime for both thick and thin devices indicates a gap in the current model. We propose further studies to investigate the impact of non-local detrapping mechanisms, taking into account the correct band profile in the vicinity of the trap, to address this discrepancy. For example, a trapped carrier near the CTL-BlOx interface does not experience the full PF barrier lowering due to the CTL – BlOx band offsets.

### B. *Variation of injection models*

Next, we show that a Modified-Fowler-Nordheim (MFN) model for injection, results in a trapped charge density that decreases as the distance from TuOx increases (Fig. 3e-h). This contrasts with the reference model that employs a Fowler-Nordheim (FN) tunneling model, similar to many (semi-)analytical compact models [2]. However, for small tunnel oxide thicknesses, it is crucial to manage energy band offsets carefully. In practice this implies injection by tunnelling through both the tunnel oxide and part of the CTL itself, a process known as MFN tunnelling (Fig. 2b) [5], [6]. This affects the transmission coefficient $T_{inj}$ in (1). MFN involves injecting thermalized carriers at a single point within the CTL, in contrast to highly energetic carriers for which energy dissipation would be described by a shape function that distributes these carriers over the CTL [2], [3]. This change in injection mechanism causes the change in the trapped charge profile. The ISPP fit (Fig. 4b) shows that thin devices do not require a detrapping model for accurate calibration, while thick devices do. This reconfirms the need for a non-local detrapping model.

### C. *Variation of ejection models*

A Fowler-Nordheim model for ejection leads to charge accumulation at the CTL-BlOx interface, unlike a purely transmitting boundary condition. A purely transmitting boundary condition on the BlOx is not appropriate for modeling retention behavior. Therefore, to unify retention and programming models, we study the effect of using a FN tunnelling model for ejection during programming (Fig. 3c, 4i-l). This affects the transmission coefficient $T_{ej}$ in (2). First, a FN boundary condition results in an increased trapping efficiency, which reduces the needed input current to program the devices (Fig. 5). Secondly, the mobility in the CTL must be decreased to avoid an ISPP slope bump, caused by charge accumulation at the CTL-BlOx interface, when the CTL is mainly uncharged (Fig. 4i-l, 6) [7]. Reducing the mobility results in more charge being trapped at the beginning of the CTL during the ISPP onset, which lowers the effect of charge accumulation at the CTL-BlOx interface on the ISPP slope (Fig. 6b).

## III. SONONOS PLANAR CAPACITORS

Starting from the SONOS calibration, the SONONOS calibration suggests a significant flowthrough from CTL1 to CTL2 and a low trapping efficiency in CTL2. First, for SONONOS devices, it is crucial to distinguish the different injection mechanisms for CTL1, thermalized, and CTL2, highly energetic (Fig. 2c). Hence, for CTL1, injection occurs at a single point (cf. SONOS devices), while for CTL2, a uniform shape function is applied to describe the energy relaxation process. Using the SONOS calibration parameters, the simulation for SONONOS devices shows coinciding ISPP onsets (Fig. 7a), which contradicts experiments. This is due to the dominance of trapped charge near the TuOx, caused by the lowered mobility. To resolve this issue, we investigate the treatment of the intermediate barrier. Using a fully transmitting boundary condition on the intermediate barrier, spaces the onsets slightly apart (Fig. 7b). Further spacing is achieved by reducing the integral of the shape function for CTL2 from 1 to 0.1, aligning better with experiments (Fig. 7c). By reducing the integral of the shape function to 0.1, one imposes that only 10% of the carriers injected from CTL1 into CTL2 dissipate enough energy to relax to the CB minimum and get trapped, while 90% of the carriers ‘fly over’ CTL2 and are emitted from the device.

These two model changes form the basis for future work. First, apart from using a transmitting barrier, the flowthrough from CTL1 to CTL2 can be enhanced by incorporating trap to band tunneling from the traps in CTL1 to the CB of CTL2. This approach has yet to be implemented. Second, the energy relaxation will be studied by a Monte-Carlo framework [4].

## IV. CONCLUSION

The combination of experimental data for SONOS and SONONOS vehicles reveals important insights into the physics behind charge trap flash memory. We conclude that boundary conditions on in- and ejection currents during programming significantly affect the trapped charge profile and hence are essential considering the subsequent retention and erase regimes. Additionally, detrapping is found to be crucial in the program saturation regime. However, none of the in- and ejection models in combination with PF detrapping accurately captures the complete programming behavior from onset untill saturation. The combination of SONOS and SONONOS results indicates a significant flowthrough out of the first nitride layer for SONONOS devices. To increase this flowthrough in the simulations and unify calibrations for thick and thin SONOS devices in the program saturation regime, we propose to investigate nonlocal detrapping mechanisms like trap to band tunneling.

## REFERENCES

[1] C. Monzio Compagnoni, A. Goda, A. S. Spinelli, P. Feeley, A. L. Lacaita, and A. Visconti, “Reviewing the Evolution of the NAND

Flash Technology,” *Proceedings of the IEEE*, vol. 105, no. 9, pp. 1609–1633, Sep. 2017, doi: 10.1109/JPROC.2017.2665781.

[2] D. Verreck *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 *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] T. Hellemans *et al.*, “Modeling the Operation of Charge Trap Flash Memory: A Monte Carlo Approach to Carrier Distribution and (De)trapping,” in *2024 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, Sep. 2024, pp. 01–04. doi: 10.1109/SISPAD62626.2024.10733042.

[5] E. Vianello *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] H. Bachhofer, H. Reisinger, E. Bertagnolli, and H. von Philipsborn, “Transient conduction in multidielectric silicon–oxide–nitride–oxide semiconductor structures,” *Journal of Applied Physics*, vol. 89, no. 5, pp. 2791–2800, Mar. 2001, doi: 10.1063/1.1343892.

[7] F. Schanovsky *et al.*, “A TCAD Compatible SONOS Trapping Layer Model for Accurate Programming Dynamics,” in *2021 IEEE International Memory Workshop (IMW)*, May 2021, pp. 1–4. doi: 10.1109/IMW51353.2021.9439598.

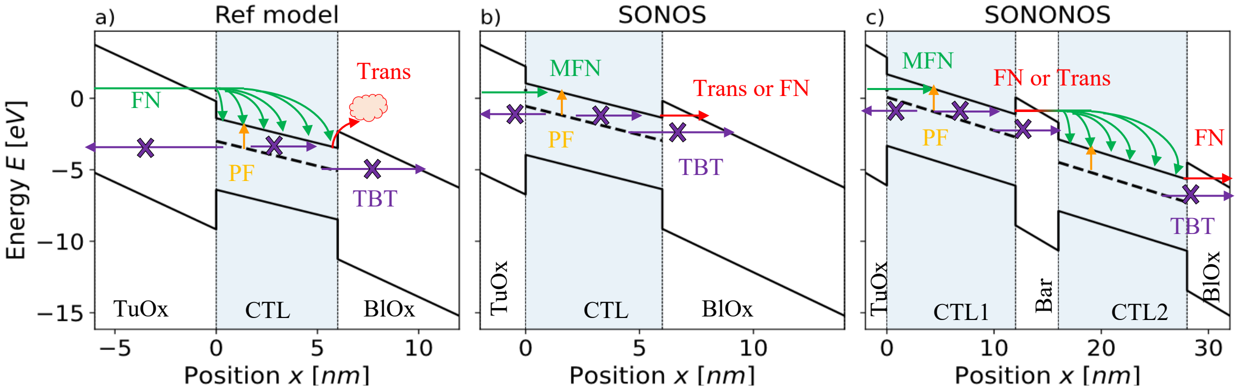

Fig. 2. Band diagram for SONOS reference case (a), SONOS with adapted in- and ejection model (b) and SONONOS devices (c). Note the distinction between the injection of highly energetic electrons, distributed according to a shape function, in (a) and CTL2 in (c), as compared to the injection of thermalized electrons in (b) and CTL1 in (c).

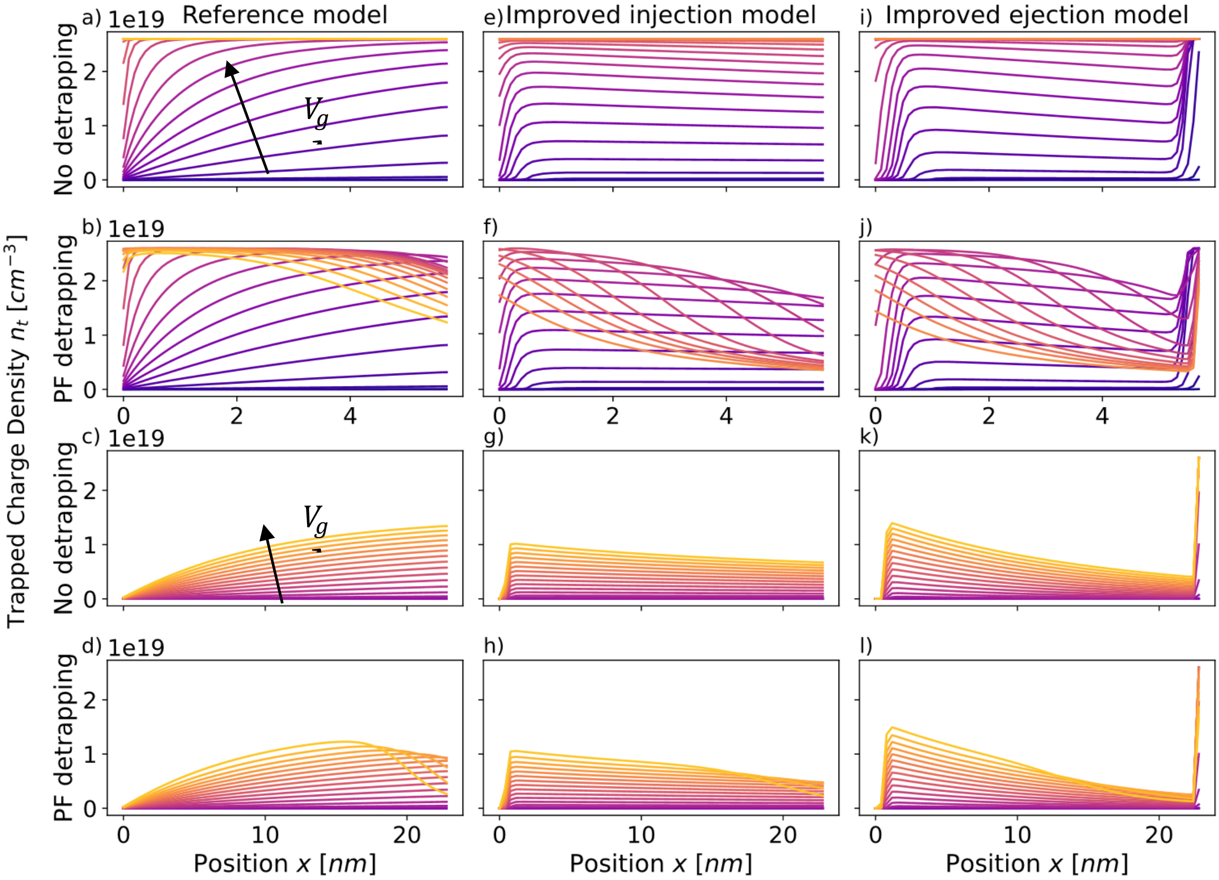

Fig. 3. Trapped charge density SONOS devices for different models (columns) with $t_{ctl}$ equal to $6nm$ (top two rows) or $24nm$ (bottom two rows). At high gate voltages, detrapping becomes significant and predominantly occurs near the BlOx due to the high electric fields in this area. Transitioning from the injection of highly energetic carriers (a-d) to the injection of thermalized carriers (e-h), the trend in trapped charge distribution reverses, shifting the trapped charge centroid from the BlOx towards the TuOx. Considering limited ejection due to the CTL-BlOx barrier causes charge accumulation at the CTL-BlOx interface.

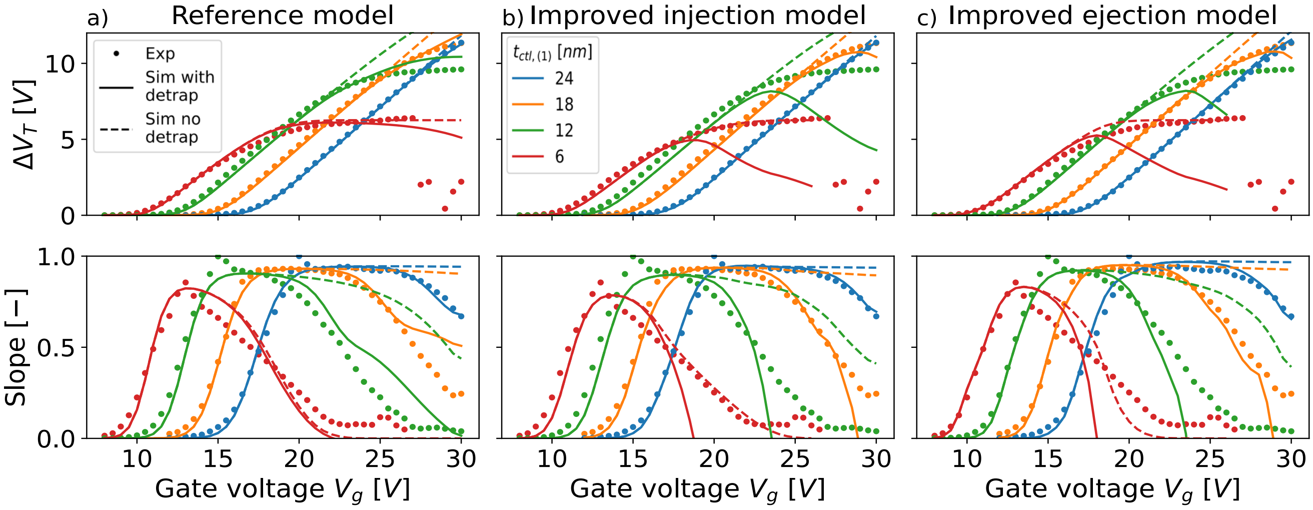

Fig. 4. ISPP calibration of SONOS devices within different models. It is important to acknowledge the discrepancy between ISPP fits for thick devices, where detrapping plays a crucial role, and thin devices, where detrapping proves to be excessively strong to achieve accurate calibration.

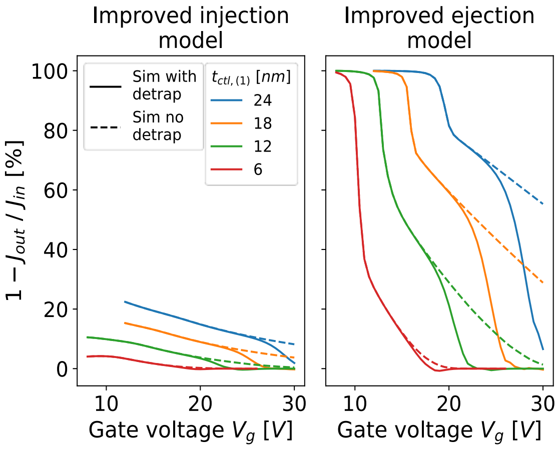

Fig. 5. The trapping efficiency $1 - J_{out}/J_{in}$ increases significantly when taking into account the CTL-BlOx CB offset. Additionally, the reduced BlOx transmission gives rise to a reduced input current in the calibration.

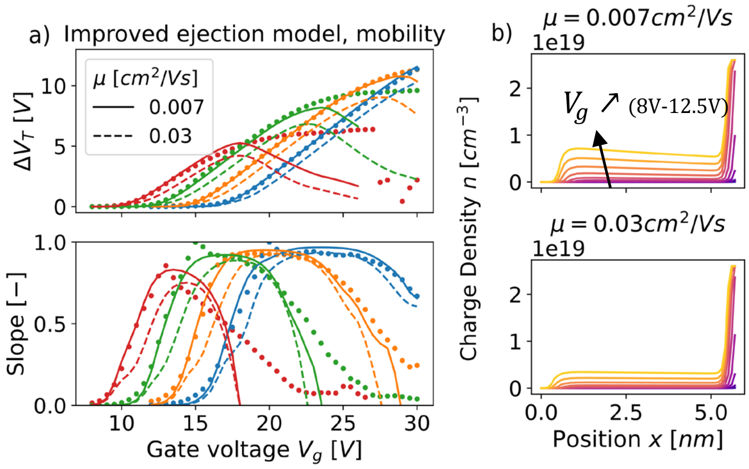

Fig. 6. To prevent a bump in the ISPP slope when using an ejection model for the BlOx that is not fully transmitting, it is essential to reduce the mobility (a). During the ISPP onset reducing the mobility, increases the trapped charge concentration near the TuOx, which decreases the impact of the accumulated interface charge on the ISPP slope (b).

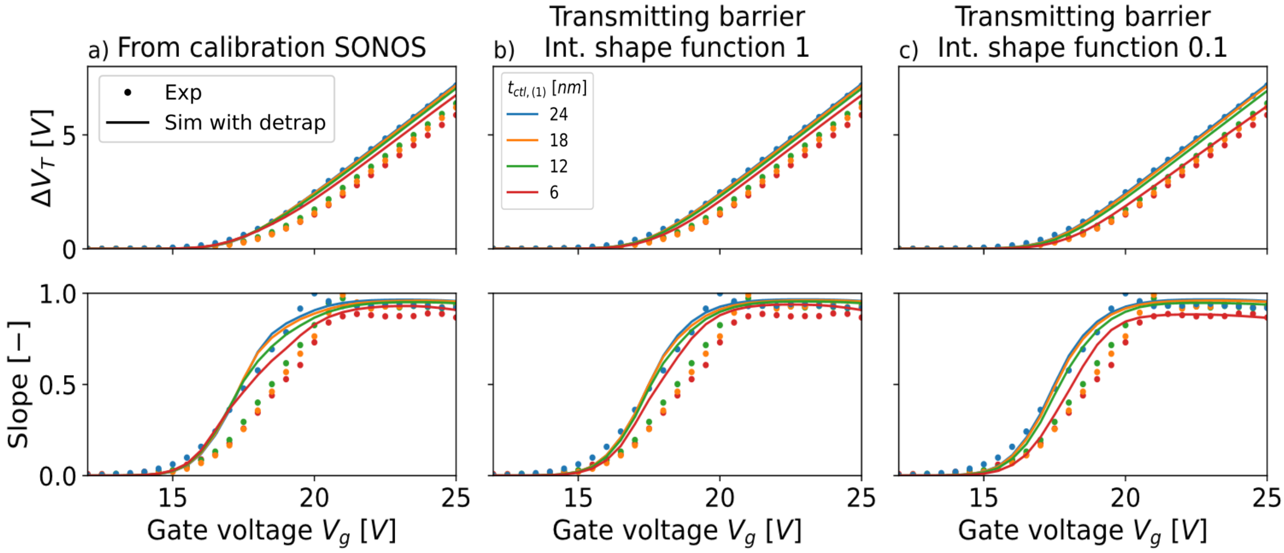

Fig. 7. (a) ISPP calibration SONONOS devices based on calibration SONOS devices. The ISPP onsets can be spaced slightly apart by considering a transmitting barrier between CTL1 and CTL2 (b). Further spacing is achieved by reducing the shape function integral from 1 (b) to 0.1 (c).