**Project:**

# MNEMOSENE

**(Grant Agreement number 780215)**

*"Computation-in-memory architecture based on resistive devices"*

Funding Scheme: Research and Innovation Action

Call: ICT-31-2017 "Development of new approaches to scale functional performance of information processing and storage substantially beyond the state-of-the-art technologies with a focus on ultra-low power and high performance"

Date of the latest version of ANNEX I: 11/10/2017

# D4.3 – Refined Models of Memristive Device


| | |
|---|---|
| **Project Coordinator (PC):** | Prof. Said Hamdioui<br>Technische Universiteit Delft - Department of Quantum and Computer Engineering (TUD)<br>Tel.: (+31) 15 27 83643<br>Email: S.Hamdioui@tudelft.nl |
| **Project website address:** | www.mnemosene.eu |
| **Lead Partner for Deliverable:** | RWTH |
| **Report Issue Date:** | 31/12/2019 |


| **Document History**<br>*(Revisions – Amendments)* | |
|---|---|
| **Version and date** | **Changes** |
| 1.0 19/12/2019 | First version |
| | |
| | |

| **Dissemination Level** | | |
|---|---|---|
| **PU** | Public | **X** |
| **PP** | Restricted to other program participants (including the EC Services) | |
| **RE** | Restricted to a group specified by the consortium (including the EC Services) | |
| **CO** | Confidential, only for members of the consortium (including the EC) | |

The MNEMOSENE project aims at demonstrating a new computation-in-memory (CIM) based on resistive devices together with its required programming flow and interface. To develop the new architecture, the following scientific and technical objectives will be targeted:

- Objective 1: Develop new algorithmic solutions for targeted applications for CIM architecture.
- Objective 2: Develop and design new mapping methods integrated in a framework for efficient compilation of the new algorithms into CIM macro-level operations; each of these is mapped to a group of CIM tiles.
- Objective 3: Develop a macro-architecture based on the integration of group of CIM tiles, including the overall scheduling of the macro-level operation, data accesses, inter-tile communication, the partitioning of the crossbar, etc.
- Objective 4: Develop and demonstrate the micro-architecture level of CIM tiles and their models, including primitive logic and arithmetic operators, the mapping of such operators on the crossbar, different circuit choices and the associated design trade-offs, etc.
- Objective 5: Design a simulator (based on calibrated models of memristor devices & building blocks) and FPGA emulator for the new architecture (CIM device combined with conventional CPU) in order demonstrate its superiority. Demonstrate the concept of CIM by performing measurements on fabricated crossbar mounted on a PCB board.

A demonstrator will be produced and tested to show that the storage and processing can be integrated in the same physical location to improve energy efficiency and also to show that the proposed accelerator is able to achieve the following measurable targets (as compared with a general purpose multi-core platform) for the considered applications:

- Improve the energy-delay product by factor of 100X to 1000X
- Improve the computational efficiency (#operations / total-energy) by factor of 10X to 100X
- Improve the performance density (# operations per area) by factor of 10X to 100X



European Commission

**The MNEMOSENE project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 780215**

# Table of Contents

# 1. Introduction

For the circuit design of the CIM architecture, compact models of the memory and computing elements are required. In this project, we employ phase-change memory (PCM) devices and redox-based resistive devices based on the valence change mechanism (VCM). Both type of devices are memristive devices. The resistance of those devices can be tuned by applying appropriate voltage pulses. The underlying physical mechanism, however, is completely different. Thus, compact models need to be developed each for PCM and VCM devices.

In the previous report D4.2 "*Initial Models of Memristive Devices*", we described initial models for PCM and VCM devices. We presented a dynamic compact model for VCM switching including programming variability, which was verified by experimental data. The next step is the development of a more refined model for the read, which includes read noise due to stochastic changes of the ionic configuration of the filament. The initial PCM model captured the accumulative behavior, conductance drift and read noise. Another approach to program a PCM device to a target conductance level is to apply a read-and-verify iterative programming algorithm. Upon application of a programming pulse, the device conductance is read, and the amplitude of the next programming pulse is adjusted according to the difference between the target and the measured device conductance. This iteration continues until the device conductance is in close vicinity of the target conductance. Owing to the feedback mechanism, the device conductance can be set more precisely in comparison to the application of successive crystallization pulses (or in other words, when exploiting the accumulation dynamics). In this report, we develop a model where we can capture the experimentally observed iterative programming distributions of PCM devices. In addition, the drift and read noise models are extended such that the drift coefficient variability and the time-dependence of the read noise are included.

# 2. Refined Models for VCM Devices

As the CIM architectures studied in MNEMOSENE are mainly utilizing read operations, the refined model takes into account read instabilities/read noise. In principle, there are two different sources for read noise. The first one is pure electronic noise, the second one is related to conduction fluctuations due to ionic processes, i.e. stochastic jumps of individual defects [1-5]. Both processes happen on different timescales and show different amplitudes. Figure 1 shows the occurrence of ionic noise during read in a Pt/$Ta_2O_5$/W device fabricated in the group of RWTH Aachen. Here, a rectangular read voltage pulse of 0.3 V was applied (gray line) and the current was measured concurrently. The blue current trace shows distinct jumps between two conduction states occurring in a time range of hundreds of microseconds to milliseconds. This current jump is about 10% of the read current and will cause severe problems if not considered in the circuit design. In addition, oscillation with smaller amplitudes and a higher frequency appear.

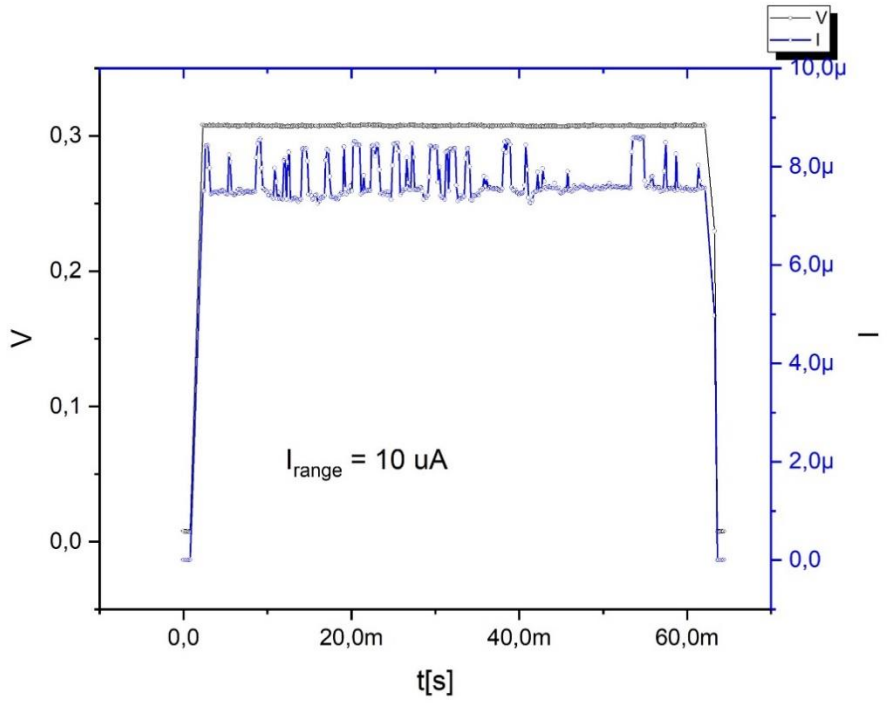


Fig. 1: Illustration of read noise of a Pt/$Ta_2O_5$/W VCM device fabricated in the group of RWTH Aachen.

In the following, two empirical models will be presented for the two sources of read instability and compared to experimental data of a Pt/$ZrO_x$/Ta device fabricated at RWTH Aachen University.

## 2.1 Model for Read Noise

The read noise model is an extension of the JART VCM v1 model, which was presented before in D4.2. It assumes a filamentary switching mechanism. The conducting filament is divided into two regimes as shown in Fig. 2. The plug region is located close to the ohmic electrode (OE), which forms an ohmic contact with the oxide materials. The disc region, in contrast, is the region close to the active electrode (AE), which forms a Schottky-type contact with the oxide materials. In the JART VCM v1 model it is assumed that the change of the oxygen vacancy concentration in the disc region dominates the total conductance change. Thus, only the Schottky-type diode at the AE/disc interface and the disc resistance $R_{disc}$ are assumed variable in the model. The plug resistance $R_{plug}$ and the interface resistance $R_{contact}$ at the OE interface are assumed to be constant during switching. The oxygen vacancy concentration $N_{disc}$ changes due to the thermally-assisted migration of the ionic defects, i.e. the ionic current $I_{ion}$, under the applied electric field. The rage of change is described by

$$\frac{dN_{\mathrm{disc}}}{dt} = \frac{-1}{zeAl_{\mathrm{disc}}} \cdot I_{\mathrm{ion}}. \qquad (1)$$

Here, *e* is the elementary charge, *z is* the charge of the oxygen vacancies, $l_{disc}$ is the thickness of the disc, and *A* the lateral cross-section of the filament. The disc concentration changes the disc resistance and influences the transport through the AE/disc interface. In general, the current increases with an increase in the disc concentration at a constant voltage. In the JART VCM v1 model, the concentration is a continuous variable and will not allow us to model discrete current jumps. Thus, we extended the model by a discrete change of the state variable

with the continuous value of the concentration representing the mean concentration. In this picture, a long-term drift (retention) or the switching process are modelled by the change of the continuous average concentration value $N_{disc}$. The random current jumps, however, are modelled by changing the concentration by an equivalent of n = 0, ±1, ±2 defects in the disc.

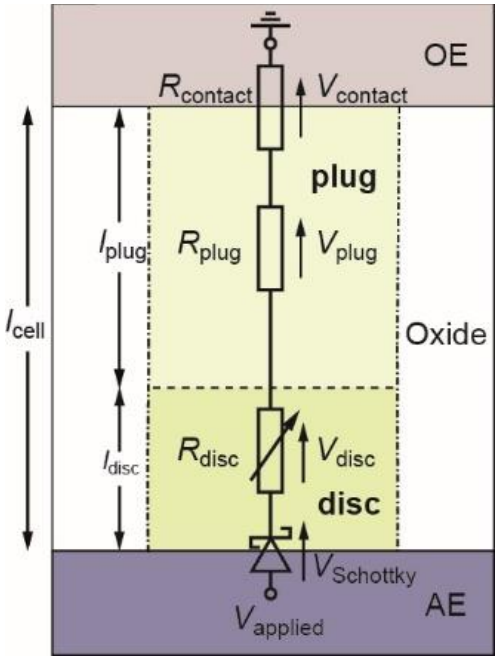


Fig. 2: Equivalent circuit diagram of the JART VCM v1 Model [6].

This number n translates into a concentration change $\Delta N_{disc}$ of

$$\Delta N_{\mathrm{disc}}(i) = \frac{n(i) - n(i-1)}{A l_{\mathrm{disc}}}. \quad (2)$$

The *update of* $\Delta N_{\mathrm{disc}}$ is calculated at a time interval of 1.5 ms. The value of *n* can only change by 1 per step. *n* is updated based on its previous value and a probability to change which is extracted from a Poisson distribution. It can change by ±1 or stay constant. Values outside of ±2 are not allowed. Choosing a Poisson distribution leads to a relaxation of n towards 0 since 0 has the highest probability and change towards 0 has a higher probability than change towards other values.

To model the electronic noise, the current value is modulated by a value $\Delta I$, which is in the range of -900…900 nA. In contrast to the ionic noise, the values for $\Delta I$ have a Gaussian distribution.

The value $\Delta I$ is updated on a shorter timescale than the ionic noise at 10 µs.

### 2.1.1 Simulation results.

The read noise model was tested with read pulses with -0.35 V and 100 ms length. Figure 3a and b show the simulated current transients during the read pulse for 10 different cells. The current traces show two type of current jumps: larger ones occurring at a longer timescale and smaller ones occurring at a shorter timescale. The large current jumps are related to the ionic noise. They occur whenever the concentration changes by discrete values of *n* as illustrated in the corresponding concentration transients in Fig. 3c, d. The small jumps are related to the

electronic noise. In addition to the jumps, the results in Fig. 3b, and d, shows a drift of the ion concentration to higher values over time. This effect is related to a read-disturb. The read voltage is biased in SET direction and the device will slowly drift towards the LRS. The difference between the results in Fig. 3b, d and Fig. 3a, c is the migration barrier. For Fig. 3b, d a migration barrier of 0.9 eV is assumed and the ion migration rate is high enough to result in a significant movement of the ionic defects on the timescale of the read pulse. When increasing the barrier to 1 eV (as in Fig. 3a,c), the drift of the state variable disappears as the migration rate becomes lower [7]. Thus, on top of the noise, the model is able to model read-disturb as well

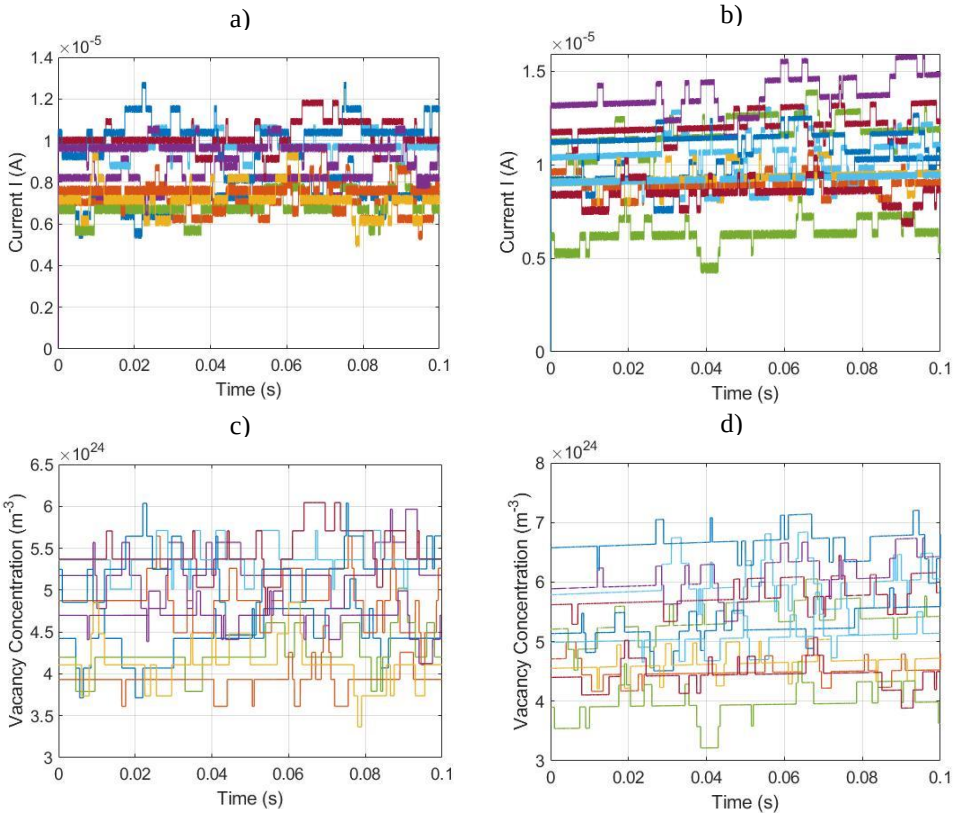


Fig. 3: Simulated current and oxygen vacancy concentration during the read pulses for 10 different cells for migration barrier of (a),(c) 1 eV and (b),(d) 0.9 eV.

Comparing Fig. 3 with Fig. 1 shows that the simulation models reproduces qualitatively the experimental behavior. In addition, we fitted our model to experimental data of Pt/$ZrO_x$/Ta VCM cell and compared the variability in terms of cumulative distribution functions. Here, the measured current after the programming of 540 devices and after 0.1 s is evaluated. For the simulation the cells were initialized within a range of oxygen vacancy concentrations according to a lognormal distribution. The resulting current values are plotted in terms of a CDF. The comparison of the measurements with the simulations shows similar results. The two most important characteristics are the distribution of currents and the change of this distribution over time. The CDFs in Fig. 4 show a lognormal distribution characterized by their linearity for a logarithmic x-axis. Although the individual cells change their current over time (see Fig. 3), the overall distribution of the cells remains almost the same. These two properties can be observed both in the measurements and in the simulation.

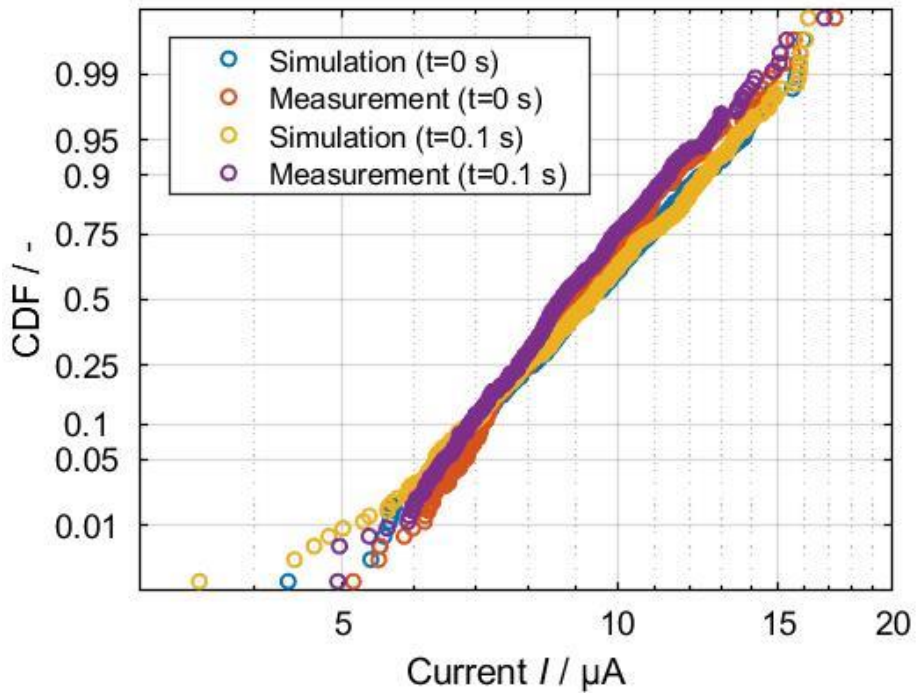


Fig. 4: Cumulative distribution function of the read current in experiment and simulation at two different times after programming.

# 3. Refined Models for PCM Devices

## 3.1 Iterative programming

The melt-quench region of the programming curve (righthand side in Fig. 5a) can be used to iteratively write the conductance values to PCM. Based on a target conductance a RESET pulse is applied, the conductance state is read and the difference between observed and desired conductance values is used to adjust the subsequent pulse amplitudes until the observed value is within an error tolerance [8]. Though ideally we may achieve arbitrary precision with such an iterative scheme, the conductance drift and 1/*f* noise characteristics cause the conductance values to deviate from the initially programmed states. Fig. 5b shows the cumulative distribution of conductance states iteratively programmed using 500 ns pulses. The conductance values reported are based on a read 23 µs after the last programming pulse and has an average standard deviation of 1.23 µS around their mean values.

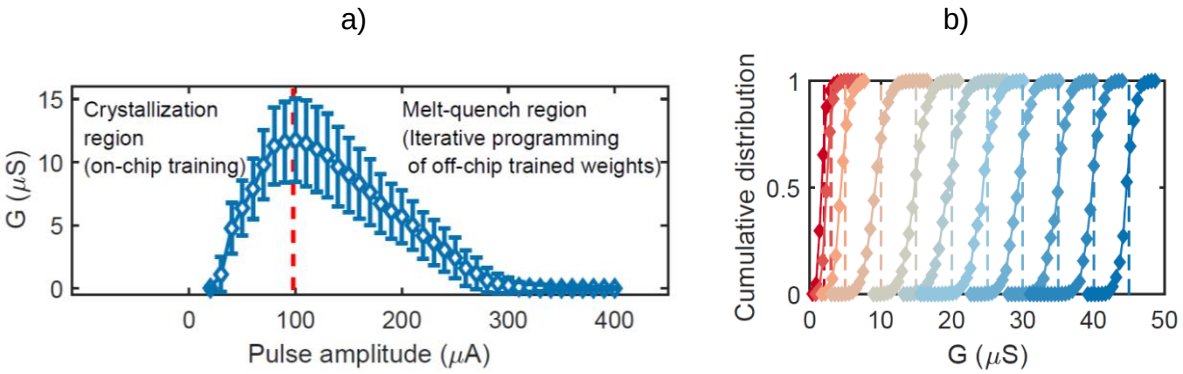


Fig. 5. (a) PCM programming curve. Conductance as a function of the amplitudes of 50 ns current pulses. The devices were RESET before each programming pulse. The conductance change on the left side of the curve is due to crystal growth (SET) and the one on the right side is via a melt-quench process (RESET). (b) Cumulative distribution of conductance states iteratively programmed using RESET pulses of 500 ns width.

## 3.2 Conductance drift

The conductance of the PCM is observed to decrease over time which is typically attributed to structural relaxation of the phase configuration formed as a result of each programming event [9]. This conductance drift is well captured by the empirical relation,

$$G(t) = G(t_0)(t/t_0)^{-\upsilon}$$

over many orders of time. Here $G(t_0)$ denote the conductance measured after time $t_0$ from a programming event and $\nu$ is a drift coefficient. For each iteratively programmed target state, 10,000 devices were read 50 times in equal intervals in a 90 s window which were repeated in logarithmic intervals until $10^5$ s. The average conductance evolution of the 10,000 devices barring a few outliers is shown in Fig. 6a. The estimated mean and standard deviation of $\nu$ is shown in Fig. 6b.

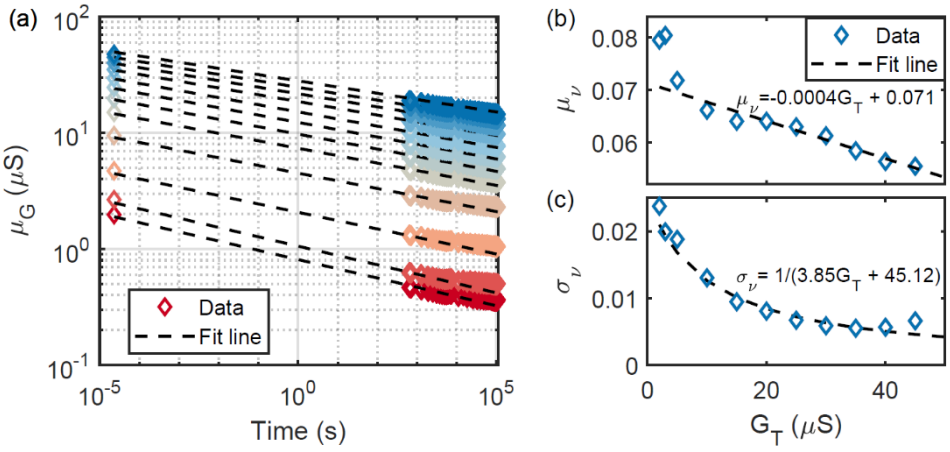


Fig. 6 (a) Drift observed from iteratively programmed conductance states (b) Mean and (c) standard deviation of the estimated drift coefficient. Equations of corresponding fit lines are also shown.

## 3.3 Read noise

Low-frequency noise could be particularly detrimental to long-term inference performance. While $1/f^{\gamma}$ noise ($\gamma$ = 0.9 to 1.1) and Random Telegraph Noise have been observed in PCM, for the modeling we assume an ideal $1/f$ behavior [10]. We use read noise estimation using samples from a fixed duration (bandwidth) to estimate the fluctuation due to read noise from the time of programming to the time of inference.

The power spectral density $S_G$ of the read noise normalized over conductance $G$ is given by,

$$\frac{S_G}{G^2} = \frac{Q}{f} \tag{3}$$

where Q is assumed to be determined by the phase configuration within the PCM. The variance of the noise can be estimated by integrating Equation 3 over the desired frequency range. Hence we have,

$$\sigma_{nG} = GQ_s\sqrt{\log(\mathrm{f}_{max}/\mathrm{f}_{min})} \tag{4}$$

where $Q_s = \sqrt{Q}$ . Read noise standard deviations, $\sigma_{nG}$, are estimated based on 50 reads from the $T$ = 90 s time window after subtracting an estimated conductance drift. $f_{max} = 1/2T_{read}$ where $T_{read}$ = 250 ns is the read pulse duration in our experimental platform and $f_{min} = 1/T$.

The estimated $Q_s$ shows dependence on both the target conductance $G_T$ and time $t$ which is illustrated in Fig. 7a, b. This behavior is well captured using the relation,

$$Q_s(G_T, t) = K/G_T^{\alpha} \times (t/t_1)^{-v_q} \tag{5}$$

where $K$ = 0.0565, $\alpha$ = 0.618 and $t_1$ = 1.089 x $10^5$. The estimated $v_q$ as a function of $G_T$ is shown in Fig. 7c. The measured read noise and the corresponding fit for two $G_T$ levels is shown in Fig. 7d. For modeling the read operation, to access the fluctuation due to read noise with respect to the initially programmed state, we have to integrate all the frequency components that can be captured over the time window starting from the point where the devices were programmed to the time at which we are performing the read. This time window determines $f_{min}$ for the projected read noise estimation (Fig. 7d).

## 3.4 Model verification with experimental PCM data

The PCM model described in Sections 3.1, 3.2 and 3.3 can capture the experimentally observed PCM behavior remarkably well. The experimental PCM data and the corresponding model fits are plotted in Fig. 8.

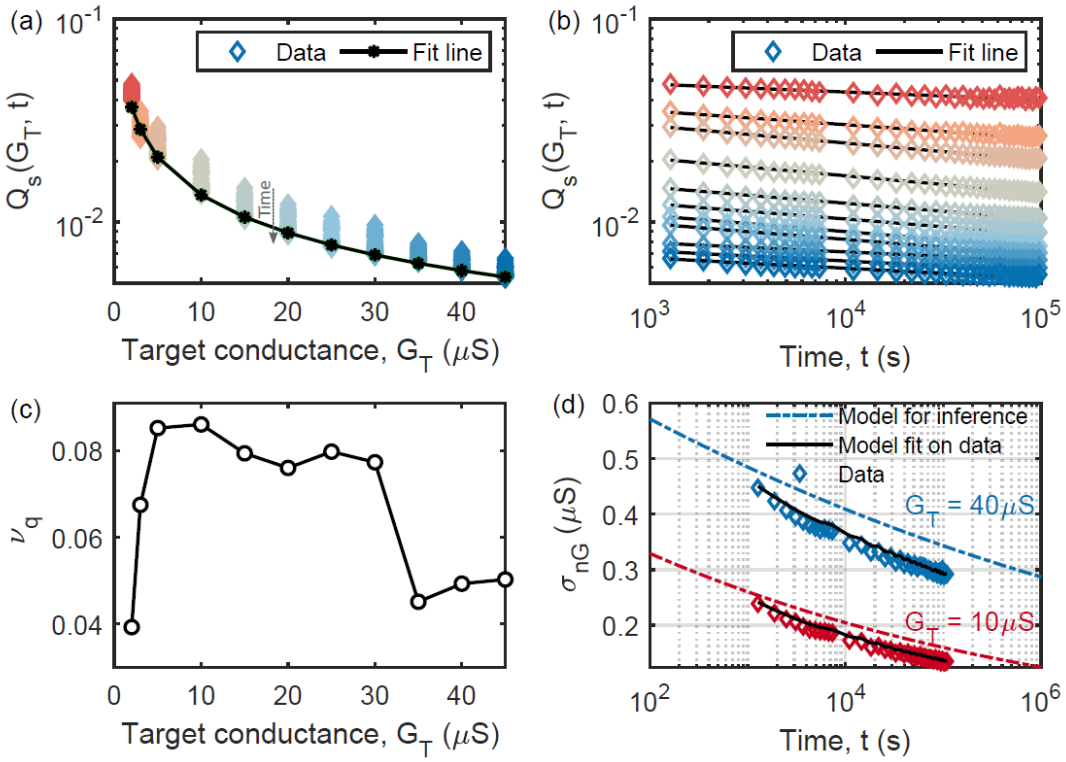


Fig. 7. $Q_s$ as a function of target conductance $G_T$ (a) and time (b). (c) Estimate for $v_q$ in Equation 5. (d) Standard deviation of read noise $\sigma_{nG}$ as a function of time. The read noise estimated from measured conductance samples over a 90 s window and corresponding fit is shown. A projected read noise estimation for inference is also shown.

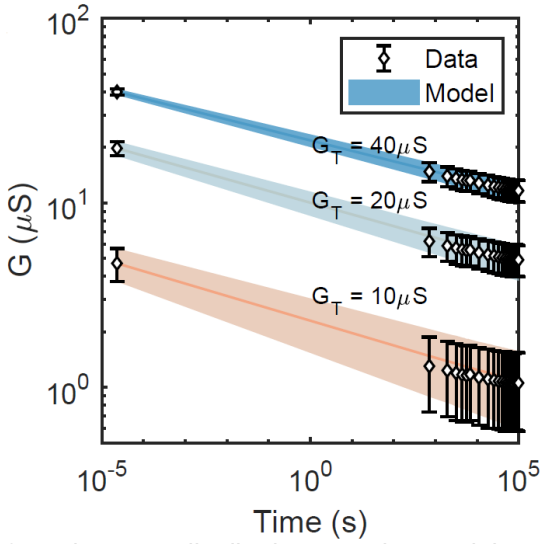


Fig. 8. Evolution of conductance distribution over time and the corresponding model response for different iteratively programmed target conductance values.

# 4. Outlook

The extended VCM and PCM models can be used to simulate the CIM primitives targeted within WP1 of the MNEMOSENE project on a CIM array with peripheral circuits. In this way, the functionality of the applications targeted in WP1 can be verified using realistic device models calibrated on experimental data. The time and energy numbers for these operations can also be extracted and used in the higher level CIM simulator.

# 5. References

[1] E. Abbaspour, S. Menzel and C. Jungemann, “Random telegraph noise analysis in Redox-based Resistive Switching Devices Using KMC Simulations (talk),” *2017 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), September 7-9, Kamakura, Japan*, 2017, pp.

[2] S. Brivio, J. Frascaroli, E. Covi and S. Spiga, “Stimulated Ionic Telegraph Noise in Filamentary Memristive Devices,” *SCIENTIFIC REPORTS*, vol. {9, pp., 2019.

[3] T. Gong, Q. Luo, X. Xu, J. Yu, D. Dong, H. Lv, P. Yuan, C. Chen, J. Yin, L. Tai, X. Zhu, Q. Liu, S. Long and M. Liu, “Classification of Three-Level Random Telegraph Noise and Its Application in Accurate Extraction of Trap Profiles in Oxide-Based Resistive Switching Memory,” *IEEE Electron Device Lett.*, vol. 39, pp. 1302-1305, 2018.

[4] F.M. Puglisi, L. Larcher, P. Pavan, A. Padovani and G. Bersuker, “Instability of HfO 2 RRAM devices: comparing RTN and cycling variability,” *2014 IEEE International Reliability Physics Symposium (IRPS)*, 2014, pp. MY.5.1-MY.5.5.

[5] F.M. Puglisi, L. Larcher, A. Padovani and P. Pavan, “A Complete Statistical Investigation of RTN in HfO2-Based RRAM in High Resistive State,” *IEEE Trans. Electron Devices*, vol. 62, pp. 2606-2613, 2015.

[6] JART, “Juelich Aachen Resistive Switching Tools (JART),”, vol., pp., 2019.

[7] S. Menzel, M. Salinga, U. Böttger and M. Wimmer, “Physics of the Switching Kinetics in Resistive Memories,” *Adv. Funct. Mater.*, vol. 25, pp. 6306-6325, 2015.

[8] N. Papandreou, H. Pozidis, A. Pantazi, A. Sebastian, M. Breitwisch, C. Lam, and E. Eleftheriou, “Programming algorithms for multilevel phase-change memory,” in IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2011, pp. 329–332.

[9] M. Le Gallo, D. Krebs, F. Zipoli, M. Salinga, and A. Sebastian, “Collective structural relaxation in phase-change memory devices,” Advanced Electronic Materials, vol. 4, no. 9, p. 1700627, 2018.

[10] M. Nardone, V. I. Kozub, I. V. Karpov, and V. G. Karpov, “Possible mechanisms for 1/f noise in chalcogenide glasses: A theoretical description,” Phys. Rev. B, vol. 79, p. 165206, Apr 2009.