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

D4.2 – Initial Models of Memristive Device

Project Coordinator (PC):	Prof. Said Hamdioui Technische Universiteit Delft - Department of Quantum and Computer Engineering (TUD) Tel.: (+31) 15 27 83643 Email: S.Hamdioui@tudelft.nl
Project website address:	www.mnemosene.eu
Lead Partner for Deliverable:	RWTH
Report Issue Date:	31/12/2018

Document History *(Revisions – Amendments)*	
Version and date	**Changes**
1.0 21/12/2018	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)	

European Commission

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

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

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.

With respect to the intended applications of the CIM architecture, the stability of the programmed resistance state and the read variability is of utmost importance for the reliability of the CIM function. Nevertheless, the control circuitry needs to be able to write the data into the CIM tile.

Both type of devices show a highly nonlinear switching kinetics, i.e. the programming time changes by orders of magnitude while the applied voltages is changed only by a few hundreds of millivolts [1]. As PCM and VCM devices are two-terminal devices the switching kinetics should show highly nonlinear switching kinetics in order to allow fast programming and low read-disturb. Furthermore, when programming the devices a half-select scheme is potentially applied. The voltage drop over the half-selected cells should not lead to unintended switching events. From this consideration, it is clear that dynamic models for the resistance switching are required in order to design the programming circuitry and optimize the voltage levels and pulse length to prevent unintended switching of non-selected cells. Eventually, the nonlinearity of the switching kinetics will also determine the maximum block size in the array, which can be addressed at the same time.

The switching of these devices are subject to cycle-to-cycle and device-to-device variability. Thus, when using identical programming conditions the LRS and HRS vary. If a high precision of the programmed state is required a write-verify algorithm may be used. To optimize such algorithm, a suitable compact model should also account for the intrinsic device variability.

As the intended applications of the CIM architecture are based on reading the resistance state in one or the other way, accurate models for state retention and read instability (e.g. random telegraph noise, RTN) are required.

2. Models for VCM Devices

Redox-based resistive switching cells consists of an ionically conducting oxide layer sandwiched between two electrodes. Typically, one electrode consists of a high work function, chemically inert metal, e.g. Pt. For the second electrode a low work function, easily oxidizable metal, e.g. Ta or Ti, is used. Resistive switching is achieved in various oxides, e.g. TiO_2, Ta_2O_5, HfO_2, ZrO_2, or $SrTiO_3$ [2]. In all of these oxides, the electronic conductivity increases if the system is substoichiometric, i.e. oxygen ions are missing in the oxide compound. The missing oxygen ion in the host lattice is termed oxygen vacancy and is typically doubly positively charged with respect to the lattice [3]. In Kröger-Vink notation, the oxygen vacancy is denoted as $V_O^{\bullet\bullet}$ [4]. The positive of the vacancy is typically compensated by electrons in the d-states of the transition metal cation, e.g. Ta. Consequently, the valence of the cation is reduced in the presence of a nearby oxygen vacancy. This process lead to the name valence change mechanism [3]. The compensating electrons increase the local conductivity of the oxides as they are typically delocalized of form large or small polarons. In addition, the oxygen vacancies act as mobile donors and can modify the shape of the electrostatic barrier at the metal/oxide

interfaces [5-8]. Before the VCM cells can be repetitively switched between a high resistive state (HRS) and a low resistive state (LRS), an electroforming process is required, which brings the device from a highly insulating state to the LRS. During the electroforming process, a filamentary region with a high amount of oxygen vacancies is formed (cf. Fig. 1). In a simplified picture, the SET and RESET processes can be understood in terms of a redistribution of the oxygen defects in this filamentary region close to one of the electrodes, as illustrated in Fig. 1 [9]. The electrode represented in the figure is the high work function inert electrode. In the HRS, the region close to this electrode has a low $V_O^{\bullet\bullet}$ concentration (A). Thus, the electrostatic barrier at the metal/oxide interface limits the current injection. In addition, the $V_O^{\bullet\bullet}$ concentration determines the local conductivity, as oxygen vacancies act as donors.[10] By applying a negative voltage to the inert electrode, oxygen vacancies drift towards it (B). Therefore, the $V_O^{\bullet\bullet}$ concentration close to this interface increases. Consequently, the electrostatic barrier is lowered and its thickness is reduced allowing easy current injection. In addition, the local conductivity close to electrode increases enabling a high current transport through the filamentary region (C). For low ohmic LRS, obtained through high SET currents, the I-V curve shows a symmetric, quite linear behavior. At lower current levels, the *I-V* characteristic becomes more asymmetric and nonlinear. When a positive potential is applied to the electrode, the oxygen vacancies move away from the interface (D). The electrostatic barrier is restored and the CF partially dissolves (A). Still, the conduction in the HRS state is filamentary.

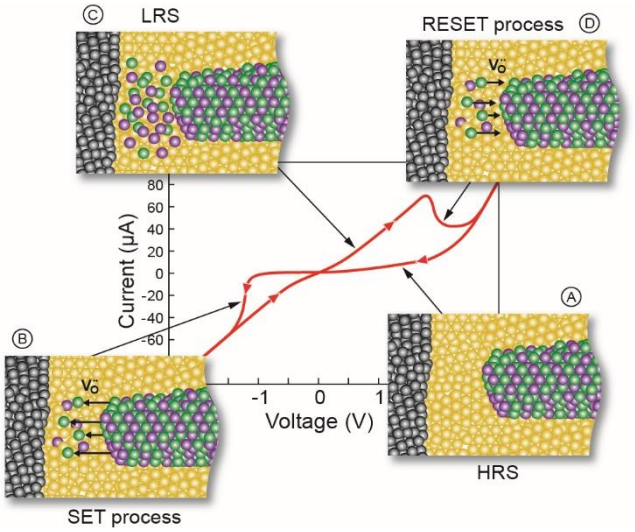

Fig. 1: Illustration of the SET and RESET processes of a filamentary switching VCM cell. The green spheres represent the oxygen vacancies, $V_O^{\bullet\bullet}$, the yellow and purple spheres represent the cations of the binary metal oxide in their standard oxidation state and the reduced state, respectively. The grey spheres on the left resemble the inert electrode. [9]

2.1 Deterministic model with variable parameter sets

The deterministic compact model is based on the physical mechanism shown in Fig. 1. A first version of the model (developed by our group) was used to model the resistive switching phenomenon in a Pt/Hf/HfO_2/Pt cell [11], a Pt/$SrTiO_3$/Ti cell [12], or a Pt/Ta_2O_5/Ta cell [13]. As VCM data on small CIM arrays is not yet available, we adapted the prior models to establish a flexible model that includes the basic properties of VCM cells and can be easily used to fit the experimental data, when available. The model is implemented in Verilog-A. The variable

parameters of the newly developed model are the resistance ratio and the nonlinearity of the switching kinetics. Using this flexible model for circuit design may already show some design constraints for the computing architecture. The model is based on the movement of ionic defects in a filamentary region within an insulating oxide layer as shown in Fig. 2a (cf. Fig 1). The defect concentration in the disc modulates the conductivity of the disc region near one of the metal/insulator interfaces. In addition, the electronic transport across the Schottky-like metal/insulator interface changes with the defect concentration in the disc. This disc concentration N_{disc} is the state variable of the implemented model. The change of the disc concentration is proportional to the ionic current I_{ion} according to

$$\frac{\mathrm{d}N_{\mathrm{disc}}}{\mathrm{d}t} = \frac{1}{z\,e\,A\,l_{\mathrm{disc}}} \cdot I_{\mathrm{ion}}. \tag{1}$$

The ionic current is calculated according to the Mott-Gurney law for ion hopping as

$$I_{\mathrm{ion}} = 2Azeca\,\nu_0 \exp\left(-\frac{\Delta W_{\mathrm{A}}}{k_{\mathrm{B}}T_{\mathrm{ion}}}\right)\sinh\left(\frac{a\,zeE}{2\,k_{\mathrm{B}}T_{\mathrm{ion}}}\right). \tag{2}$$

In Eq. (1)-(2), z denotes the charge of the moving ion, e is the elementary charge, A is the cross-section of the filament region, l_{disc} is the disc length, a is the mean ion hopping distance, ν_0 is the attempt frequency, and k_{B} is the Boltzmann constant. The concentration c is the mean value of the disc and plug concentration, i.e. $c = (N_{\mathrm{plug}}+N_{\mathrm{disc}})/2$. The ionic current is temperature-activated with an activation energy ΔW_{A}. The temperature T_{ion} can increase during operation due to Joule heating. It is calculated according to

$$T_{\mathrm{ion}} = \left(V_{\mathrm{Schottky}} + V_{\mathrm{disc}} + V_{\mathrm{plug}}\right) I R_{\mathrm{th,eff}} + T_0. \tag{3}$$

It depends on the ambient temperature T_0, the dissipated electrical power and an effective thermal resistance $R_{\mathrm{th,eff}}$. The power dissipated in the device is the product of the current I and the sum of the voltage drops V_{Schottky}, V_{disc} and V_{plug}. The electric field E in eq. (2) is modeled according to

$$\begin{aligned} E &= V_{\mathrm{disc}}/l_{\mathrm{disc}} && \mathrm{SET}(V>0), \\ E &= \left(V_{\mathrm{Schottky}} + V_{\mathrm{disc}} + V_{\mathrm{plug}}\right)/l_{\mathrm{cell}} && \mathrm{RESET}(V<0). \end{aligned} \tag{4}$$

For positive voltages, the electric field is determined by the voltage drop V_{disc} over the disc region. In contrast, the sum of the voltages V_{Schottky}, V_{disc} and V_{plug} is used to determine the electric field during reset. This asymmetry is motivated by the asymmetry of the band bending under forward and reverse bias. For negative voltages $V < 0$, the Schottky-like contact is biased in forward direction and the current I_{Sc} is modeled according to thermionic emission theory as

$$I_{\mathrm{Sc},V<0\,\mathrm{V}} = -AA^{*}T^{2}\exp\left(-\frac{e\phi_{\mathrm{Bn}}}{k_{\mathrm{B}}T}\right)\cdot\left(\exp\left(\frac{e(-V_{\mathrm{Schottky}})}{k_{\mathrm{B}}T}\right)-1\right), \tag{5}$$

where A^{*} is the Richardson constant and T is the temperature. For positive voltages $V > 0$, thermionic-field emission is assumed as dominant current contribution resulting in

$$I_{\mathrm{Sc},\,V>0\ \mathrm{V}} = AA^{*}\frac{T}{k_{\mathrm{B}}}\sqrt{\pi W_{00}e\left(V_{\mathrm{Schottky}}+\frac{\phi_{\mathrm{Bn}}}{\cosh^{2}\left(\frac{W_{00}}{k_{\mathrm{B}}T}\right)}\right)}\cdot \exp\left(\frac{-e\phi_{\mathrm{Bn}}}{W_{0}}\right)\left(\exp\left(\frac{eV_{\mathrm{Schottky}}}{\xi}\right)-1\right). \tag{6}$$

The effective electrostatic barrier height $e\phi_{\mathrm{Bn}}$ is smaller than the nominal barrier height $e\phi_{\mathrm{Bn0}}$ due to the Schottky barrier lowering effect. It is calculated according to

$$e\phi_{\mathrm{Bn}} = e\phi_{\mathrm{Bn0}} - e\sqrt[4]{\frac{e^{3}zN_{\mathrm{disc}}(\phi_{\mathrm{Bn0}}-\phi_{\mathrm{n}}+V_{\mathrm{Schottky}})}{8\pi^{2}\varepsilon_{\phi_{B}}^{3}}}, \tag{7}$$

and is thus a function of the disc concentration. In Eq. (7), $e\phi_{\mathrm{n}}$ is the energy difference between conduction band edge and Fermi level in the insulator. The dielectric permittivity related to the barrier lowering is $\varepsilon_{\phi\mathrm{B}}$. The parameters W_{00}, W_0 and ξ in Eq. (6) are calculated as

$$W_{00} = \frac{eh}{4\pi}\sqrt{\frac{zN_{\mathrm{disc}}}{m^{*}\varepsilon \mathrm{e}}}, \tag{8}$$

$$W_{0} = W_{00}\coth\left(\frac{W_{00}}{k_{\mathrm{B}}T}\right), \tag{9}$$

and,

$$\xi = \frac{W_{00}}{(W_{00}/k_{\mathrm{B}}T)-\tanh(W_{00}/k_{\mathrm{B}}T)}. \tag{10}$$

In Eq. (8), ε denotes the dielectric permittivity of the insulator, h is Planck's constant and m^* is the effective electron mass. The resistance R_{plug} of the plug region is

$$R_{\mathrm{plug}} = \frac{l_{\mathrm{plug}}}{ezN_{\mathrm{plug}}\mu_{\mathrm{n}}A} \tag{11}$$

and R_{disc} is given by

$$R_{\mathrm{disc}} = \frac{l_{\mathrm{disc}}}{ezN_{\mathrm{disc}}\mu_{\mathrm{n}}A}\cdot \tag{12}$$

In Eq. (11), (12) μ_{n} denotes the electron mobility and N_{plug} is the defect concentration in the plug region.

To obtain different resistance ratios and nonlinearities of the switching kinetics the parameters need to be adapted as follows. The resistance ratio $r = R_{\mathrm{HRS}}/R_{\mathrm{LRS}}$ is determined by the definition of a minimum concentration N_{min} and a maximum concentration N_{max} for the state variable. Here, the LRS resistance is constant as we use a constant N_{max}. Four different values of N_{min} are chosen to obtain resistance ratios of 10^1, 10^2, 10^3, and 10^4 at a read voltage of V_{read} = 0.1 V. The nonlinearity of the switching kinetics can be tuned using two parameters: the activation energy ΔW_{A} and the effective thermal resistance R_{th}. The values were chosen in a way that the set switching occurs in about 1 µs if a rectangular voltage pulse of V_{Set} is applied. The magnitude of the voltage $V_{\mathrm{Set}}/V_{\mathrm{Reset}}$ is a function of the chosen resistance ratio as experimentally reported [14]. To define the variable parameters, we also set $T = T_0$ in Eqs. (5)-(10). This does not affect the temperature T_{ion}, which is calculated by Eq. (3).

Four different nonlinearities *SL* are chosen, which are consistent with the values reported in [1]. The nonlinearity is equivalent to the slope in a $\log(t_{sw})$-V_{app}, which defines how many orders of magnitude the device switches faster/slower due to an increase/decrease by 1 V. The variable parameters used in the simulation studies are summarized in Table I, whereas the constant simulation parameters are given in Table II.

TABLE I. VARIABLE SIMULATION PARAMETER

R_{th} [K/W]/ ΔW_A [eV]	$r = 1 \cdot 10^1$ $N_{min} = 9.0 \cdot 10^{25}$ m^{-3}	$r = 1 \cdot 10^2$ $N_{min} = 5.0 \cdot 10^{25}$ m^{-3}	$r = 1 \cdot 10^3$ $N_{min} = 3.4 \cdot 10^{25}$ m^{-3}	$r = 1 \cdot 10^4$ $N_{min} = 2.4 \cdot 10^{25}$ m^{-3}
$SL = 4.76$ V^{-1}	$6.9 \cdot 10^5$/0.6	$6.4 \cdot 10^5$/0.75	$7.6 \cdot 10^5$/0.85	$9.2 \cdot 10^5$/0.95
$SL = 5.56$ V^{-1}	$9.3 \cdot 10^5$/0.65	$1 \cdot 10^6$/0.85	$1.1 \cdot 10^6$/0.95	$1.3 \cdot 10^6$/1.05
$SL = 6.67$ V^{-1}	$1.2 \cdot 10^6$/0.7	$1.4 \cdot 10^6$/0.95	$1.4 \cdot 10^6$/1.05	$1.6 \cdot 10^6$/1.15
$SL = 8.83$ V^{-1}	$1.7 \cdot 10^6$/0.8	$1.8 \cdot 10^6$/1.05	$2 \cdot 10^6$/1.2	$2.1 \cdot 10^6$/1.3

TABLE II. CONSTANT SIMULATION PARAMETER

Parameter	*Value*	*Parameter*	*Value*	*Parameter*	*Value*
A [m^2]	$2.8 \cdot 10^{-15}$	z	2	μ_n [m^2/(Vs)]	$9.4 \cdot 10^{-6}$
l_{disc} [nm]	4	a [nm]	0.4	$e\phi_n$ [eV]	0.1
l_{cell} [nm]	6	ν_0 [s^{-1}]	$1 \cdot 10^{14}$	$e\phi_{Bn0}$ [eV]	0.8
m^* [kg]	$9.1 \cdot 10^{-31}$	k_B [J/K]	$1.38 \cdot 10^{-23}$	A^* [A/(mK^2)]	$6.01 \cdot 10^5$
$\varepsilon_{\phi B}/\varepsilon_0$	5.5	h [Js]	$6.63 \cdot 10^{-36}$	T_0 [K]	300
$\varepsilon/\varepsilon_0$	21.46	e [C]	$1.6 \cdot 10^{-19}$	R_S [Ω]	1160
N_{max} [m^{-3}]	$1 \cdot 10^{27}$	N_{plug} [m^{-3}]	$25 \cdot 10^{27}$		

2.1.1 Simulation results

Figure 2b shows the simulated *I-V* characteristics for the four different R_{HRS}/R_{LRS} ratios. A triangular voltage with 2 V amplitude and a sweep rate of 1 V/s was used. The simulation result shows that the resistance ratio can be adjusted by tuning N_{min}. To this end, higher voltages need to be applied to switch completely between N_{min} and N_{max}. Using lower voltages would lead to lower resistance ratios. Thus, the choice of N_{min} and N_{max} only defines the maximum possible R_{HRS}/R_{LRS} ratio. In a second simulation study, the switching kinetics are studied. For this, a constant voltage with variable amplitude is applied and the switching time is determined. For the set/reset switching study, the initial state is N_{min}/N_{max}. The set/reset switching time is the time required for complete switching between N_{min} and N_{max}. The resulting reset and set switching times are shown as a function of the applied voltage in Fig. 2c and 2d, respectively. As intended, the four *SL* parameter sets per *r* cross each other in one point close 1 µs switching time and exhibit different slopes.

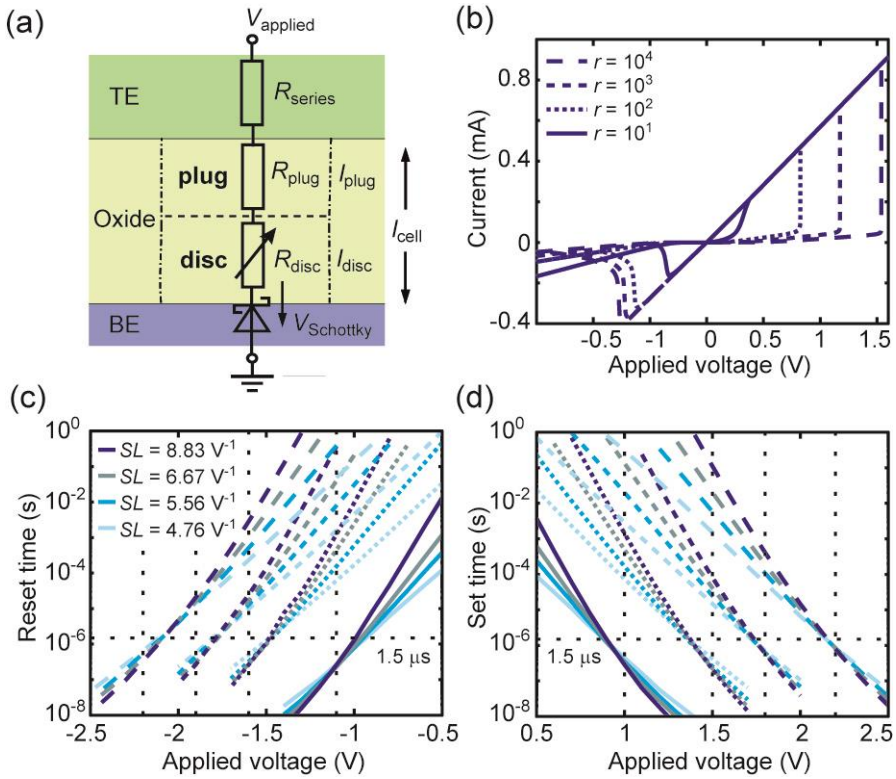

Fig. 2: (a) Equivalent circuit diagram of the simulation model showing the top electrode (TE), the bottom electrode (BE) and the active oxide layer. The filamentary region is divided into a plug region and a disc region. (b) Simulated *I-V* characteristics all resistance ratios *r* and the nonlinearity *SL* = 8.83 V^{-1}. (c) Reset and (d) set times vs. applied voltage for all combinations of *r* and *SL*. The corresponding legends are shown in (b) and (c) for *r* and *SL*, respectively.

3. Models for PCM Devices

Phase-change memory (PCM) is arguably the most advanced emerging non-volatile memory technology [16]. PCM is based on the property of certain materials such as $Ge_2Sb_2Te_5$ that exhibit a significant difference in resistivity depending on whether they are in the ordered crystalline phase or the disordered amorphous phase. In a PCM device, a tiny volume of such a material is sandwiched between two metal electrodes. A typical device structure is shown in the cross-sectional TEM image in Fig. 5(a). By the application of suitable electrical pulses and subsequent Joule heating, it is possible to reversibly alter the phase-configuration of the material within the device. Pulses that result in an increase in the size of the amorphous region are typically referred to as RESET pulses. In this case, the application of the pulse results in melting of a critical volume of the material and which is then rapidly quenched to induce glass transition. The pulses that reduce the size of the amorphous region are referred to as SET pulses. Here, the temperature reached within the device is favorable for crystallization (see Fig. 5(b)) [17]. Typically, the SET pulses that induce partial crystallization of the material is referred to as partial SET pulses and all these pulses are collectively referred to as programming pulses.

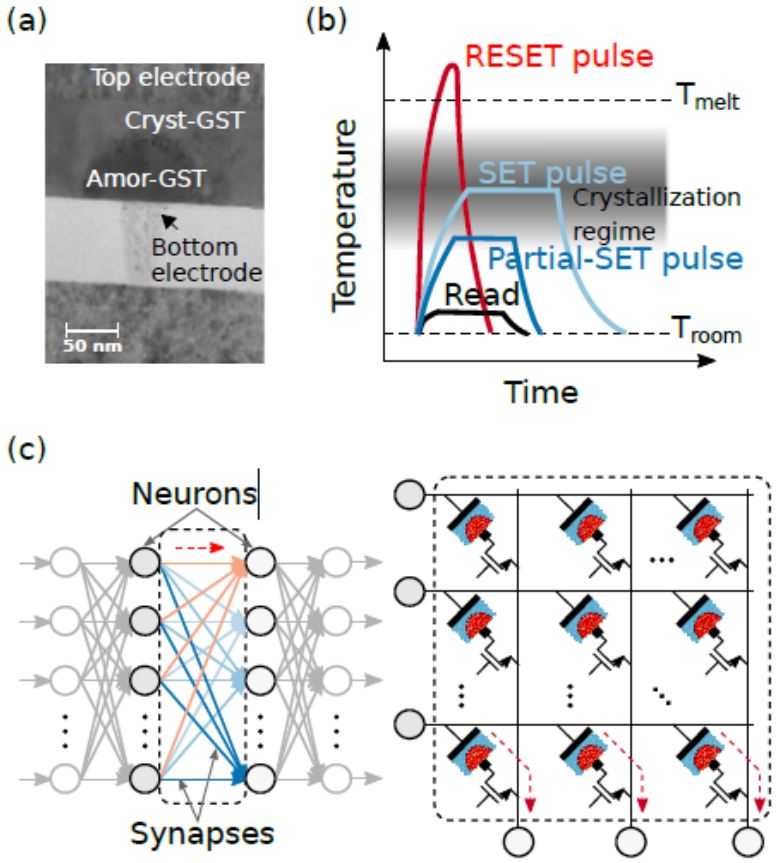

Fig. 5: (a) TEM image of a mushroom-type PCM device. Amorphous dome (Amor-GST) inside the crystalline (Cryst-GST) dielectric is visible (b) Pictorial representation of the programming pulses and the resulting relative temperature for RESET, SET, partial SET and read operation in PCM (c) Schematic illustration of the application of PCM devices as synaptic elements in neuromorphic computing. A crossbar array of PCM devices could be used to represent the connection strengths in a neural network layer.

The electrical resistance/conductance of the device will depend on the resulting phase-configuration. In fact, it is possible to achieve a continuum of resistance values in a single device and this can be exploited for computation-in-memory applications such as neuromorphic computing. For example, as shown in Fig. 5(c), PCM devices organized in a cross-bar configuration can be used to emulate the synaptic elements in an artificial neural network [18, 19]. The synaptic weights are captured by the conductance values of the PCM devices. The inputs from one layer of neurons are weighted by these conductance values (via Ohm's law) and the resulting current along the columns serve as inputs to the next layer of neurons. During the training of a neural network, the initial conductance values are typically chosen randomly, which is then modified (synaptic plasticity) via some appropriate learning rule. The programming pulses can be used to alter the conductance values during the training process. Unlike RESET pulses, which cause in an abrupt transition to lower conductance values, successive application of a partial SET pulse results in a more progressive increase in the conductance value. This cumulative evolution of conductance is highly beneficial for neuromorphic applications. Hence, often in PCM, only the partial SET pulses are used to implement synaptic plasticity rules [19]. To avoid the use of RESET pulses, PCM devices are organized in a differential configuration [20]. A comprehensive understanding of this

accumulative behavior across a large number of devices is central to the realization of large-scale neural networks. Besides crystallization, there are other structural dynamics at play in PCM devices. These devices exhibit a temporal evolution of conductance values after the application of each programming pulse. This is attributed to a spontaneous structural relaxation of the material [21] and could also play a key role in neuromorphic computing.

In this report, we present a comprehensive model of PCM devices that captures the accumulative behavior, conductance drift and read noise. Extensive experimental characterization of 10,000 PCM devices has been performed to develop this statistical model. Finally, we demonstrate the efficacy of this model by using it to match experimentally observed array level characteristics.

3.1 PCM device characterization and modeling

For device characterization, we used mushroom-type PCM devices fabricated in the 90nm technology node [22]. The phase-change material is doped $Ge_2Sb_2Te_5$ (GST). A prototype chip comprising 3 million devices was used in the study [23] . Individual devices are addressed via word lines and bit lines and the devices have access transistors in series. The devices are programmed using current pulses of designated amplitude and width generated in the peripheral circuits. The conductances are read by applying a 0.3 V read pulse and the resulting current is read using an 8-bit ADC. The ADC is calibrated to span a conductance range between 0.1 µS and 27 µS.

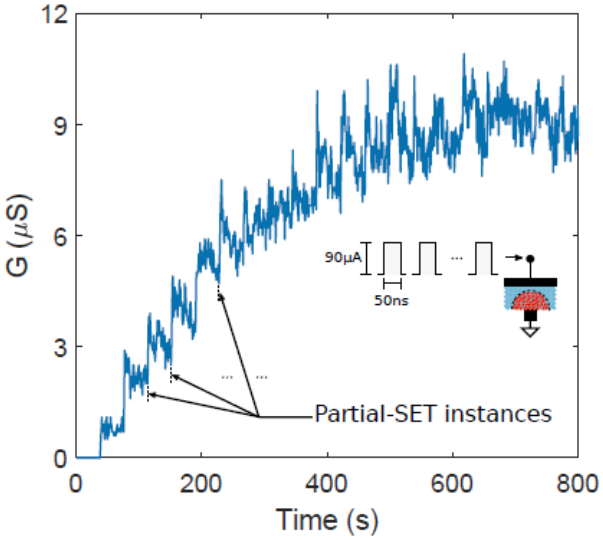

Fig. 6: The measured conductance evolution in a single device in accordance with the application of 20 consecutive partial SET pulses. The time instances of programming are indicated by the spike sequence displayed at the top. After the application of each pulse, the device conductance is measured 50 times.

First, the device conductances were initialized to a distribution close to 0.1 µS using iterative programming [24]. Subsequently, we applied 20 partial SET pulses of 90 µA amplitude and 50 ns duration. After the application of each pulse, devices are read 50 times. In addition, an immediate conductance measurement is performed approximately 100 ns after the programming pulse. However, subsequent measurements are obtained at time intervals in the order of seconds. As a result, consecutive programming pulses were applied with an average interval of 38.6 s for the 10,000 devices. The resulting conductance evolution, except for the

immediate read after 100 ns, for one such representative device is shown in Fig. 6. In subsequent sections, we will use all the measurements from the 10,000 devices (barring a few whose conductances where outside the ADC limits) to develop the statistical model.

3.1.1 Accumulative behavior

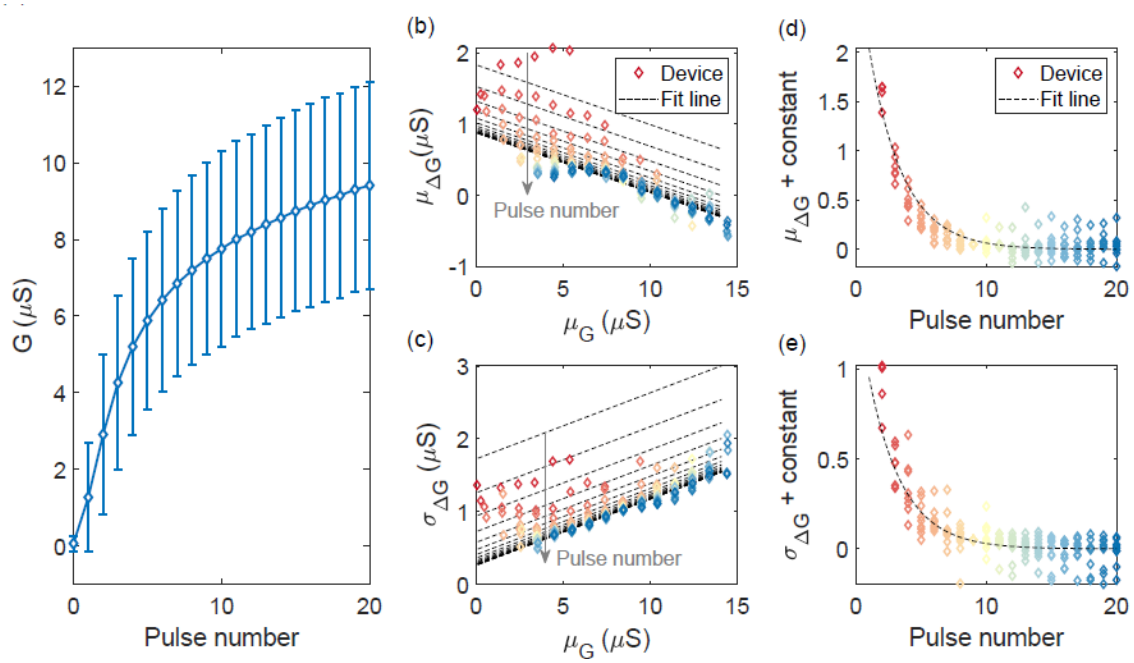

Fig. 7: (a) The statistics of cumulative conductance evolution as a function of the number of partial SET pulses. The error bars indicate the standard deviation. (b) The mean $\mu_{\Delta G}$ and (c) standard deviation $\sigma_{\Delta G}$ of conductance change as a function of the average initial conductance μ_G for each programming pulse. The initial conductance distribution for each programming pulse is divided into smaller intervals and μ_G, $\mu_{\Delta G}$ and $\sigma_{\Delta G}$ is determined separately for each interval. Each data point in (b) and (c) corresponds to an average of measurements from at least 100 devices. Also depicted are the fit lines used to obtain the model parameters. (d), (e) The same data points of $\mu_{\Delta G}$ and $\sigma_{\Delta G}$ are plotted as a function of the pulse number with a constant added for data points corresponding to a single μ_G interval. The dependency of $\mu_{\Delta G}$ and $\sigma_{\Delta G}$ on pulse number is approximated using an exponential function with a decay constant of 2.6.

First, we characterized the accumulative behavior arising from the successive application of partial SET pulses. To decouple the accumulative behavior from conductance drift, the 50th read measurement was used. The distribution of the conductance values as a function of the pulse number is shown in Fig. 7(a). It can be seen that the average conductance change is high at low conductance values and it gradually reduces as the conductance values increase. It can also be seen that there is significant randomness associated with the conductance values. This is mostly attributed to the inherent randomness associated with the crystallization process [25, 26]. In fact, the inter- and intra-device variability in the array has been observed to be of comparable magnitude [27-29].

To obtain a quantitative description of this behavior, we studied how the conductance change arising from the application of a single SET pulse depends on the conductance state of the device prior to the application of the pulse as well as the programming history of the device. The devices were split into different groups based on their conductance values. Each group

corresponds to a conductance interval of 1 μS. For each group, the mean ($\mu_{\Delta G}$) and standard deviation ($\sigma_{\Delta G}$) of the conductance change due to the application of a single programming pulse is plotted against the mean conductance (μ_G) of each group (see Fig. 7(b) and 7(c)). The data points are generated only for those groups with 100 or more devices. This is repeated for the conductance values measured after the application of each programming pulse. In Fig. 7(b) and 7(c), each color corresponds to a single programming pulse with the red color indicating the first pulse and the blue color the 20th pulse. We observe that there is a negative correlation between $\mu_{\Delta G}$ and μ_G that suggests a linear decrease in the conductance change as the device conductance increases. In addition, in a particular conductance range, the conductance change observed seem to decrease with increasing number of applied pulses. This behavior can be captured using a linear fit of a negative slope to map the relation between $\mu_{\Delta G}$ and μ_G f for any particular pulse number. Further, the dependency on the pulse number is encoded in the *y*-intercept of this linear fit. It can be seen that for any given conductance value, the extent of conductance change induced by a single partial SET pulse reduces significantly with increasing number of applied pulses. This could be captured using an exponential empirical relation (Fig. 7(d) and (e)).

It can be seen that the behavior of $\sigma_{\Delta G}$ is also very similar to that of $\mu_{\Delta G}$ except that there is a positive correlation with the μ_G in this case. Therefore, the mean and standard deviation of the ΔG is modeled respectively using lines of negative and positive slopes and with an intercept which is an exponential function of the pulse number (p) as in the following equations (also in Fig. 7(b) and (c)):

$$\mu_{\Delta G} = m_1 G + \left(c_1 + A_1 e^{-p/\alpha}\right), \tag{18}$$

$$\sigma_{\Delta G} = m_2 G + \left(c_2 + A_2 e^{-p/\alpha}\right), \tag{19}$$

where the fit parameters m_1, m_2, c_1, c_2, A_1, A_2, and α are -0.084, 0.091, 0.880, 0.260, 1.40, 2.15, and 2.6, respectively.

3.1.2 Conductance drift and read noise

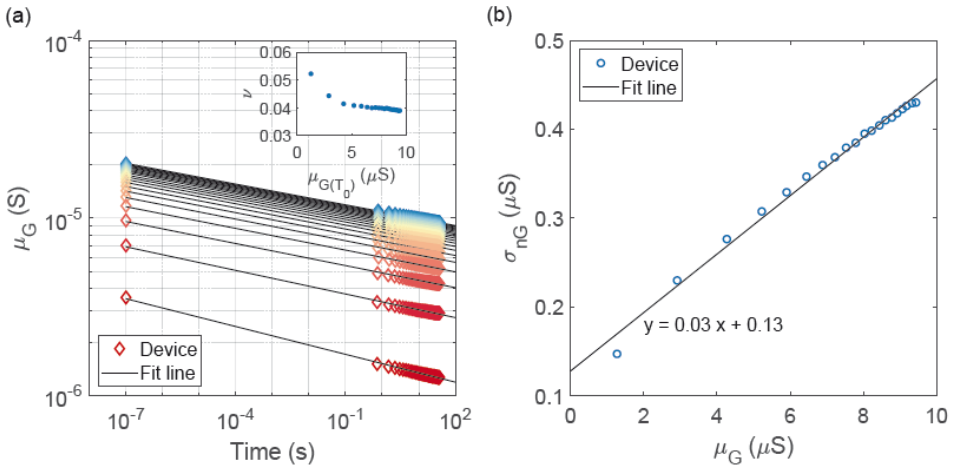

Fig. 8: (a) The average conductance evolution after each programming pulse obtained by the 50 read operation is fitted using equation (20). The estimated drift-coefficient, ν, is shown in the inset (b) The read noise measured from the device array is plotted as a function of the average device conductance. The linear fit used to estimate read noise for the model in a state-dependent manner is also shown.

In this section, we model the conductance drift in the devices arising from structural relaxation. For this, we use the 50 read measurements obtained after the application of each SET pulse. The mean conductance evolution after each programming event as a function of time is plotted in Fig. 8(a). The response is fitted using the model [30, 31]

$$G(t) = G(T_0)\left(\frac{t}{T_0}\right)^{-\nu}. \quad (20)$$

According to equation (20), if the device conductance, $G(T_0)$ is known at time T_0 after programming, the conductance at any time t can be estimated with the knowledge of the drift coefficient, ν. The estimated ν from the fit lines has a mean value of 0.04 (Fig. 8(a) inset). Note that the logarithmic dependence on time suggests that after programming, the conductance drift slows down with time. We observe that the partial SET pulses result in a state that drifts, with a drift coefficient that decreases with increasing conductance $\mu_G(T_0)$. The application of a partial SET pulse re-initiates structural relaxation and conductance drift. Hence, we speculate that each partial SET pulse creates a new unstable glass state because of the atomic rearrangement that occurs upon its application, which then structurally relaxes to an energetically more favorable amorphous state [21, 32].

In addition to the conductance drift, there are also significant fluctuations in the conductance values (read noise) mostly arising from the $1/f$ noise exhibited by amorphous phase-change materials [33]. To model this, we estimated the noise from the last ten reads from the fifty read measurements. The objective was to decouple the read noise from the conductance drift. The standard deviation of the zero-mean read noise is plotted as a function of the mean conductance (see Fig. 8(b)). It can be seen that the read noise increases with the device conductance. The read noise standard deviation, σ_{nG}, for the device conductance range is fitted using the linear relation

$$\sigma_{nG} = m_3 G + c_3, \quad (21)$$

where m_3 = 0.03 and c_3 = 0.13.

3.1.3 The overall model description and validation

In this section, we combine the various elements of the model describing the accumulative behavior, conductance drift, and read noise to generate a complete statistical model and validate it based on the experimental data. The objective is to capture the evolution of conductance values for a large collection of devices after a certain time T_0 after programming with an arbitrary number of partial SET pulses. More specifically, we would like to determine the device conductance $G(t)$ at any time t, which has been initialized to approximately 0.1 μS, and is subjected to a sequence of 90 μA, 50 ns programming pulses with arbitrary time intervals between them.

Table III: PCM model parameters

Symbol	Value	Symbol	Value	Symbol	Value
m_1	-0.084	c_1	0.880	A_1	1.40
m_2	0.091	c_2	0.260	A_2	2.15
α	2.6	T_0	38.6 s	ν	0.04
m_3	0.03	c_3	0.13		

To simulate this, three quantities are recorded per device: (a) $G_i(T_0)$, the conductance after time T_0 after the i^{th} programming pulse for $i = 0, 1, 2 \ldots$, (b) P_{mem}, a quantity that captures the programming history, and (c) t_p, the time of the last programming event. t_p is initialized to zero. Based on the chosen initial conductance value $G_0(T_0)$, P_{mem} is initialized to $P_{mem} = \exp(-p_0/\alpha)$, where p_0 is the effective number of pulses applied to reach the initial conductance $G_0(T_0)$. p_0 is zero for initialization around 0.1 μS and p_0 for higher values of conductance is determined from the average conductance evolution curve shown in Fig. 7(a). The effective number of pulses versus conductance can be approximated empirically as

$$p_0 = 0.027\mu_G^3 - 0.15\mu_G^2 + 0.81\mu_G \tag{22}$$

for conductances ranging from 0.1 μS to around 8 μS. After initialization, for the N^{th} programming event, P_{mem} is first updated as $P_{mem,N} = P_{mem,N-1}\exp(-1/\alpha)$ for $N = 1, 2 \ldots$. Then, $G(t)$, which has seen N programming pulses can be determined as follows:

$$\mu_{\Delta G_N} = m_1 G_{N-1}(T_0) + (c_1 + A_1 P_{mem}) \tag{23}$$

$$\sigma_{\Delta G_N} = m_2 G_{N-1}(T_0) + (c_2 + A_2 P_{mem}) \tag{24}$$

$$\Delta G_N = \mu_{\Delta G_N} + \sigma_{\Delta G_N}\chi \tag{25}$$

$$G_N(T_0) = G_{N-1}(T_0) + \Delta G_N \tag{26}$$

$$G(t) = G_N(T_0)\left(\frac{t - t_p}{T_0}\right)^{-\nu} + n_G \tag{27}$$

Here, χ represents a Gaussian random number of mean zero and variance 1. Another Gaussian random variable with mean zero, n_G, captures the conductance fluctuations arising from PCM noise, whose standard deviation is calculated based on the instantaneous conductance state as dictated by the linear fit in equation (21) (also in Fig. 8(b)). All the model parameters are listed in Table III. Please note that the conductance values predicted by the model are in μS. First, the model is used to validate the same experimental data that was used to generate the model parameters. In particular, the model is used to generate the distribution of conductance values as a function of the number of programming pulses. As shown in Fig. 9, the mean and variance match remarkably well with experimental data. It can also be seen that the distributions themselves are remarkably similar. Fig. 10(a)-(c) show the conductance distribution from the 50^{th} read, after initialization, after the application of 5 programming pulses, and after the application of 20 programming pulses, respectively. The model also matches the correlation coefficient observed between G and ΔG for the pulses applied. From Fig. 10(e) and (f), it can be seen that the statistical model also captures the individual device behavior remarkably well.

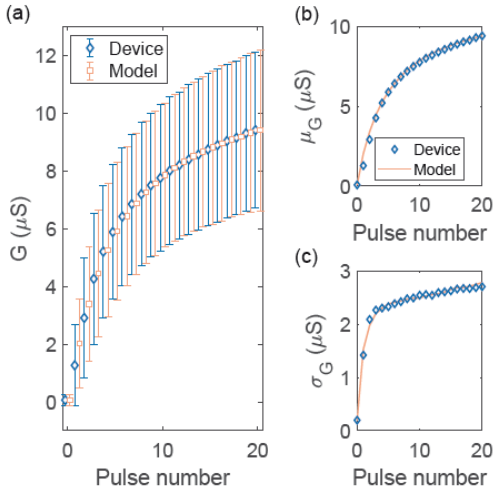

Fig. 9: (a) The distribution of conductance values obtained using the model as a function of the number of partial SET pulses and match with experiments. (b) The mean of the conductance as a function of the pulse number. (c) The standard deviation of the conductance as a function of the pulse number.

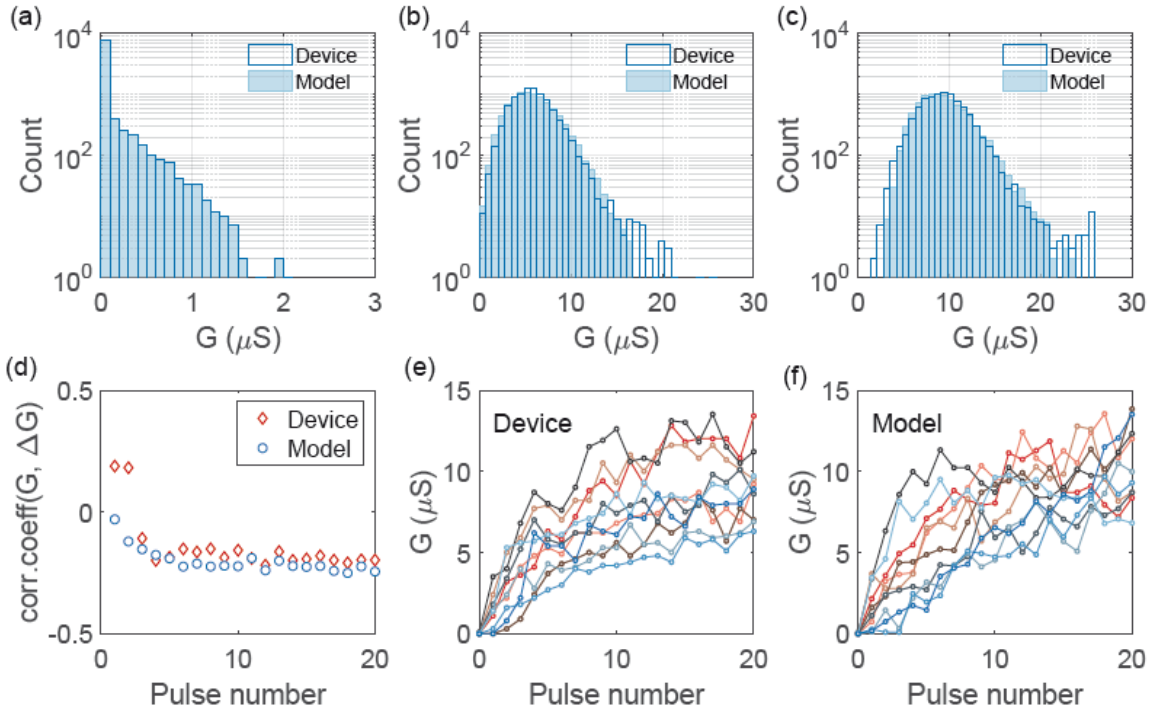

Fig. 10: The distribution of conductance values after (a) initialization (b) the application of 5 programming pulses and (c) the application of 20 programming pulses. It can be seen that there is a remarkable agreement between the experimental distribution and that predicted by the model. (d) The correlation coefficient between the G and ΔG after the application of each programming pulse calculated based on the model and is compared with the experimental measurement. The conductance evolution of individual devices as measured experimentally (e) and as predicted by the model (f).

Additional measurements were performed where the devices are programmed with 20 programming pulses, however, with varying time intervals between the application of each pulse. The time interval was determined based on the number of reads performed and in the current experiment, each read process took approximately 1 s for the 10,000 devices. Fig. 11 shows the programming events in time (top) and the resulting evolution of the mean conductance of the 10,000 devices (bottom). The spikes in the programming event plot correspond to the application of partial SET pulse and the device conductances are read at all other time instances. As discussed earlier, it can be seen that with the application of each programming pulse, the drift process is re-initiated. Another interesting observation is that the net change in conductance seems to be independent of structural relaxation. There is some evidence that structural relaxation slows down crystal growth rate [17]. But at least in these devices and at these time scales, this does not seem to be significant. The final conductance values at the end of programming seem to converge to similar conductance levels independent of the rate of programming. Hence, our proposed model is able to capture this behavior remarkably well with the additional incorporation of equation (20).

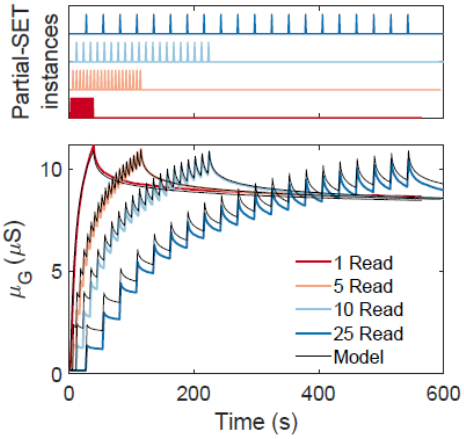

Fig. 11: The 10,000 PCM devices are programmed using sequences of 20 current pulses of 90 µA and 50 ns width. The number of read operations performed after each programming event is varied resulting in different time intervals between the programming events. The resulting conductance evolution during all the read operations is illustrated. The proposed model captures the experimentally observed behavior remarkably well.

4. Outlook

For the VCM modeling, the next step is the development of a read model including read noise due to stochastic changes of the configuration. Moreover, we intend to fit the developed models to the new set of data obtained from the array measurements. For the new devices, the models need to be verified, too. A further task is the development of a VCM retention model, which could be similar to the PCM retention model. To this end, experimental data needs to be collected first.

On the PCM side, the presented model will be extended further to more accurately capture the drift characteristics. For this, we will incorporate to the model the drift coefficient variability across devices and the conductance dependency of the drift coefficient. Another way to obtain a desired conductance on a PCM device is through iterative programming. A sequence of

pulses with varying amplitudes, determined with a feedback mechanism involving conductance reads, can set the conductance of a device more precisely (i.e. with a tighter conductance distribution) compared to the blind application (i.e. without any feedback mechanism) of successive partial SET pulses A further task would be to experimentally obtain and subsequently model the iterative programming distributions of PCM.

5. References

[1] 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.

[2] F. Pan, S. Gao, C. Chen, C. Song and F. Zeng, “Recent progress in resistive random access memories: Materials, switching mechanisms, and performance,” *Mater. Sci. Eng. R-Rep.*, vol. 83, pp. 1-59, 2014.

[3] R. Waser, R. Dittmann, G. Staikov and K. Szot, “Redox-Based Resistive Switching Memories - Nanoionic Mechanisms, Prospects, and Challenges,” *Adv. Mater.*, vol. 21, pp. 2632-2663, 2009.

[4] F.A. Kröger and H.J. Vink,*Solid State Physics* Academic Press, 1956, pp. 307.

[5] J. J. Yang, M. D. Pickett, X. Li, D. A. A. Ohlberg, D. R. Stewart and R. S. Williams, “Memristive switching mechanism for metal/oxide/metal nanodevices,” *Nat. Nanotechnol.*, vol. 3, pp. 429-433, 2008.

[6] K. Szot, W. Speier, G. Bihlmayer and R. Waser, “Switching the electrical resistance of individual dislocations in single-crystalline $SrTiO_3$,” *Nat. Mater.*, vol. 5, pp. 312-320, 2006.

[7] A. Marchewka, B. Roesgen, K. Skaja, H. Du, C. L. Jia, J. Mayer, V. Rana, R. Waser and S. Menzel, “Nanoionic Resistive Switching Memories: On the Physical Nature of the Dynamic Reset Process,” *Adv. Electron. Mater.*, vol. 2, pp. 1500233/1-13, 2016.

[8] A. Marchewka, R. Waser and S. Menzel, “Physical Simulation of Dynamic Resistive Switching in Metal Oxides Using a Schottky Contact Barrier Model,” *2015 International Conference On Simulation of Semiconductor Processes and Devices (SISPAD), 9-11 September, Washington D.C, USA*, 2015, pp. 297-300.

[9] R. Waser, R. Bruchhaus and S. Menzel, “Redox-based Resistive Switching Memories,” *Wiley-VCH*, vol., pp. 683-710, 2012.

[10] 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.

[11] A. Hardtdegen, C. La Torre, F. Cüppers, S. Menzel, R. Waser and S. Hoffmann-Eifert, “Improved Switching Stability and the Effect of an Internal Series Resistor in HfO_2/TiO_x Bilayer ReRAM Cells,” *IEEE Trans. Electron Devices*, vol. 65, pp. 3229-3236, 2018.

[12] K. Fleck, C. La Torre, N. Aslam, S. Hoffmann-Eifert, U. Böttger and S. Menzel, “Uniting Gradual and Abrupt SET Processes in Resistive Switching Oxides,” *Phys. Rev. Applied*, vol. 6, pp. 064015, 2016.

[13] A. Siemon, S. Menzel, A. Marchewka, Y. Nishi, R. Waser and E. Linn, “Simulation of TaO_x-based Complementary Resistive Switches by a Physics-based Memristive Model,” *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, vol., pp. 1420-1423, 2014.

[14] Y. Nishi, S. Menzel, K. Fleck, U. Boettger and R. Waser, “Origin of the SET Kinetics of the Resistive Switching in Tantalum Oxide Thin Films,” *IEEE Electron Device Lett.*, vol. 35, pp. 259-261, 2013.

[15] C. Baeumer, R. Valenta, C. Schmitz, A. Locatelli, T. O. Mentes, S. P. Rogers, A. Sala, N. Raab, S. Nemsak, M. Shim, C. M. Schneider, S. Menzel, R. Waser and R. Dittmann, “Subfilamentary Networks Cause Cycle-to-Cycle Variability in Memristive Devices,” *ACS Nano*, vol. 11, pp. 6921-6929, 2017.

[16] G. W. Burr, M. J. Brightsky, A. Sebastian, H. Cheng, J. Wu, S. Kim, N. E. Sosa, N. Papandreou, H. Lung, H. Pozidis, E. Eleftheriou and C. H. Lam, “Recent Progress in Phase-Change Memory Technology,” *IEEE J. Emerging Sel. Top. Circuits Syst.*, vol. 6, pp. 146-162, 2016.

[17] A. Sebastian, M. Le Gallo and D. Krebs, “Crystal growth within a phase change memory cell,” *Nature Communications*, vol. 5, pp. 4314/1-, 2014.

[18] B. L. Jackson, B. Rajendran, G. S. Corrado, M. Breitwisch, G. W. Burr, R. Cheek, K. Gopalakrishnan, S. Raoux, C. T. Rettner, A. Padilla, A. G. Schrott, R. S. Shenoy, B. N. Kurdi, C. H. Lam and D. S. Modha, “Nanoscale electronic synapses using phase change devices,” *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 9, pp. 12, 2013.

[19] G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola, L. L. Sanches, I. Boybat, M. Le Gallo, K. Moon, J. Woo, H. Hwang and Y. Leblebici, “Neuromorphic computing using non-volatile memory,” *Adv. Phys.-X*, vol. 2, pp. 89-124, 2017.

[20] Manan Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat and B. DeSalvo, “Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction,” *Proceedings of the International Electron Devices Meeting (2011)*, vol., pp., 2011.

[21] M. Le Gallo, D. Krebs, F. Zipoli, M. Salinga and A. Sebastian, “Collective structural relaxation in phase-change memory devices,” *Adv. Electron. Mater.*, vol., pp. 1700627, 2018.

[22] M. Breitwisch, T. Nirschl, C.F. Chen, Y. Zhu, M.H. Lee, M. Lamorey, G.W. Burr, E. Joseph, A. Schrott, J.B. Philipp, R. Cheek, T.D. Happ, S.H. Chen, S. Zaidi, P. Flaitz, J. Bruley, R. Dasaka, B. Rajendran, S. Rossnagel, M. Yang, Y.C. Chen, R. Bergmann, H.L. Lung and C. Lam, “Novel Lithography-Independent Pore Phase Change Memory,” *Proceedings of the IEEE Symposium on VLSI Technology 2007*, vol., pp. 100-101, 2007.

[23] G. F. Close, U. Frey, M. Breitwisch, H. L. Lung, C. Lam, C. Hagleitner and E. Eleftheriou, “Device, circuit and system-level analysis of noise in multi-bit phase-change memory,” *Proceedings of the 2010 International Electron Devices Meeting, 6-8 December 2010, San Francisco, USA*, 2010, pp.

[24] N. Papandreou, H. Pozidis, A. Pantazi, A. Sebastian, M. Breitwisch, C. Lam and E. Eleftheriou, “Programming algorithms for multilevel phase-change memory,” *2011 IEEE International Symposium of Circuits and Systems, 15-18 May 2011, Rio de Janeiro, Brazil*, 2011, pp. 329-332.

[25] Tomas Tuma, Angeliki Pantazi, Manuel Le Gallo, Abu Sebastian and Evangelos Eleftheriou, “Stochastic phase-change neurons,” *Nat. Nanotechnol.*, vol., pp., 2016.

[26] M. Le Gallo, T. Tuma, F. Zipoli, A. Sebastian and E. Eleftheriou, “Inherent stochasticity in phase-change memory devices,” *2016 46th European Solid-State Device Research Conference (essderc)*, 2016, pp. 373-376.

[27] I. Boybat, M. Le Gallo, T. Moraitis, Y. Leblebici, A. Sebastian and E. Eleftheriou, “Stochastic weight updates in phase-change memory-based synapses and their influence on artificial neural networks,” *2017 13th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME), 12-15 June 2017, Giardini Naxos, Italy*, 2017, pp. 13-16.

[28] N. Gong, T. Idé, S. Kim, I. Boybat, A. Sebastian, V. Narayanan and T. Ando, “Signal and noise extraction from analog memory elements for neuromorphic computing,” *Nature Communications*, vol. 9, pp. 2102, 2018.

[29] I. Boybat, M. Le Gallo, S. R. Nandakumar, T. Moraitis, T. Parnell, T. Tuma, B. Rajendran, Y. Leblebici, A. Sebastian and E. Eleftheriou, “Neuromorphic computing with multi-memristive synapses,” *Nature Communications*, vol. 9, pp. 2514, 2018.

[30] A. Pirovano, A. L. Lacaita, F. Pellizzer, S. A. Kostylev, A. Benvenuti and R. Bez, “Low-field amorphous state resistance and threshold voltage drift in chalcogenide materials,” *IEEE Trans. Electron Devices*, vol. 51, pp. 714-719, 2004.

[31] Daniele Ielmini, Deepak Sharma, Simone Lavizzari and Andrea. L. Lacaita, “Reliability Impact of Chalcogenide-Structure Relaxation in Phase-Change Memory (PCM) Cells—Part I: Experimental Study,” *IEEE Trans. Electron Devices*, vol. 56, pp. 1070-1077, 2009.

[32] M. Boniardi and D. Ielmini, “Physical origin of the resistance drift exponent in amorphous phase change materials,” *Appl. Phys. Lett.*, vol. 98, pp. 243506, 2011.

[33] 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: Condens. Matter*, vol. 79, pp. 1654206, 2009.