Compact modeling of resistive switching memory (RRAM) with voltage and temperature dependences

A. Glukhov[1], D. Bridarolli[1], S. Ricci[1], R. Li[2], S. Shreya[2], H. Farkhani[2], F. Moradi[2] and D. Ielmini[1]
[1]Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano and IU.NET, Milano, Italy.
[2]Department of Electrical and Computer Engineering, Aarhus University, Denmark.
Email: daniele.ielmini@polimi.it

***Abstract*—Resistive-switching memory (RRAM) devices are an attractive technology for in-memory computing (IMC) to accelerate data-intensive tasks, such as deep neural network (DNN) inference. However, the RRAM characteristics are significantly affected by voltage and temperature dependences, as well as by statistical variability of the programmed state, leading to accuracy degradations of the computed results. Compact models of RRAM need to account for these dependencies and variations to allow for accurate simulations of IMC circuits and design-technology co-optimization (DTCO). This work presents a compact model of RRAM devices that accounts for the conduction characteristics of various programmed states as a function of voltage and temperature. The compact model is validated through experimental measurements, and simulations of a simple use case of matrix-vector multiplication in a passive crosspoint array. The model can serve for the simulation and design of neural network accelerators based on RRAM technology.**

***Index Terms*—Resistive-switching memory (RRAM), compact model, in-memory computing (IMC), deep neural network (DNN), matrix-vector multiplication (MVM)**

I. Introduction

Resistive-switching memory (RRAM) devices are receiving increasing interest as a viable and promising technology for in-memory computing (IMC) systems capable of higher parallelism and lower energy consumption [1]. RRAM devices display nonvolatile storage, multilevel cell (MLC) programming, low-power operation, and compatibility with the standard CMOS technology. When organized in crosspoint arrays, RRAM can perform matrix-vector multiplication (MVM) in the analog domain based on physical Ohm's and Kirchhoff's laws [2]. However, RRAM devices suffer from several nonidealities, such as statistical variability of the programmed state [3], time-dependent fluctuations [4] and $1/f$ noise [5]. Also, the RRAM characteristics generally display a non-linear dependence on voltage and temperature. These nonidealities significantly impact the reliability and accuracy of the RRAM-based IMC systems, such as hardware accelerators of deep neural networks (DNN). To optimize the accuracy of IMC-based accelerators, accurate compact models for RRAM are essential. Several models have been developed for RRAM devices, including ab-initio models [6], numerical finite-element-method (FEM) models [7], kinetic Monte Carlo models [8], and compact models for circuit simulations [9]–[11] to allow for the simulation of RRAM circuits.

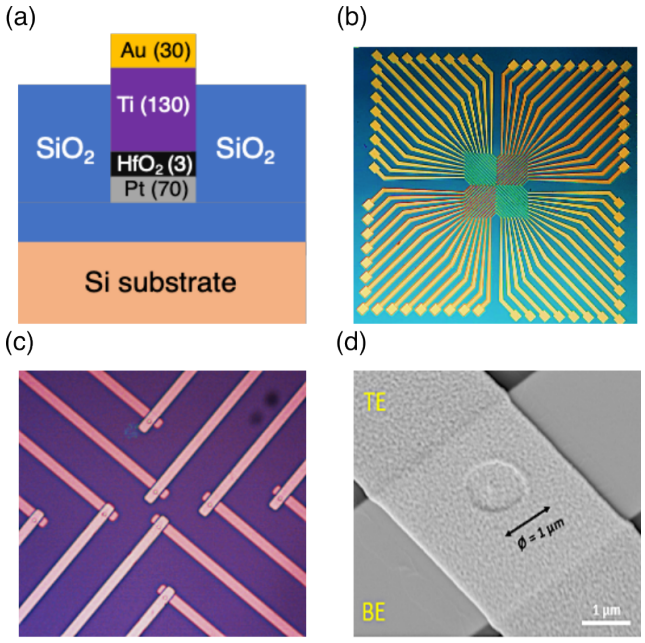

Fig. 1. RRAM devices used in the experiment to develop the compact model. (a) Sketch of the materials stack in the RRAM device. (b) Microphotograph of the test chip. (c) Zoom of the microphotograph on the crosspoint. (d) SEM image of a single RRAM cell.

This work presents a physics-based compact model of RRAM able to describe the nonlinear $I-V$ characteristics at low voltages (below 0.5 V) for various temperatures. Experimental $I-V$ curves for HfO_2-based RRAM devices are first studied for various programmed states and temperatures. Then, the compact model is developed on the concept of Poole-Frenkel (PF) transport, featuring only three free parameters. The model is validated against experimental $I-V$ curves and applied to the SPICE-level simulations of IMC circuits.

II. RRAM devices and experiments

The RRAM devices consist of a HfO_2 switching layer with a thickness $t_{ox} = 3$ nm, a bottom electrode (BE) made

This work received funding from the Italian Ministry of University and Research (MUR) and the European Union (EU) under the PON/REACT project and from the European Union's Horizon 2020 research and innovation program (grant agreement No. 899559).

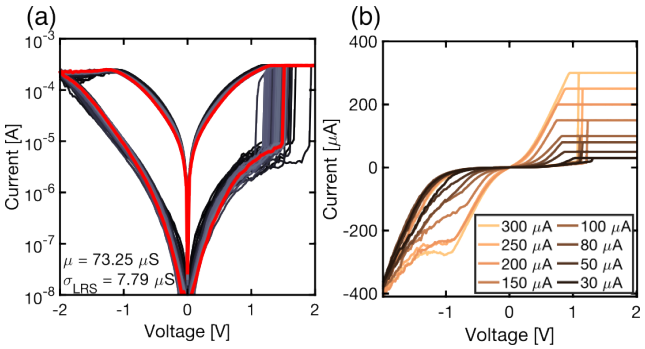

Fig. 2. (a) Experimental $I-V$ characteristics of 50 set and reset cycles in logarithmic scale at fixed compliance current. (b) $I-V$ characteristics for several compliance currents in linear scale. The different slope at low voltage underlies the programmability of the cell.

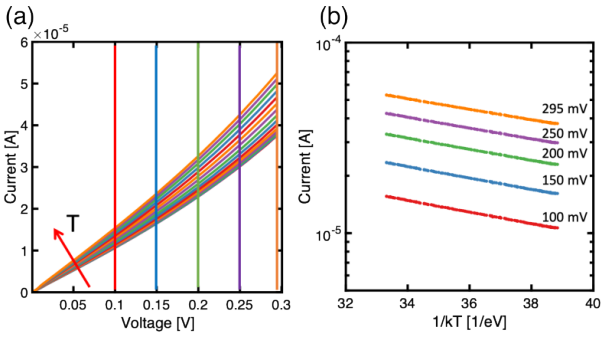

Fig. 3. Experimental characteristics of the memristor. (a) Measured I-V curve for G_0 = 150 μS at increasing temperature, from room temperature (25 °C) to 100 °C. (b) Arrhenius plot of measured current for increasing applied voltage.

of Pt, and a top electrode (TE) made of Ti. Fig. 1a shows a sketch of the device stack. All layers were deposited by e-beam evaporation and capped by Au metallization to minimize the electrode resistance. Details of the fabrication process can be found elsewhere [12]. Fig. 1b and c show optical images of the fabricated cells arranged as single independent crosspoint devices, while Fig. 1d shows a scanning electrode microscope (SEM) image of the crosspoint region, where the active device area is defined by a via hole across the insulating SiO_2 layer (see Fig. 1a).

Fig. 2a shows the measured current-voltage $(I-V)$ curves of set and reset cycles obtained by quasistatic measurements. Prior to these measurements, the device was formed to initialize the conductive filament across the HfO_2 layer. The application of a positive voltage sweep to the TE leads to a set transition to the low resistance state (LRS), while the application of a negative voltage results in a reset transition to the high resistance state (HRS). Note that the current during the set transition was limited by a compliance current I_C. Fig. 2b shows measured $I-V$ curves for increasing I_C, resulting in an increasing conductance G of the programmed LRS state. Note the non-linear shape of the $I-V$ curve in the read voltage range, namely $|V| < 0.5$ V.

To study the transport process on the basis of the nonlinear behavior in Fig. 2, the RRAM devices were programmed in a target state at room temperature followed by $I-V$ measurement at elevated temperature in the read domain for multiple programmed states. Fig. 3a shows the measured $I-V$ characteristic for a device programmed at G_0 = 150 μS at increasing T (from 25 to 100 °C). The $I-V$ curves show a linear current relation for voltages lower than 0.1 V, indicating Ohmic conduction. At higher voltage, the current increases nonlinearly. As the temperature T increases, the measured current increases substantially. To better grasp the T-dependence of the current, Fig. 3b shows the Arrhenius plot of the current at increasing V from Fig. 3a. Data indicate an Arrhenius dependence of the measured current according to

$$I = I_0 e^{-\frac{E_A(V)}{kT}}, \tag{1}$$

where k is the Boltzmann constant and E_A is the activation energy which depends on the voltage. Fig. 4a shows the extracted activation energy E_A as a function of the read voltage for the device in Fig. 3. The measured E_A decreases at increasing V.

III. Compact Model

The experimental $I-V$ curves show an ohmic relationship between the current and the voltage for $V < 0.1$ V, followed by an exponential dependence at higher voltages. This conduction mechanism can be explained by the Poole-Frenkel conduction model, where carriers (*e.g.*, electrons) must be thermally emitted from a trapped state at localized defects, *e.g.*, oxygen vacancies, to the band of extended states (*e.g.*, the conduction band) to contribute to the electrical transport. The non-linear increase of current can be explained by the barrier-lowering effect, where the Coulombic barrier decreases for increasing V. At relatively low voltages, the linear behavior occurs due to the non-negligible probability for an electron to jump in the direction opposite to the applied field, generating a reverse current that cancels out the forward current for $V = 0$, thus resulting in zero net current [13]. The current can thus be expressed as:

$$I = I_0 e^{-\frac{E_{C0}}{kT}} sinh\left(\alpha \frac{qV}{KT}\right) \tag{2}$$

where I_0 is a pre-exponential factor, proportional to the programmed conductance of the cell, E_{C0} is the energy barrier for the thermally-induced emission of the carriers, q is the electron charge and α is a barrier-lowering coefficient.

Fig. 4 shows the measured E_A and I as a function of the applied voltage compared to calculation results from Eq. (2). In particular, by taking the derivative of $log(I)$ with respect to $1/kT$, one can obtain an analytical expression for the activation energy, given by:

$$E_A = -\frac{\partial(logI)}{\partial(1/kT)} = E_{C0} - \alpha q V coth\left(\alpha \frac{qV}{KT}\right) \tag{3}$$

By accurately fitting the experimental $I-V$ curves with the simulation results, it is possible to extract the values of parameters E_{C0}, α, and I_0 in Eq. (2), for any given temperature T.

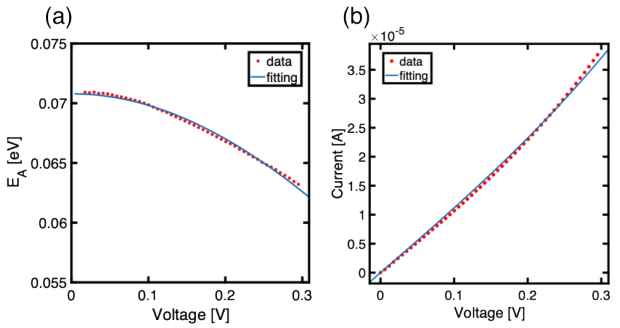

Fig. 4. Experimental and calculated I and E_A as a function of V. (a) Activation energy E_A as a function of the applied voltage. (b) Measured and calculated $I-V$ characteristic of a RRAM cell at room temperature.

IV. Simulation Results

Fig. 5 shows the measured and calculated characteristics for increasing G_0 from 2 μS to 106 μS. The model provides an accurate description of the curves in Fig. 3 and Fig. 4, thus corroborating the physical picture and allowing to capture the device behavior in both V and T with just three parameters, namely I_0, E_{C0}, and α. Fig. 6 shows the extracted parameters E_{C0} (a), α (b) and I_0 (c) as a function of the resistance $R_0 = 1/G_0$ for all available data. All parameters were obtained by accurately fitting the measured current as a function of both V and T as shown by Fig. 3. Despite some scattering, the extracted parameters show a qualitative trend where E_{C0} increases linearly with R_0, α is approximately constant, and I_0 decreases with R_0 according to a power law. The observed behavior was summarized by the analytical formulae:

$$E_{C0} = 0.02\ log(R_0) - 0.1\ [eV] \tag{4}$$

$$\alpha = 0.12 \tag{5}$$

$$I_0 = 0.01\ R_0^{-0.3}\ [A] \tag{6}$$

Eqs. (4-6) allow for a straightforward simulation of the $I-V$ curve for any given T and G_0 across a large range of conductance values over voltage and temperature variation. Based on the measurement of G_0, one can generate the characteristic parameters, namely I_0, E_{C0}, and α, and simulate the current for any given T and V. The compactness of the analytical model allows for circuit simulations of memory arrays, neural networks, and in-memory computing circuits for evaluating, *e.g.*, the energy efficiency of the computation, the impact of parasitic effect (*e.g.*, the IR drop arising from the non-negligible wire resistance) and the impact of noise, drift, and variation of the memristor conductance.

V. SPICE simulation of RRAM arrays

To assess the impact of temperature and voltage dependence on the compute operations of RRAM arrays, we integrated the previously developed model into a SPICE environment as a SPICE subcomponent, as reported in Fig. 7a. Fig. 7b depicts a sample circuit that was assumed for the MVM simulation. The RRAM elements are

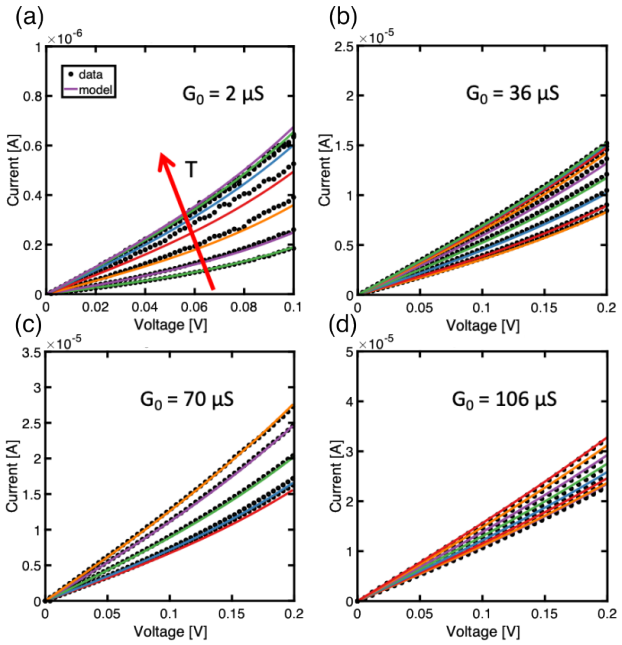

Fig. 5. Measured and calculated $I-V$ curves for increasing conductance levels and increasing temperature (from 25 to 100 °C).

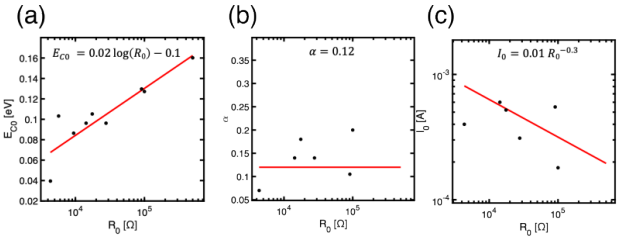

Fig. 6. Extracted parameters for the compact model of Eqs. (2-3) and their simplified state dependence according to Eqs. (4-6). (a) Parameters E_{C0}, (b) parameter α and (c) parameter I_0.

represented by the conventional memristor symbol, with their TEs connected to horizontal wordlines (WL) and BEs connected to vertical bitlines (BL). To execute fully-parallel MVM operation in the RRAM array, analog voltages are applied to the WLs while the total BL current is measured at each grounded BL [1]. Simulations were performed on a RRAM array with a size of 128x128 devices, where the conductance values were chosen randomly between 10 and 100 μS.

Fig. 8 shows the simulated current normalized to the ideal current I_0, where we assumed a linear (ohmic) $I-V$ curve and no dependence on temperature. Fig. 8a shows the current deviation as a function of temperature with a fixed input voltage of 0.25 V. The simulated current increases by a factor of 2 due to the temperature activation of PF transport. Fig. 8b shows the simulated I/I_0 as a function of the read voltage, where I_0 was evaluated at $V = 0.25$ V at room temperature. As V increases/decreases from the nominal read voltage, the error increases due to the $I-V$ curve non-linearity. Each BL displays a different nonlinearity as a result of the different compositions of MLC states.

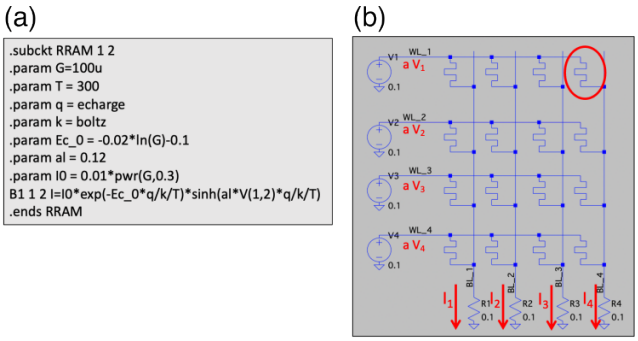

Fig. 7. (a) SPICE subcomponent definition. (b) Schematic of SPICE simulation circuit with the modeled RRAM devices (circled in red) displaced in a crosspoint structure.

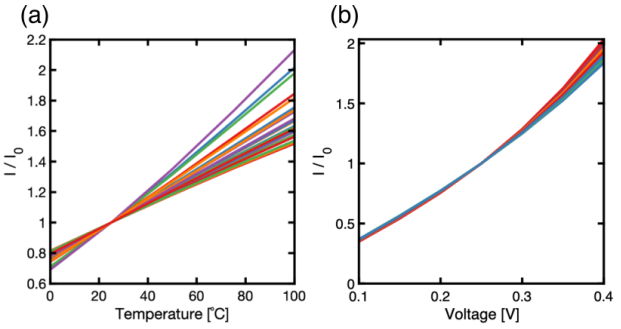

Fig. 8. Simulation results with normalized current as a function of (a) read temperature and (b) applied voltage.

VI. DNN SIMULATION

The flexibility and the compactness of the model enable its implementation also in scaled and complex scenarios such as DNN inference. As a proof of concept, we considered a small DNN with a single hidden layer trained to recognize images of handwritten digits from a cropped MNIST dataset [14]. The DNN features a total of 8000 parameters and is capable of correctly classifying the test images with an inference accuracy of 95.5 % when parameters are trained with floating-point precisions. The accuracy drops to 92 % when quantization-aware training is performed to fine-tune the network with low-resolution weights. We implemented each synaptic weight as a differential pair of two RRAM elements so that the desired target weight can be obtained as $W = G_+ - G_-$ [15]. The finite RRAM conductance levels were chosen in the range between 0 and 100 μS with 10 μS step, thus resulting in 21 programmable weights, *i.e.*, higher than 4-bit precision. After the off-line training, we assumed to map the resulting weights into the RRAM array at a reference voltage $V =$ 0.2 V and reference temperature $T = 300$ K. The simulation included a conductance variability of σ_G =3 μS and the results were extracted by averaging the final accuracy over 100 separate runs. In particular, several simulations were performed, varying the applied voltage amplitude and the operating temperature of the network. Fig. 9 shows the simulation results for the DNN inference: the accuracy suffers from degradation as the voltage and temperature deviate from the reference programming values. Overall,

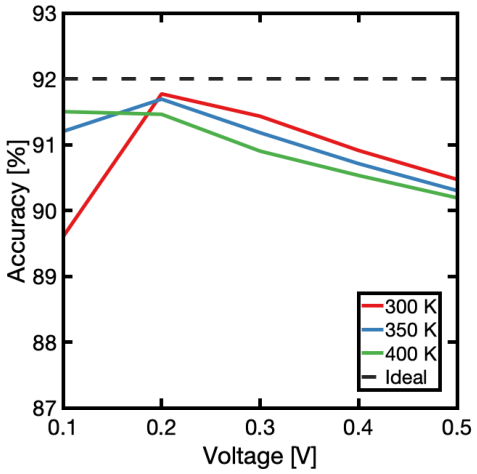

Fig. 9. Simulated DNN accuracy prediction varying voltage amplitude and operating temperature.

the compact model allows us to predict the impact of conductance variation, T- and V-dependence which is an essential tool for the DNN accelerator design and the design-technology co-optimization (DCTO) of IMC systems.

VII. CONCLUSIONS

The paper presents a compact model for RRAM devices that predicts the deviation from the linear I-V characteristic of the read conductance due to voltage and temperature changes. The model's implementation is supported by and fine-tuned over experimental measurements performed on fabricated RRAM devices. Due to its compactness and flexibility, the model allows accurate simulation and assessment of the impact of RRAM nonlinearities, even in complex computational tasks such as MVM and DNN inference. In this work we proposed a simple use-case for the model, simulating an implementation of a small neural network and predicting the classification accuracy drop caused by voltage and temperature variations, thus enabling a more accurate optimization of DNN architecture to increase real-world reliability of embedded and edge neural accelerators.

REFERENCES

[1] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," Nature Electronics 1, 333-343 (2018).
[2] N. Lepri, A. Glukhov, L. Cattaneo, M. Farronato, P. Mannocci, and D. Ielmini, "In-memory computing for machine learning and deep learning," J. Electron Device Soc. (2023).
[3] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Statistical fluctuations in HfOx resistive-switching memory (RRAM): Part I – Set/Reset variability," IEEE Trans. Electron Devices 61, 2912-2919 (2014).
[4] S. Ambrogio, S. Balatti, V. McCaffrey, D. Wang, and D. Ielmini, "Noise-induced resistance broadening in resistive switching memory (RRAM) - Part II: Array statistics," IEEE Trans. Electron Devices 62, 3812-3819 (2015).
[5] S. Ambrogio, S. Balatti, V. McCaffrey, D. Wang, and D. Ielmini, "Noise-induced resistance broadening in resistive switching memory (RRAM) - Part I: Intrinsic cell behavior," IEEE Trans. Electron Devices 62, 3805-3811 (2015).

[6] L. Zhao, S. Clima, B. Magyari-Kope, M. Jurczak, and Y. Nishi, “Ab initio modeling of oxygen-vacancy formation in doped-HfOx RRAM: Effects of oxide phases, stoichiometry, and dopant concentrations,” Appl. Phys. Lett. 107, 013504 (2015).
[7] S. Larentis, F. Nardi, S. Balatti, D. C. Gilmer and D. Ielmini, “Resistive switching by voltage-driven ion migration in bipolar RRAM – Part II: Modeling,” IEEE Trans. Electron Devices 59, 2468–2475 (2012).
[8] X. Guan, S. Yu, and H.-S. P. Wong, “On the Switching Parameter Variation of Metal-Oxide RRAM—Part I: Physical Modeling and Simulation Methodology,” IEEE Trans. Electron Devices 59, 1172-1182 (2012).
[9] S. Ambrogio, S. Balatti, D. C. Gilmer, and D. Ielmini, “Analytical modeling of oxide-based bipolar resistive memories and complementary resistive switches,” IEEE Trans. Electron Devices 61, 2378-2386 (2014).
[10] P.-Y. Chen and S. Yu, “Compact Modeling of RRAM Devices and Its Applications in 1T1R and 1S1R Array Design,” IEEE Trans. Electron Devices 62, 4022-4028 (2015).
[11] A. Glukhov et al., ”Statistical model of program/verify algorithms in resistive-switching memories for in-memory neural network accelerators,” 2022 IEEE International Reliability Physics Symposium (IRPS), pp. 3C.3-1-3C.3-7, (2022).
[12] S. Ricci, P. Mannocci, M. Farronato, S. Hashemkhani, D. Ielmini, “Forming-free resistive switching memory crosspoint arrays for in-memory machine learning,” Adv. Intell. Syst. 4, 2200053 (2022).
[13] D. Ielmini and Y. Zhang, “Analytical model for subthreshold conduction and threshold switching in chalcogenide-based memory devices,” J. Appl. Phys. 102, 054517 (2007).
[14] Y. LeCun, Y. B. L. Bottou and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, 86(11):2278-2324, (1998).
[15] V. Milo, A. Glukhov, E. Perez, C. Zambelli, N. Lepri, M. K. Mahadevaiah, E. P.-B. Quesada, P. Olivo, C. Wenger, and D. Ielmini, “Accurate program/verify schemes of resistive switching memory (RRAM) for in-memory neural network circuits,” IEEE Transactions on Electron Devices, vol. 68, no. 8, pp. 3832–3837, 2021.