Compact modeling and mitigation of parasitics in crosspoint accelerators of neural networks

N. Lepri, *Graduate Student Member, IEEE*, A. Glukhov, *Graduate Student Member, IEEE*, P. Mannocci, *Member, IEEE*, M. Porzani, *Graduate Student Member, IEEE* and D. Ielmini, *Fellow, IEEE*.

***Abstract*— In-memory computing (IMC) can accelerate data-intensive tasks, such as matrix-vector multiplication (MVM) or artificial neural networks (ANN) inference, by means of the crosspoint memory array, allowing to reduce time and energy consumption. IMC accuracy, however, is affected by nonidealities, such as variability of the conductive weights or IR drop along wires due to parasitic resistances, whose impact steeply increases with the increase of array size. This work proposes a compact model to assess the impact of nonidealities for various circuital implementations, together with architectural schemes for their mitigation based on replicated arrays. The proposed mitigation techniques allow to restore the ANN accuracy from 72.7% to 94.9%, close to the software accuracy of 96.9%, in view of an increased area and energy consumption.**

***Index Terms*— In-memory computing, deep learning, emerging memory technologies, hardware accelerator, resistive switching memory.**

I. Introduction

The spread of artificial intelligence (AI) is increasing the request for data-intensive and decentralized computing tasks, colliding with the memory bottleneck of traditional computing architectures. In-memory computing (IMC) overcomes this limitation by merging the memory and computational units, as in the case of crosspoint memory array for the acceleration of matrix-vector multiplication (MVM), sketched in Fig. 1a [1]. By applying a voltage vector V at the bitlines (rows), the resistive memory elements produce current contributions that are collected at the sourceline (column) grounds resulting in an output current vector $I = GV$, where G is the conductance matrix. Exploiting Ohm's and Kirchhoff's laws, crosspoint array is thus capable of performing a one-step *in situ* MVM. Multilevel programming of the memory devices, compatibility with the back-end-of-line (BEOL) process, and 3D stackability further boost the performances of the accelerator in terms of energy efficiency, area occupation, and integration density. Moreover, thanks to the inherent parallelism of the architecture [2], crosspoint-based MVM shows a computational complexity of $O(1)$. Several data-intensive applications can benefit from IMC approach, such as Artificial Neural Networks (ANN) inference [3], [4] and training [2], combinatorial optimization [5], and image compression [6].

This work received funding from ECSEL Joint Undertaking (JU) under grant agreement no. 101007321. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Belgium, Czech Republic, Germany, Italy, Sweden, Switzerland, Turkey. N. Lepri, A. Glukhov, P. Mannocci, M. Porzani, and D. Ielmini are with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano and IU.NET, 20133 Milan, Italy (e-mail: daniele.ielmini@polimi.it).

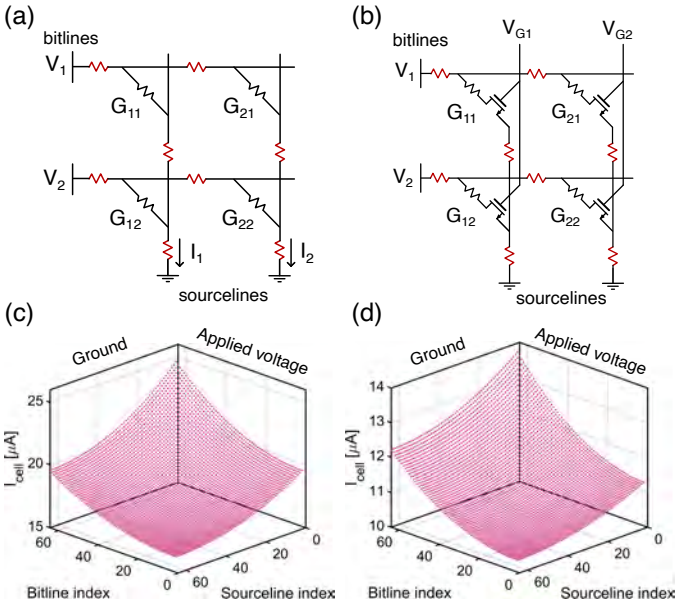

Fig. 1. Schematics of a (a) one-resistor (1R) and (b) one-transistor/one-resistor (1T1R) crosspoint array affected by IR drop due to parasitic wire resistances. Assuming a uniform voltage vector ($\boldsymbol{V_0} = \mathbf{0.2}$ V), uniform conductance pattern ($\boldsymbol{G} = \mathbf{125}$ μS) and a parasitic resistance $\boldsymbol{r} = \mathbf{1}\ \boldsymbol{\Omega}$, the device current profiles for (c) 1R and (d) 1T1R configuration result strongly non-uniform. In particular, current profile in 1T1R array is asymmetric because of the presence of the transistor, whose conductance is modified by the IR drop at the sourcelines.

Various parasitic effects limit the actual possibilities of crosspoint array. Device nonidealities, such as conductance drift [7] and variability [8], [9], can affect the MVM computation, degrading the accuracy of the implemented task. For instance, in the application of ANN inference acceleration, the combination of limited weight resolution and conductance variability due to programming error can significantly decrease the classification accuracy [9]. Furthermore, nonideal conductors give rise to IR drop along wires, which can be schematically modeled by means of lumped parasitic resistances, as shown in Fig. 1a. IR drop causes a distortion of voltages along bitlines and sourcelines, which increasingly deviate from the nominal input signals or ground, affecting the MVM computation. IR drop is particularly detrimental for IMC because of the parallel readout of numerous devices, resulting in a large amount of current flowing through the

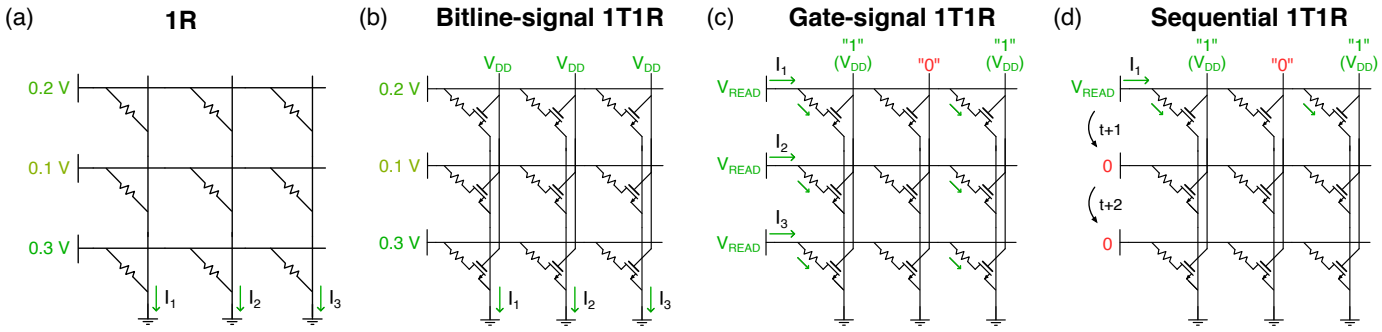

Fig. 2. Schematic diagrams depicting the working principle of the various binary MVM implementations, namely (a) 1R, (b) 1T1R with application of input signals at the bitline, (c) 1T1R with application of input signals at the transistor gates and (d) 1T1R with application of input signals at the transistor gates and sequential readout of one bitline at a time.

wires, and of the increased wire resistivity for sub-50 nm nodes, due to the increased electron surface scattering [10]. Furthermore, as will be discussed in the following, error due to IR drop scales approximately linearly with the average device conductance and quadratically with the size of the array, preventing the up-scaling of the computing cores and limiting the actual computational complexity of the architecture. To cope with parasitic IR drop, several techniques were proposed at device level [3], [4], at algorithmic level via *ad hoc* scaling vectors or activation functions [11], [12], and at training level via parasitic-aware training schemes [13], [14].

In this work we present a compact model for IR drop and variability, a closed-form relationship for minimizing the error and identifying the optimum array size, and a mitigation technique based on replicated arrays. While the replication schemes for parasitics mitigation were already proposed for purely resistive arrays [15], here we expand the scheme set and the application portfolio to more complex array configurations that involve access transistors. Finally, we demonstrate their effectiveness by simulating a crosspoint accelerator of ANN inference for MNIST image classification [16].

II. Crosspoint Array Configurations

The MVM concept can be implemented in various array configurations. In particular, Fig. 1a shows the one-resistor (1R) configuration, which allows to perform MVM while keeping a compact architecture and an integration density up to $4F^2/n$, where F is the lithographic feature size of the process and n the number of stacked layers. A significant drawback of the 1R structure is the presence of sneakpath currents in undesired cells or other leakage currents during the programming phase [17]. To overcome these programming issues, an access transistor is typically added in series to the resistive device, obtaining the one-transistor/one-resistor (1T1R) configuration, shown in Fig. 1b. This configuration allows for a finer current control, at the cost of a bulkier footprint and a more complex architecture that needs to accommodate a third terminal.

In 1T1R arrays, MVM can be carried out through two different approaches, depending on where the input signal is applied. Fig. 2b shows the first approach, which we will refer to as the bitline-signal approach, where the input voltages are applied at the bitlines and the output currents are collected at the grounded sourcelines. In this implementation, conceptually similar to 1R array MVM, the transistors are simply enabled during the computing phase. Alternatively, in the second approach depicted in Fig. 2c, which will be referred to as the gate-signal approach, the input voltages are applied at the transistor gate terminals, while supplying a fixed voltage V_{READ} and collecting the output currents at the bitlines. Because of the nonlinear trans-characteristic of the transistor, the latter approach typically involves binary conductances and gate voltages. Gate-signal implementation can also involve temporal encoding of the input signal, consisting of modulating the input gate signal in pulse width and acquiring the MVM output through analog integration. Finally, both 1R and 1T1R structures can make use of sequential readout, consisting of supplying the input voltage vector (or V_{READ}) and collecting currents one sourceline (or bitline) at a time. Fig. 2d shows the sequential approach in the case of 1T1R array. To investigate the impact of the various nonidealities on the proposed array configurations, we adopted a simulation framework based on an *ad hoc* fast numerical algorithm for the nodal analysis of crosspoint arrays affected by parasitic effects [11].

Figs. 1c-d show the current contribution I_{cell} of the memory devices in presence of IR drop for 1R and 1T1R configurations, respectively, where we assumed a N×N square crosspoint array ($N = 64$) based on HfO_2 resistive switching memory (RRAM) devices [18], a uniform voltage vector ($V_0 = 0.2$ V), a uniform conductance pattern ($G = 125$ µS), and a parasitic wire resistances of value $r = 1\ \Omega$. While the device current profile should ideally be uniform across the array, current profiles are strongly non-uniform because of IR drop, with the current error that increases as the cell position becomes farther from grounds and voltage signals. Also, in the case of 1T1R array the curve in Fig. 1d is asymmetric because IR drop at the sourcelines increases the source voltages, thus decreasing the transistor overdrive voltages and the cell currents. As a result, cells distant from the grounds and close to the input voltages experience a stronger current reduction compared to cells distant from the input voltages and close to the grounds. Nevertheless, 1T1R configuration is typically less affected from IR drop than 1R configuration, since the access transistors can reduce the average device conductance and thus IR drop.

Fig. 3 shows the average IR drop impact on the output

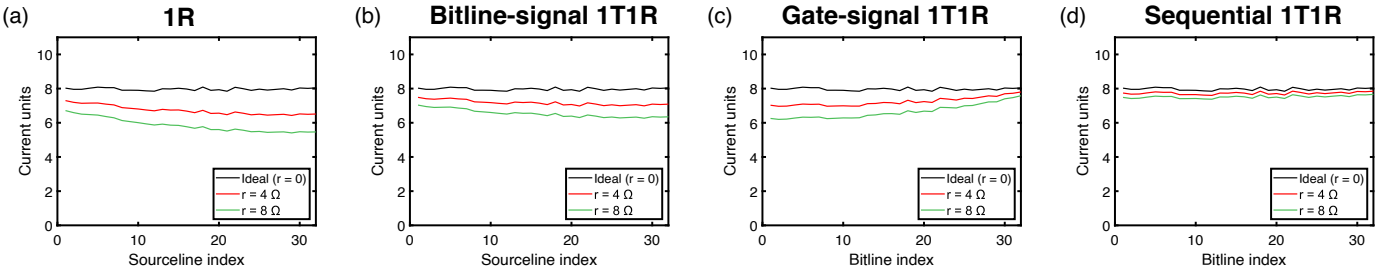

Fig. 3. Normalized average current output in presence of IR drop for various binary MVM implementations, namely (a) 1R, (b) 1T1R with application of input signals at the bitline, (c) 1T1R with application of input signals at the transistor gates and (d) 1T1R with application of input signals at the transistor gates and sequential readout of one bitline at a time. IR drop impact decreases thanks to the adoption of the transistor, which reduces the average cell conductance (b)-(c)-(d), to the gate-signal approach, where all zero-input columns do not represent possible leakage paths (c)-(d), and to the sequential readout, where IR drop along sourcelines is practically negligible (d). Notice that 1R scenario applies also for 1T1R arrays having highly conductive transistors. Generally, the nominal gate voltage influences IR drop.

current vector for various binary MVM implementations normalized in current units, where a current unit is the current flowing through an ideal cell in the low resistive state (LRS) considering the specific array configuration. Figs. 3a-b, different only in the presence of the access transistor, confirm the beneficial effect of 1T1R configuration in decreasing IR drop. Compared to bitline-signal approach, gate-signal approach reduces slightly further the error (Fig. 3c), since all array columns receiving zero as input do not contribute to possible leakage paths. When adopting sequential readout, IR drop along sourcelines is practically negligible since at maximum one current unit is flowing through each sourceline at a time, resulting in a significant reduction of the error at the cost of a strongly increased latency (Fig. 3d). In all cases, the average error increases with the parasitic wire resistance value r.

From a spatial standpoint, the error increases with increasing distance from voltage generators (Fig. 3a-b) or sourceline grounds (Fig. 3c), resulting in a different output current orientation because of the different position of the acquisition chain. The sequential approach (Fig. 3d) does not show a clear spatial trend, since IR drop is significant only along one axis.

III. Compact Modeling of IR Drop and Variability

The compact model of IR drop relies on two main assumptions that have been verified through Monte Carlo simulations. The first assumption is that, given two 1R arrays, one with a random pattern G and one with a uniform pattern with average conductance $\overline{G}$, they experience a similar average IR drop that depends neither on the input voltage vector nor on the average input voltage. Thus, regardless of the actual patterns, we can consider arrays with uniform conductance $\overline{G}$ and uniform applied voltage V_0. The second assumption is that the overall impact of IR drop is well described by the IR drop affecting a particular reference cell, located at sourceline index $\approx 0.425 \cdot N$ and bitline index $\approx 0.575 \cdot N$. Indeed, Monte Carlo simulations and approximated solutions of the differential equation system describing IR drop in the uniform scenario [11] indicate such reference cell as the one experiencing the average IR drop for a wide range of applications and array parameters. Hence, the compact model aims to assess the IR drop impact at the reference cell, which will result to be an accurate estimation of the average error across the whole array.

Afterward, we assume that in each device flows an ideal current $I = V_0 \cdot \overline{G}$ contributing to the current flowing through bitline wire resistances, which increases linearly from I to $N \cdot I$ approaching the voltage generators, as depicted in Fig. 4a. Considering a wire resistance r, the voltage at the top electrode of the reference cell is $V \simeq V_0 - Ir \cdot 0.335N^2$, where IR drop is computed as the sum of consecutive integer currents. Since the same applies to the bottom electrode, the voltage across the cell is $V_\Delta \simeq V_0 - Ir \cdot 0.67N^2$, generating a current

$$I_{IR,0} \simeq I - Ir\overline{G} \cdot 0.67N^2 \tag{1}$$

and hence an error

$$\varepsilon \simeq r\overline{G} \cdot 0.67N^2 \tag{2}$$

Error in Eq. (2), however, does not take into account the reduction of cell current I due to IR drop. To account for this second-order effect, we proceed with the second iteration of the numerical estimation, deriving how much is the IR drop-induced error when considering the cell current $I_{IR,0}$, described in Eq. (1). By solving the equation system, the error at the reference memory cell can be estimated as

$$\varepsilon_{1R} \simeq \frac{\alpha r\overline{G}N^2}{1 + \alpha r\overline{G}N^2} \tag{3}$$

where $\alpha = 0.67$.

In the general case of a rectangular array of size $N_1 \times N_2$, the model can be applied by adopting as array size N the normalized diagonal size, namely $N = \sqrt{(N_1^2 + N_2^2)/2}$, while in the case of different wire resistances between bitlines and sourcelines, it is sufficient to consider r as the weighted average resistance. Furthermore, by replacing $\overline{G}$ with an effective conductance G_{eff}, the model can also be extended to squared arrays implemented in the various 1T1R configurations. Fig. 4b shows the estimated G_{eff} for the various MVM implementations, empirically extracted as the best-fitting parameter for the simulation results. In 1T1R arrays, the presence of a transistor in series to the memory device inherently decreases the average conductance, while IR drop at sourcelines additionally reduces transistor conductance because of the reduced gate-source voltage V_{GS}. G_{eff} for the gate-signal 1T1R configuration is slightly lower than in the

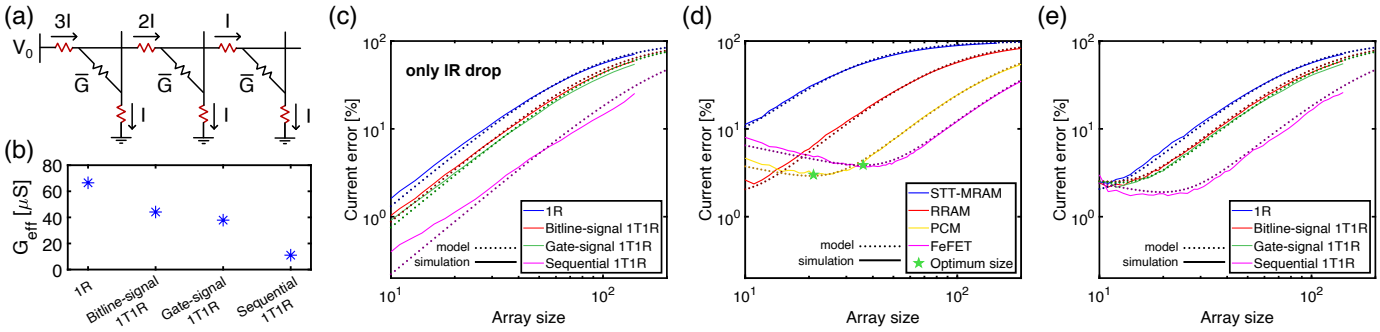

Fig. 4. (a) Approximation for IR drop modeling. (b) Empirically extracted G_{eff} for IR drop modeling, for various MVM implementations. (c) Simulation results and compact model of IR drop-induced error as a function of the array size N for various RRAM-based MVM implementations. (d) Simulation results and compact model of error due to IR drop and conductance variability in 1R arrays based on various memory technologies. By including variability, error curves become non-monotonic, identifying an optimum array size N_{OPT} that minimizes the error. (e) Simulation results and compact model in presence of IR drop and conductance variability for various RRAM-based MVM implementations.

TABLE I

CONDUCTANCE AND CONDUCTANCE VARIABILITY OF VARIOUS MEMORY DEVICES (INDICATIVE VALUES, NOT REPRESENTING THE MEDIAN).

	STT-MRAM [19]	RRAM [18]	PCM [20]	FEFET [21]
G_{LRS} (σ_{LRS})	605 µS (103 µS)	200 µS (20 µS)	30 µS (2 µS)	35 µS (2 µS)
G_{HRS} (σ_{HRS})	380 µS (30 µS)	10 µS (5 µS)	2 µS (1.6 µS)	1 µS (0.3 µS)

bitline-signal case, with a reduction that is weakly dependent on the gate signal sparsity, since all array columns receiving zero as input do not contribute to possible leakage paths. Finally, G_{eff} for the sequential MVM is significantly lower, because of the negligible IR drop along sourcelines.

Fig. 4c shows the calculated IR drop-induced current error as a function of the array size for the various MVM implementations, compared with the average simulation output. The simulation assumed random binary MVM having low resistive state at $G_{LRS} = 125$ µS and high resistive state (HRS) at $G_{HRS} = 8$ µS occurring with equal probabilities, namely with a pattern density of 50%. Activation density was set at 50%, while parasitic resistances were set at $r = 3\ \Omega$. Simulation results, reported in normalized current units to facilitate the comparison, are in good agreement with the proposed compact modeling based on Eq. (3) and G_{eff}, with a partial exception for the sequential MVM. Indeed, in sequential MVM the average sourceline resistance produces a minor error which cannot be properly included in G_{eff}, since its value quickly increases at low array sizes and then saturates.

While IR drop-induced error quickly increases when increasing the array size, error due to device or read variability decreases, because of the averaging of more conductive weights. In particular, the variability-induced error for a 1R configuration can be estimated as

$$\varepsilon_{var} = \sqrt{\frac{2}{\pi}} \cdot \frac{\overline{\sigma}}{\overline{G} \cdot \sqrt{N}} \tag{4}$$

where $\overline{\sigma}$ is root sum square of the HRS and LRS conductance variability, and the coefficient $\sqrt{2/\pi}$ stems from the computation of the error in absolute value.

Fig. 4d shows the root sum square of IR drop and variability induced errors as a function of the array size, assuming 1R configuration based on various emerging memory technologies, such as spin-transfer torque magnetic random-access memory (STT-MRAM) [19], resistive random access memory (RRAM) [18], phase change memory (PCM) [20], and ferroelectric field-effect transistor (FeFET) [21]. The devices were incorporated in the compact model by considering their binary conductance values with the corresponding standard deviations for variability, according to the values reported in Tab. I.

For small array size N, the error is dominated by variability, while IR drop becomes the most significant contribution when increasing the array size. This non-monotonic behavior allows to identify the point of minimum error in correspondence of an optimum array size N_{OPT}, which can be approximately derived as

$$N_{OPT} \simeq \sqrt[5]{\frac{\overline{\sigma}^2}{2\pi \cdot \alpha^2 \overline{G}^4 r^2}} \tag{5}$$

Fig. 4d also includes the compact model estimations, highlighting the good agreement with simulation results. Finally, Fig. 4e shows the impact of both RRAM variability and IR drop in the various RRAM-based MVM implementations. In 1T1R configurations, Eq. 4 allows us to assess the variability impact by considering the actual cell conductance, namely the series of the transistor and device conductances.

IV. REPLICATION SCHEMES

Fig. 5a shows the simulated device current profile for a 1R array with uniform voltage input and conductance pattern in presence of IR drop. As confirmed by the device current profile, impact of IR drop along the array is strongly non-uniform, which is particularly detrimental for IMC applications.

If the application of voltage signals and the acquisition of the output currents take place on the opposite array edges, the current profile results mirrored, as shown in Fig. 5b. The two current profiles can be combined by averaging the output currents of the two MVMs, to improve the uniformity of the error due to IR drop. Such approach can be implemented through the adoption of a second array with remapped conductive weights and input signals, with the additional advantage of providing

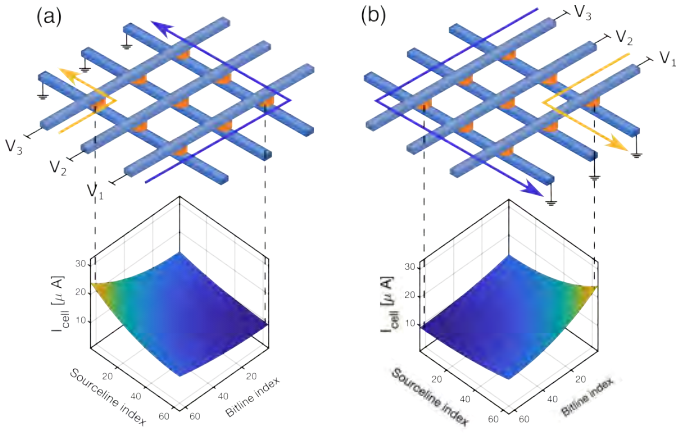

Fig. 5. (a) The device current profiles in presence of IR drop are strongly non-uniform. (b) Assuming a uniform pattern and input, if the application of voltage signals and the acquisition of currents take place on the opposite edges of the array, the current profile results mirrored. By averaging the two current outputs, one can distribute IR drop-induced error more uniformly across the devices.

a mitigation of conductance variability thanks to the weight redundancy [22].

Figs. 6a-d display the standard array (R1) for the MVM and the proposed replication schemes for 1R and 1T1R configurations. The replication schemes vary depending on the number of replication arrays to be implemented, such as 2 (R2, in Fig. 6b), 4 (R4, in Fig. 6c), or 8 (R8, in Fig. 6d). In particular, R2 scheme is the direct physical mapping of the mirroring concept proposed in Fig. 5. Fig. 6b displays its patterns, with voltage and current vectors represented in different shades of brown and green, respectively, and conductive weights represented with letters from A to P and colors ranging from red to blue. In addition to the original array (R1, in Fig. 6a), R2 involves a second array whose conductance matrix has undergone a 180° rotation and whose voltage and current vector indices have been inverted.

As mentioned, the concept can be extended to more complex replication schemes, based on more arrays, at an increased cost of area and power consumption. By comparing various mapping patterns, the best options are the ones proposed in Fig. 6c-d, based on 4 (R4) or 8 arrays (R8), respectively, featuring various 180° rotations and bitline/sourceline-wise exchange. After averaging the output current of each replicated array, equivalent normalized current profiles are obtained, as reported in Fig. 6e for the case of uniform voltage and conductance patterns. The profiles become increasingly uniform for increasing number of replicated arrays. This means that we are increasingly approximating the ideal MVM current profile, which in this uniform scenario would be flat, except for a multiplicative factor. To validate the concept, Fig. 6f illustrates the resulting output current vector for the various proposed schemes, highlighting an increasing similarity to the ideal output, except for a scalar factor. Furthermore, Fig. 6g shows the cumulative distributions of output current for the various replication schemes compared to the ideal case, confirming that weight redundancy inherently mitigates error due to variability.

While so far we have considered 1R array, the proposed techniques apply also for bitline-signal 1T1R and gate-signal 1T1R configurations. *Ad hoc* mapping strategy can be adopted when the transistor conductance is kept particularly low and thus the current profiles result exceptionally asymmetric, while in most cases the proposed schemes result to be the most beneficial ones. Instead, gate-signal 1T1R configuration with sequential readout shows a drastically different current profile, since IR drop along the sourcelines is practically negligible. This characteristic simplifies the mapping strategies of the replication schemes, that can avoid bitline-wise exchange and thus the inversion of the voltage vector. The resulting best-performing schemes are sketched in Fig. 7a-d for the case of 2-array (R2), 4-array (R4), and 8-array (R8) systems, respectively. Fig. 7e shows the equivalent normalized current profiles, confirming the trend already seen for 1R configuration. The curves are increasingly uniform for increasing number of replicated arrays, thus more similar to the ideal MVM except for a multiplicative factor.

V. Application in Neural Network Inference

Acceleration of ANN inference is a typical application of crosspoint-based IMC, since the workload is predominantly composed of MVM with stationary weights, thus with no need to reprogram the device conductances. However, parasitic effects have a detrimental impact on the accuracy, forcing to use small computing tiles in order to maintain low IR drop and consequently requiring an increased peripheral circuitry. To validate the proposed replication schemes for ANN inference acceleration, we trained a 2-layer fully-connected neural network (FCNN) for image classification of the MNIST dataset [16], adopting the Adamax optimizer [23] to minimize the cross-entropy loss function along 50 epochs. To explore application cases more complex than binarized neural networks, the quantization was set to 4 bits for the synaptic weights and 6 bits for the activations. The IMC accelerator is sketched in Fig. 8a, based on 64×64 1R crosspoint arrays implemented in a differential architecture, enabling the mapping of negative weights through the equivalent weight $W = G_+ - G_-$. To this aim, two sets of arrays were adopted to store positive (G_+) and negative (G_-) conductance values. The peripheral circuitry consists of transimpedance amplifier (TIA), 6-bit analog-to-digital converters (ADC) with fixed full-scale range, 6-bit digital-to-analog converters (DAC), and a digital signal processor (DSP) to perform subtractions and activation functions.

To validate the replication schemes, we simulated the accelerator assuming a wire resistance $r = 8\ \Omega$ and a conductance variability $\sigma_G = 20$ μS. While RRAM variability is typically in the range of 3-10 μS [18], we deliberately overstated its value to better comprehend its impact on classification accuracy. Fig. 8b shows the simulated accuracy for the replication schemes compared to the standard case (R1), with and without a gain calibration of the TIAs. Thanks to a more uniform distribution of IR drop and to information redundancy, replication schemes manage to recover the accuracy loss due to parasitics from 72.7% to 91.8% (R4) and 92.6% (R8), despite a significant scalar factor loss in the MVM computation, as

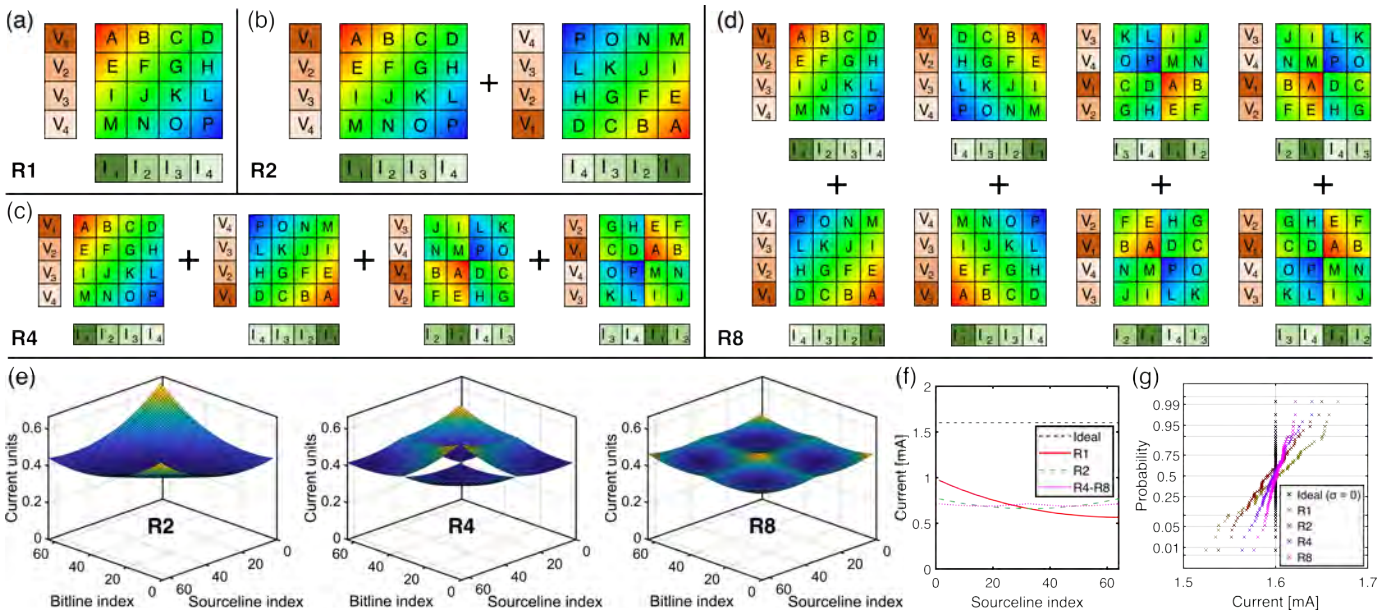

Fig. 6. Graphical sketches of the replication schemes for 1R and 1T1R configurations operated with parallel readout, namely when IR drop affects both sourcelines and bitlines. (a) Sketches show the standard MVM based on a single array (R1), (b) the R2 scheme, where the second array features a 180° rotation of the conductance matrix and an inversion of voltage and current vectors, (c) the R4 scheme and (d) the R8 scheme, based on 4 and 8 arrays, respectively, featuring various 180° rotations and bitline/sourceline-wise exchange. Assuming 1R arrays having $r = 4\,\Omega$ (e) the normalized average current profiles and (f) the output current vectors for the proposed schemes become increasingly uniform when increasing the number of arrays. Therefore, the operation is increasingly similar to the ideal MVM except for a multiplicative factor. (g) Cumulative distribution of the output current vectors, whose variability is strongly reduced when implementing the replication schemes.

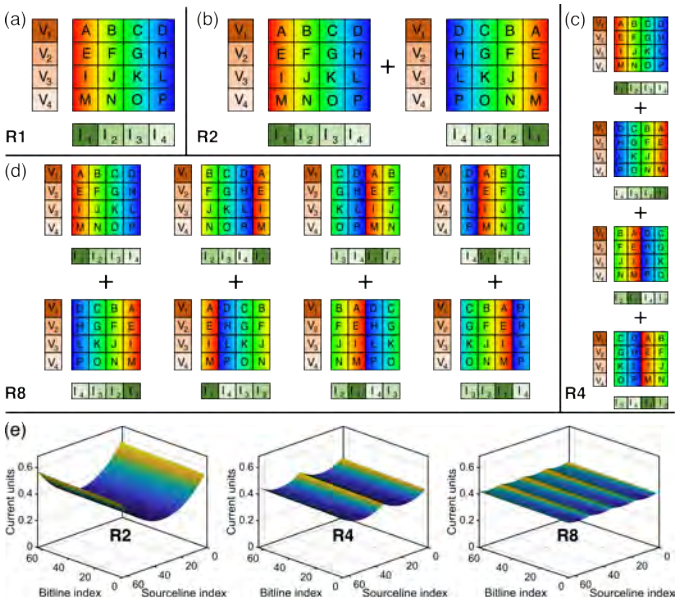

Fig. 7. Graphical sketches of the replication schemes for 1T1R configuration with sequential readout, namely when IR drop mainly affects the bitlines or, more generally, the direction of current accumulation. The proposed schemes are based on (b) 2, (c) 4, and (d) 8 arrays, featuring various sourceline-wise exchange. They are referred to as R2, R4, and R8 scheme, respectively. (e) Normalized average current profiles for the proposed schemes, becoming increasingly uniform when increasing the number of replicated arrays. r was set to $20\,\Omega$, higher than in Fig. 6, for a better visualization.

seen in Fig. 8b. As a further improvement, TIA gain can be calibrated to recover the scalar factor loss, obtaining accuracy up to 94.7% (R4) and 94.9% (R8), close to the ideal accuracy of 96.9%. TIA gain calibration was computed offline with a separate validation set, considering one fixed gain for all TIAs of a neuron layer.

Notice that the proposed replication schemes are agnostic on the device, on the network topology and training, and they can be implemented directly at the inference stage, except for the gain calibration procedure. Furthermore, since the main portion of energy and area budget is typically consumed by peripheral circuitry and thanks to the possibility of sharing common TIAs, ADCs, or DACs among the replicated arrays, energy and area consumption is expected to increase less than linearly with the number of arrays. The actual energy and area overheads due to replication schemes, however, depend on the layout design constraints, that can allow or not to share common peripheral circuits in the various schemes. As an example, the R2 scheme can likely be implemented by sharing readout periphery and summing the two analog signals before a single ADC conversion, while the R8 scheme will surely require additional circuitry. Alternatively, in the case of generic in-memory ANN accelerators, our proposal can leverage the presence of multiplexing stages between the array output and the readout stages, usually fewer in number, not reducing the energy consumption but limiting the area increase. With the proposed replication schemes, the user can choose to trade off performances, in terms of increased area consumption, increased energy consumption, and decreased throughput, with an increase in the network accuracy. Power consumption is maintained. In this scenario, the compensation can be flexibly adopted *a posteriori* in case the network accuracy is not sufficiently high.

Simulation results support the efficacy of replication

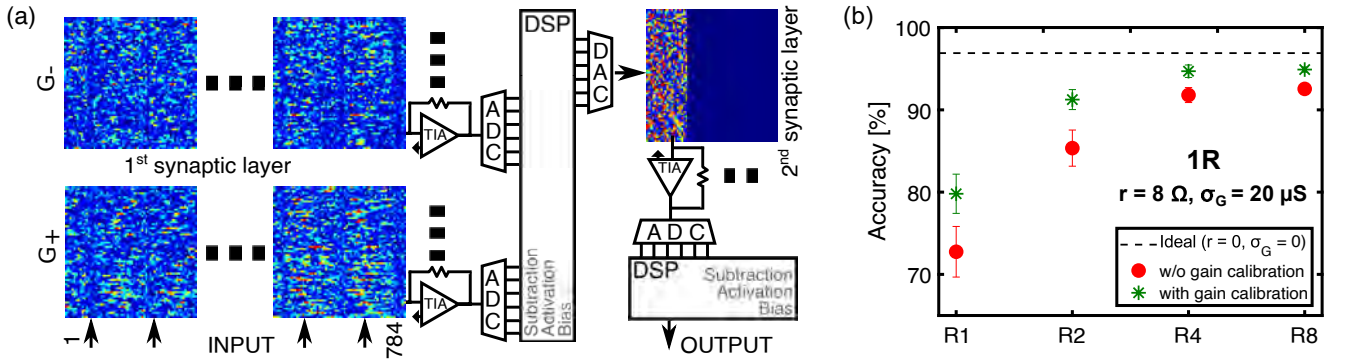

Fig. 8. Replication schemes were validated in the application of FCNN inference for MNIST classification, (a) simulating a crosspoint-based accelerator. (b) Classification accuracy in presence of IR drop and conductance variability implementing the replication schemes without or with a gain calibration of the first layer TIAs. The same gain is adopted for all the TIAs.

schemes for IR drop and variability mitigation in crosspoint accelerator of ANN inference, obtaining significant increase in classification accuracy, close to the full recovery of software accuracy.

VI. Conclusions

We presented a compact modeling framework for assessing the impact of IR drop and conductance variability effects, applicable to various MVM implementations and to different memory technologies. Also, we presented architectural schemes for nonidealities mitigation based on replicated and remapped arrays, which benefit from a more uniform distribution of IR drop and from the redundancy of conductive weights. In the scenario of ANN inference acceleration, the proposed replication schemes achieve a significant increase of the classification accuracy, in front of an increased area and energy consumption.

References

[1] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nature Electron.*, vol. 1, no. 6, pp. 333–343, Jun. 2018. doi: 10.1038/s41928-018-0092-2

[2] T. Gokmen and Y. Vlasov, "Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices: Design Considerations," *Frontiers in Neuroscience*, vol. 10, 2016. doi: 10.3389/fnins.2016.00333

[3] Q. Liu *et al.*, "33.2 A Fully Integrated Analog ReRAM Based 78.4TOPS/W Compute-In-Memory Chip with Fully Parallel MAC Computing," in *ISSCC*, Feb. 2020, pp. 500–502. doi: 10.1109/ISSCC19947.2020.9062953

[4] S. Cosemans *et al.*, "Towards 10000TOPS/W DNN Inference with Analog in-Memory Computing – A Circuit Blueprint, Device Options and Requirements," in *IEDM Tech. Dig.*, Dec. 2019, pp. 22.2.1–22.2.4. doi: 10.1109/IEDM19573.2019.8993599

[5] M. R. Mahmoodi *et al.*, "An Analog Neuro-Optimizer with Adaptable Annealing Based on 64×64 0T1R Crossbar Circuit," in *IEDM Tech. Dig.*, Dec. 2019, pp. 14.7.1–14.7.4. doi: 10.1109/IEDM19573.2019.8993442

[6] C. Li *et al.*, "Analogue signal and image processing with large memristor crossbars," *Nature Electron.*, vol. 1, no. 1, pp. 52–59, Jan. 2018. doi: 10.1038/s41928-017-0002-z

[7] D. Ielmini, S. Lavizzari, D. Sharma, and A. L. Lacaita, "Physical interpretation, modeling and impact on phase change memory (PCM) reliability of resistance drift due to chalcogenide structural relaxation," in *IEDM Tech. Dig.*, Dec. 2007, pp. 939–942. doi: 10.1109/IEDM.2007.4419107

[8] E. Pérez *et al.*, "Analysis of the statistics of device-to-device and cycle-to-cycle variability in TiN/Ti/Al:HfO2/TiN RRAMs," *Microel. Eng.*, vol. 214, pp. 104–109, Jun. 2019. doi: 10.1016/j.mee.2019.05.004

[9] T.-H. Kim *et al.*, "Effect of Program Error in Memristive Neural Network With Weight Quantization," *IEEE Transactions on Electron Devices*, vol. 69, no. 6, pp. 3151–3157, Jun. 2022, conference Name: IEEE Transactions on Electron Devices. doi: 10.1109/TED.2022.3169112

[10] S. M. Rossnagel and T. S. Kuan, "Alteration of Cu conductivity in the size effect regime," *J. of Vacuum Science & Tech. B*, vol. 22, no. 1, pp. 240–247, Jan. 2004. doi: 10.1116/1.1642639

[11] N. Lepri, M. Baldo, P. Mannocci, A. Glukhov, V. Milo, and D. Ielmini, "Modeling and Compensation of IR Drop in Crosspoint Accelerators of Neural Networks," *IEEE Trans. Electron Devices*, vol. 69, no. 3, pp. 1575–1581, Mar. 2022. doi: 10.1109/TED.2022.3141987

[12] D. Song *et al.*, "Mitigate IR-Drop Effect by Modulating Neuron Activation Functions for Implementing Neural Networks on Memristor Crossbar Arrays," *IEEE Electron Device Letters*, vol. 44, no. 8, pp. 1280–1283, Aug. 2023, conference Name: IEEE Electron Device Letters. doi: 10.1109/LED.2023.3285916

[13] D. Joksas *et al.*, "Nonideality-Aware Training for Accurate and Robust Low-Power Memristive Neural Networks," *Adv. Sci.*, vol. 9, no. 17, p. 2105784, 2022. doi: 10.1002/advs.202105784

[14] T. Cao, C. Liu, Y. Gao, and W. L. Goh, "Parasitic-Aware Modeling and Neural Network Training Scheme for Energy-Efficient Processing-in-Memory With Resistive Crossbar Array," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 12, no. 2, pp. 436–444, Jun. 2022, conference Name: IEEE Journal on Emerging and Selected Topics in Circuits and Systems. doi: 10.1109/JETCAS.2022.3172170

[15] N. Lepri, A. Glukhov, and D. Ielmini, "Mitigating read-program variation and IR drop by circuit architecture in RRAM-based neural network accelerators," in *IEEE IRPS*, Mar. 2022, pp. 3C.2–1–3C.2–6, iSSN: 1938-1891. doi: 10.1109/IRPS48227.2022.9764486

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998. doi: 10.1109/5.726791

[17] A. Flocke and T. G. Noll, "Fundamental analysis of resistive nano-crossbars for the use in hybrid Nano/CMOS-memory," in *ESSCIRC*, Sep. 2007, pp. 328–331. doi: 10.1109/ESSCIRC.2007.4430310

[18] V. Milo *et al.*, "Accurate Program/Verify Schemes of Resistive Switching Memory (RRAM) for In-Memory Neural Network Circuits," *IEEE Trans. Electron Devices*, vol. 68, no. 8, pp. 3832–3837, Aug. 2021. doi: 10.1109/TED.2021.3089995

[19] C.-C. Chang *et al.*, "NV-BNN: An Accurate Deep Convolutional Neural Network Based on Binary STT-MRAM for Adaptive AI Edge," in *Proc. DAC*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–6. doi: 10.1145/3316781.3317872

[20] M. Le Gallo, A. Sebastian, G. Cherubini, H. Giefers, and E. Eleftheriou, "Compressed Sensing With Approximate Message Passing Using In-Memory Computing," *IEEE Trans. Electron Devices*, vol. 65, no. 10, pp. 4304–4312, Oct. 2018. doi: 10.1109/TED.2018.2865352

[21] S. Dutta, C. Schafer, J. Gomez, K. Ni, S. Joshi, and S. Datta, "Supervised Learning in All FeFET-Based Spiking Neural Network: Opportunities and Challenges," *Frontiers in Neuroscience*, vol. 14, 2020. doi: 10.3389/fnins.2020.00634

[22] G. Pedretti, P. Mannocci, C. Li, Z. Sun, J. P. Strachan, and D. Ielmini, "Redundancy and Analog Slicing for Precise In-Memory Machine Learning—Part I: Programming Techniques," *IEEE Trans. Electron Devices*, vol. 68, no. 9, pp. 4373–4378, Sep. 2021. doi: 10.1109/TED.2021.3095433

[23] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Jan. 2017, arXiv:1412.6980 [cs].