

Analysis and Mitigation of IR Drop in Memristor-based AI Hardware Accelerators

Emmanouil Arapidis Theofilos Spyrou Konstantinos Stavrakakis
Emmanouil Anastasios Serlis Moritz Fieback Said Hamdioui Anteneh Gebregiorgis
Computer Engineering Lab, Delft University of Technology, Delft, The Netherlands

Abstract—Although offering great potential for energy-efficient edge-AI, memristor-based CIM accelerators are severely hindered by IR drop induced errors. To tackle this, we propose a low-cost mitigation technique by first quantifying the impact of IR drop on the accuracy. Then, a mitigation strategy is developed to compensate for IR drop-induced inference accuracy reduction by combining an optimized mapping scheme with a fine-tuned calibration of the ADC. Results show the proposed solution can effectively mitigate IR drop with a negligible overhead.

Index Terms—IR Drop, Memristors, CIM, Edge AI, Reliability.

I. INTRODUCTION

Computing-In-Memory (CIM) accelerators implemented with emerging memristive devices, such as Resistive RAM (RRAM), offer high density, non-volatility, and low power consumption [1]. However, various *non-ideality* issues affect the computational accuracy of CIM and limit its widespread adoption [2]–[4]. A major non-ideality is IR drop, i.e. the voltage drop across the wordlines and bitlines of a CIM crossbar array due to parasitic resistances, leading to an erroneous output [5]. Existing mitigation strategies include algorithmic solutions [6]–[8] that model IR drop and inject it through a retraining phase, and hardware-level solutions [9], [10] exploring different cell structures or employing various flavors of redundancy schemes to deal with IR drop. However, these solutions are often implementation dependent and incur area, and power overheads.

To address this issue, we propose an accurate, and low-cost IR drop mitigation technique for memristor-based AI hardware accelerators. First, perform a SPICE-level analysis to quantify the impact of IR drop. Then, a mitigation strategy is proposed to compensate for IR drop-induced inference accuracy reduction. The strategy combines an optimized mapping scheme with a fine-tuned calibration of the current-sensing periphery circuitry to deal with output discrepancies caused by IR drop. Simulation results show that the proposed solution can effectively mitigate IR drop and restore accuracy with a negligible overhead.

II. IR DROP IMPACT ANALYSIS

To demonstrate the impact of IR drop we performed SPICE Monte Carlo simulations on a 32×32 1T1R crossbar array using the JART VCM v1b memristive device [11]. The parasitic resistance was set to 1Ω , consistent with typical CIM implementations [12]–[14]. The impact of IR drop varies based

This work was funded by the Dutch Organization for Scientific Research (NWO) under grant agreement No KICH1.ST04.22.021 for the project Self-Healing Neuromorphic Systems (SNS).

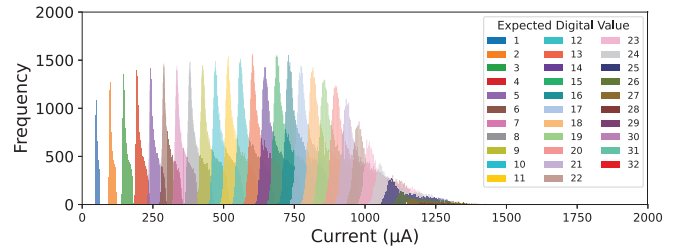


Fig. 1: Distributions of output MAC currents.

on the amount of current flowing through the crossbar which is dependent on the percentage of cells in LRS, i.e., P_{LRS} , and the percentage of simultaneously active rows, i.e., P_{AR} . As illustrated in Fig. 1(a), each distribution I_{MAC}^D represents all measured MAC currents that, ideally, should be converted to the same digital value D , i.e., all the cumulative currents produced by D active LRS cells (regardless of their column position). Due to IR Drop, as the targeted digital value D increases (higher P_{LRS} and/or P_{AR}), the distributions I_{MAC}^D widen and scale non-linearly to the number of active LRS cells. As a result, distributions start to overlap for $D \geq 5$ without any clear distinction, leading to erroneous digital conversions.

III. IR DROP MITIGATION STRATEGY

The values of P_{LRS} and P_{AR} vary from crossbar to crossbar, making it difficult to compensate for IR drop. Balanced mapping aims to balance the P_{LRS} and P_{AR} distributions, so the effect of IR drop is better anticipated. The uniformity is leveraged to calibrate the ADCs and mitigate IR drop effects. Fig. 2(a) presents the overview of the proposed IR drop mitigation strategy.

A. Balanced mapping scheme

The goal of balanced mapping is to homogenize P_{LRS} and P_{AR} values across all the C crossbars ($M \times N$) required to map a layer, where M and N is the number of rows and of columns, respectively. To quantify this, the load of a crossbar is defined as $L = \sum_{i=1}^M L^i$, where $L^i = P_{LRS}^i \cdot P_{AR}^i$ is the load of the i^{th} row, P_{LRS}^i is the percentage of cells in LRS at row i , and P_{AR}^i is the probability that row i will be activated calculated throughout a complete forward pass of the DNN. For an optimal mapping, the goal is to make the loads of all crossbars equal or identical to each other.

Firstly, the weights are grouped in vectors of size equal to N bits. Then, the row load L^i is calculated for every vector, and the vectors are sorted in descending order. In iterative rounds,

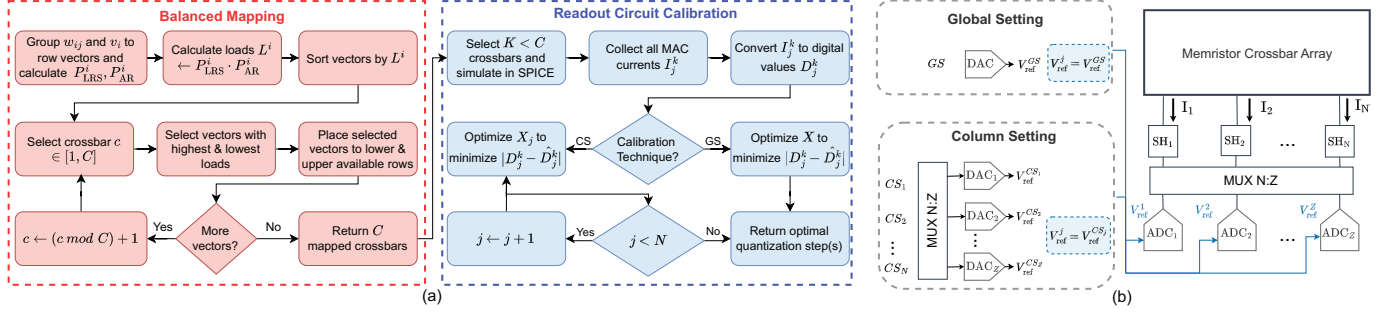


Fig. 2: (a) Proposed IR drop mitigation strategy: balanced mapping scheme (red) and readout calibration techniques (blue); (b) circuit implementation of GS and CS calibration techniques.

the vectors with the highest and lowest loads are then extracted and placed in the lower and upper available rows of the next c crossbar, respectively. Placing the high-load vectors closer to the readout circuitry minimizes the impact of IR drop. In the case where crossbars contain spare columns, the checkerboard pattern is applied to fill cells with LRS and HRS alternately, ensuring that the load of those crossbars will not deviate.

B. Readout circuitry calibration

Re-calibration of the ADC-based readout circuitry is critical in order to tackle digital conversion errors. To do so, the reference voltage V_{ref} of an n -bit ADC is adjusted in a way that its quantization step $X = V_{ref}/2^n$ matches closer the length of the output MAC current distributions. Finding the optimal quantization step X , first, requires the profiling of the DNN workload and a subset of K out of all mapped crossbars is selected to be simulated in SPICE. The resulting output currents I_j^k for each column j and each crossbar k are converted to the corresponding digital values following the ADC operation described in Eq. 1:

$$D_j^k = \left\lfloor \frac{I_j^k + 0.5 \cdot X}{X} \right\rfloor, \quad \forall j \in [1, N], k \in [1, K] \quad (1)$$

The optimal X is the one that minimizes the absolute difference $|D_j^k - \hat{D}_j^k|$ between the actual D_j^k and the expected ideal \hat{D}_j^k digital value. Once X is determined, a DAC-based control circuit generates the optimal V_{ref} and feeds it to the ADCs. The DAC is configured by a register that can be programmed externally. Two calibration techniques are proposed based on how X is applied to the ADCs:

- **Global Setting (GS):** A single X is shared by all ADCs. For calculating the global X the results from all crossbars K and all columns N are considered. One DAC and register are needed to drive the common voltage V_{ref}^{GS} .
- **Column Setting (CS):** A separate X_j is used to account for different IR drop impact on each column j . To calculate X_j , only the currents that correspond to column j are used. Each ADC is tuned with $V_{ref}^{CS_j}$ driven by a dedicated DAC and register.

The circuit implementation of GS and CS is shown in Fig 2(b). The effectiveness of the scheme is demonstrated in Fig 3(a)-(c) where the distribution I_{MAC}^6 is mapped against the ADC ladder. Both CS and GS achieve 100% and 80% accuracy, respectively, while the accuracy of the uncalibrated ADC is 11%.

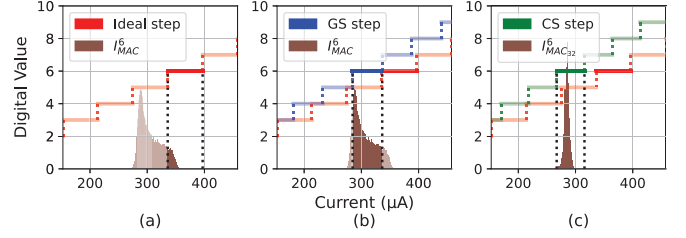


Fig. 3: (a) Uncalibrated (b) Global- (c) Column-Setting calibration techniques.

TABLE I: Drop in classification accuracy.

Calibration Technique	Mapping Scheme	Accuracy Drop (%)		
		MLP	LeNet-5	ECG
Global Setting ^{6b}	Unbalanced	22.21	13.33	15.62
	Balanced	7.56	9.42	13.12
Column Setting ^{6c}	Unbalanced	0.98	0.45	0.63
	Balanced	0.04	0.01	0.01

IV. RESULTS

A Python-SPICE framework was designed that implements the balanced mapping algorithm in Python and simulates crossbar operation in SPICE. Results are evaluated across BNN implementations of Multi-Layer Perceptron (MLP), LeNet-5 and a six-layer convolutional network on the MNIST [15] and MIT-BIH dataset [16], with ideal accuracy 95.53%, 96.72% and 89.53%, respectively. Experiments applying different calibration and mapping schemes are summarized in Table I. When only GS is applied, the classification accuracy drop is the worst; however, adding the balanced mapping scheme improves results. The CS technique improves significantly with $< 1\%$ drop in accuracy, while combining CS and balanced mapping achieves $\leq 0.04\%$ loss, restoring accuracy close to ideal IR drop-free levels. While CS is designed for accurate computation, GS offers a more lightweight approach with a tolerable cost in accuracy.

V. CONCLUSIONS

IR drop can compromise the computation accuracy in CIM-based analog designs. This work analyzed the impact of IR drop and proposed a mitigation strategy composed of a balanced mapping scheme and a calibration technique of the readout circuitry, to reclaim ideal IR drop-free accuracy with an error $\leq 0.04\%$ by introducing a minimal hardware overhead.

REFERENCES

- [1] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nature Electronics*, vol. 1, no. 6, pp. 333–343, Jun. 2018.
- [2] A. Singh, S. Diware, A. Gebregiorgis, R. Bishnoi, F. Catthoor, R. V. Joshi, and S. Hamdioui, "Low-Power Memristor-Based Computing for Edge-AI Applications," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, Feb. 2021, pp. 1–5.
- [3] A. Gebregiorgis, A. Singh, S. Diware, R. Bishnoi, and S. Hamdioui, "Dealing with Non-Idealities in Memristor Based Computation-In-Memory Designs," in *2022 IFIP/IEEE 30th International Conference on Very Large Scale Integration (VLSI-SoC)*, Jul. 2022, pp. 1–6.
- [4] T. Spyrou, H.-G. Stratigopoulos, I. Alouani, S. Hamdioui, and A. Gebregiorgis, "On the Trustworthiness of Spiking Neural Networks and Neuromorphic Systems," in *2025 IEEE European Test Symposium (ETS)*, Feb. 2025, pp. 1–10.
- [5] Y. Jeong, M. A. Zidan, and W. D. Lu, "Parasitic Effect Analysis in Memristor-Array-Based Neuromorphic Systems," *IEEE Transactions on Nanotechnology*, vol. 17, no. 1, pp. 184–193, Jan. 2018.
- [6] Z. He, J. Lin, R. Ewertz, J.-S. Yuan, and D. Fan, "Noise Injection Adaption: End-to-End ReRAM Crossbar Non-ideal Effect Adaption for Neural Network Mapping," in *Proceedings of the 56th Annual Design Automation Conference 2019*, Mar. 2019, pp. 1–6.
- [7] M. E. Fouda, S. Lee, J. Lee, G. H. Kim, F. Kurdahi, and A. M. Eltawi, "IR-QNN Framework: An IR Drop-Aware Offline Training of Quantized Crossbar Arrays," *IEEE Access*, vol. 8, pp. 228 392–228 408, 2020.
- [8] B. Liu, H. Li, Y. Chen, X. Li, T. Huang, Q. Wu, and M. Barnell, "Reduction and IR-drop compensations techniques for reliable neuromorphic computing systems," in *2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Aug. 2014, pp. 63–70.
- [9] 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 *2020 IEEE International Solid-State Circuits Conference - (ISSCC)*, Oct. 2020, pp. 500–502.
- [10] C. Huang, N. Xu, K. Qiu, Y. Zhu, D. Ma, and L. Fang, "Efficient and Optimized Methods for Alleviating the Impacts of IR-Drop and Fault in RRAM Based Neural Computing Systems," *IEEE Journal of the Electron Devices Society*, vol. 9, pp. 645–652, 2021.
- [11] F. Cüppers *et al.*, "Exploiting the switching dynamics of HfO₂-based ReRAM devices for reliable analog memristive behavior," *APL Materials*, vol. 7, no. 9, p. 091105, Sep. 2019.
- [12] G. Pedretti and D. Ielmini, "In-Memory Computing with Resistive Memory Circuits: Status and Outlook," *Electronics*, vol. 10, no. 9, p. 1063, Jan. 2021.
- [13] Y. Liao *et al.*, "A Compact Model of Analog RRAM With Device and Array Nonideal Effects for Neuromorphic Systems," *IEEE Transactions on Electron Devices*, vol. 67, no. 4, pp. 1593–1599, Apr. 2020.
- [14] C.-W. S. Yeh and S. S. Wong, "Compact One-Transistor-N-RRAM Array Architecture for Advanced CMOS Technology," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 5, pp. 1299–1309, Feb. 2015.
- [15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [16] N. Pu, Z. Wu, A. Wang, H. Sun, Z. Liu, and H. Liu, "Arrhythmia Classifier Based on Ultra-Lightweight Binary Neural Network," in *2023 15th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, 2023, pp. 1–7.