28M Weights×TOPs/W/mm² PCM-Based Analog in-Memory Computing Core with 8 512×512-Weight Layers in 28nm FD-SOI CMOS

M. Pasotti*, R. Zurla†, J.J. Bertolini Agnoletto*, E. Calvetti*, A. Antolini‡, L. Croce*, G. Desoli*, L. Iannelli§, A. Lico‡, R. Vignali§, F. Zavalloni‡, E. Franchi Scarselli‡, A. Cabrini§

* STMicroelectronics, Agrate Brianza, Italy; † STMicroelectronics, Pavia, Italy;

‡ ARCES-DEI, University of Bologna, Italy; § Dept. of Electrical, Computer and Biomedical Eng., University of Pavia, Italy.

***Abstract*—In-Memory Computing (IMC) hardware accelerators for Deep Neural Networks (DNNs) require massive quantity of coefficients stored in a single device to limit performance losses in multicore clusters. This aspect, generally neglected by prevalent Figure of Merits (FoM), can be addressed by Phase-Change Memory (PCM) technology thanks to its high density, scalability, non-volatility, and analog storage capability. This paper presents a PCM-based (Ge-rich GST) Analog-IMC macro designed for multilayer, drift-tolerant, and temperature-resilient computations. Fabricated in a 28nm FD-SOI CMOS process with 4M-cells array, the accelerator achieves a $<$ 2.14% error for Matrix-Vector-Multiplication (MVM) across a wide temperature range (from -40°C to 125°C), offering an improvement, with respect to other state-of-the-art hardware accelerators, by a factor 3.5 estimated by means of a FoM defined as No. of Weights×TOPS/W/mm² (or storage-energy efficiency per area).**

***Index Terms*—Analog in-Memory Computing (AiMC), Phase-Change Memory (PCM), drift and temperature compensation.**

I. Introduction

DATA transfers are the most power-demanding operations in multicore clusters. For this reason, it is fundamental to maximize, in each core, density and size of the memory used to accommodate the large amount of parameters required for Deep Neural Network (DNN) [1]. When considering DNNs implemented in Analog in-Memory Computing (AiMC) cores based on Phase-Change Memory (PCM), the Matrix-Vector-Multiplication (MVM) accuracy is primarily impacted by two factors: voltage drops across the array (known as IR drop); and variations in matrix coefficients due to the physical properties of the memory cells [2]. Due to IR drop, the effective voltage applied to a memory cell changes depending on its position inside the array thus leading to cell current variations which limit the precision of the MVM operation. Analogously, variations of matrix coefficients, determined by conductance drift over time and/or temperature-induced fluctuations, affect programmed conductances thus degrading MVM reliability and accuracy.

Recent hardware solutions to enhance resilience to conductance drift include Reference-Cell Conductance Tracking (RCCT) [3] and differential weight architectures [4]. RCCT uses reference cells to track and compensate for conductance drift, ensuring more stable computations, but sacrifices constant latency and energy efficiency. Differential weight architectures use pairs of PCM cells to represent weights differentially, mitigating conductance drift impact but requiring a narrow, high-conductance range for matrix coefficients, limiting system-level efficiency.

Post-processing strategies, such as those in [5], focus on models to capture temperature impacts on drift and conductance. These strategies show that simple array-level scaling can correct conductance shifts due to temperature and drift, preventing accuracy drops during inference in neural networks. However, these strategies are constrained by operating temperature and requires recurrent evaluation of the rescaling factor, making it unsuitable for environments with significant temperature fluctuations. Relying solely on rescaling reduces the effective ADC dynamic range during computation.

This work introduces several solutions to address these challenges. It allows 512×512 signed MVMs to be executed across 8 distinct coefficient layers managed through column decoding, significantly increasing weight density to approximately 2 million weights on a single core. A Bit-Line Biasing Circuit (BL-BC) compensates for resistive drop caused by layer decoding, improving precision by 20% as compared to conventional methods.

To minimize coefficient variations, the computing architecture employs current-biased RCCT with matched computing elements for temperature compensation and Output Digital Rescaling (ODR) to recover drift-induced errors. This ensures constant computation time, compensates for frequency variations, and is effective over a wide range of conductance values, providing higher energy efficiency. Results show consistent core performance across a wider temperature range compared to state-of-the-art solutions.

II. AiMC architecture and MVM computation

The AiMC core (Fig. 1 and Fig. 2) incorporates 4 PCM Reference Local Arrays (RLAs) for RCCT and binary storage, along with 16 PCM Computational Local Arrays (CLAs), each consisting of 256 rows and 1024 columns. At the bottom of the arrays, 32 Digital Sense Amplifiers (DSAs) are employed

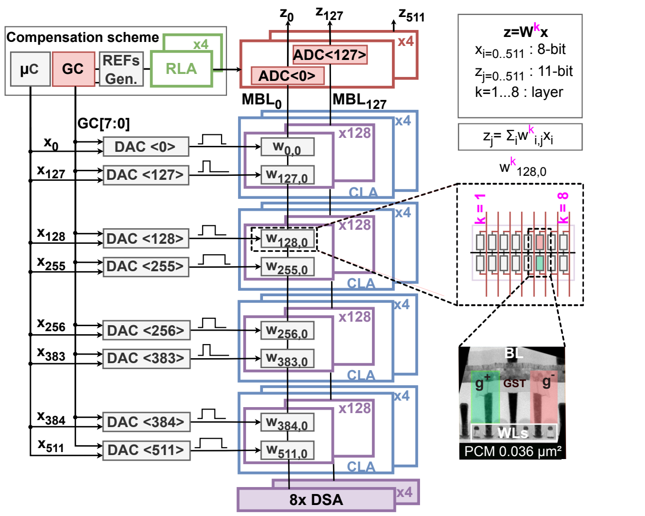

Fig. 1. Block diagram of the AiMC core and, bottom-right, TEM microphotograph of "Wall" PCM 2T2R cell [6].

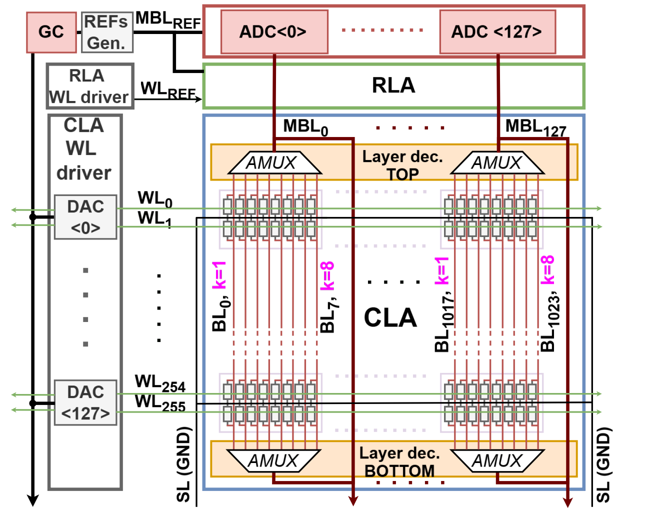

Fig. 2. Detailed block diagram of a Computational Local Array (CLA) with peripheral blocks. Each CLA consists of 256 rows and 1024 columns.

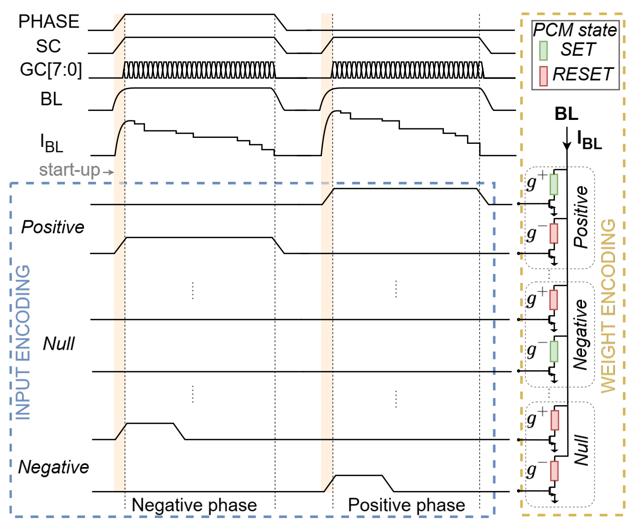

Fig. 3. Sketch of signed MVM operation performed inside the AiMC Core with detailed input and weight encoding strategy. Each phase is initiated by Start Count (SC) signal. For the sake of clarity, only 3 out of 9 possible combinations of input and weight signs are represented.

to read RLAs. Additionally, 512 local 8-to-1-layer decoders connect the Bit-Lines (BLs) from top and bottom edges of the CLAs to 512 Main BLs (MBLs), each one connected to an 11-bit ADC integrated in the pitch of the MBLs. A microcontroller is included to manage all operations. PCM cells are Ge-rich GST with MOS selector [6], and a single signed weight $w_{i,j}^k$ of the k-th layer is stored in two cells, whose conductance difference gives the absolute value of weights. A Global Counter (GC) synchronizes MVM computations. Input data, encoded in 8-bit sign-magnitude format, are converted by Word-Line (WL) drivers into activation times using Pulse-Width Modulation. Each BL current is processed by the ADC Current-Controlled Oscillator (CCO) and a ripple counter, which accumulates contributions from individual WLs. A signed MVM $\mathbf{z} = \mathbf{W^k x}$ is performed in two phases (Fig. 3): in the negative phase the contributions come from the combination of inputs and weights with the same sign, while in the positive phase from combination of inputs and weights with the opposite sign. The two contributions are then summed on the ripple counter. WL DACs alternately select positive or negative cells to ensure that the output is consistent with the input signs and phase. Weights sign is encoded as the position of the most conductive PCM cell.

III. Drift and Temperature Compensation

To address and limit output variations due to cells drift and temperature, the BL voltage is generated by injecting a scaled replica of a reference current, I_{REF}, into the RLAs (Fig. 4). As a result, the generated voltage, MBL_{REF} tracks averaged weight variations, achieving RCCT compensation and ensuring a constant analog output swing to prevent A/D precision loss. Moreover, the GC and all ADCs leverage identical CCOs, ensuring matched computation times thus enabling further compensation for residual variations in A/D conversion, which may arise due to temperature-induced frequency mismatches.

In addition to RCCT, to recover the residual errors, an ODR factor is independently applied to each ADC. The ODR factor is calculated as the ratio of the ideal MVM and a calibration MVM, both obtained using a predefined input vector. The ideal MVM result is permanently stored in the RLAs (one for each BL) and regularly read using DSAs, thus allowing ODR factor updates. This ensures the periodical tracking of any conductance variation at single ADC level, enhancing MVM precision compared to [5] and to sole RCCT.

IV. Bit Line Biasing Circuit (BL-BC) for Layer Compensation

The proposed BL-BC (Fig. 5) is employed to force the constant voltage V_{REF} to the BL node, independently of the

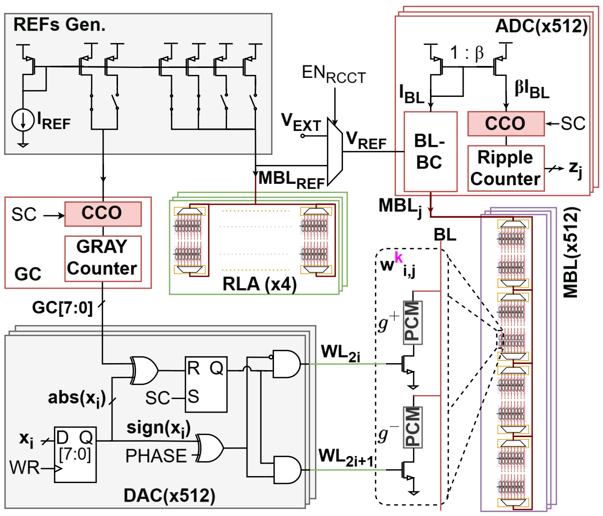

Fig. 4. Schematic solution for temperature and drift compensation: RCCT tracks the mean weight variation, while DAC-ADCs matched CCOs compensate for temperature-induced frequency mismatches.

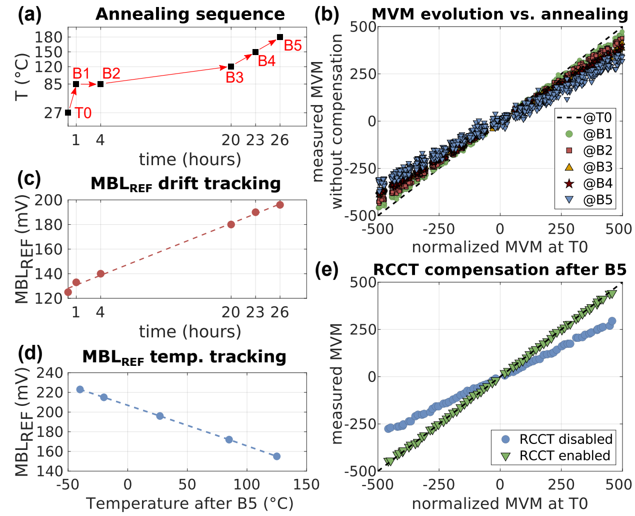

Fig. 6. Drift tolerance and temperature resilience tests. Used annealing sequence (a), MVM drift-induced variation (b), MBL_{REF} drift tracking (c), MBL_{REF} temperature tracking (d), RCCT drift compensation (e).

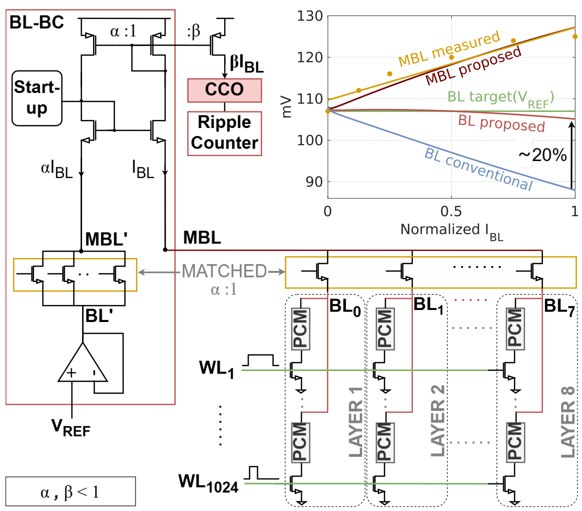

Fig. 5. BL-BC schematic for layer-decoder and parasitic voltage drop compensation, comparative simulation, and measurement results.

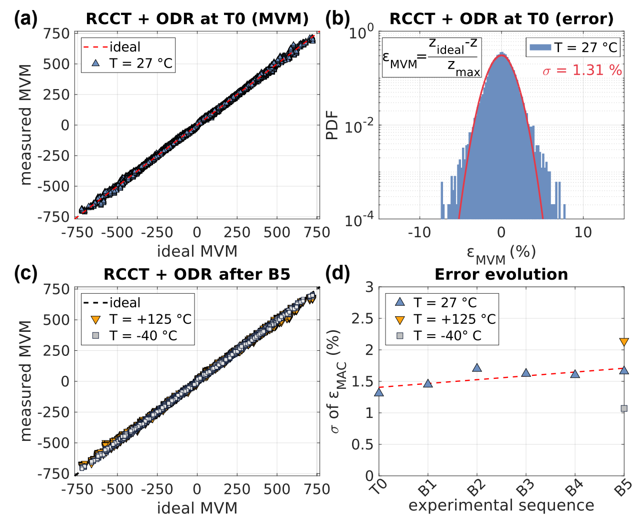

Fig. 7. Initially measured results of the MVM operation (a), corresponding normalized error ε_{MVM} (which includes drift random variations and is calculated as specified in the inset with z defined as in Fig. 1) (b), measured MVM results after annealing sequence (c), standard deviation of ε_{MVM} at different annealing sequences and temperatures (d).

current flowing through BL, thereby avoiding the I-V non-linearity of PCM cells. The BL-BC employs a scaled copy of the BL current (αI_{BL}) to equalize MBL and MBL$'$ nodes. A scaled and matched replica of the layer decoder is included in the left branch to compensate for the mean voltage drop of the layer decoders. Moreover, the metal line resistance of MBL$'$ in the BL-BC is laid out to match the scaled average resistance of MBL. Overall, the BL-BC provides a maximum biasing error of 1.8mV across the full current range relative to the target voltage V_{REF}, and a 20% improvement in precision compared to conventional biasing without voltage drops compensation.

V. Measurements

Functionality and robustness of the device (fabricated in a 28nm FD-SOI CMOS process) have been experimentally evaluated, at room temperature, at the end of different annealing steps (T0 to B5) shown in Fig. 6a. This has been done to accelerate stresses and drift on the stored conductances. Fig. 6b shows the evolution of the measured results of MVM operations as a function of the applied annealing sequence. The variation of MBL_{REF} during the applied annealing sequence is represented in Fig. 6c demonstrating the capability of the proposed circuits to track drift-induced variations of the conductance of the cells (drift tolerance test). Fig. 6d provides the measured temperature dependence of MBL_{REF} voltage, after the last annealing step (B5 of Fig. 6a) from –40°C to 125°C (temperature resilience test). Finally, Fig. 6e shows the effects of RCCT compensation (green triangles) on the

TABLE I
COMPARISON WITH STATE-OF-THE-ART.

	[7]	[8]	[9]	[10]	**This work**
CMOS Technology	22nm	28nm	28nm	14nm	**28nm**
Memory Technology	ReRAM	Digital	MRAM	PCM	**PCM**
Weight Unit Cell	1T1C	Register based	Pseudo 2T 2MTJ	8T4R	**2T2R**
Input/Output precision	FP16/FP32	8/8 bit	9/5 bit	8/8 bit	**8/11 bit**
Weights precision	FP16	8 bit	9 bit	analog	**analog**
Operating Temperature	Not specified	Not specified	-25/+85°C	Not specified	**-40/+125°C**
MVM error	Not specified	NA	Not specified	1.94 / NA % (27/125°C)	**1.7 / 2.14% (27/125°C)**
Throughput [TOPs]	0.86	3.86	Not specified	1	**1.7***
Energy Efficiency [TOPs/W]	65.5	83.23	41.5	10.5	**24*,†**
Core Area [mm^2]	8.2	1.41	4.5	0.64	**1.7**
Number of weights (NW)	1M	16K	223K	65.5K	**2M**
FoM [NW×TOPs/W/mm^2]	8M	1M	2.05M	1.1M	**28M*,†**

* Evaluated @ GC = 1GHz; † Including DACs/ADCs power consumption.

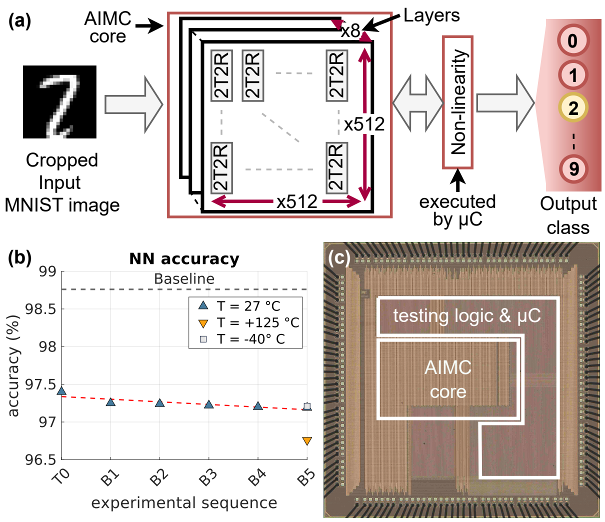

Fig. 8. Implementation of a DNN for MNIST image classification (a). NN accuracy evolution over time and temperature (b). Test-chip photo (c).

MVM operation, which demonstrates the effectiveness of the proposed solution.

The combined effects of RCCT and ODR are shown in Fig. 7. More specifically, Fig. 7a provides the results of the MVM operations (collected at T0, see Fig. 6a) with RCCT and ODR enabled. The corresponding normalized errors ε_{MVM} (calculated as specified in the figure inset with z defined as in Fig. 1) are given in Fig. 6b which shows a measured σ of the normalized errors ε_{MVM} approximately equal to 1.3%. The results of MVM operations (collected after B5, see Fig. 6a) with RCCT and ODR enabled are provided (for two different temperatures) in Fig. 7c. Additionally, the evolution of the measured σ of the normalized errors ε_{MVM} during the annealing sequence is finally given in Fig. 7d. To evaluate the worst cases of the normalized errors after completing the annealing sequence (i.e., after B5) Fig. 7d also provides the same measurements at -40°C and 125°C.

The presented AiMC core has been used to map a multilayer DNN for MNIST dataset classification, achieving drift- and temperature- quasi-invariant accuracy (Fig. 8). As reported in Table I, our solution combines high MVM resilience with a storage-energy efficiency per area (a suitable FoM has been introduced to this end in Table I) larger than 3.5 times when compared to state-of-the-art hardware accelerators [7]–[10].

VI. Conclusions

In this paper, a test chip for AiMC using PCM cells for multilayer neural network, drift-tolerant, and temperature-resilient computations has been presented. The proposed AiMC core has been manufactured in a 28nm FD-SOI CMOS process with an array of 4M-cells, achieving a less than 2.14% MVM error over a temperature range from –40°C to 125°C. Measured data show an improvement, with respect to other state-of-the-art hardware accelerators, by a factor 3.5 estimated by means of a FoM defined as No. of Weights×TOPS/W/mm^2.

References

[1] S. Ambrogio, *et al.*, "An analog-AI chip for energy-efficient speech recognition and transcription," *Nature*, vol. 620, pp. 986–992, 2023.
[2] M. Bertuletti, *et al.*, "A multilayer neural accelerator with binary activations based on phase-change memory," *IEEE Trans. on Electron Devices*, vol. 70, no. 3, pp. 986–992, 2023.
[3] A. Antolini, *et al.*, "A readout scheme for PCM-based analog in-memory computing with drift compensation through reference conductance tracking," *IEEE Open J. of the Solid-State Circ. Soc.*, vol. 4, pp. 69–82, 2024.
[4] L. Pistolesi, *et al.*, "Differential phase change memory (PCM) cell for drift-compensated in-memory computing," *IEEE Trans. on Electron Devices*, vol. 71, no. 12, pp. 7447–7453, 2024.
[5] I. Boybat, *et al.*, "Temperature sensitivity of analog in-memory computing using phase-change memory," in *2021 IEEE Int. Electron Dev. Meeting (IEDM)*, 2021, pp. 28.3.1–28.3.4.
[6] F. Arnaud, *et al.*, "Truly innovative 28nm FDSOI technology for automotive micro-controller applications embedding 16MB phase change memory," in *2018 IEEE Int. Electron Dev. Meeting (IEDM)*, 2018, pp. 18.4.1–18.4.4.
[7] T.-H. Wen, *et al.*, "34.8 a 22nm 16Mb floating-point ReRAM compute-in-memory macro with 31.2TFLOPS/W for AI edge devices," in *2024 IEEE Int. Solid-State Circ. Conf. (ISSCC)*, vol. 67, 2024, pp. 580–582.
[8] Y. Wang, *et al.*, "34.1 a 28nm 83.23TFLOPS/W POSIT-based compute-in-memory macro for high-accuracy AI applications," in *2024 IEEE Int. Solid-State Circ. Conf. (ISSCC)*, vol. 67, 2024, pp. 566–568.
[9] H. Cai, *et al.*, "33.4 a 28nm 2Mb STT-MRAM computing-in-memory macro with a refined bit-cell and 22.4 - 41.5TOPS/W for AI inference," in *2023 IEEE Int. Solid-State Circ. Conf. (ISSCC)*, 2023, pp. 500–502.
[10] R. Khaddam-Aljameh, *et al.*, "HERMES core – a 14nm CMOS and PCM-based in-memory compute core using an array of 300ps/lsb linearized cco-based adcs and local digital processing," in *2021 Symp. on VLSI Circ.*, 2021, pp. 1–2.