3-D Crosspoint (3DXP) Memory Arrays With Subthreshold Operation for Low-Energy, High-Accuracy Neural Network Accelerators

F. Carletti, *Graduate Student Member, IEEE*, M. Farronato, *Member, IEEE*, G. Y. C. Hu, N. Lepri, *Graduate Student Member, IEEE*, I. Tortorelli, *Member, IEEE*, A. Pirovano, *Member, IEEE*, P. Fantini, *Member, IEEE*, and D. Ielmini, *Fellow, IEEE*

***Abstract*— In-memory computing (IMC) has been identified as a promising paradigm for hardware neural network accelerators thanks to the reduced data movement and improved parallelism. A known issue of IMC is the relatively large summation current within the memory array, which causes energy inefficiency and computing inaccuracy due to IR drop. An additional burden is the area and energy-demanding readout circuits, which limit the density and energy efficiency of computing. This work reports a hardware demonstration of IMC using 3-D crosspoint (3DXP) arrays of ovonic threshold switches and phase change memories (PCMs). We demonstrate a precise program–verify (PV) algorithm optimized for the subthreshold regime, allowing for a reduction of the operating currents by more than two orders of magnitude with respect to the conventional 3DXP technology. We experimentally demonstrate vector–vector multiplication (VVM) and feature extractions, which are key operations of convolutional neural networks (CNNs). Simulation study of LeNet5 with binary and ternary quantization, including device variability, 1/f noise, and drift, demonstrates high accuracy and low-energy inference thanks to precise programming, subthreshold operation, and careful drift compensation.**

***Index Terms*— 3-D crosspoint (3DXP), binary neural networks (BNNs), convolutional neural network (CNN), in-memory computing (IMC), ovonic threshold switch (OTS), phase change memory (PCM), ternary neural network (TNN).**

I. Introduction

ARTIFICIAL intelligence (AI) experienced a rapid growth, booming in all sectors, including industry, finance, health, and society. Recent advances in large language models (LLMs) and agentic AI [1], [2] are revolutionizing the AI landscape, although the rapid increase of the computational complexity of inference and training raises key concerns about the energy sustainability of AI [3]. The energy efficiency of AI systems can be significantly improved by in-memory computing (IMC). The latter enables in situ data processing through massively parallel matrix-vector multiplications (MVMs) [4]. Nonvolatile memories (NVMs) such as phase change memory (PCM) and resistive switching memory (RRAM) are currently leading technologies for high-precision IMC thanks to their high integration density, multilevel cell operation, and back-end-of-line (BEOL) integration [5]. However, a fundamental issue of these two NVMs is the relatively large read current that causes energy inefficiency and accuracy loss due to IR drop [6]. Recently, the 3-D crosspoint (3DXP) technology has been proposed for IMC, benefiting from a high density thanks to the one-selector/one-resistor (1S1R) structure [7], [8] and the low current in the subthreshold regime [9], [10], [11], [12]. However, high accuracy and robustness against variation, off-state leakage, noise, and drift still need to be systematically understood and demonstrated to fully support 3DXP-based IMC systems for AI accelerators.

This work reports an experimental demonstration of binary and ternary convolutional neural networks (CNNs) with 3DXP arrays. We demonstrate high-precision tuning of binary states in the subthreshold regime thanks to the program–verify (PV) algorithm. A differential weight scheme allows to compensate for the nonnegligible read current of the high-resistive state (HRS) and to map positive and negative weights [13], [14], [15], [16]. Feature extraction is experimentally carried out in 4×4 3DXP arrays, while a full CNN is demonstrated by noise, drift, and variation-aware simulations of a LeNet5 model for MNIST recognition. High classification accuracies are obtained by quantization-aware training for both weights and activations. Our results indicate that the 3DXP technology allows to reduce the summation current during MVM by more than two orders of magnitude, thus paving the way for energy- and area-efficient AI hardware accelerators.

II. 3DXP Device Characterization

Fig. 1(a) shows a sketch of a 4×4 mini-array of 3DXP memory cells [7], [12]. The 1S1R structure consists of an

Received 9 September 2025; revised 13 November 2025; accepted 17 November 2025. Date of publication 3 December 2025; date of current version 6 January 2026. This work was supported by European Research Council (ERC) through European Union's Horizon Europe Research and Innovation Program under Grant 101054098. The review of this article was arranged by Editor X. Liu. *(Corresponding author: F. Carletti.)*

F. Carletti, M. Farronato, G. Y. C. Hu, N. Lepri, and D. Ielmini are with the Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), IU.NET, Politecnico di Milano, 20133 Milan, Italy (e-mail: fabio.carletti@polimi.it; daniele.ielmini@polimi.it).

I. Tortorelli, A. Pirovano, and P. Fantini are with Micron Technology Inc., 20871 Vimercate, Italy.

Digital Object Identifier 10.1109/TED.2025.3635643

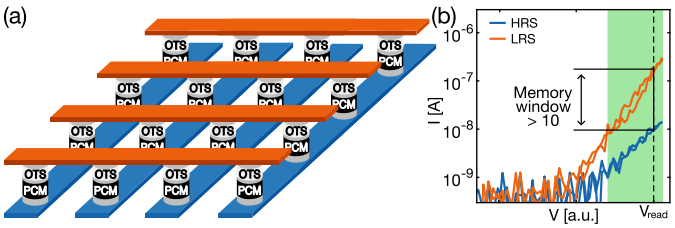


Fig. 1. 3DXP arrays and characteristics. (a) Sketch of a 4 × 4 mini-array containing test devices externally available from larger 512 × 512 arrays. (b) Subthreshold characteristic (I–V) of a device. A memory window larger than 10 at the selected V_{read} is visible. Green area highlights the interval of possible reading voltages.

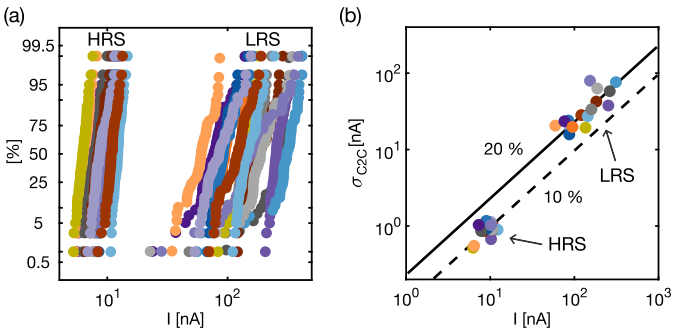


Fig. 2. Read current distributions and their variations. (a) C2C distributions for 16 devices programmed multiple times in both LRS and HRS. (b) C2C standard deviation σ_{C2C} as a function of the median value of the read current.

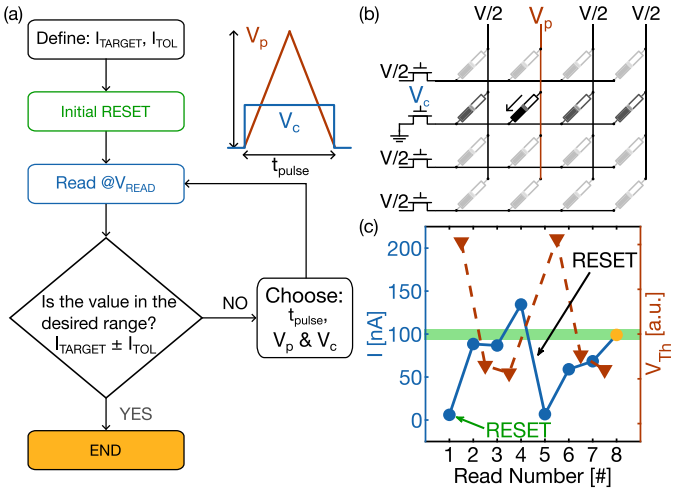


Fig. 3. Program verify explained (a) Flowchart of the write verify. (b) Half-bias scheme applied to an array, avoiding switching of unselected cells during write phase and reducing leakage in the verify phase. Programming current is tuned with the row-selection transistor. (c) Evolution of the current and threshold voltage of a device during the PV.

ovonic threshold switching (OTS) selector device and a PCM resistive device for storing a binary state. Thanks to its highly nonlinear characteristic, the OTS selector allows the selection of the memory devices for programming and reading operations. For instance, programming of individual cells is enabled by the so-called V/2 biasing scheme [17], where voltages +V/2 and −V/2 are applied to the row and column, respectively, of the device to be programmed, while all other rows/columns are left grounded. 3DXP devices were programmed by the application of pulses with voltage higher than the threshold voltage V_{Th} to enable a set operation, namely, a transition from high-resistance state (HRS) to low-resistance state (LRS), or a reset operation, namely, a transition from LRS to HRS. The reset operation consists of the amorphization of the active material within the PCM, while the set operation must favor the crystallization of the active material. The row-selection transistor was properly biased during set and reset operation to limit the current to a compliance level suitable for the desired transition.

Fig. 1(b) shows the measured current–voltage (I–V) characteristics for the LRS and HRS of a selected 3DXP element. The characteristics were measured in the subthreshold regime, namely, for voltages below V_{Th}, thus resulting in extremely low currents in the range below 1 μA. As a result, read currents $I_{LRS} \approx 100$ nA and $I_{HRS} \approx 10$ nA were obtained for LRS and HRS, respectively, at a convenient read voltage V_{read}, with a memory window given by $I_{LRS}/I_{HRS} \approx 10$. Larger memory windows can be obtained at higher currents by increasing V_{read}. However, higher read currents also impact energy consumption, IR drop, and the probability of read disturbs potentially induced by threshold switching [5].

III. PV Technique

To provide a comprehensive statistical characterization of the memory behavior, Fig. 2(a) presents the cumulative distributions of the read currents for LRS and HRS of 16 devices over 100 set/reset cycles. Each data point was obtained as the average of 100 read operations to mitigate read-to-read (R2R) variations while preserving the device-to-device (D2D) and cycle-to-cycle (C2C) variations. Data indicate D2D standard deviations $\sigma_{D2D} = 1.7$ and 74 nA for HRS and LRS, respectively. In particular, the D2D variation for LRS is about 75% of the average read current, which cannot satisfy the accuracy requirements for IMC. Fig. 2(b) shows the C2C standard deviation σ_{C2C} as a function of the median read current for devices in Fig. 2(a), indicating a relative σ_{C2C} of about 10% and 20% for HRS and LRS, respectively, which is in line with other NVM technologies [18].

To mitigate the D2D and C2C variations, a program/verify (PV) algorithm was developed, as illustrated in the flowchart of Fig. 3(a). In the PV algorithm, each set/reset operation is followed by a read phase to verify that the cell current is within a target window with tolerance $\Delta I = 10$ nA. Fig. 3(b) shows the adopted V/2 scheme for cell programming, where a voltage V/2 is applied to all rows/columns, except for a voltage V_p applied to the selected column with the selected row tied to ground. For proper control of the HRS/LRS state, a gate voltage V_c is applied to the gate of the row-select transistor. Fig. 3(c) shows the measured read current and threshold voltage during a PV algorithm for a device. Applying consecutive set pulses gradually increases the read current. If the current increases above the target value, a reset pulse is applied. Depending on the difference with respect to the target, we select V_p, t_p, and more importantly V_c. After eight iterations, the device converges to the desired current;

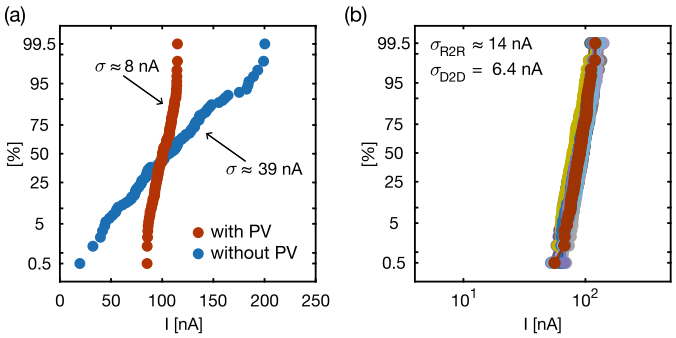


Fig. 4. PV impact on LRS. (a) C2C distributions of a device in the LRS with and without PV algorithm. (b) Read cycle distributions using read pulses of 80 μs. All devices of the 4 $\times$ 4 array have been programmed in the LRS, evidencing the strong reduction of σ_{D2D} and the leftover σ_{R2R}.

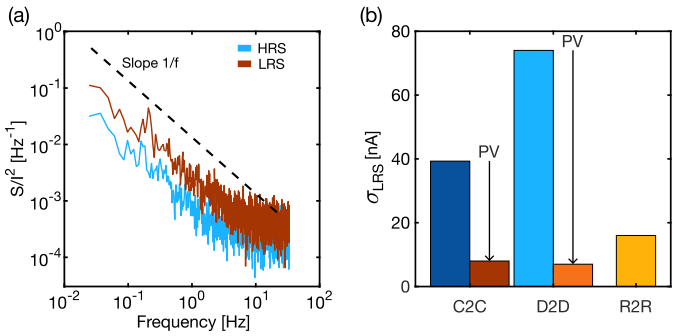


Fig. 5. 3DXP noise and variations. (a) Normalized PSD for both states. A dominant 1/f noise and higher intensity for the LRS are shown. (b) Summary of C2C, D2D, and R2R variations for the LRS, indicating the benefit of PV for C2C and D2D variations.

however, due to σ_{R2R}, the current read in the last iteration is not sufficient to confirm convergence. Therefore, to strengthen the PV precision, the device is measured with a sequence of ten read operations. PV success is achieved only if the resulting average read current falls within the desired range (yellow circle); otherwise, more iterations are applied until convergence.

Fig. 4(a) shows the C2C distributions for a single device in LRS with and without PV, indicating a suppression of variation to a residual $\sigma_{C2C} \approx 8$ nA after PV. Fig. 4(b) shows the current distributions for LRS after PV obtained from 100 consecutive read operations applied to each device in the array. Each read cycle has been implemented by a rectangular pulse of height V_{read} and a time-width of 80 μs. The D2D variation is strongly reduced to $\sigma_{D2D} \approx 6.4$ nA, while the overall variation is dominated by $\sigma_{R2R} \approx 14$ nA, corresponding to a relative noise $\sigma_{R2R}/I_{LRS} = 15\%$, which can be attributed to read noise. To investigate the origin of the read noise, Fig. 5(a) shows the normalized power spectral density (PSD) of the read current, namely, S_I/I^2, for both LRS and HRS. Data indicate a dominant 1/f noise which can be attributed to defect relaxation in the amorphous phase of the active regions of OTS and PCM elements after set/reset [19], [20], [21]. Read noise might be mitigated by material or algorithm optimization. Fig. 5(b) summarizes the various sources of read current variation, highlighting the ability to minimize C2C and D2D variations via precise PV.

IV. Vector–Vector Multiplication

To demonstrate IMC using 3DXP, we conducted experiments implementing vector–vector multiplication (VVM) operations on individual columns and rows of the 4 $\times$ 4 arrays, as illustrated in Fig. 6(a). First, the selected devices were programmed to the target conductance states by the PV algorithm of Fig. 3(a). Then, a binary voltage vector V_j, with j = 1, 2, 3, and 4, was applied simultaneously to the top electrodes of the devices. Every single element of the voltage vector can have two possible values, namely, $V_j = 0$ V or V_{read}. Finally, we measured the summation current at the grounded common bottom electrode, given by

$$I = \sum_j v_j I_j \tag{1}$$

where v_j is the normalized voltage being either 0 or 1 for $V_j = 0$ or $V_j = V_{read}$, respectively [4].

Fig. 6(b) shows the measured I–V curves for a 3DXP 4 $\times$ 1 vector, obtained by applying the same voltage V_{read} to all TEs, while the summation BE current was collected for various combinations of the programmed states, indicated by $G = 0$ and $G = 1$ for HRS and LRS, respectively. The current was averaged among ten different read operations, each spaced by a time period $T_{int} = 30$ s. Note the nonnegligible off-state current when all the four devices are in the HRS, which is due to the leakage current of the PCM in the HRS, corresponding to the amorphous phase. Fig. 6(c) shows the measured BE current as a function of the combination of the input voltages v_j at increasing V_{read} for programmed states in the 4 $\times$ 1 vector being either HRS, i.e., $\mathbf{G}$ = (0 0 0 0), or LRS, i.e., $\mathbf{G}$ = (1 1 1 1). As the read voltage V_{read} increases, the summation current increases the energy consumption associated with analog computation within the array. On the other hand, a larger V_{read} can enhance the sensing margin between different VVM products, which is beneficial to the accuracy of computation.

To better support the accuracy of the VVM operation, Fig. 6(d) shows the correlation plot of the measured summation current as a function of the ideal current based on (1), spanning all combinations of input v_j and programmed states I_j. It is worth mentioning that, even if only five results are mathematically possible according to (1), the contribution of the HRS current is nonnegligible, thus obtaining slightly different BE current values for the same ideal results.

To compensate for the nonzero HRS read current, differential VVM experiments were carried out. The concept is schematically illustrated in Fig. 7(a), where two array columns of devices are used to map a VVM operation given by

$$I = \sum_j v_j I_j = \sum_j v_j \left(I_j^+ - I_j^-\right) \tag{2}$$

where I_j^+ and I_j^- are the positive and negative currents to map a certain positive or negative coefficient in the memory array, respectively. Thanks to the differential scheme in the figure, a zero current is obtained by mapping two HRS cells within a differential pair since $I_j = I_{HRS} - I_{HRS} = 0$. Fig. 7(b) shows the correlation plot of the measured summation current in the differential scheme as a function of the ideal value, demonstrating

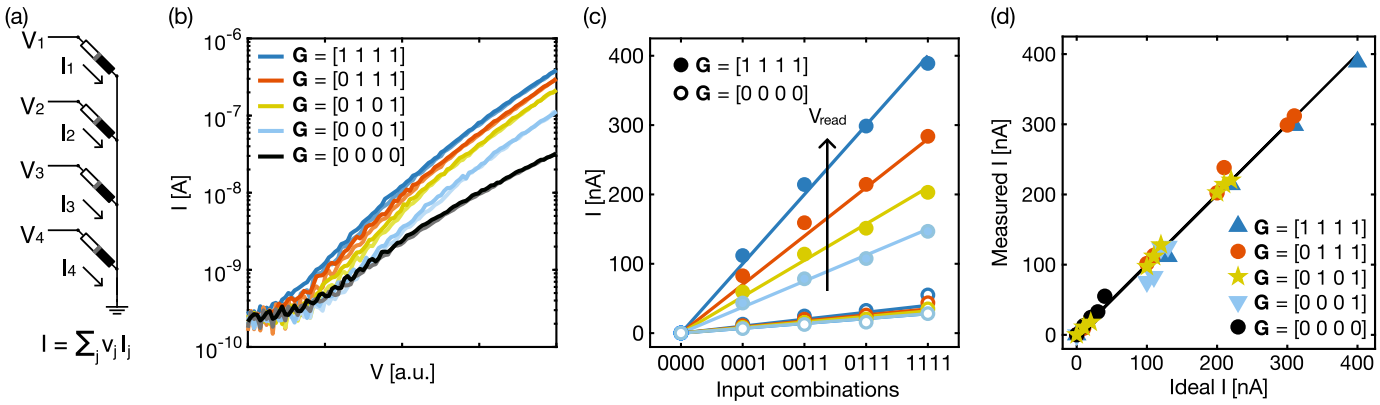


Fig. 6. VVM. (a) Schematic of the parallel VVM obtained by accessing with a proper input voltage the top electrodes of a column of devices. (b) $I-V$ curves collected with all the elements of the input voltage vector set at V_{read} while changing the programmed weights of the conductance vector **G**. (c) Median VVM obtained by randomly changing the input vector and the V_{read} value. The linearity is preserved at every V_{read}. (d) Correlation plot of the measured I as a function of the ideal I while varying both the input voltage and the weight vector.

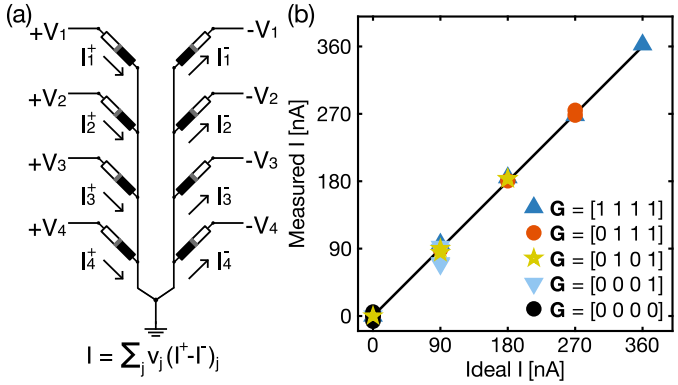


Fig. 7. Differential VVM. (a) Schematic of the parallel VVM using one column for positive weights and one column for negative weights, i.e., I_+ *and* I_-. (b) Correlation plot of the measured $I = I_+ - I_-$ while varying both **V** and **G**. With respect to Fig. 6(d), we correctly have only five possible results, as it should be in the ideal case.

that only five possible outcomes are now obtained, thanks to compensation of the HRS leakage. The excellent results in Figs. 6(d) and 7(b) support the accuracy of the binary VVM operations thanks to the precise PV algorithm and differential mapping of computational coefficients in the array.

V. Feature Extraction

To realistically support 3DXP for IMC, we considered a CNN for image classification [18], [22]. In a convolutional layer, features are extracted by dividing the input activation into multiple slices as big as the size of the network filter. Then, the VVM between each filter and these portions of the input activation is carried out. Finally, the sum of all these contributions is submitted as the input of a nonlinear activation function to obtain the output activation. Feature extraction was experimentally validated by convolution of a 3×3 filter programmed in our 3DXP array and a portion of a 15×15 input image as shown in Fig. 8(a). The differential scheme of Fig. 7(a) was adopted to suppress the nonzero current of the HRS. Fig. 8(a) shows the input image submitted to the binarized filters, where the positive weight $\delta_j = I_j^+ - I_j^-$ at the jth element of the filter is obtained by $I_j^+ = I_{LRS}$ and $I_j^- = I_{HRS}$ in (2), while a negative weight is obtained by $I_j^+ = I_{HRS}$ and $I_j^- = I_{LRS}$. As a result, differential filters with complementary states were programmed, yielding a binary weight $\delta_j = \pm(I_{LRS} - I_{HRS}) \approx \pm 90$ nA in our experimental setup. To collect the current in (2), a 3×3 portion of the 3DXP was used with shortened BE lines. In addition to binary weights, ternary weights can also be mapped in the differential filter by adopting for the intermediate level either $\delta_j = I_{HRS} - I_{HRS} \approx 0$ nA or $\delta_j = I_{LRS} - I_{LRS} \approx 0$ nA.

Fig. 8(b) shows the combinations of differential weights δ_j, where the third configuration $\delta_j = \pm(I_{LRS} - I_{LRS})$ is discarded to minimize variations and power consumption. Fig. 8(c) shows the experimentally extracted feature map for the MNIST image with a 3×3 diagonal filter. The experimental results show good agreement with simulations, thus confirming the accuracy of binary feature extraction using 3DXP arrays. Fig. 8(d) shows the correlations of the measured I_{BE} as a function of the ideal ones of (2). Data are obtained from several experiments, i.e., convolutions, with variable input image and filters, including anti-diagonal (d), diagonal (e), horizontal (f), and vertical filters (g). The good correlation between experimental and theoretical features supports the accuracy of feature extraction enabled by the combination of precise PV algorithm and accurate IMC circuit.

VI. CNN Demonstration

To demonstrate that IMC within a 3DXP technology can complete a relatively large computational task, we studied the accuracy of a LeNet5 model trained offline and mapped into a set of 3DXP arrays [22], [23]. Both binary and ternary LeNet5 were trained with TensorFlow assuming ideal 3DXP devices. MATLAB has been adopted for the inference to properly account for all nonidealities. D2D variations were added according to the measured distribution after mitigation by the PV algorithm. R2R fluctuations due to 1/f noise and time-dependent drift were also included in the simulations.

A. Impact of IR-Drop

IMC with crossbar arrays suffers from accuracy degradation caused by the metal-line resistance across both the bitline

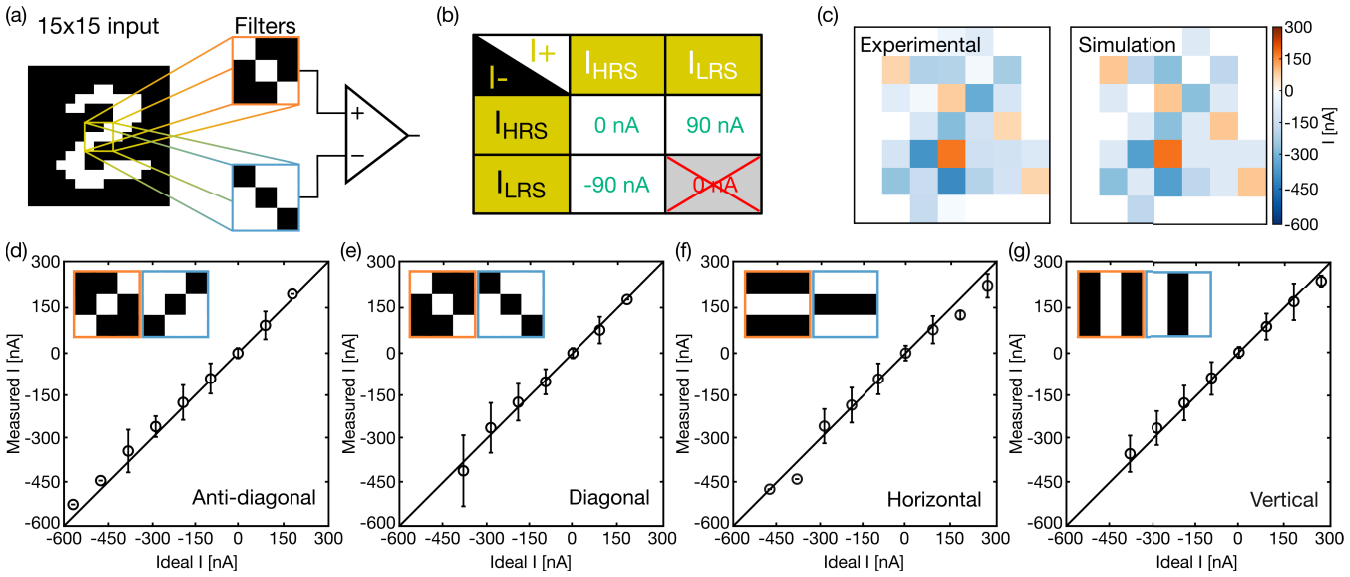


Fig. 8. Feature extraction. (a) Sketch of the convolutional layer, where a given filter is multiplied to a portion of the input image. Feature extraction was experimentally carried out by VVM in the 3DXP array for each position in the image and various filters, mapped with a differential approach to compensate for I_{HRS}. (b) Summary of the possible $I_+ - I_-$ in the differential weights yielding the binary and ternary weights. (c) Experimental and simulated feature extraction. (d)–(g) Correlation plots of the experimentally extracted features for various filters and MNIST images.

and the wordline. Because of the IR drop, when voltages are applied to several lines at the same time, the voltage across each memory cell is lower than expected; thus, the individual cell current is lower compared to that obtained by individual readout. The IR drop effect depends on the array size, the parasitic line resistance, and the programmed conductance levels. Several models have been proposed to quantify this effect [6], [24], and various compensation strategies have been explored to mitigate it, including optimal weight mapping techniques [25], design technology co-optimization (DTCO) [24], and local current cancellation schemes [26]. To assess possible IR drop effects in our networks, we carried out simulations of our crossbar arrays at increasing size from 32×32 to 512×512 and we found that the resulting accuracy degradation is always below 0.1% for a line resistance $r_{par} \leq 20\,\Omega$. This is thanks to the low current in the subthreshold regime of our devices, which thus strongly reduces the design complexity required for IR-drop compensation.

B. Impact of Variations

Fig. 9(a) shows the computed accuracy of the network in the case of ideal, i.e., zero variation, and nonideal as a function of the standard deviation σ_{R2R} of R2R noise, normalized by the LRS current I_{LRS}. A constant $\sigma_{D2D} = 6.4$ nA was assumed according to data in Fig. 4(b). The calculated accuracy in the figure decreases from the ideal value at increasing relative variations, with the ternary neural network (TNN) showing generally higher accuracy due to the higher number of levels per weights and the higher use of HRS cells benefiting from a lower R2R variation. For the measured $\sigma_{R2R}/I_{LRS} = 15\%$ in our samples, the accuracy is around 93.5% for the TNN and around 91.5% for the binary neural network (BNN), indicating a minor drop with respect to the ideal accuracy. Simulations were also carried out for R2R variation exceeding

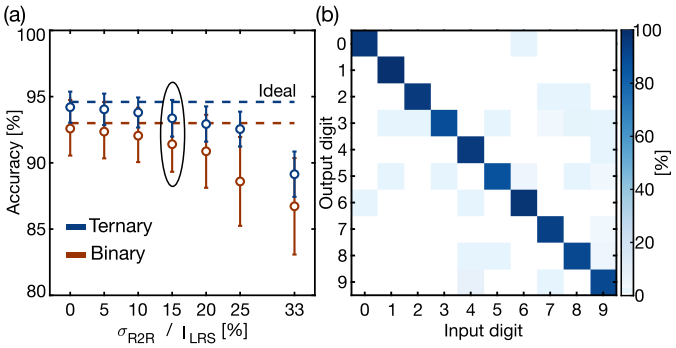


Fig. 9. Inference simulation results. (a) Accuracy against σ_{R2R}. Ternary CNN has higher accuracy and lower spread thanks to: 1) lower content of I_{LRS}, which suffers from a higher 1/f noise, and 2) zero activation value. (b) Confusion matrix at $\sigma_{R2R} = 15\%$ for TNN.

the experimental data in Fig. 3(b) to account for a possible increase of σ_{R2R} at decreased device size [27]. Fig. 5(b) shows the confusion matrix for the TNN assuming the experimental $\sigma_{R2R}/I_{LRS} = 15\%$, highlighting the good accuracy of the neural network.

C. Impact of Drift

Both PCM and OTS devices are affected by drift, that is, a time-dependent decrease of conductance due to structural relaxation of defects in the amorphous structure of the active material [28], [29]. Fig. 10(a) shows the measured current for various devices programmed in either LRS or HRS. Data indicate a power-law decay according to $I \sim t^{\nu}$, where the drift exponent ν is about 0.08 for the HRS and about 0.04 for the LRS. The smaller ν for LRS can be explained by the crystalline state of the PCM showing negligible drift, compared to the amorphous state. To assess the impact of drift on IMC, Fig. 10(b) shows the accuracy of the networks

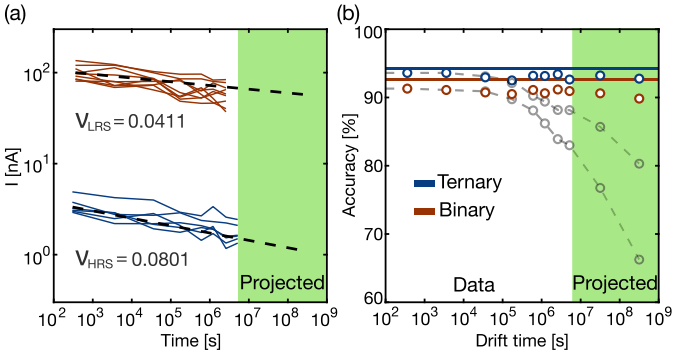
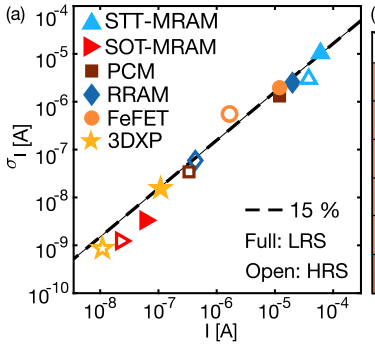


Fig. 10. Drift characterization. (a) HRS and LRS characterized for a drift time $>10^6$ s. Drift coefficient ν extracted from data and drift projected up to ten years at room temperature. (b) Impact on accuracy when compensating through the batch normalization layer is negligible.



(b) Device comparison

	IR drop	Endurance	Number of bits
PCM	Yes	10^7	3
RRAM	Yes	10^7	3
SOT	No	$>10^{12}$	1
STT	Yes	10^{12}	1
FeFET	Yes	10^{10}	>3
3DXP	No	$>10^7$	1

Fig. 11. Benchmarking. (a) Correlation plot of σ_I as a function of the average current for 3DXP and other NVMs. 3DXP reduces the read current and its variation by at least 100x compared to state-of-the-art NVMs. (b) Comparison of additional properties among NVMs.

	PCM	RRAM	FeFET	STT	SOT	3DXP (This work)
μ_{LRS} [µA]	10	20	2.2	60	0.075	0.100
μ_{HRS} [µA]	1	1	0.8	20	0.025	0.010
μ_{AVG} [µA]	5.5	10.5	1.5	40	0.050	0.055
V_{read} [V]	0.3	0.2	0.2	0.2	1.2	1.8
E_{read} [fJ] (t_{read} = 50 ns)	82.5	105	15	400	3	4.95

Fig. 12. Energy consumption estimation. The same reading time of 50 ns has been considered for all devices. Our V_{read} has been adapted from [34]. It is possible to see how our device provides almost best energy consumption per read cycle while providing the lowest area occupation.

as a function of time after programming with $\sigma_{R2R} = 15\,\%$. The accuracy significantly drops at increasing time, due to the weights decreasing significantly from the trained values. This performance drop can be mitigated by rescaling the output summation current acting on normalization layers without any weight reprogramming. The figure also shows the accuracy with the modified batch normalization, indicating maintained accuracy even after a long time.

VII. Memory Benchmarking

To assess the low current operation of 3DXP in comparison to other NVM devices, Fig. 11(a) shows the correlation plot of read current variation σ_I as a function of the average read current I_{read} for various NVM technologies, including spin-transfer-torque magnetic-random access-memory (STT-MRAM), spin-orbit-torque magnetic-random access-memory (SOT-MRAM), resistive random access-memory (RRAM), ferroelectric field-effect-transistor (FeFET), and transistor selected PCM [16], [30], [31], [32], [33]. Low read currents are beneficial to IMC as they result in a relatively high energy efficiency and in a relatively low IR drop, hence a low distortion of the MVM [5]. On the other hand, low variations are essential to ensure accurate MVM and inference operation, as shown in Fig. 9(a). According to data in Fig. 11(a), 3DXP and most of the other NVMs show a relative variation of about 15% of the mean current. Fig. 11(b) compares the number of bits and the endurance of the aforementioned devices. NVM such as PCM and RRAM can store more than 1 bit of information; however, this is not strictly required for BNN and TNN. Finally, Fig. 12 compares the energy consumption arising from a single-device single-read operation for all the devices in Fig. 11(b). As a representative V_{read}, we used a suitable value [34], while the read time was assumed equal to $t_{read} = 50$ ns. These results indicate that our 3DXP devices provide one of the lowest energy consumptions while achieving the highest area efficiency among all of the proposed emerging memories. Overall, 3DXP technology shows better performance in terms of low currents, thanks to the subthreshold operation, and high density, thanks to the 1S1R structure and the 3D stacking architecture. These results support the 3DXP technology for energy-efficient, high-accuracy IMC.

VIII. Conclusion

This work demonstrates IMC with 3DXP arrays operated in the subthreshold regime. Convolution is experimentally demonstrated by VMM with 3×3 subarrays, while full CNN with binary and ternary weights is studied via simulations. 3DXP appears as a good candidate for IMC, thanks to its precise subthreshold programming of 3DXP, the low operating current, and the high integration density.

References

[1] A. Grattafiori et al., "The Llama 3 herd of models," 2024, *arXiv:2407.21783*.
[2] D. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey," *IEEE Access*, vol. 13, pp. 18912–18936, 2025.
[3] K. Crawford, "World view," *Nature*, vol. 626, p. 693, Feb. 2024.
[4] 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.
[5] N. Lepri, A. Glukhov, L. Cattaneo, M. Farronato, P. Mannocci, and D. Ielmini, "In-memory computing for machine learning and deep learning," *IEEE J. Electron Devices Soc.*, vol. 11, pp. 587–601, 2023.
[6] 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.
[7] D. Kau et al., "A stackable cross point phase change memory," in *IEDM Tech. Dig.*, Dec. 2009, pp. 1–4.
[8] A. Fazio, "Advanced technology and systems of cross point memory," in *IEDM Tech. Dig.*, Dec. 2020, pp. 24.1.1–24.1.4.
[9] J. M. Lopez et al., "1S1R sub-threshold operation in crossbar arrays for low power BNN inference computing," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2022, pp. 1–4.

[10] D. A. Robayo et al., "Reliability and variability of 1S1R OxRAM-OTS for high density crossbar integration," in *IEDM Tech. Dig.*, Dec. 2019, pp. 35.3.1–35.3.4.
[11] N. Lepri et al., "In-memory neural network accelerator based on phase change memory (PCM) with one-selector/one-resistor (1S1R) structure operated in the subthreshold regime," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2023, pp. 1–4.
[12] F. Carletti et al., "Low-energy, high-accuracy convolutional network inference in 3D crosspoint (3DXP) arrays," in *Proc. IEEE Eur. Solid-State Electron. Res. Conf. (ESSERC)*, Sep. 2024, pp. 412–415.
[13] V. Milo et al., "Multilevel HfO_2-based RRAM devices for low-power neuromorphic networks," *APL Mater.*, vol. 7, no. 8, Aug. 2019, Art. no. 081120.
[14] 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.
[15] L. Pistolesi et al., "Differential phase change memory (PCM) cell for drift-compensated in-memory computing," *IEEE Trans. Electron Devices*, vol. 71, no. 12, pp. 7447–7453, Dec. 2024.
[16] L. Pistolesi et al., "Drift compensation in multilevel PCM for in-memory computing accelerators," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2024, pp. 1–4.
[17] D. Ielmini and G. Pedretti, "Device and circuit architectures for in-memory computing," *Adv. Intell. Syst.*, vol. 2, no. 7, Jul. 2020, Art. no. 2000040.
[18] D. Ielmini, N. Lepri, P. Mannocci, and A. Glukhov, "Status and challenges of in-memory computing for neural accelerators," in *Proc. Int. Symp. VLSI Technol., Syst. Appl. (VLSI-TSA)*, Apr. 2022, pp. 1–2.
[19] P. Fantini, N. Polino, A. Ghetti, and D. Ielmini, "Threshold switching by bipolar avalanche multiplication in ovonic chalcogenide glasses," *Adv. Electron. Mater.*, vol. 9, no. 7, Jul. 2023, Art. no. 2300037.
[20] D. Fugazza, D. Ielmini, S. Lavizzari, and A. L. Lacaita, "Random telegraph signal noise in phase change memory devices," in *Proc. IEEE Int. Rel. Phys. Symp.*, May 2010, pp. 743–749.
[21] P. Fantini, G. Betti Beneventi, A. Calderoni, L. Larcher, P. Pavan, and F. Pellizzer, "Characterization and modelling of low-frequency noise in PCM devices," in *IEDM Tech. Dig.*, Dec. 2008, pp. 1–4.
[22] 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, 1998.
[23] Y. Li, "Research and application of deep learning in image recognition," in *Proc. IEEE 2nd Int. Conf. Power, Electron. Comput. Appl. (ICPECA)*, Jan. 2022, pp. 994–999.
[24] F. Yang et al., "Fast IR-drop model of memristor crossbars and circuit compensation utilizing DTCO," *IEEE Trans. Electron Devices*, vol. 72, no. 8, pp. 4063–4069, Aug. 2025.
[25] L. X. Han et al., "Novel weight mapping method for reliable NVM based neural network," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Mar. 2021, pp. 1–6.
[26] Q. Liu et al., "33.2 A fully integrated analog ReRAM based 78.4 TOPS/W compute-in-memory chip with fully parallel MAC computing," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 500–502.
[27] G. B. Beneventi, M. Ferro, and P. Fantini, "1f noise in 45-nm RESET-state phase-change memory devices: Characterization, impact on memory readout operation, and scaling perspectives," *IEEE Electron Device Lett.*, vol. 33, no. 11, pp. 1559–1561, Nov. 2012.
[28] D. Ielmini, D. Sharma, S. Lavizzari, and A. L. Lacaita, "Reliability impact of chalcogenide-structure relaxation in phase-change memory (PCM) cells—Part I: Experimental study," *IEEE Trans. Electron Devices*, vol. 56, no. 5, pp. 1070–1077, May 2009.
[29] S. Lavizzari, D. Ielmini, D. Sharma, and A. L. Lacaita, "Reliability impact of chalcogenide-structure relaxation in phase-change memory (PCM) cells—Part II: Physics-based modeling," *IEEE Trans. Electron Devices*, vol. 56, no. 5, pp. 1078–1085, May 2009.
[30] C.-C. Chang et al., "NV-BNN: An accurate deep convolutional neural network based on binary STT-MRAM for adaptive AI edge," in *Proc. 56th ACM/IEEE Design Autom. Conf. (DAC)*, Jun. 2019, pp. 1–6.
[31] M. Trentzsch et al., "A 28nm HKMG super low power embedded NVM technology based on ferroelectric FETs," in *IEDM Tech. Dig.*, Dec. 2016, pp. 11.5.1–11.5.4.
[32] D. Bridarolli et al., "High-density multilevel 3D vertical resistive switching memory (VRRAM) for massively parallel in-memory computing," in *IEDM Tech. Dig.*, Dec. 2024, pp. 1–4.
[33] M. Y. Song et al., "High RA dual-MTJ SOT-MRAM devices for high speed (10ns) compute-in-memory applications," in *IEDM Tech. Dig.*, Dec. 2023, pp. 1–4.
[34] W.-C. Chien et al., "A study on OTS-PCM pillar cell for 3-D stackable memory," *IEEE Trans. Electron Devices*, vol. 65, no. 11, pp. 5172–5179, Nov. 2018.