A 115.1 TOPS/W, 12.1 TOPS/mm^{2}Computation-in-Memory using Ring-Oscillator based ADC for Edge AI

Singh, Abhairaj; Bishnoi, Rajendra; Kaichouhi, Ali; Diware, Sumit; Joshi, Rajiv V.; Hamdioui, Said

DOI
10.1109/AICAS57966.2023.10168647

Publication date
2023

Document Version
Final published version

Published in
AICAS 2023 - IEEE International Conference on Artificial Intelligence Circuits and Systems, Proceeding

2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS) | 979-8-3503-3267-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/AICAS57966.2023.10168647

A 115.1 TOPS/W, 12.1 TOPS/mm^2 Computation-in-Memory using Ring-Oscillator based ADC for Edge AI

Abhairaj Singh[1], Rajendra Bishnoi[1], Ali Kaichouhi[1], Sumit Diware[1], Rajiv V. Joshi[2], Said Hamdioui[1]
[1]Computer Engineering Laboratory, Delft University of Technology, Delft, The Netherlands
[1]Email: (a.singh-5, r.k.bishnoi, a.kaichouhi, s.s.diware, s.hamdioui)@tudelft.nl
[2]IBM Thomas J. Watson Research Centre, Yorktown Heights, NY 10598 USA: rvjoshi@ibm.us

***Abstract*—Analog *computation-in-memory* (CIM) architecture alleviates massive data movement between the memory and the processor, thus promising great prospects to accelerate certain computational tasks in an energy-efficient manner. However, data converters involved in these architectures typically achieve the required computing accuracy at the expense of high area and energy footprint which can potentially determine CIM candidacy for low-power and compact edge-AI devices. In this work, we present a memory-periphery co-design to perform accurate A/D conversions of analog matrix-vector-multiplication (MVM) outputs. Here, we introduce a scheme where select-lines and bit-lines in the memory are virtually fixed to improve conversion accuracy and aid a ring-oscillator-based A/D conversion, equipped with component sharing and inter-matching of the reference blocks. In addition, we deploy a self-timed technique to further ensure high robustness addressing global design and cycle-to-cycle variations. Based on measurement results of a 4Kb CIM chip prototype equipped with TSMC 40nm, a relative accuracy of up to 99.71% is achieved with an energy efficiency of 115.1 TOPS/W and computational density of 12.1 TOPS/mm^2 for the MNIST dataset. Thus, an improvement of up to 11.3X and 7.5X compared to the state-of-the-art, respectively.**

***Index Terms*—Computation-in-memory, analog-to-digital converters, analog computing, ring-oscillator**

I. Introduction

Analog *computation-in-memory* (CIM) has the potential to improve energy efficiency, provide massive parallelism and make the design compact for edge AI devices [1, 2]. CIM performs in-situ matrix-vector-multiplication (MVM) operations by leveraging circuit laws, such as Ohm's law and Kirchhoff's current law to realize analog computation with $\mathcal{O}(1)$ time complexity [3]. CIM produces column current I_{BL} proportional to the aggregated product of input data (IN) and neuron weights (W) represented by word-line (WL) voltages (V) and bitcell conductance (G) states, respectively [4]. In Fig. 1a, we show a crossbar that can be programmed to store W matrix as G states and CIM-based in-situ MVM operation details. However, the computational accuracy and efficiency greatly depend on the analog periphery, in particular, analog-to-digital converter (ADC) that converts an I_{BL} current into digital output for data communications among different CIM cores. The key challenges pertaining to CIM-based ADC design are the physical dimension and the energy efficiency while maintaining high conversion accuracy [5].

Recent works on ADC designs can be classified into three classes based on their intermediate physical quantity used for conversion: 1) voltage (V-ADC), 2) current (C-ADC), and 3) time-based ADCs (T-ADC). V-ADCs and T-ADCs are expensive as they typically consist of large components such as large hold capacitors [4, 6–10] or/and a series of sense amplifiers [4, 7–9] and large time-digital converters (TDCs) [10, 11], respectively, along with digital-to-analog converters (DACs) [8, 10–13] for providing reference signals to compute intermediate output calculations. In addition, these ADC classes require several power-hungry comparison cycles for their final digital output conversions. Furthermore, they require an additional conversion from a current (I_{BL}) to a voltage or time domain, thereby, introducing an additional source of inaccuracy. On the other hand, C-ADCs alleviate the need for an additional conversion and typically occupy less area, however, encounter the following serious challenges, as shown in Fig. 1b: (a) large I_{BL} range causes a variable differential voltage ($\Delta V = V_{SL} - V_{BL}$) leading to non-linear input-output (INxW - I_{BL}) characteristics and in addition, increases the energy consumption [4, 14, 15], (b) inaccuracies due to non-idealities such as process variations and wire parasitic delay mismatch [14, 16] necessitates the need to keep larger quantization margins, and thereby leading to a reduced dynamic range of the ADC. Therefore, to ensure the

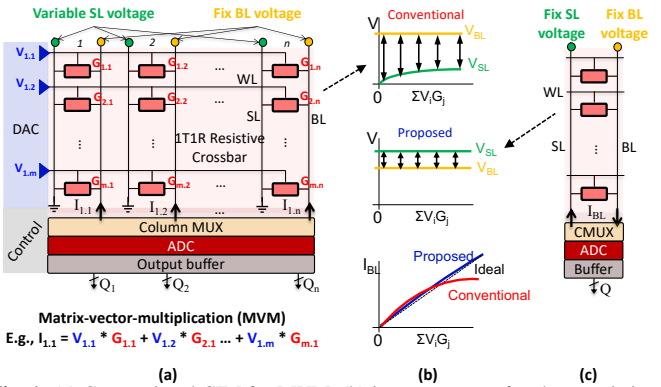

Fig. 1: (a) Conventional CIM for MVM, (b) impact on transfer characteristics, and (c) overview of the proposed scheme.

This work was supported by EU H2020 grant "DAIS" that received funding from ECSEL Joint Undertaking (JU) under grant agreement No 101007273.

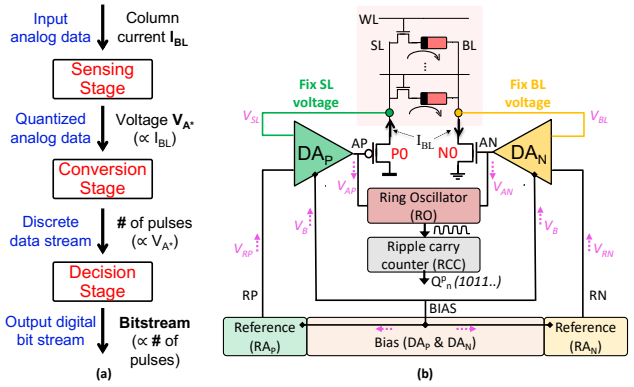

Fig. 2: (a) Proposed ADC methodology and (b) ADC design concept.

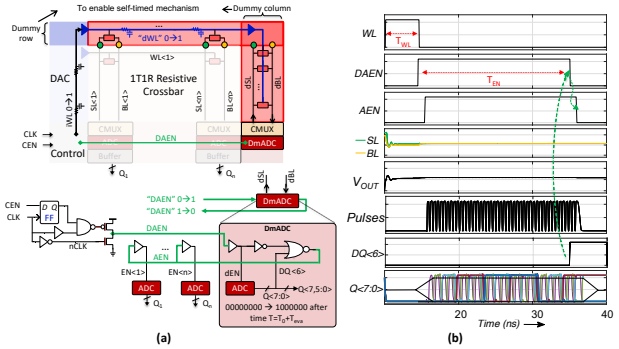

Fig. 3: (a) Proposed self-timed scheme and (b) timing-diagram of the MVM operation.

required computational accuracy ADCs contribute to major area overhead and energy consumption, severely degrading the overall efficiency of CIM.

To address these challenges, this paper presents an optimized co-design of the CIM array and current-based ADC to build an ultra-low power and compact CIM-based MVM engine, as summarized in Fig. 1b and Fig. 1c. The contributions of this paper are:

- A novel biasing scheme to obtain a linear activation function by virtually fixing the difference of select-line (SL) and bit-line (BL) to a constant voltage with precise inter-matching design techniques for accurate A/D conversion.
- An approach to improve the area/energy efficiency of the A/D conversion by introducing a voltage-controlled ring-oscillator (RO) and an asynchronous ripple carry counter (RCC) arranged in such a way that it captures high-speed RO pulse generations.
- A novel self-timed technique to address the impact of global design variations, cycle-to-cycle mismatch, and CIM array wire delay mismatch on computing accuracy by regulating the duration of the A/D conversion in each cycle.

Measurement results based on our 4Kb CIM chip prototype equipped with TSMC 40nm CMOS technology show that relative accuracy up to 99.71% and 94.74% realizing image classifications can be achieved for the MNIST and E-MNIST datasets, respectively, with an energy-efficiency of 115.1 TOPS/W and computational density of 12.1 TOPS/mm^2.

II. CIM Design

Next, we present our proposed design and A/D conversion characteristics for CIM-based MVM operations.

A. ADC Design Methodology

Fig. 2 provides an overview of our ADC methodology and working principle. Our ADC design comprises of three stages; (1) Sensing stage (SS): stabilizes nodes SL and BL and converts MVM output I_{BL} to proportional voltages V_{AP} and V_{AN}, (2) Conversion stage (CS): converts analog V_{AP} and V_{AN} values to a discrete number of pulses, and (3) Decision stage (DS): converts the discrete pulses to digital bit-streams. SS is realized using a combination of high-gain differential amplifiers DA_P and DA_N, SL driver PMOS P0, and BL driver NMOS N0. The combinations of DA_P/P0 and DA_N/N0 enable negative feedback to virtually fix SL and BL, respectively, such that $\Delta V = V_{SL} - V_{BL} = 50mV$ for all possible INxW combinations. In CS, the amplifier outputs V_{AP} and V_{AN} bias the header and footer, respectively, of the ring-oscillator RO that generates pulses proportional to I_{BL}. DS is implemented using an asynchronous ripple carry counter (RCC) that converts these pulses into bit-streams.

1) Self-timed ADC: To adapt to global variations and RC degradation, we introduce a dummy row and column that normalize the duration of the conversion in each cycle. This duration determines the time ADCs are allowed to generate pulses, hence implying a variation-prone duration parameter that requires a normalization step. To achieve this in each operating cycle, we introduce a self-timed mechanism in which a dummy ADC (DmADC) is allowed to capture a pre-determined MVM output of a dummy column programmed with known G states. This column is programmed to have all ON devices while all pass transistors are enabled, independent of IN. Fig. 3a shows the implementation of the self-timed scheme and Fig. 3b presents the timing diagram to illustrate its working. As the amplifiers establish $\Delta V = 50mV$, INxW is performed to generate I_{BL} in the selected columns and in the dummy column during the WL activation time T_{WL}. After an adequate settling time, signal DAEN enables the RO in the DmADC and triggers AEN signals to enable all ROs in the array ADCs to generate pulses. The normalization period T_{EN} i.e., the regulation of the activation time of the RO units in each ADC is achieved by disabling them when DmADC reaches a pre-determined count at its output. For instance, for a 6-bit resolution presented in Fig. 3b, DmADC resets DAEN to disable array ADCs at the time instant when DQ<7:0>=64 i.e., DQ<6> toggles to 1.

2) ADC Design Components: Fig. 4a shows the detailed implementation of our ADC design. The aim is to ensure linear input-output characteristics at each of the three ADC stages. In SS, with WL activation, DA_P and DA_N fixes $\Delta V = 50mV$ to generate $I_{BL} \propto$ INxW. Drivers P0/N0 always operate in a linear mode to produce linear $|V_{GS}|$ (i.e., V_{AP} and V_{AN}) $\propto$ INxW. Both DA_P and DA_N are designed using common-centroid matching techniques to accurately match load impedance and tolerate variation. Corresponding reference voltages V_{RP} (for

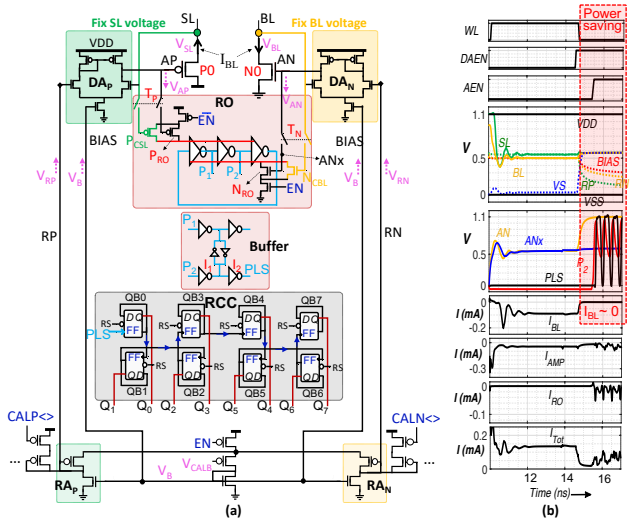

Fig. 4: (a) Detailed circuit design of the proposed ADC and (b) timing diagram of the critical signals.

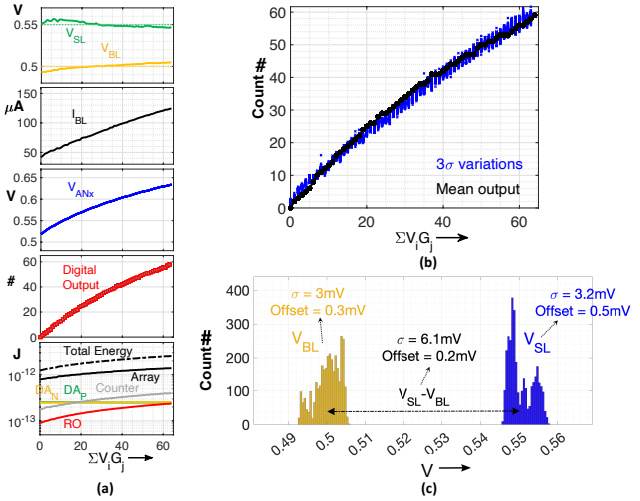

Fig. 5: (a) Input-output characteristics of the ADC, (b) variation analysis of MVM, and (c) of settled voltages V_{SL} and V_{BL}.

DA_P) and V_{RN} (for DA_N), bias voltage V_B share transistors and have a common bias signal to minimize systematic and random offsets. An inevitable variation in ΔV occurs due to the finite gain of DA_P and DA_N; a higher ΔV value at low currents and vice-versa, incurring a degradation in the linearity of I_{BL} (and V_{AP} and V_{AN}). This is compensated by current sources P_{CSL} and N_{CBL}, which introduce additional current in the RO proportional to the settled V_{SL} and V_{BL}, respectively. Calibration of the bias signal is performed using external bias voltage V_{CALB} and of reference voltages, V_{RP} and V_{RN}, using a series of diode-connected PMOS drivers driven by signals CALP<> and CALN<>, respectively.

Fig. 4b shows the timing diagram capturing the behavior of various critical signals to illustrate the working of the ADC. During WL activation, V_{SL} and V_{BL} are settled to produce proportional V_{AP} and V_{AN}, respectively. V_{AP} and V_{AN} are captured at nodes APx and ANx, respectively. Thereafter, enable signal AEN (and DAEN for DmADC) disconnects the AP and AN nodes (also, SL and BL nodes) from the RO through T_P and T_N transmission gates, respectively, and simultaneously activates the RO. The RO produces pulses $P_2 \propto V_{AP}$ and V_{AN} as these voltages bias the header P_{RO} and footer N_{RO}, respectively, of the RO in a current-mirroring configuration while always operating in the linear mode. Post-buffering improves the dynamic range of P_2, thereby generating a PLS signal. To count these high-frequency PLS pulses, adjacent flip-flops FF of RCC are arranged in such a way that it allows minimum path delays. To save power, WL disconnects the array after 5ns, a sufficient period T_{WL} determined by the settling time of SL/BL and AP/AN nodes as required by the amplifiers. This significantly reduces the duration of the flow of I_{BL} in the array, thereby, reducing massive dynamic power at the cost of additional hold caps at APx and ANx. Fig. 4b highlights the impact of this power saving *i.e.*, reduction of $I_{BL} \sim 0$ after WL deactivation. Note that WLs are deactivated after the EN signal disconnects the AP and AN nodes from the RO.

B. ADC Input-Output Characteristics

Fig. 5a shows the input-output characteristics for the accumulation of 64 rows with 1-bit IN and 1-bit W elements. Virtually fixing ΔV=50mV allows $I_{BL} \propto$INxW (or $\Sigma V_i \cdot G_j$) and V_{AP} and V_{AN} follow I_{BL} to generate proportional digital outputs. The compensation scheme described previously linearizes the number of pulses generated by the RO, albeit with a slight variation in ΔV, as depicted in the final digital output. The total energy consumption increases with $\Sigma V_i \cdot G_j$ corresponding to the increased currents in the array (I_{BL}), RO, and counter, while DA_P and DA_N consume nearly constant energy. Fig. 5b presents 3σ variation analysis of the MVM output. The spread (σ) of the settled ΔV is 6.2mV and the combined offset is 0.2mV.

III. CIM Implementation and Characterisation

A. Chip Prototype and Experimental Setup

Fig. 6a presents the microscopic view of the CIM implementation equipped with TSMC 40nm CMOS technology and Fig. 6b shows the experimental setup. The chip prototype comprises a 64x64 crossbar array built using 1-transistor-1-resistor (1T1R) bitcell configuration. Here, to conduct the characterisation of our ADC, the conductance of the bitcell (in 1R) is pre-programmed using fixed NMOS-based resistors to mimic the resistive properties of a 1-bit storage device i.e., a low (LCS) and a high conductance state (HCS) corresponding to logic 0 and 1, respectively. The programming sequence is such that all possible 64 combinations of W vector i.e., from the minimum to the maximum number of ON devices in a column are pre-programmed sequentially to the 64 columns in the crossbar array. This allows a complete range of conductance values possible for the MVM operation.

B. ADC Characterisation Results

The chip prototype is characterized to determine the functional voltage boundaries and measure energy consumption and latency per conversion cycle of the ADC. Fig. 7a presents

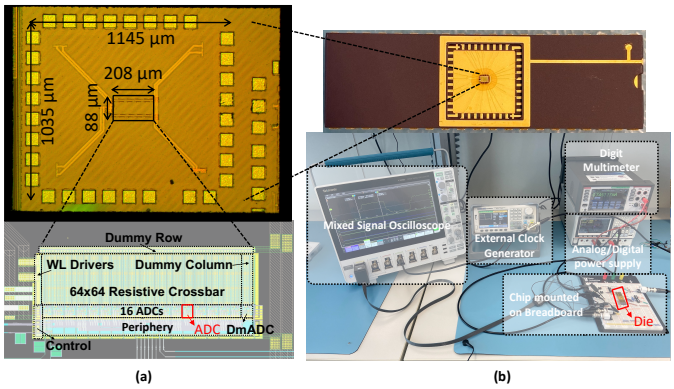

Fig. 6: (a) Microscopic view of the fabricated chip with CIM layout and (b) experimental setup with the prototype die.

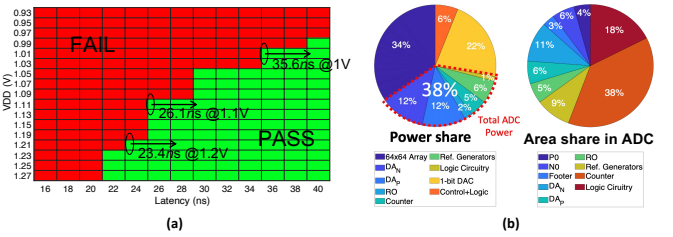

Fig. 7: (a) Shmoo plot and (b) component-wise energy and area.

the shmoo plot along with the conversion speed for a 6-bit resolution. As expected, the conversion speed increases with the voltage of operation with a conversion time of 26.1 *n*s at 1.1V. Fig. 7b presents the component-wise energy breakdown of the MVM operation per conversion cycle, where ADC consumes an average energy of 3.3 *p*J which is roughly 38% of the total energy. In addition, the figure shows component-wise area utilization in our ADC.

IV. System-level Results

A. System-level Validation

We evaluate the benefits of our CIM design on image classification applications, using CNN-based Lenet-5 for MNIST and E-MNIST datasets.

Fig. 8a presents the validation of our design for different ADC bit-resolutions. The resolution is adjusted by varying the maximum number of allowed pulse generation that corresponds to the dummy column. For instance, in a 4-bit ADC, toggling DQ<4> to 1 disables DAEN and subsequently, AEN signals. It can be seen that a low-resolution ADC allows better energy efficiency at the expense of accuracy and latency. We introduce a figure of merit (FoM) which describes the energy efficiency of a system for accurate computations and show the comparative merits of different ADC bit-resolutions. Fig. 8b shows that a mid-range resolution offers a good trade-off between accuracy and energy efficiency for image classification applications. Based on simulation results equipped with measured latency and energy consumption during an MVM operation, this work can provide as high as 10.9X FoM improvement compared to [17] when evaluated for the MNIST dataset, as shown in Fig. 8c.

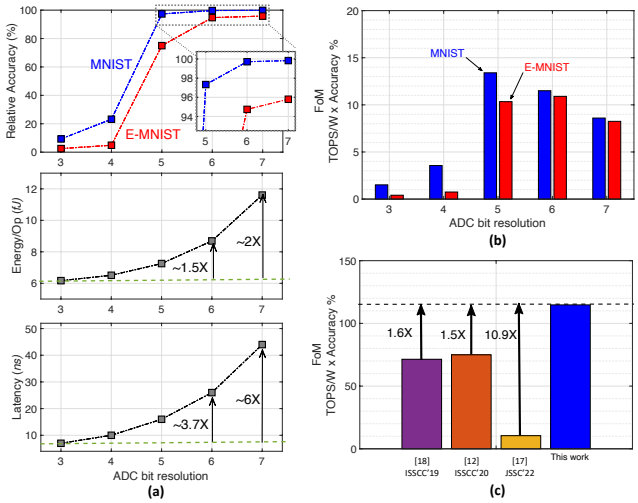

Fig. 8: (a) Accuracy and efficiency, and (b) FoM with different ADC resolutions. (c) FoM comparison with MNIST dataset.

Parameters	[18] ISSCC-'19	[12] ISSCC-'20	[4] ISSCC-'21	[17] JSSC-'22	[19] ISSCC-'22	**This work (6b ADC)**
Technology	55nm	130nm	22nm	14nm	40nm	40nm
Voltage	1V	4.2V	0.8V	0.8V	0.9V	1.1V
Storage device	SRAM	RRAM	RRAM	PCM	PCM	Resistive
Storage	2b	analog	1b	analog	analog	1b
Bitcell	Twin 8T	2T2R	1T1R	8T4R	1T1R	1T1R
Capacity	4Kb	158Kb	4Mb	65.6Kb	2Mb	4Kb
Accumulation	9	-	1024	256	256	64
DOUT	3b	8b	10b	8b	11b	6b
ADC resolution	1b (SA)	1b-8b	1b (VSA)	8b	4b	3b-7b
Latency	3.2	51.1	10.3	130	8.6	26.1
TOPS/W	72.1	78.4	47.3	10.5	57.6	**115.1**
TOPS/mm^2	-	-	-	1.6	-	**12.1**
Accuracy						
MNIST	99.02%	95.6%	-	99.7%	-	99.7%
E-MNIST	-	-	-	-	-	94.7%

TABLE I: Comparison table. - implies data is not reported.

B. Comparison Results

A detailed comparison is presented in Table I. We show that for our 6-bit resolution ADC, an improvement of 11.3X and 7.5X can be achieved in terms of TOPS/W and TOPS/mm^2, respectively, compared to [17] with comparable accuracy on the MNIST database. The conversion time is longer as compared to [18], [4] and [19]. However, the energy efficiency is improved by 1.6X, 2.4X, and 2X, respectively, owing to the reduced number of components used in our proposed memory array-periphery co-design scheme.

V. Conclusion

This work presents a novel memory-periphery co-design to perform accurate A/D conversions of analog MVM outputs with high energy efficiency and computational density. The paper introduces a scheme where array access lines can be virtually fixed to improve the accuracy of the MVM operation and paves the path for a compact and ultra-low power ADC design. In addition, the measurement results of the chip prototype validate the ADC design and derive its input-output characteristics. A relative accuracy of up to 99.71% is achieved with an energy efficiency of 115.1 TOPS/W and computational density of 12.1 TOPS/mm^2 for the MNIST dataset, thus, making it a suitable implementation for executing MVM operations in low-power edge AI devices.

REFERENCES

[1] S. Hamdioui *et al.*, "Applications of computation-in-memory architectures based on memristive devices," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2019, pp. 486–491.

[2] A. Sebastian *et al.*, "Memory devices and applications for in-memory computing," *Nature nanotechnology*, vol. 15, no. 7, pp. 529–544, 2020.

[3] A. Shafiee *et al.*, "ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, pp. 14–26, 2016.

[4] C.-X. Xue *et al.*, "16.1 A 22nm 4Mb 8b-precision ReRAM computing-in-memory macro with 11.91 to 195.7 TOPS/W for tiny AI edge devices," in *International Solid-State Circuits Conference-(ISSCC)*, vol. 64, 2021, pp. 245–247.

[5] A. Singh *et al.*, "Low-power Memristor-based Computing for Edge-AI Applications," in *ISCAS*, 2021, pp. 1–5.

[6] C. Liu *et al.*, "A spiking neuromorphic design with resistive crossbar," in *Proceedings of the 52nd Annual Design Automation Conference*, 2015, pp. 1–6.

[7] C. Liu *et al.*, "A memristor crossbar based computing engine optimized for high speed and accuracy," in *Computer Society Annual Symposium on VLSI (ISVLSI)*, 2016, pp. 110–115.

[8] M. E. Sinangil *et al.*, "A 7-nm compute-in-memory SRAM macro supporting multi-bit input, weight and output and achieving 351 TOPS/W and 372.4 GOPS," *Journal of Solid-State Circuits*, vol. 56, no. 1, pp. 188–198, 2020.

[9] S. Xie *et al.*, "16.2 eDRAM-CIM: Compute-in-memory design with reconfigurable embedded-dynamic-memory array realizing adaptive data converters and charge-domain computing," in *International Solid-State Circuits Conference-(ISSCC)*, vol. 64, 2021, pp. 248–250.

[10] S. Hong *et al.*, "Low voltage time-based matrix multiplier-and-accumulator for neural computing system," *Electronics*, vol. 9, no. 12, p. 2138, 2020.

[11] D. S. Kang *et al.*, "Time-Based Compute-in-Memory for Cryogenic Neural Network With Successive Approximation Register Time-to-Digital Converter," *Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 8, no. 2, pp. 128–133, 2022.

[12] 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 *International Solid-State Circuits Conference-(ISSCC)*, 2020, pp. 500–502.

[13] X. Si *et al.*, "15.5 A 28nm 64Kb 6T SRAM computing-in-memory macro with 8b MAC operation for AI edge chips," in *International Solid-State Circuits Conference-(ISSCC)*, 2020, pp. 246–248.

[14] A. Singh *et al.*, "SRIF: Scalable and reliable integrate and fire circuit ADC for memristor-based CIM architectures," *Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 5, pp. 1917–1930, 2021.

[15] W.-H. Chen *et al.*, "A 65nm 1Mb nonvolatile computing-in-memory ReRAM macro with sub-16ns multiply-and-accumulate for binary DNN AI edge processors," in *International Solid-State Circuits Conference-(ISSCC)*, 2018, pp. 494–496.

[16] A. Kneip *et al.*, "Impact of analog non-idealities on the design space of 6T-SRAM current-domain dot-product operators for in-memory computing," *Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 5, pp. 1931–1944, 2021.

[17] R. Khaddam-Aljameh *et al.*, "HERMES-core—A 1.59-TOPS/mm 2 PCM on 14-nm CMOS in-memory compute core using 300-ps/LSB linearized CCO-based ADCs," *Journal of Solid-State Circuits*, vol. 57, no. 4, pp. 1027–1038, 2022.

[18] X. Si *et al.*, "24.5 A twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning," in *International Solid-State Circuits Conference-(ISSCC)*, 2019, pp. 396–398.

[19] W.-S. Khwa *et al.*, "A 40-nm, 2M-cell, 8b-precision, hybrid SLC-MLC PCM computing-in-memory macro with 20.5-65.0 TOPS/W for tiny-Al edge devices," in *International Solid-State Circuits Conference-(ISSCC)*, vol. 65, 2022, pp. 1–3.