A Charge-Sharing-Based Ternary 17T1C SRAM for Energy-Efficient CIM Realization with MAC Operations

Pradyumna Vellanki*, Chi-Lin Lee*, L S S Pavan Kumar Chodisetti*, and Chua-Chin Wang†
*Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan 80424
†Institute of Integrated Circuit Design, National Sun Yat-sen University, Kaohsiung, Taiwan 80424
Email: ccwang@ee.nsysu.edu.tw

***Abstract*—This work presents a ternary compute-in-memory (T-CIM) architecture based on a novel ternary 17T1C SRAM cell implemented in TSMC 40-nm CMOS technology. A 1-kb (32 × 32) ternary SRAM array performs in-memory ternary multiply-and-accumulate (MAC) operations using charge-sharing-based accumulation scheme along the computing line. The proposed bit-cell supports reliable ternary write and read/multiplication operations with stacked metal–oxide–metal (MOM) capacitors, enabling area-efficient computation and high linearity. The complete T-CIM system occupies a core area of 0.211 mm^2, and dissipates 0.543 mW at 1 GHz system clock. Post-layout simulations across all PVT corners demonstrate robust operation, making the proposed T-CIM suitable for energy-efficient edge AI and neural network acceleration.**

***Index Terms*—Ternary, charge-sharing, SRAM, T-CIM, MAC**

I. Introduction

Traditional AI and neural-network hardware architectures are predominantly based on the von Neumann computing model, where memory and computational units such as the arithmetic logic unit (ALU) are physically separated. This separation leads to the well-known von Neumann bottleneck, where massive data movement between memory and the ALU increases latency and degrades throughput and power efficiency. To address these limitations, researchers have increasingly focused on computing-in-memory (CIM) architectures, e.g., [1], [2]. By performing operations directly within memory arrays, CIM systems significantly reduce data-transfer overhead. In these architectures, data storage, retrieval, and computation are integrated within the same memory structure, supported by peripheral circuits that manage logic operations and coordinate functional transitions. Among various CIM implementations, static random-access memory (SRAM) based CIM architectures are particularly attractive due to their compatibility with standard CMOS processes, high operating speed, and robust read/write characteristics [3]–[5]. However, most existing SRAM-CIM designs primarily target binary neural networks (BNNs), which simplify hardware implementation by restricting weights and activations to binary values. Despite their hardware efficiency, the limited numerical precision of BNNs often leads to noticeable accuracy degradation in practical applications, especially for complex inference tasks [6]. Ternary neural networks (TNNs), which utilize ternary weights and activations from the set (+1, -1, 0), provide an effective trade-off between computational accuracy and hardware efficiency. Compared to binary networks, TNNs achieve higher inference accuracy while maintaining relatively low computational complexity. This characteristic makes ternary representations particularly suitable for CIM architectures, where parallel multiply-and-accumulate (MAC) operations can be efficiently executed within memory arrays.

In this work, a charge-sharing-based ternary computing-in-memory (T-CIM) architecture is proposed, featuring a novel 17-Transistor–1-Capacitor (17T1C) ternary SRAM bit-cell. The proposed bit-cell is constructed using dual symmetric single-ended 6T-SRAM cells combined with four low-threshold-voltage computing transistors, enabling direct ternary multiplication and accumulation within the memory array. A metal–oxide–metal (MOM) capacitor is employed as the computation capacitor and is vertically stacked above the SRAM cell, effectively eliminating additional area overhead while preserving high linearity in MAC operations.

II. Energy-efficient Ternary CIM Architecture

Fig. 1 illustrates the proposed T-CIM architecture. The system integrates a 1-kb (32 × 32) ternary SRAM array with peripheral control circuits to support in-memory ternary MAC operations. Ternary weight values are stored in the SRAM array, while ternary input activations are applied simultaneously, enabling parallel MAC computations along each column.

The overall architecture consists of the 1-kb Ternary SRAM array, Row and Column decoders, an SRAM Control circuit, MAC Charge sharing control circuit, a Column selector, and an Output buffer. The SRAM control circuit manages write and read/multiplication operations, while the MAC control circuit coordinates charge-sharing-based accumulation. The Output Buffer follows the design principles of previous works [7]–[10] to ensure reliable signal readout. The structure of the ternary SRAM array is illustrated in Fig. 2.

Taiwan Semiconductor Research Institute, National Institutes of Applied Research is acknowledged for the assistance of EDA tool support.

This work was partially funded by National Science and Technology Council (NSTC), Taiwan, under grant NSTC 115-2923-E-110-006-, 115-2221-E-110-075-, 114-2218-E-110-009-, and 112-2923-E-006-003-MY3.

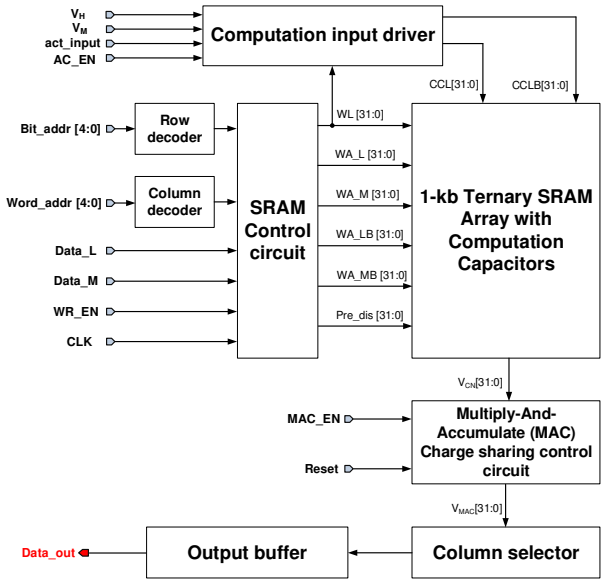


Fig. 1. Architecture of the proposed T-CIM system

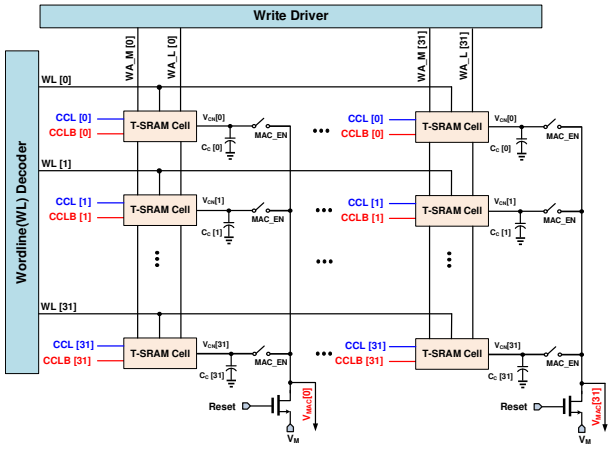


Fig. 2. Illustration of the proposed 1-kb ternary SRAM array

TABLE I
CONTROL INPUTS FOR WRITE OPERATIONS

Modes	Write +1	Write -1	Write 0
WL	1	1	1
WR_EN	1	1	1
WA_L	0	1	1
WA_LB	1	0	0
WA_M	1	1	0
WA_MB	0	0	1

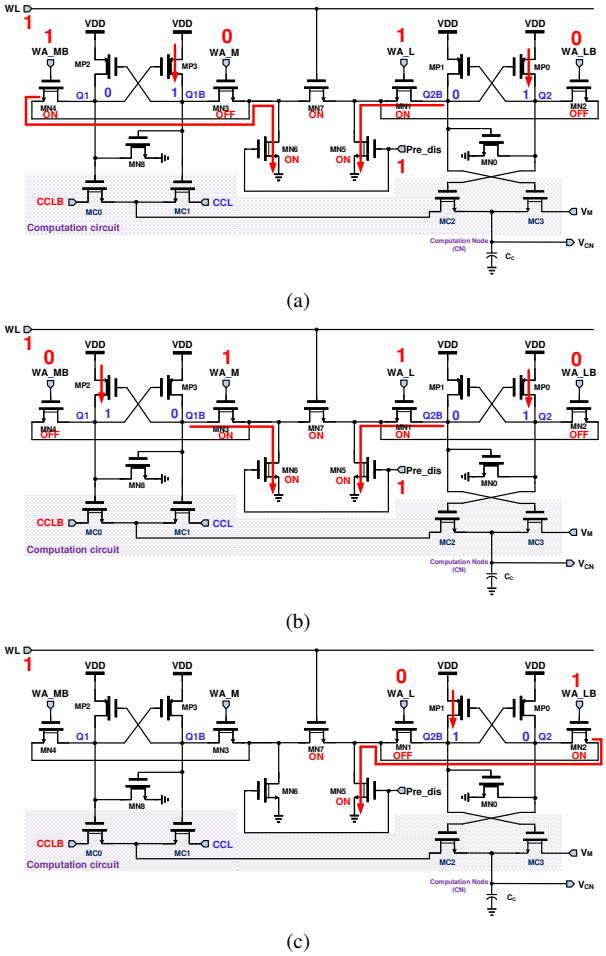


Fig. 3. Schematic of 17T1C Ternary SRAM cellcell and write operations: (a) Write +1, (b) Write -1, (c) Write 0

A. Ternary 17T1C SRAM

The proposed 17T1C SRAM cell is capable of storing ternary weight values of +1, -1, and 0. The voltages VH, VL, and VM represent the ternary logic levels +1, -1, and 0, and are chosen as 0.6 V, 0.3 V, and 0 V due to VDD = 0.9 V for 40-nm CMOS node. The 17T1C SRAM cell supports two primary operations: write and multiplication/read (Read/MULT) operations.

1) Write operations: The schematic diagram of the proposed ternary SRAM cell, along with the write operations for the three ternary weight values, is shown in Fig. 3(a), 3(b), 3(c). Table I details the control signals provided by the SRAM control circuit during the write operation. Fig. 4(a) presents the corresponding logic states within the 17T1C SRAM cell for each ternary weight.

2) Read/Multiplication operations: Read and multiplication operations are performed by transistors T0–T3 in Fig. 3, which execute the Read/MULT function on the stored ternary weight using ternary activation inputs applied through the computation control line (CCL) and its complement (CCLB). The control inputs for computation in a single cell are listed in Fig. 4(b). The resulting multiplication output is represented as an analog voltage at the computation node (CN) and is stored on the computation capacitor C_C. In this design, C_C is selected as 6 fF to achieve a balance between energy efficiency and computation latency. The corresponding multiplication results for all combinations of ternary weights and activations are summarized in Table II.

To reduce area overhead, MOM capacitors are employed as

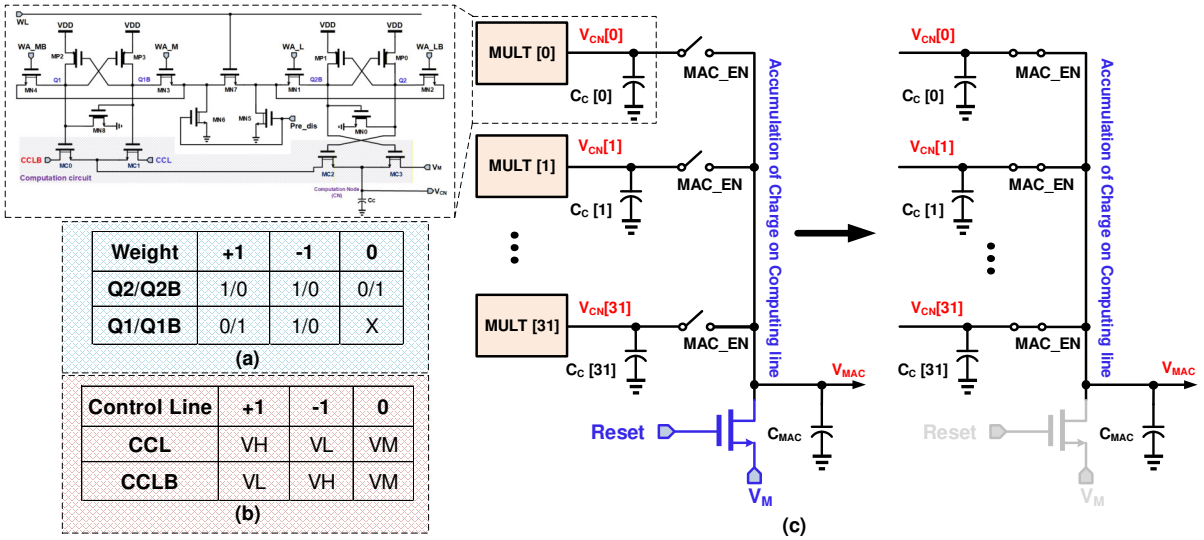


Fig. 4. (a) State table for ternary weights; (b) Control inputs for computations in the cell; (c) Charge-sharing-based accumulation on the computing line

TABLE II
MULTIPLICATION OUTPUTS FOR TERNARY WEIGHT-ACTIVATION COMBINATIONS

Weight * Activation = V_{CN}	+1	-1	0
+1	VH	VL	VM
-1	VL	VH	VM
0	VM	VM	VM

computation capacitors. These MOM capacitors are vertically stacked on top of the ternary SRAM cells, requiring no additional layout area. Moreover, MOM capacitors exhibit excellent linearity and matching characteristics, making them well suited for accurate charge-domain MAC operations.

B. Accumulation of the charge on the computing line

The accumulation operation is implemented using a charge-sharing mechanism controlled by dedicated switches. When the multiplication result of a bit-cell is +1, the voltage V_{MAC} on the computing line increases; when the result is -1, V_{MAC} decreases; and when the result is 0, V_{MAC} emains unchanged. Prior to each MAC operation, the Reset signal initializes the computing line voltage to VM, corresponding to a MAC value of zero. Fig. 4(c) illustrates the charge-sharing based accumulation process. After each memory cell completes the multiplication of the stored weight and the applied activation, the MAC_EN signal enables charge sharing by turning on the switches along the computing line. The accumulated voltage V_{MAC} is determined according to the charge conservation principle and is given by

$$V_{MAC} = \frac{\sum_{i=0}^{31}(C_C \times V_{CN}[i])}{32 \times C_C + C_{MAC}} \quad (1)$$

where $V_{CN}[i]$ the computation-node voltage of the i-th bit-cell, and C_{MAC} represents the parasitic capacitance of the computing line. The resulting MAC output voltage exhibits high linearity and corresponds to a range from -32 to +32.

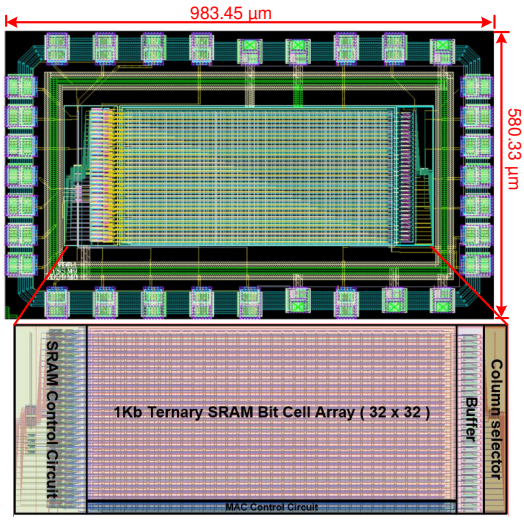


Fig. 5. Layout of the proposed T-CIM

III. POST-LAYOUT SIMULATION ANALYSIS

The proposed Ternary CIM is designed using TSMC 40-nm technology. The layout of the proposed design is shown in Fig. 5. The total chip area and core area are 0.57 mm^2 and 0.211 mm^2, respectively. The power dissipation of the chip is 0.543 mW at a system clock frequency of 1 GHz.

The power consumption of the proposed 17T1C ternary SRAM cell was evaluated under all-PVT-corner post-layout simulations. During write operations, the power consumption for storing ternary weights of +1, -1, and 0 is 1.1472 μW, 1.1472 μW, and 0.5634 μW, respectively. During Read/MULT operations, the corresponding power consumption values are 1.7643 μW, 1.6234 μW, and 1.3895 μW for +1, -1, and 0, respectively. The write operation latency for storing +1, -1, and 0 is 0.01221 ns, 0.0112 ns, and 0.01049 ns, respectively. The read/multiplication operation latency for +1, -1, and 0 is

TABLE III
PERFORMANCE COMPARISON OF CIM

	[11]	[12]	[13]	[14]	Our work
Activation	Ternary	Ternary	Ternary	Binary	Ternary
Weight	Binary	Binary	Ternary	Binary	Ternary
Technology	65-nm	12-nm	12-nm	28-nm	40-nm
Bit cell type	SRAM (type N.A.)	17T SRAM (2-bit)	4T2M SRAM (Multi-bit)	1T1C+3T1C eDRAM	17T1C SRAM (2-bit)
Cell Area (μm^2)	N.A.	N.A.	27.6	N.A.	4.495
Core area (mm^2)	3.9	9	N.A.	14	0.211
Clock frequency (MHz)	400	100	1000	8250	1000
Supply volatge	1.0	0.8	0.8	1.0	0.9
Energy efficiency (TOPS/W)	2.3	8.8	33.7	46	112.76
Area efficiency (TOPS/mm^2)	0.365	0.019	41.62	N.A.	40.15

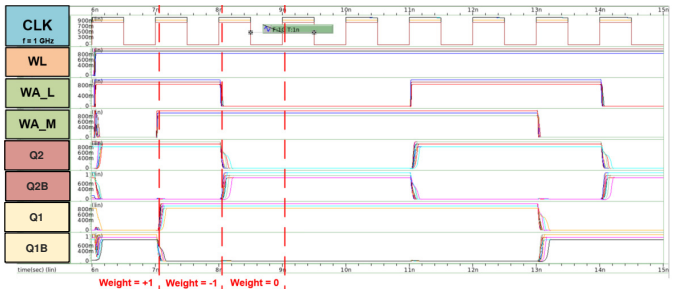


Fig. 6. Write operation for ternary weights across 45 PVT corners

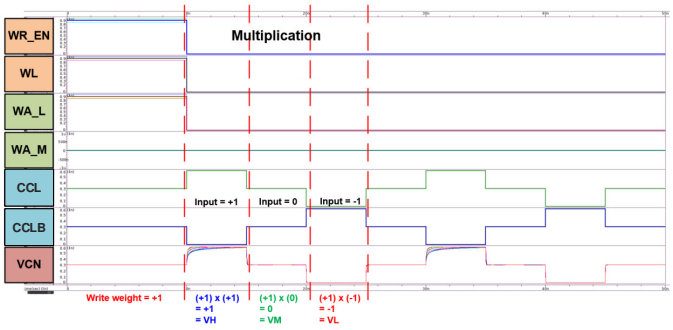


Fig. 7. Read/multiplication operation for +1 weight with +1, 0, and -1 inputs across 45 PVT corners

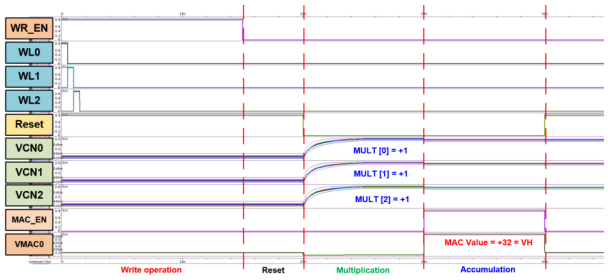


Fig. 8. Charge-sharing accumulation results across 45 PVT corners

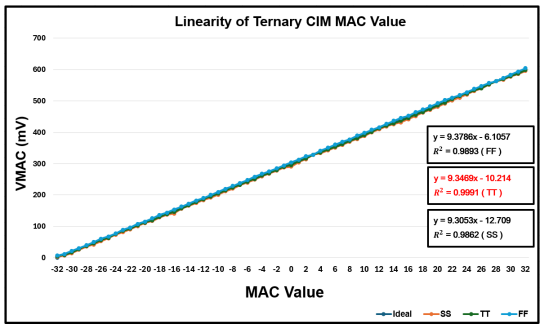


Fig. 9. Linearity of charge-sharing accumulation under FF, TT, and SS PVT corners

3.12 ns, 0.007 ns, and 0.003 ns, respectively, evaluated under the worst-case process–voltage–temperature (PVT) corner.

Fig. 6 illustrates the write operation waveforms for all ternary weight values across 45 PVT corners, validating the correctness of the weight state table shown in Fig. 4(a). Fig. 7 shows an example of the Read/MULT operation results for +1 weight and input activations, confirming the correct multiplication behavior summarized in Table. II. The charge-sharing-based accumulation operation described in Section II-B was also verified through post-layout simulations. Fig. 8 illustrates the accumulation operation across 45 PVT corners. The linearity of the accumulated output voltage V_{MAC}, derived from Eqn. (1) , was evaluated under three critical PVT corners: FF, TT, and SS, over a MAC range from -32 to +32. The extracted coefficients of determination (R^2) are 0.9893, 0.9991, and 0.9862 for the FF, TT, and SS corners, respectively. The fitted trend lines and corresponding simulation results, shown in Fig. 9, demonstrate high linearity across all evaluated PVT corners. These results confirm the robustness of the proposed charge-sharing accumulation scheme against process and operating variations. Table. III summarizes the performance comparison between the proposed T-CIM architecture and recently reported CIM designs.

IV. CONCLUSION

This work presented a charge-sharing-based T-CIM architecture using a 17T1C ternary SRAM bit-cell to enable in-memory ternary multiply-and-accumulate operations with high linearity and area efficiency. Implemented in a 40-nm CMOS process, a 1-kb (32 × 32) T-CIM operates at 1 GHz with 0.543 mW power consumption, achieving 112.76 TOPS/W energy efficiency and 40.15 TOPS/mm^2 area efficiency, while post-layout simulations across process–voltage–temperature corners confirm robust ternary write, read/multiplication, and charge-sharing accumulation over a MAC range of -32 to +32.

REFERENCES

[1] Q. Dong, S. Jeloka, M. Saligane, Y. Kim, M. Kawaminami, and A. Harada, "A 4 + 2T SRAM for searching and in-memory computing with 0.3-V VDDmin," *IEEE J. of Solid-State Circuits*, vol. 53, no. 4, pp. 1006-1015, April 2018.

[2] A. Agrawal, A. Jaiswal, C. Lee, and K. Roy, "X-SRAM: enabling in-memory boolean computations in CMOS static random access Memories," *IEEE Trans. on Circuits and Syst. I: Regular Papers*, vol. 65, no. 12, pp. 4219-4232, Dec. 2018.

[3] C.-C. Wang, C.-Y. Huang and C.-H. Yeh, "SRAM-Based computation in memory architecture to realize single command of add-multiply operation and multifunction," in *Proc. 2021 IEEE Int. Symp. on Circuits and Syst. (ISCAS)*, Korea, 2021, pp. 1-4.

[4] L. S. S. P. K. Chodisetti, C.-Y. Lo, J. S. Walling, L. C. Gunnam, and C.-C. Wang, "A 75.47 TOPS/W 4-kb single-ended SRAM-based compute-in-memory architecture using 16-nm FinFET technology," *J. of Signal Processing Syst.*, vol. 98, no. 26, 2026, doi: 10.1007/s11265-026-01998-7.

[5] L. S. S. P. K. Chodisetti, Y.-C. Lin, Y.-C. Chung, J. S. Walling, Y. Yi, and C.-C. Wang, "A high-efficiency 16-kb SRAM-based CIM architecture with automatic write-back using 28-nm CMOS technology," in *Proc. 2026 14th International Electrical Engineering Congress (iEECON)*, Pattaya, Thailand, 2026, pp. 1-5.

[6] C.-C. Wang, L. K. S. Tolentino, C.-Y. Huang, and C.-H. Yeh, "A 40-nm CMOS multifunctional computing-in-memory (CIM) using single-ended disturb-free 7T 1-Kb SRAM," *IEEE Trans. on Very Large Scale Integration (VLSI) Syst.*, vol. 29, no. 12, pp. 2172-2185, Dec. 2021.

[7] C.-C. Wang, L. S. S. P. K. Chodisetti, B.-H. Liao, P. Vellanki, and T.-J. Lee, "A 1–6.5 Gbps dual-loop CDR design with coarse-fine Tuning VCO and modified DQFD," *Microelectronics Journal*, vol. 151, pp. 106355, Sep. 2024, doi: 10.1016/j.mejo.2024.106355.

[8] C.-C. Wang, L. S. S. P. K. Chodisetti, D. S. Kamarajugadda, O. L. J. A. Jose, and P. Vellanki, "A 15.13 mW 3.2 GHz 8-bit carry look-ahead adder using single-phase all-N-transistor logic," *Integration*, vol. 98, pp. 102234, Sep. 2024, doi: 10.1016/j.vlsi.2024.102234.

[9] L. S. S. P. K. Chodisetti, W.-C. Cheng, T.-J. Lee, and C.-C. Wang, "A dual-loop clock and data recovery system with HLD for extended lock-in range demand," in *Proc. 2025 International Conference on Electronics, AI and Computing (EAIC)*, India, pp. 1-4, 2025.

[10] L. S. S. P. K. Chodisetti, J.-Y. Li, P. Vellanki, Y.-J. Hung and C.-C. Wang, "A wide bandwidth TIA with low power consumption for PNN prototyping," in *Proc. 2025 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, Republic of Korea, 2025.

[11] K. Ando, K. Ueyoshi, K. Orimo, H. Yonekawa, S. Sato, H. Nakahara, M. Ikebe, T. Asai, S. Takamaeda-Yamazaki, T. Kuroda, and M. Motomura, "BRein memory: A 13-layer 4.2 K neuron/0.8 M synapse binary/ternary reconfigurable in-memory deep neural network accelerator in 65 nm CMOS," in *Proc. 2017 Symp. on VLSI Circuits*, Kyoto, Japan, pp. C24-C25, 2017.

[12] S. Okumura, M. Yabuuchi, K. Hijioka, and K. Nose, "A ternary based bit scalable, 8.80 TOPS/W CNN accelerator with many-core processing-in-memory architecture with 896K synapses/mm2," in *Proc. 2019 Symp. on VLSI Tech.*, Japan, pp. C248-C249, 2019.

[13] Z. Tong, C. Li, C. Wang, S. Zhao, Q. Peng, Z. Yan, S. Zhang, D. Zhou, Z. Wang, X. Lin, and W. Zhao "BSTCIM: A balanced symmetry ternary fully digital in-MRAM computing macro for energy efficiency neural network," *IEEE Trans. on Circuits and Syst. I: Regular Papers*, vol. 71, no. 12, pp. 6114-6127, Dec. 2024.

[14] S. An, H. Jeong, S. Kim, K. Park, D. Kim, and K. Lee, "A 46 TOPS/W in-/near-memory computing processor for large language model with extended sparse attention," in *Proc. 2025 IEEE Int. Symp. on Circuits and Syst. (ISCAS)*, United Kingdom, pp. 1-5, 2025.