# A 30.7 TOPS/W Sparsity-Aware Analog-Digital Hybrid eDRAM CIM by Effective Row Activation with Simultaneous Multi-Row-Multi-Task Control

*Hoichang Jeong[1], *Seungbin Kim[1], Jueun Jung[1], and Kyuho Jason Lee[2]
*Equally-Credited Authors; [1]Department of Electrical Engineering, UNIST, Republic of Korea
[2]Department of Electrical and Electronic Engineering, Yonsei University, Republic of Korea
jhc3261@unist.ac.kr / kyuho.jsn.lee@yonsei.ac.kr

***Abstract*—This paper presents a sparsity-aware analog-digital hybrid eDRAM computing-in-memory (CIM) processor for high energy efficiency with three key innovations: (1) input activation grouping convolution that *completely skips zero computations* by activating only effective rows, achieving 4.59× higher computation efficiency; (2) a hybrid-CIM macro which integrates reversed-MAC near-memory logic and a SAR-Flash ADC for *energy-efficient MAC operations* in both digital and analog domains, increasing macro efficiency by 2.39×; and (3) in-macro multi-row-multi-task for *simultaneous refresh/update operations* during in-memory computation, enhancing system efficiency by 1.28×. Fabricated in 28 nm CMOS technology, the proposed CIM achieves the highest macro energy efficiency of 47.5 TOPS/W. Furthermore, it outperforms state-of-the-art CIM processors with 1.55× and 10.37× improvements in benchmark energy efficiency on ResNet-18 and VGGNet-16, respectively.**

***Keywords—computing-in-memory (CIM), embedded DRAM (eDRAM), hybrid computing, sparsity-aware.***

## I. Introduction

Deep neural networks (DNNs) are widely deployed in everyday applications. However, conventional von Neumann architecture suffers from the memory wall when accelerating DNNs, which require massive computation and memory bandwidth. Computing-in-memory (CIM) has emerged as a promising alternative, offering high throughput and improved energy efficiency [1-8].

Nevertheless, previous CIMs face several critical challenges as depicted in Fig. 1. First, input activations (IAs) and weights in DNNs exhibit tremendous sparsity with random patterns. Thus, without considering sparsity, direct mapping of convolution into the CIM macro and entire row activation for multiply-and-accumulate (MAC) operations results in significant energy waste, with 83% of zero computation [2–8], indicating a low element-wise effective computation ratio (EECR). Second, the distinct architectures of analog CIM (A-CIM) and digital CIM (D-CIM) have notable drawbacks. A-CIMs [2-3] ensure high throughput by activating entire rows simultaneously but require high-resolution ADCs to maintain computational accuracy, leading to enormous power consumption. In contrast, D-CIMs [4-5] consume less power without ADC, but suffer from significantly low throughput due to sequential row-by-row operations. Third, embedded DRAM (eDRAM) necessitates periodic refresh operations to prevent bit-flip errors. Therefore, previous eDRAM-based CIMs [2-6] encounter inevitable stalls to avoid collisions among three mutually-exclusive operations (*computation*, *refresh*, and *update*), which reduce energy efficiency and increase latency.

To address these challenges, a sparsity-aware analog-digital hybrid eDRAM CIM is proposed with three features (Fig. 2): 1)

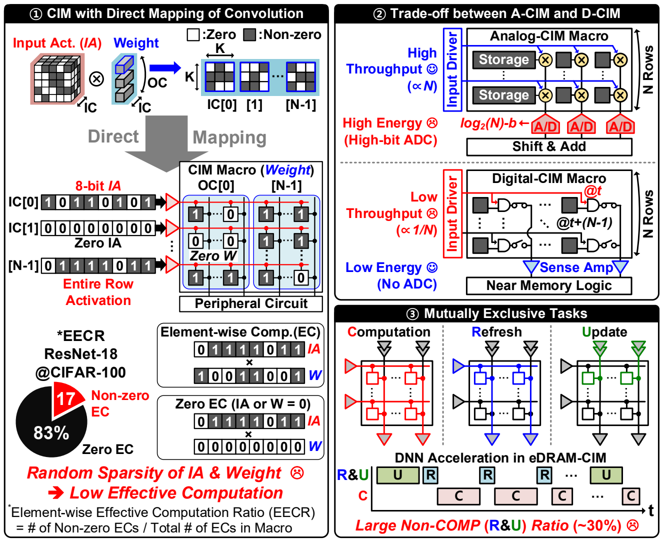


Fig. 1. Challenges of the previous CIM processors.

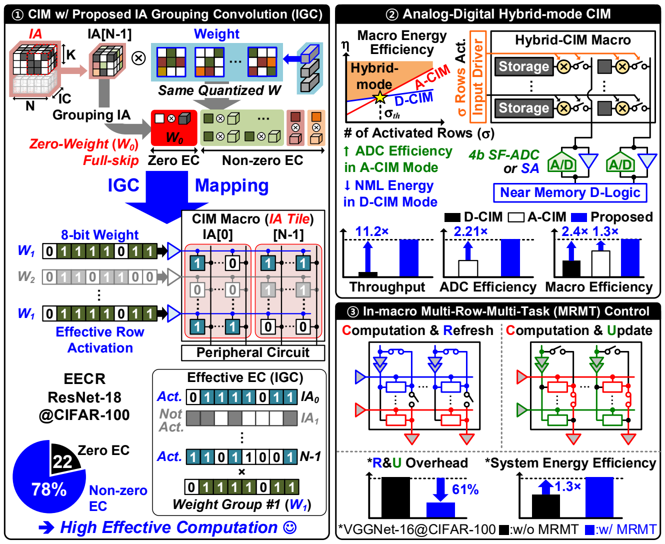


Fig. 2. Solutions for the proposed eDRAM CIM processor.

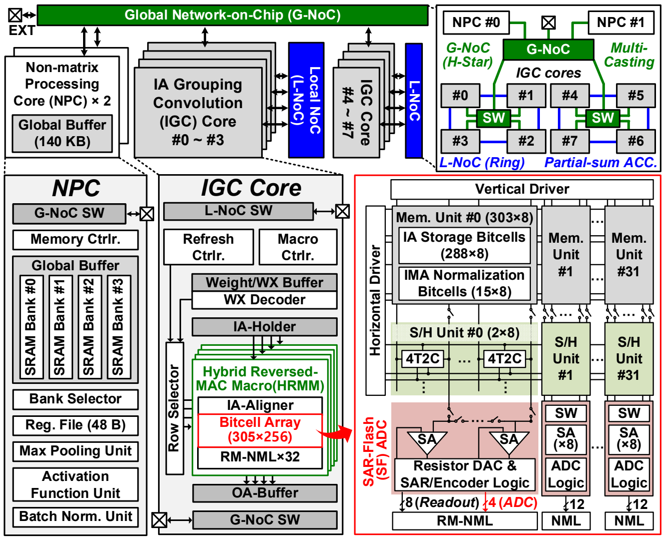


Fig. 3. Overall architecture of proposed highly energy-efficient sparsity-aware analog-digital hybrid eDRAM CIM processor.

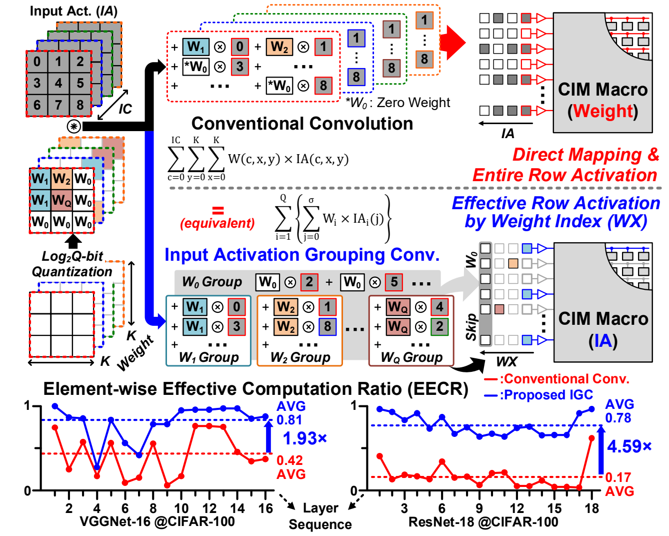


Fig. 4. Sparsity-aware computation by IGC and EECR improvement.

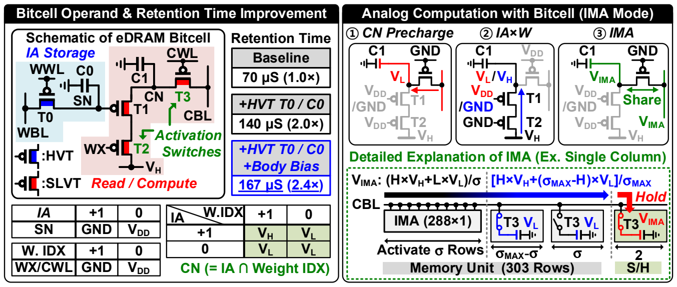


Fig. 5. Schematic of the proposed bitcell with retention time improvement and bitcell operation in IMA mode.

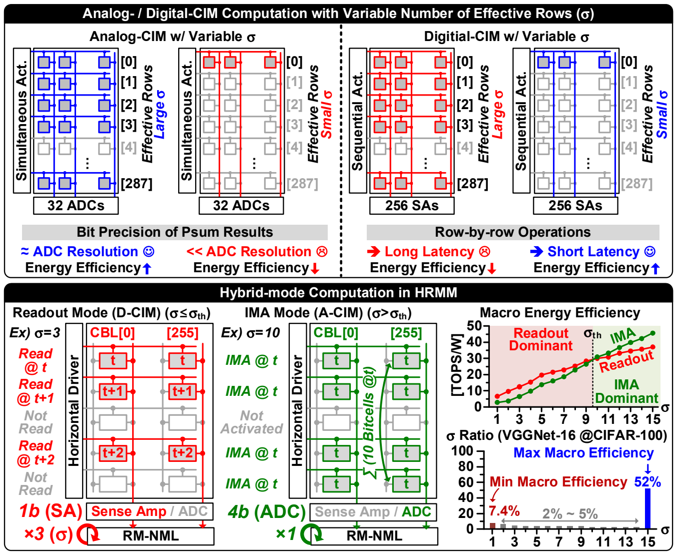


Fig. 6. Hybrid-mode computation in HRMM with various range of σ.

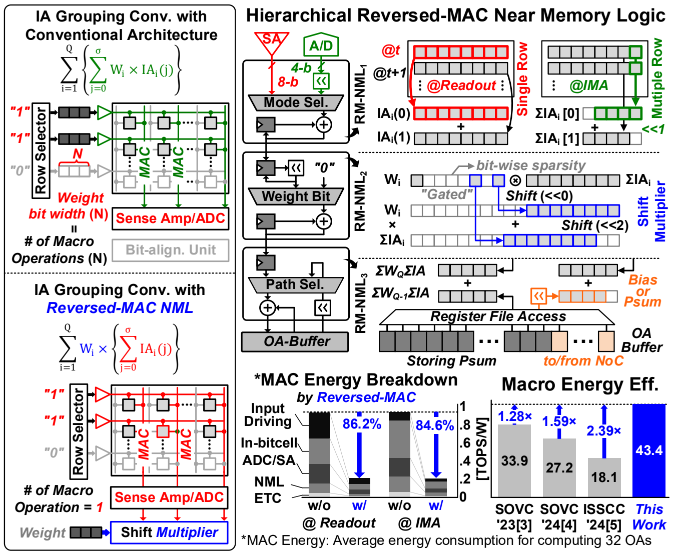
IA grouping convolution (IGC) with effective row activation to skip computations associated with zero-weight ($W_0$) by reordering convolution, thereby enhancing EECR; 2) hybrid-CIM macro with SAR-Flash ADC (SF-ADC) and reversed-MAC near-memory logic (RM-NML) to leverage advantages of A-/D-CIMs; and 3) in-macro multi-row-multi-task (MRMT) control to mitigate system performance degradation from refresh/update overhead. The rest of this paper is organized as follows. Section II introduces overall architecture with detailed explanations of the proposed key features. Implementation results and evaluation are presented in Section III, followed by conclusion in Section IV.

## II. Sparsity-aware Hybrid-CIM Processor

Fig. 3 presents the overall architecture of the proposed CIM. It consists of 8 IGC cores and 2 non-matrix processing cores. Each core is interconnected by hierarchical network-on-chips: global (G-NoC) and local (L-NoC). G-NoC multicasts data from global buffers to 8 IGC cores. L-NoC facilitates partial-sum accumulation from 4 IGC cores. Each IGC core comprises a row selector, weight/weight index (WX) buffer with WX decoder, output activation (OA)-buffer, and 4 hybrid reversed-MAC macros (HRMM) with 305×256 array and 32 RM-NMLs. The 288 rows of array store IAs and perform either in-memory accumulation (IMA) in analog domain or readout in digital domain through compute bit-line (CBL), maintaining computation linearity with 15 rows. The remaining 2 rows are used to sample and hold (S/H) the IMA result. SF-ADCs are configured for every 8 columns, sharing 8 sense amplifiers and a resistor DAC (RDAC).

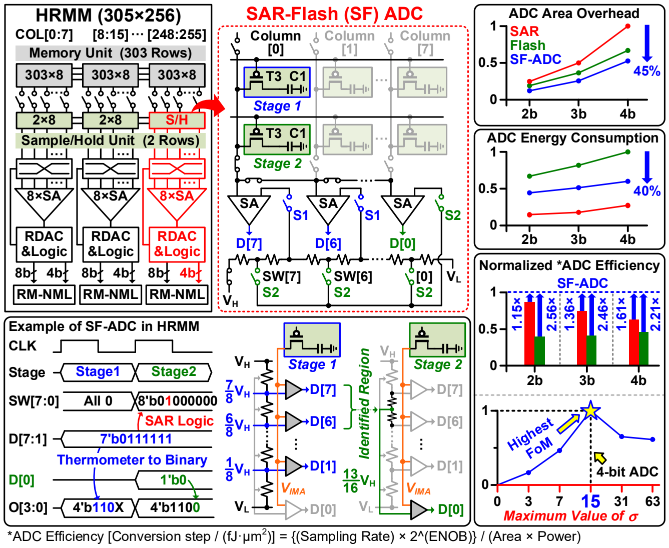
Fig. 4 describes the proposed IGC. Direct mapping of convolution onto CIM macro neglects sparsity, indiscriminately forcing entire row activation for MAC across all weights and IAs. This results in a low EECR, thereby causing significant energy waste due to zero computations. In contrast, IGC groups IAs that share the same quantized weight value ($W_i$) by tagging their coordinates with WX. In CIM macro, only effective rows are activated by referencing WX. It eliminates zero computations by *never activating* rows that correspond to the $W_0$ group, retaining algebraic equivalence to conventional convolution. The proposed IGC improves EECR by 1.93× and 4.59× on VGGNet-16 and ResNet-18, respectively.

Fig. 5 illustrates the schematic of the proposed eDRAM bitcell, where the storage node is isolated (T1) from the read/compute paths to enable non-destructive operations. The activation switches (T2/T3) are utilized for activating only



*MAC Energy: Average energy consumption for computing 32 OAs

Fig. 7. Detailed operation of hierarchical RM-NML for in-macro MAC energy reduction and macro energy efficiency improvement.



*ADC Efficiency [Conversion step / (fJ·μm²)] = {(Sampling Rate) × 2^(ENOB)} / (Area × Power)

Fig. 8. Operation of SF-ADC and comparison with SAR and Flash ADCs.

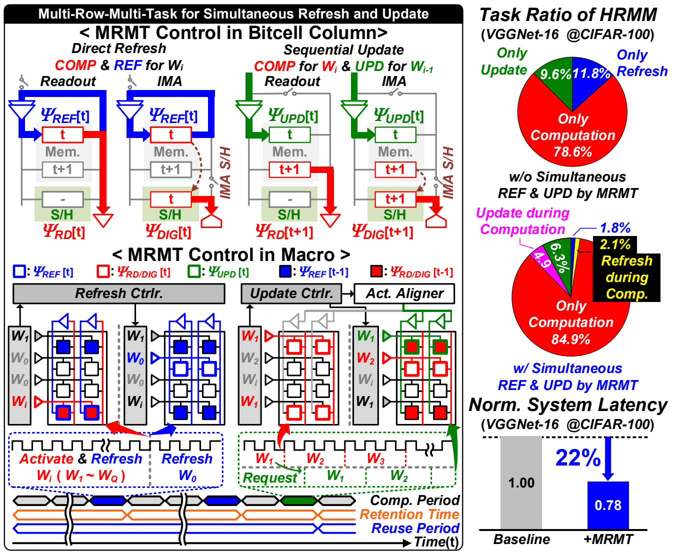


Fig. 9. MRMT control in bitcell- and macro-level for system latency reduction.

effective rows, whose number is represented as $\sigma$. It supports hybrid-CIM by facilitating both digital IA readout through CBL and linear analog computation using charge-sharing (C1). To ensure ideal IMA results, the proposed bitcell can be utilized to normalize dynamically varying $\sigma$. Also, retention time is increased by 2.4× through HVT PMOS (T0), MOMCAP (C0), and body biasing.

Fig. 6 illustrates the proposed HRMM. IGC dynamically varies $\sigma$ across $W_i$ groups. For small $\sigma$, A-CIM wastes ADC resolutions, whereas large $\sigma$ causes long latency in D-CIM due to repetitive row-by-row operations. To address this, HRMM supports hybrid-mode computations. When $\sigma \leq \sigma_{th}$, HRMM sequentially reads IAs of effective rows, accumulating them in RM-NML. For $\sigma > \sigma_{th}$, it simultaneously activates up to 15 effective rows for IMA. The maximum value of $\sigma$ is determined to optimize the ADC efficiency (Fig. 8) since ADC is the most power-hungry circuit in the macro. The $\sigma_{th}$ is determined as the point at which energy efficiency of IMA mode exceeds that of the readout mode. The peak macro efficiency is 47.5 TOPS/W at $\sigma$=15, which occupies >50% of $W_i$ groups for VGGNet-16.

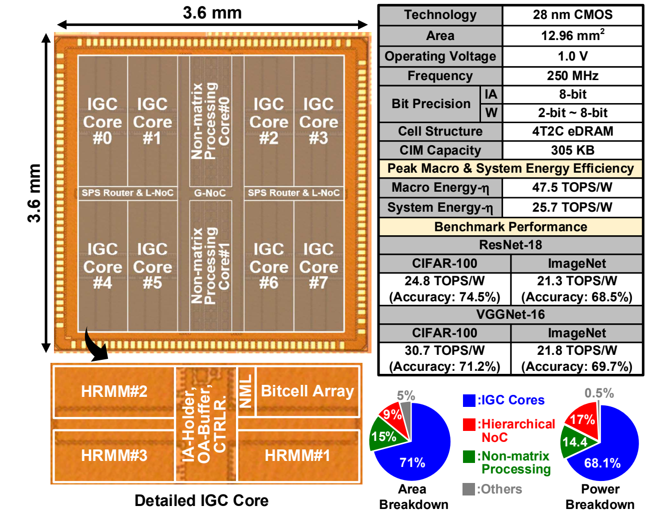
Mapping IGC onto conventional CIM requires iterative macro-operations depending on weight bitwidth, consuming massive energy. However, the proposed 3-level hierarchical RM-NML enables a single macro-operation for MAC, supporting dynamic bit precision of weight, as depicted in Fig. 7. RM-NML$_1$ accumulates IAs in readout mode, whereas it bit-aligns and adds the bitwise accumulated IAs in IMA mode. Then, RM-NML$_2$ multiplies the accumulated IAs with weight using shift multiplication to generate the MAC results. These are aggregated in RM-NML$_3$ with partial-sum from OA-buffer or L-NoC to calculate final OAs. Therefore, MAC energy is reduced by 86.2% and 84.6% for readout and IMA modes.

Fig. 8 shows the operation of SF-ADC, which digitizes the IMA result in two clock stages. In stage 1, RDAC produces 7 distinct reference voltages, and the IMA result in the S/H unit is compared with the references, generating a 7-bit code, which is encoded into 3-bit MSB. In stage 2, the middle voltage of the identified voltage region is used as the final reference voltage for 1-bit LSB. The SF-ADC reduces area by 45% compared to SAR ADCs [2-4] and has 40% lower energy consumption than Flash ADCs [7-8], which are adopted in prior CIMs. As a result, SF-ADC enhances ADC FoM by 2.21×.

As depicted in Fig. 9, the proposed HRMM supports the simultaneous refresh/update ($\Psi_{REF}/\Psi_{UPD}$), while reading data from bitcells ($\Psi_{RD}$) in readout mode or digitizing IMA results in S/H units ($\Psi_{DIG}$) in IMA mode. In the last computation period before retention time, effective rows are activated and directly refreshed. Non-activated rows by $W_0$ are refreshed in additional clock cycles to ensure that every row is refreshed. Updates are conducted when the stored IAs are completely reused. The rows activated for previous $W_{i-1}$ group are requested for update, which is performed once the new IAs for these rows are ready. Consequently, MRMT control reduces system latency by 22%.

## III. Implementation Results and Evaluation

Fig. 10 shows the chip photograph and performance specifications of the proposed CIM processor. Fabricated in 28



| Technology | | 28 nm CMOS | |
|---|---|---|---|
| Area | | 12.96 mm$^2$ | |
| Operating Voltage | | 1.0 V | |
| Frequency | | 250 MHz | |
| Bit Precision | IA | 8-bit | |
| | W | 2-bit ~ 8-bit | |
| Cell Structure | | 4T2C eDRAM | |
| CIM Capacity | | 305 KB | |
| Peak Macro & System Energy Efficiency | | | |
| Macro Energy-η | | 47.5 TOPS/W | |
| System Energy-η | | 25.7 TOPS/W | |
| Benchmark Performance | | | |
| ResNet-18 | | | |
| CIFAR-100 | | ImageNet | |
| 24.8 TOPS/W<br>(Accuracy: 74.5%) | | 21.3 TOPS/W<br>(Accuracy: 68.5%) | |
| VGGNet-16 | | | |
| CIFAR-100 | | ImageNet | |
| 30.7 TOPS/W<br>(Accuracy: 71.2%) | | 21.8 TOPS/W<br>(Accuracy: 69.7%) | |

Fig. 10. Chip photograph and performance specification.

TABLE I. COMPARISON TABLE

| | | ISSCC'20[1] | ISSCC'23[2] | SOVC'23[3] | SOVC'24[4] | ISSCC'24[5] | *This Work* |
|---|---|---|---|---|---|---|---|
| Technology | | 65 nm | 28 nm | 28 nm | 28 nm | 28 nm | 28 nm |
| Cell Structure | | 8T (SRAM) | 3T2C | 3T2C | 1T1C | 3T | 4T2C |
| Bit Precision | IA | 2b/4b/6b/8b | 4b/8b | 4b/8b | 8b | 1b~8b | 8b |
| | W | 4b/8b | 5b/9b | 5b/9b | 8b | 8b | 2b~8b |
| CIM Capacity [KB] | | 0.5 | 1200 | 100 | 3375 | 250 | 305 |
| REF/UPD in COMP. | | × | × | × | × | O | O |
| Sparsity-aware CIM | | O | × | × | × | × | O |
| Peak Energy Efficiency [TOPS/W] | *Macro (IA/W) | 15.8 (4b/8b) | 30.1 (8b/9b) | 33.9 (8b/9b) | 27.2 (8b/8b) | 18.1 (8b/8b) | 47.5 (8b/2b)<br>43.4 (8b/8b) |
| | *System (IA/W) | 3.56 (4b/8b) | 13.8 (8b/9b) | 15.5 (8b/9b) | - | - | 25.7 (8b/2b)<br>24.6 (8b/8b) |
| **Benchmark Energy Efficiency (Practical Energy Efficiency in Network Acceleration) [TOPS/W] | | | | | | | |
| ResNet-18 (CIFAR-100)(IA/W) | | - | - | 16 (8b/9b) | - | - | 24.8 (8b/3b~8b) |
| ResNet-18 (ImageNet)(IA/W) | | - | 10.8 (8b/9b) | 15.6 (8b/9b) | - | - | 21.3 (8b/5b~8b) |
| VGGNet-16 (CIFAR-10/100)(IA/W) | | 2.96 (4b/8b) (CIFAR-10) | - | - | - | - | 30.7 (8b/2b~7b) (CIFAR-100) |

*For fair comparison, peak macro and system energy efficiency are reported without IA and weight sparsity, except [1]. In [1], due to limited information, results for [IA = 4-bit/Weight = 8-bit] on ResNet-18 are reported.

**Benchmark energy efficiency includes the sparsity with layer-wise dynamic weight bit precision.

**Power consumption and latency related to external DRAM access are excluded for fair comparison. The measurements account for ***On-Chip*** computation/data transactions/memory accesses, including memory refresh/update operations.

nm CMOS technology, the chip occupies a 12.96 mm² die area and integrates 305 KB of the proposed eDRAM bitcells. The macro supports sparsity-aware convolution with the dynamic range of weight precisions from 2-bit to 8-bit. It achieves 47.5 TOPS/W of a peak macro energy efficiency and 25.7 TOPS/W of a system energy efficiency at 250 MHz and 1.0 V. Under benchmark evaluation, the proposed CIM processor achieves 24.8 TOPS/W of system energy efficiency on ResNet-18 and 30.7 TOPS/W on VGGNet-16 with CIFAR-100 dataset.

Table I compares the proposed CIM processor with previous CIM processors, including both a sparsity-aware SRAM-based design [1] and eDRAM-based designs [2-5]. The proposed processor supports effective row activation and simultaneous MRMT control to achieve the highest macro and system energy efficiency among all previous CIMs. Furthermore, it is the first solution to support sparsity-aware convolution using an eDRAM bitcell in an analog-digital hybrid CIM architecture. As a result, the proposed HRMM achieves 1.28× higher energy efficiency compared to the previous eDRAM CIM macro [3]. Additionally, by leveraging MRMT control, the proposed CIM processor demonstrates a 1.59× improvement in system energy efficiency compared to [3]. It shows a 1.97× and 1.55× improvement in energy efficiency on ResNet-18 with ImageNet [2] and CIFAR-100 [3], respectively. Compared to a previous sparsity-aware CIM processor [1], the proposed processor achieves 10.37× higher energy efficiency on VGGNet-16 with the larger dataset, CIFAR-100, even without any dedicated retraining process to enforce a specific sparsity pattern.

## IV. CONCLUSION

This paper presents the most energy-efficient analog-digital hybrid eDRAM CIM processor. IGC eliminates zero-weight computations, improving the effective computation ratio by 4.59×. HRMM supports both digital readout mode and analog in-memory accumulation mode, reducing MAC energy consumption by 85.4%. In addition, the proposed SF-ADC achieves 2.21× higher ADC efficiency compared to conventional ADCs. MRMT control enables simultaneous refresh and update operations during in-memory computation, reducing system latency by 22%. As a result, the proposed processor achieves 47.5 TOPS/W of macro energy efficiency and 10.37× improvement in benchmark energy efficiency compared to previous state-of-the-art CIM processors [1], [3].

## ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC support program (IITP-2025-RS-2020-II201847) supervised by the IITP. The chip fabrication and EDA tool were supported by the IC Design Education Center (IDEC), South Korea.

## REFERENCES

[1] J. Yue et al., "14.3 A 65nm Computing-in-Memory-Based CNN Processor with 2.9-to-35.8TOPS/W System Energy Efficiency Using Dynamic-Sparsity Performance-Scaling Architecture and Energy-Efficient Inter/Intra-Macro Data Reuse," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. papers,* Feb, 2020, pp. 234-236.

[2] S. Kim et al., "16.5 DynaPlasia: An eDRAM In-Memory-Computing-Based Reconfigurable Spatial Accelerator with Triple-Mode Cell for Dynamic Resource Switching," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. papers,* Feb, 2020, pp. 256-258.

[3] S. Kim et al., "Scaling-CIM: An eDRAM-based In-Memory-Computing Accelerator with Dynamic-Scaling ADC for SQNR-Boosting and Layer-wise Adaptive Bit-Truncation," in *Proc. Symp. VLSI Circuits,* Jun. 2023, pp. 1-2.

[4] S. Hong et al., "Dyamond: A 1T1C DRAM In-memory Computing Accelerator with Compact MAC-SIMD and Adaptive Column Addition Dataflow," in *Proc. Symp. VLSI Circuits,* Jun. 2024, pp. 1-2.

[5] Y. He et al., "34.7 A 28nm 2.4Mb/mm2 6.9 - 16.3TOPS/mm2 eDRAM-LUT-Based Digital-Computing-in-Memory Macro with In-Memory Encoding and Refreshing," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. papers,* Feb, 2024, pp. 578-580.

[6] S. Xie et al., "Gain-Cell CIM: Leakage and Bitline Swing Aware 2T1C Gain-Cell eDRAM Compute in Memory Design with Bitline Precharge DACs and Compact Schmitt Trigger ADCs," in *Proc. Symp. VLSI Circuits,* Jun. 2022, pp. 112-113.

[7] Z. Jiang et al., "C3SRAM: An In-Memory-Computing SRAM Macro Based on Robust Capacitive Coupling Computing Mechanism," *IEEE J. Solid-State Circuits,* vol. 55, no. 7, pp. 1888-1897, July 2020.

[8] S. Yin et al., "XNOR-SRAM: In-Memory Computing SRAM Macro for Binary/Ternary Deep Neural Networks," *IEEE J. Solid-State Circuits,* vol. 55, no. 6, pp. 1733-1743, June 2020.