A 28nm 51.6TOPS/W 2.22Mb/mm^2 System-in-One-Macro Computing-in-Memory Chip Utilizing Leakage-Eliminated 2T1C and Capacitor-Over-Logic 1T1C eDRAM

Zhaori Cong[1,2], Zhihang Qian[1,2], Shengzhe Yan[1,2], Zeyu Guo[1,2], Zhuoyu Dai[1,2], Yifan He[3], Wenyu Sun[3],
Chunmeng Dou[1,2], Feng Zhang[1,2], Jinshan Yue[1], Di Geng[1,2], Yongpan Liu[3]
[1]Institute of Microelectronics of the Chinese Academy of Sciences, Beijing, China,
[2]University of Chinese Academy of Sciences, Beijing, China, [3]Tsinghua University, Beijing, China
*Corresponding email: {yuejinshan, digeng}@ime.ac.cn

***Abstract*—This work presents a computing-in-memory (CIM) chip that integrates all system-level components (except control) into one single macro, achieving system-level high energy efficiency, area efficiency, and density. The main contributions include: 1) A system-in-one-macro CIM architecture with leakage-eliminated high-density 2T1C eDRAM storage; 2) A cap-over-logic 1T1C eDRAM CIM array that stacks capacitors over MUX-based CIM circuits to achieve both high density and long retention time; 3) A near-memory-computing module to reduce memory access of high-bit-position accumulation. The fabricated 28nm system-in-one-macro CIM chip achieves 51.6TOPS/W energy efficiency (1.24×), 1.53TOPS/mm^2 area efficiency (3.56×), and 2.22Mb/mm^2 storage density (5.20×).**

***Index Terms*—Computing-in-memory, eDRAM, energy efficiency, area efficiency, storage density, system-in-one-macro**

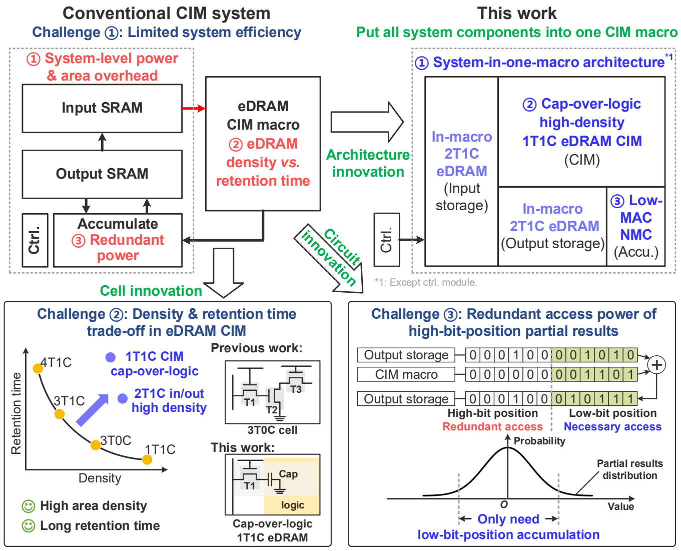

Conventional CIM system
Challenge ①: Limited system efficiency
① System-level power & area overhead
Input SRAM
Output SRAM
Ctrl.
Accumulate
③ Redundant power
eDRAM CIM macro
② eDRAM density vs. retention time
Architecture innovation
This work
Put all system components into one CIM macro
① System-in-one-macro architecture*1
In-macro 2T1C eDRAM (Input storage)
② Cap-over-logic high-density 1T1C eDRAM CIM (CIM)
In-macro 2T1C eDRAM (Output storage)
③ Low-MAC NMC (Accu.)
Ctrl.
*1: Except ctrl. module.
Cell innovation
Circuit innovation
Challenge ②: Density & retention time trade-off in eDRAM CIM
Retention time
Density
4T1C
3T1C
3T0C
1T1C
1T1C CIM cap-over-logic
2T1C in/out high density
Previous work:
T1
T2
T3
3T0C cell
This work:
T1
Cap
logic
Cap-over-logic 1T1C eDRAM
☺ High area density
☺ Long retention time
Challenge ③: Redundant access power of high-bit-position partial results
Output storage 0 0 0 1 0 0 0 0 1 0 1 0
CIM macro 0 0 0 0 0 0 0 0 1 1 0 1
Output storage 0 0 0 1 0 0 0 1 0 1 1 1
+
High-bit position
Redundant access
Low-bit position
Necessary access
Probability
Partial results distribution
O
Value
Only need
low-bit-position accumulation

Fig. 1. Challenges in improving system-level CIM metrics (system-level memory access, eDRAM CIM density vs. retention, accu. power).

I. Introduction

The general matrix multiplication (GEMM) is the dominant operation in emerging machine learning (ML) algorithms, including convolutional neural network, Transformer, etc. To this end, computing-in-memory (CIM) is a promising paradigm to reduce the computation and memory access power in the heavy GEMM operations. Though macro-level CIM chips [1]–[4] have demonstrated high area/energy efficiency for ML algorithms, several challenges still exist to reflect these high metrics onto system-level CIM chips [5]–[11], as shown in Fig. 1.

Firstly, the system-level components, including input/output SRAMs, accumulation, etc., are necessary to build a complete CIM system, which consume a large proportion of power/area outside the CIM macro and degrade the system-level area/energy efficiency. Previous work [6] has optimized the system architecture to reduce SRAM access and accumulation power. However, bottlenecked by the intrinsic feature of SRAM, the optimized minimal access power still takes a large proportion. Meanwhile, the SRAM area is considerable. Besides, the separated physical layout between CIM macros and the other system-level components leads to additional power/area overhead (e.g. clock trees, long data paths).

Secondly, to improve the area/energy efficiency, eDRAM-based CIM chips [1]–[5] have demonstrated high density and high energy efficiency. However, the frequent refresh of eDRAM circuits leads to a trade-off between density and retention time. The remaining space over the CIM circuits (e.g. adder tree) is not explored, which can be adopted as additional capacitors to improve retention time under a limited eDRAM cell area. Thirdly, the partial results from the CIM macro still require external accumulation, where the high-bit-position data barely changes considering the data characteristics of ML algorithms, leading to redundant memory access power.

To overcome these challenges, this work presents a system-in-one-macro CIM chip to achieve system-level high area/energy efficiency for GEMM in ML algorithms. To address the system-level power/area overhead, the pivotal idea is to integrate all system-level components into one single macro, achieving a compact layout without large clock trees and long datapaths. Further, at the system level, a high-density leakage-eliminated 2T1C eDRAM is proposed to replace the power/area-consuming SRAMs. At the macro level, the capacitor(cap)-over-logic 1T1C MUX-based CIM achieves both high density and long retention time. Besides, a low-MAC-aware near-memory-computing (NMC) circuit is implemented to reduce high-bit-position accumulation power.

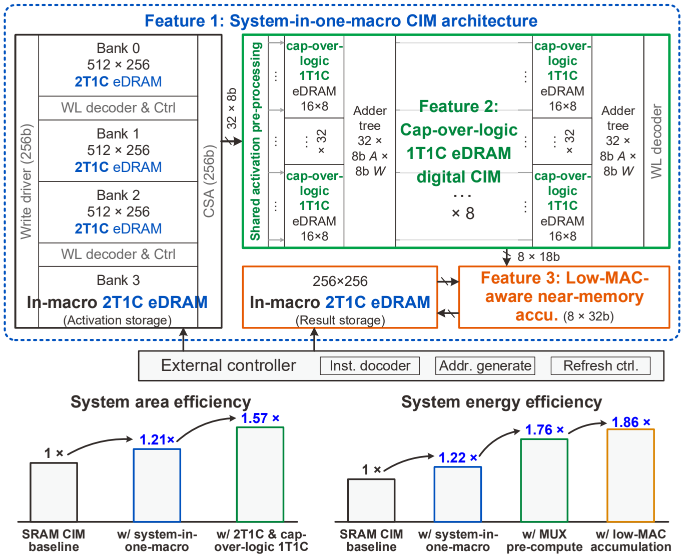


Fig. 2. The proposed system-in-one-macro CIM architecture with 2T1C storage, cap-over-logic 1T1C CIM, and low-MAC-aware NMC.

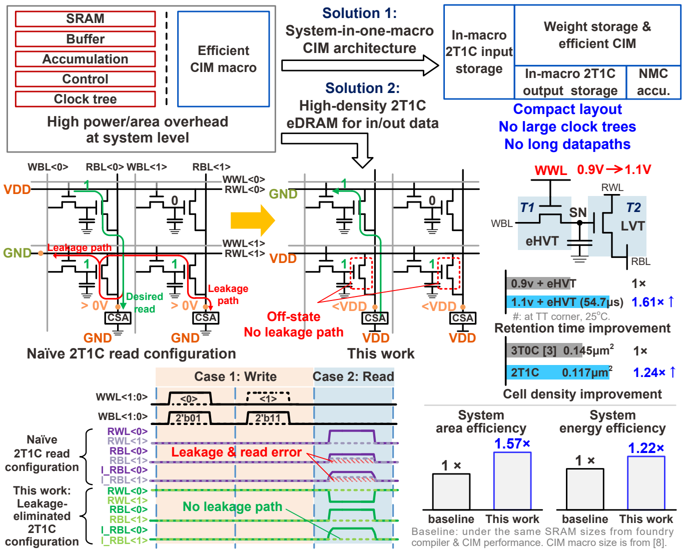


Fig. 3. Compact system-in-one-macro CIM chip layout paradigm with high-density leakage-eliminated 2T1C eDRAM input/output storage.

II. Proposed System-in-One-Macro CIM Chip

Fig. 2 presents the overall system-in-one-macro CIM architecture with the 2T1C eDRAM storage, cap-over-logic 1T1C CIM, and low-MAC-aware NMC. All the system-level components (except external controller) are tightly integrated into one CIM macro, including the 2048×256/256×256 2T1C eDRAM for input/output storage, the cap-over-logic 1T1C CIM array, and NMC accumulation. In each cycle, a 32×8bit vector is fetched from 2T1C activation eDRAM, then performs multiply-accumulation (MAC) with a 32-row×8-bank×8bit weight matrix. The CIM array pre-reads 8bit weight from each 16×8 cap-over-logic 1T1C sub-array, and latches it for many CIM cycles. The 8×18bit partial results (Psums) after CIM operations are accumulated in the low-MAC-aware NMC circuits with previously stored Psums from the 2T1C result eDRAM. The external controller is currently difficult to be integrated into the CIM macro, thus is left outside and performs instruction fetch, decoding, and eDRAM refresh.

A. Leakage-Eliminated High-Density 2T1C eDRAM

Fig. 3 illustrates the system-in-one-macro CIM schematic with the leakage-eliminated high-density 2T1C eDRAM. To improve system-level area/energy efficiency, this work first proposes the system-in-one-macro paradigm. It features a compact layout without power-consuming clock trees and long data paths. A leakage-eliminated high-density 2T1C eDRAM is designed to further reduce input/output access power/area. Each 2T1C cell consists of an eHVT write transistor and an LVT read transistor. In a naïve 2T1C read operation (similar to [12]), it applies VDD on the target read wordline (RWL) and GND on current-sensing read bitlines (RBLs), which would cause leakage paths due to the ‘1’ storage on the other rows.

To avoid this problem, this work proposes a leakage-eliminated voltage configuration, by applying GND on the desired RWL and applying VDD on RBLs and the unselected RWLs. The read transistors in the unselected rows are at off state to eliminate the leakage paths since $V_{gs}<0$ (storage node $SN<0.9\text{V}$). The WWL port adopts a higher voltage domain (VDDH=1.1V) while the others are at normal voltage domain, enabling 1.61× longer retention time. The 2T1C cell shows 1.24× density compared to the previous 3T0C [3]. The compact system-in-one-macro layout and high-density 2T1C eDRAM achieve 1.57×/1.22× area/energy efficiency compared to the system-level SRAM CIM baseline from [8].

B. Capacitor-Over-Logic 1T1C eDRAM CIM Array

Fig. 4 presents the overall architecture of the cap-over-logic 1T1C array with shared-preprocessing MUX-based CIM, and Fig. 5 illustrates the detailed circuits and waveform. The whole CIM array consists of 32×8 sub-arrays, a shared pre-processing unit, and eight adder trees (32-row×8bit×8bit).

Cap-over-logic 1T1C eDRAM. To achieve both high density and long retention time, this work proposes to utilize the remained metal layers over the CIM circuits. The capacitor of the 1T1C cell is stacked over the 1st stage CIM circuits in the metal layers (M5-M7) with a much larger MOM area. In the 1T1C sub-array, the WLs adopt a similar high VDDH as the 2T1C array to achieve a high SN voltage. In the read process, the read transmission gate (TG) turns on, sends data to the latch node (LN), and turns on the refresh TG for an immediate self-refresh in the same cycle. Then, all the corresponding activation data are traversed from the 2T1C activation eDRAM and perform multiple cycles of CIM operations with the latched weight.

MUX-based shared-preprocessing CIM. Each 8bit weight is pre-read from a 16×8 1T1C sub-array. The CIM operation contains two stages: **1)** Multiplication between 8bit activation (A) and four 2bit parts of 8bit weight (W), **2)** The subsequent accumulation. In the 1st stage, since each part of weight is 2bit,

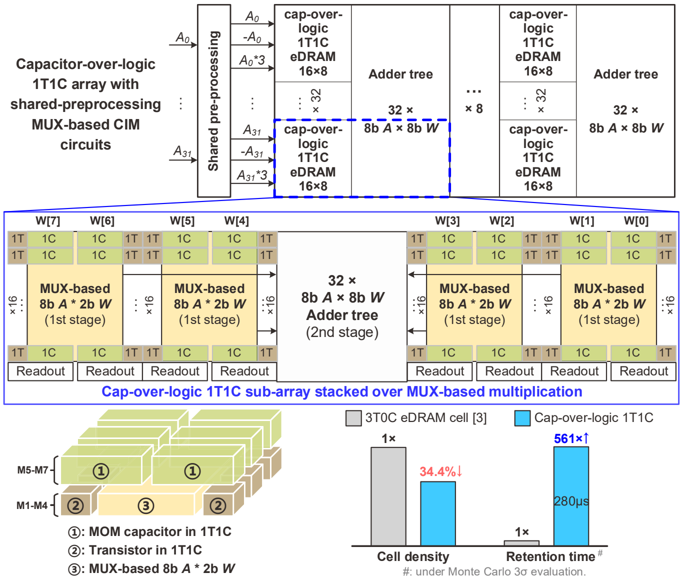

Capacitor-over-logic 1T1C array with shared-preprocessing MUX-based CIM circuits
Shared pre-processing
cap-over-logic 1T1C eDRAM 16×8
Adder tree
32 × 8b A × 8b W
× 8
Cap-over-logic 1T1C sub-array stacked over MUX-based multiplication
MUX-based 8b A * 2b W (1st stage)
32 × 8b A × 8b W Adder tree (2nd stage)
Readout
M5-M7
M1-M4
①: MOM capacitor in 1T1C
②: Transistor in 1T1C
③: MUX-based 8b A * 2b W
3T0C eDRAM cell [3]
Cap-over-logic 1T1C
1×
34.4%↓
561×↑
280µs
Cell density
Retention time#
#: under Monte Carlo 3σ evaluation.

Fig. 4. The capacitor-over-logic 1T1C array with shared-preprocessing MUX-based CIM circuits.

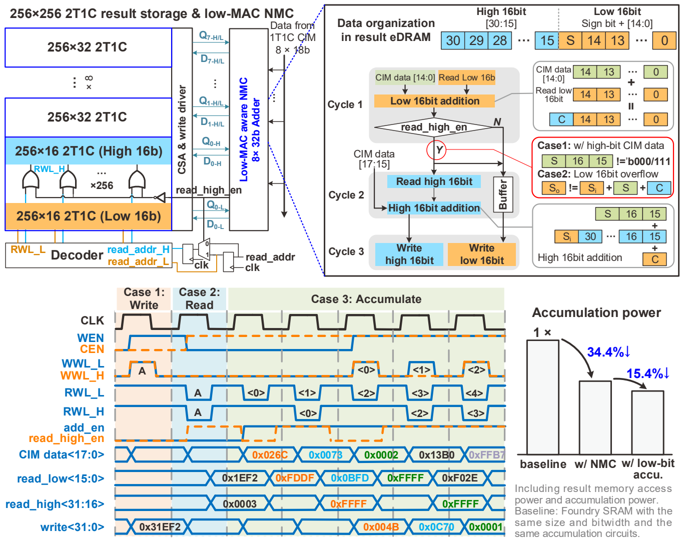
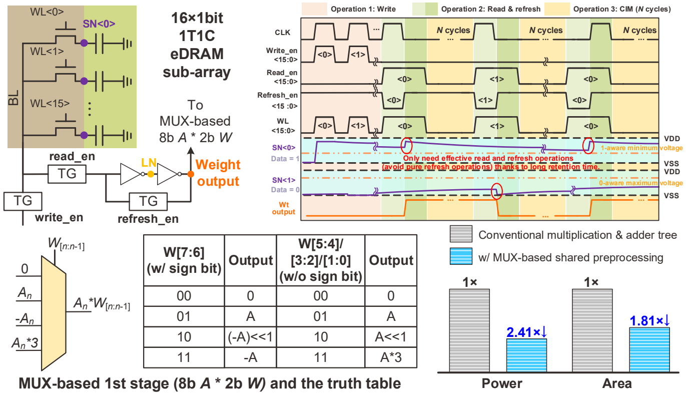


Fig. 6. The 2T1C result eDRAM and low-MAC-aware NMC circuits to avoid redundant high-bit accumulation.


16×1bit 1T1C eDRAM sub-array
To MUX-based 8b A * 2b W
WL<0>
WL<1>
WL<15>
SN<0>
BL
read_en
TG
write_en
refresh_en
LN
Weight output
Operation 1: Write
Operation 2: Read & refresh
Operation 3: CIM (N cycles)
Only need effective read and refresh operations
MUX-based 1st stage (8b A * 2b W) and the truth table
Conventional multiplication & adder tree
w/ MUX-based shared preprocessing
1×
2.41×↓
1.81×↓
Power
Area

W[7:6] (w/ sign bit)	Output	W[5:4]/[3:2]/[1:0] (w/o sign bit)	Output
00	0	00	0
01	A	01	A
10	(-A)<<1	10	A<<1
11	-A	11	A*3

Fig. 5. Details of the cap-over-logic 1T1C subarray and shared-preprocessing MUX-based CIM circuits.

the activations can be pre-processed to generate A, -A, and A*3. The 8b(A)×2b(W) turns into simple MUX circuits based on the signed/unsigned 2bit W[n:n-1]. Therefore, the common part of the computation can be extracted to the preprocessing circuits and shared by all the CIM columns to save power/area.

Thanks to the long retention time (280μs) of cap-over-logic 1T1C, the storage node (SN) keeps the weight data until the next effective read and refresh operation (after thousands of cycles), avoiding intermediate pure refresh power. Compared to the state-of-the-art 3T0C CIM [3], the cap-over-logic 1T1C requires 34.4% larger cell area, but shows 561× higher retention time. The shared-preprocessing MUX-based CIM circuits achieve 2.41× power reduction and 1.81× area reduction.

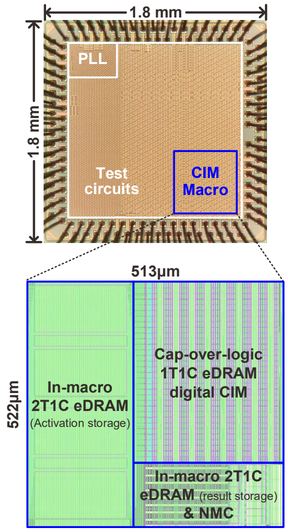
C. Low-MAC-Aware NMC Accumulation

Fig. 6 shows the 2T1C result eDRAM and corresponding low-MAC-aware NMC circuits. Since a large proportion of accumulations in the ML algorithms only change the low-bit-position results, this work reduces the accumulation and memory access power by avoiding unnecessary high-bit-position computation. The 2T1C result array consists of eight banks, each bank owning two separate 256×16 subarrays for high/low 16bit Psum storage. For each 32bit storage, the sign bit (S) and low-15bit data [14:0] are stored in the low-16bit bank while the high-16bit [30:15] are stored in the high-16bit bank. In the three pipelined cycles of each NMC accumulation, it performs (1) low-16bit addition and high-16bit check, (2) optional high-16bit read and addition, (3) low-16bit and optional high-16bit write-back. The high-16 partial result only needs to be readout in two cases: 1) The 18bit CIM result has effective high-bit data (i.e. [S:15]!='b000/111); 2) The low-16bit accumulation overflow, with a carry bit to high-16bit or an inversed sign-bit. An OR-gate/AND-gate is inserted before the high-bit RWL/WWL (RWL_H/WWL_H) to turn off the unnecessary high-bit read/write operations. The 2T1C NMC accumulation shows 34.4% power reduction compared to conventional SRAM-based accumulation, while the low-MAC-aware accumulation further saves 15.4% power.

III. Measurement Results

Fig. 7 presents the die photo and metrics of the fabricated 28nm system-in-one-macro CIM chip. This chip can work at 50-800MHz under 0.6-1.05V voltage (VDDC) for CIM and NMC, with a consistent 0.9V 2T1C/1T1C memory voltage (VDDM) and 1.1V VDDH. The CIM array and NMC circuits take a 58.9% high power proportion since the power of other system-level components is significantly suppressed. Fig. 8 presents the 1T1C/2T1C eDRAM layout with 0.7×0.39μm^2 and 0.36×0.4μm^2 area (before shrink 0.9).

Fig. 9 shows the comparison with previous macro-level and system-level CIM chips. Utilizing the cap-over-logic 1T1C and shared-preprocessing MUX-based CIM circuits, this work achieves 117.3TOPS/W macro-level (weight storage + CIM) energy efficiency at the peak efficiency point (200MHz, 0.7V).


1.8 mm
1.8 mm
PLL
Test circuits
CIM Macro
513μm
522μm
In-macro 2T1C eDRAM (Activation storage)
Cap-over-logic 1T1C eDRAM digital CIM
In-macro 2T1C eDRAM (result storage) & NMC

Technology	28nm
System-in-one-macro area *1	513μm × 522μm
Act / W / Res+accu. area	0.122 / 0.106 / 0.031 mm^2
Act / W / Res size	512Kbit, 32Kbit, 64Kbit
Bit cell *2	Act/Res: 2T1C, 0.117μm^2 W: 1T1C, 0.221μm^2
Act / W bits (per cycle)	8b / 8b
CIM voltage (VDDC) *3	0.6-1.05 V
Frequency	50 - 800 MHz
CIM + accumulation power	0.36 - 7.57 mW
System power	0.62 - 11.39 mW
Performance (INT8) *4	0.41 TOPS
System area efficiency (INT8) *4	1.53 TOPS/mm^2
System density	2.22 Mb/mm^2
System energy efficiency (INT8) *5	51.6 TOPS/W

*1: For 1T1C, capacitor area over logic is not included.
*2: Except for the external control logic.
*3: Memory voltage (VDDM) keeps at 0.9V. VDDH keeps at 1.1V.
*4: Peak performance point (800MHz, 1.05V).
*5: Peak efficiency point (200MHz, 0.7V), 50% weight bit sparsity, 10% input toggle rate.

Fig. 7. The die photo and measured metrics of the fabricated 28nm system-in-one-macro CIM chip.

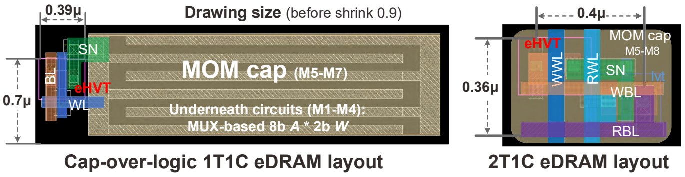

0.39μ
Drawing size (before shrink 0.9)
0.7μ
SN
BL
eHVT
WL
MOM cap (M5-M7)
Underneath circuits (M1-M4): MUX-based 8b A * 2b W
Cap-over-logic 1T1C eDRAM layout
0.4μ
0.36μ
eHVT
WWL
RWL
MOM cap
M5-M8
SN
lvt
WBL
RBL
2T1C eDRAM layout

Fig. 8. The layout of the cap-over-logic 1T1C and 2T1C eDRAM cells.

This work aims at high system-level area/energy efficiency and storage density. Therefore, the designed CIM array supports 8b×8b MAC in one cycle, which degrades the macro-level area efficiency and density, but helps to improve the system-level metrics. With the system-in-one-macro compact layout, 2T1C high-density eDRAM, and low-MAC-aware NMC accumulation, this work achieves system-level 51.6TOPS/W energy efficiency (1.24×), 1.53TOPS/mm^2 area efficiency (3.56×), and 2.22Mb/mm^2 storage density (5.20×) compared to the state-of-the-art system-level CIM chip [6], and highlights the first CIM chip that integrates all the system-level components (except control) into one CIM macro.

IV. Conclusion

This work presents an area/energy-efficiency system-in-one-macro CIM chip. By integrating all the system-level components into one compact CIM macro, the system-level power/area is significantly reduced. The leakage-eliminated high-density 2T1C eDRAM further reduces the dominant area of the input/output SRAMs, while the cap-over-logic 1T1C eDRAM array achieves both high density and long retention time. Besides, the low-MAC-aware NMC circuits reduce the memory access and computation power of the high-bit-position accumulation. The fabricated 28nm CIM chip achieves system-level 51.6TOPS/W energy efficiency, 1.53TOPS/mm^2 area efficiency, and 2.22Mb/mm^2 storage density, and highlights the system-in-one-macro paradigm to achieve system-level high area/energy efficiency.

		Macro-level CIM				System-level CIM		
		JSSC'24 [1]	ISSCC'24 [2]	ISSCC'24 [3]	This work	VLSI'24 [6]	ISSCC'23 [7]	ISSCC'23 [8]
Technology		28nm	16nm	28nm	28nm	28nm	28nm	28nm
Cell structure		3T1C eDRAM	4T gain cell	3T eDRAM	2T1C in/out + 1T1C CIM	6T SRAM	8T SRAM	6T SRAM
Cell size (μm^2)		0.178	/	0.145	2T1C: 0.117 1T1C: 0.221 *1	/	/	/
Retention time (μs)		36.22	/	1.3	2T1C: 84.5 1T1C: 447.1	/	/	/
CIM type		Analog Charge-based	Digital	Digital	Digital	Digital	Digital	Digital
Activation/weight bits		A/W: 1-8	A/W:8, BF16	A:1-8 W:8	A:8 W:8	A/W:4,8	A/W:8,16	A/W:4,8,FP16
Perf. (TOPS@8b)		1.02	0.40	0.20	0.41 *2	1.84	3.55	0.41
Macro-level	Energy efficiency (TOPS/W@8b)	234.6	163.3	18.1	117.3 *3	94.6	/	68.7
	Area efficiency (TOPS/mm^2@8b)	18.68	11.07	16.2	3.86 *2	1.47	/	1.53
	Density (Mb/mm^2)	2.28	/	2.4	0.295	0.22	/	0.24
System-level	Energy efficiency (TOPS/W@8b)	/	/	/	51.6 (1.24×) *3	41.7 (1×)	48.4	12.8
	Area efficiency (TOPS/mm^2@8b)	/	/	/	1.53 (3.56×) *2	0.43 (1×)	0.25	0.09
	Density (Mb/mm^2)	/	/	/	2.22 (5.20×)	0.43 (1×)	0.17	0.94
System-in-one-macro		No	No	No	Yes *4	No	No	No

*1: For 1T1C, capacitor area over logic is not included. *2: Peak performance point (800MHz, 1.05V).
*3: Peak efficiency point (200MHz, 0.7V), 50% weight bit sparsity, 10% input toggle rate. *4: Except for the external control logic.

Fig. 9. Comparison with the state-of-the-art macro/system-level CIM chips.

Acknowledgment

This work was supported in part by National Key Research and Development Program of China under Grant 2023YFB4402400; National Natural Science Foundation of China under Grants 92464203, 62488101, 92364202, 92464201, U2341218, and 62204256.

References

[1] Y. Zhan *et al.*, "A 28-nm 18.7 TOPS/mm2 89.4-to-234.6 TOPS/W 8b single-finger eDRAM compute-in-memory macro With bit-wise sparsity aware and kernel-wise weight update/refresh," *IEEE Journal of Solid-State Circuits*, vol. 59, no. 11, pp. 3866–3876, 2024.

[2] W.-S. Khwa *et al.*, "A 16nm 96Kb integer/floating-point dual-mode-gain-cell-computing-in-memory macro achieving 73.3-163.3TOPS/W and 33.2-91.2TFLOPS/W for AI-edge devices," in *2024 International Solid-State Circuits Conference (ISSCC)*. IEEE, 2024, pp. 568–569.

[3] Y. He *et al.*, "A 28nm 2.4 Mb/mm2 6.9-16.3 TOPS/mm2 eDRAM-LUT-based digital-computing-in-memory macro with in-memory encoding and refreshing," in *ISSCC*. IEEE, 2024, pp. 578–579.

[4] S. Hong *et al.*, "Dyamond: A 1T1C DRAM in-memory computing accelerator with compact MAC-SIMD and adaptive column addition dataflow," in *VLSI Technology and Circuits*. IEEE, 2024, pp. 1–2.

[5] D. Kim *et al.*, "DPIM: A 19.36 TOPS/W 2T1C eDRAM Transformer-in-memory chip with sparsity-aware quantization and heterogeneous dense-sparse core," in *2024 IEEE European Solid-State Electronics Research Conference (ESSERC)*. IEEE, 2024, pp. 141–144.

[6] Z. Dai *et al.*, "A 41.7TOPS/W@INT8 computing-in-memory processor with zig-zag backbone-systolic CIM and block/self-gating CAM for NN/recommendation applications," in *VLSI Technology and Circuits*, 2024.

[7] F. Tu *et al.*, "MulTCIM: A 28nm 2.24 μJ/Token attention-token-bit hybrid sparse digital CIM-based accelerator for multimodal Transformers," in *ISSCC*. IEEE, 2023, pp. 248–239.

[8] J. Yue *et al.*, "A 28nm 16.9-300TOPS/W computing-in-memory processor supporting floating-point NN inference/training with intensive-CIM sparse-digital architecture," in *ISSCC*. IEEE, 2023, pp. 252–253.

[9] K. Ueyoshi *et al.*, "DIANA: An end-to-end energy-efficient digital and analog hybrid neural network SoC," in *ISSCC*, 2022, pp. 256–257.

[10] J. Yue *et al.*, "A 5.6-89.9TOPS/W heterogeneous computing-in-memory SoC with high-utilization producer-consumer architecture and high-frequency read-free CIM macro," in *VLSI Technology and Circuits*, 2023.

[11] W. Jiang *et al.*, "HUNBN, a 1.77MB digital in-memory-compute SoC for edge applications achieving 126 TOPs/W (4b) at macro level and 24 TOPs/W at SoC level," in *2024 IEEE European Solid-State Electronics Research Conference (ESSERC)*. IEEE, 2024, pp. 137–140.

[12] D. Somasekhar *et al.*, "2GHz 2Mb 2T gain cell memory macro with 128 GBytes/sec bandwidth in a 65 nm logic process technology," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 1, pp. 174–185, 2009.