D3TA: 38.9TOPS/W Transformer Accelerator with Dual-Port 3T-eDRAM Digital Compute-In-Memory using HyperAttention and Triple-Sparsity-Handling

Donghyuk Kim*, In-Jun Jung*, Geonwoo Ko, Seri Ham, Cuong Duong Manh, Gyeongrok Yang, and Joo-Young Kim
School of Electrical Engineering,
Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea
{kar02040, injun, geonwooko, seri1215, cuongdm1410, toddlerf, jooyoung1203}@kaist.ac.kr
*Equal Contribution

***Abstract*—This paper presents D3TA that accelerates transformers with high system performance. It introduces three key features: 1) HyperAttention (HA) that supports maximum pipelining in memory to minimize the external memory access (EMA) and increase CIM utilization, 2) triple sparsity handling CIM engine that skips redundant near-zero values for energy and latency reduction, and 3) high throughput Dual-port 3T-eDRAM Digital CIM (D3CIM) macro with charge-recycling (CR) and bit-parallel dual radix-4 booth (BPDB) MAC tree group to increase energy- and area-efficiency. The D3TA achieves 2.6× latency reduction in BERT-Base with 2.0× EMA reduction and 1.28-to-9.75× higher system FoM than previous works.**

***Index Terms*—Transformer, Dataflow, Sparsity, 3T-eDRAM cell, Computing-in-Memory**

I. Introduction

Transformers demonstrate remarkable performance in AI domains, and digital CIM based transformer accelerators [1]–[3] have emerged to address memory bandwidth (BW) bottlenecks. However, their excessive multi-head attention (MHA) and matrix multiplication kernels cause three major system-level challenges, as shown in Fig. 1. First, as token length increases, MHA exponentially generates intermediate data of Q, K, V, score (S), and probability matrices (P), increasing EMA and lowering CIM utilization. Second, transformers perform redundant compute due to numerous near-zeros in activations (A) and weights (W) [4], as well as weakly related tokens for output (O), increasing inefficiency in CIM macros. Third, achieving high throughput in CIM macros incurs substantial energy consumption and area overhead due to multi-row activation, multi-bit-line switching, and large bit-parallel MAC. We propose D3TA, D3CIM based Transformer Accelerator equipped with three key features: 1) HA that reduces EMA and enhances CIM utilization, 2) CIM engine (CE) with triple sparsity handler (TSH) for A/W/O sparsity, and 3) D3CIM macro supporting CR and BPDB MAC tree group to reduce energy and area overhead.

II. D3TA Architecture

Fig. 2 shows the overall D3TA architecture, consisting of 2 HA clusters (HACs). Each HAC includes 8 CEs, a streaming vector processing unit (SVPU), an 88KB attention

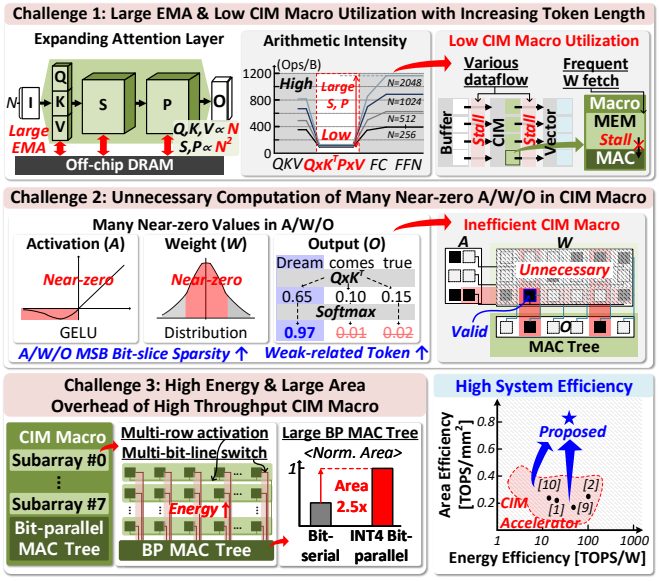


Fig. 1. System-level challenges for transformer accelerator

buffer, a reconfigurable HAC network (RHCN) for efficient data movement, and an HAC controller. A CE consists of a CE controller, input buffer, D3CIM with 544×320 cells, D3CIM accumulator for partial sum (PSUM), PSUM buffer for output of D3CIM, and TSH. A TSH consists of a triple sparsity controller (TSC), early zero-gating (EZG) bit-mask (BM) generator, and triple BM buffer for managing triple sparsity. The SVPU supports inter-CE aggregation (ICEA) for 8 CEs, quantization for 2's complement (2C) and sign-magnitude (SM), softmax, GELU, transpose, and compressor for activations.

A. HyperAttention (HA)

Fig. 3 illustrates HA dataflow that fully retains Q, K, V, S, and P on chip using a tile-basis operation via attention buffer, and pipelines QKV generation and MHA with two optimized computations. Q, K, and V are generated on a per-head basis, then a group of S, P, and attention output rows are computed. Furthermore, RHCN and in-memory pipeline (IMP) provide

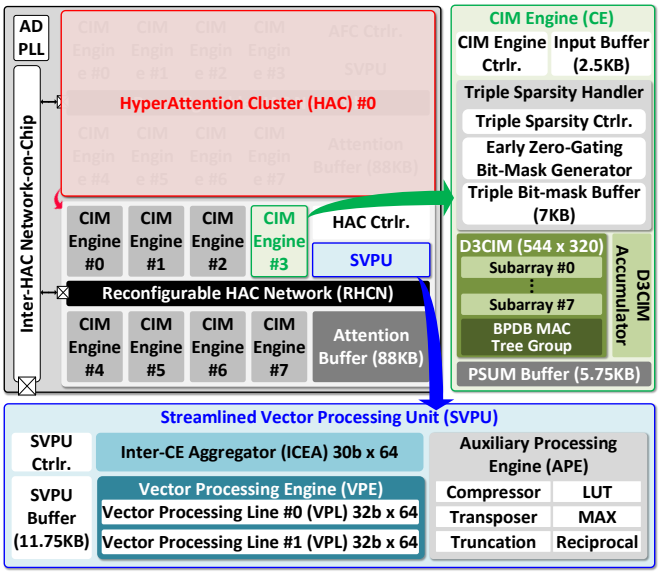


Fig. 2. Overall architecture of D3TA

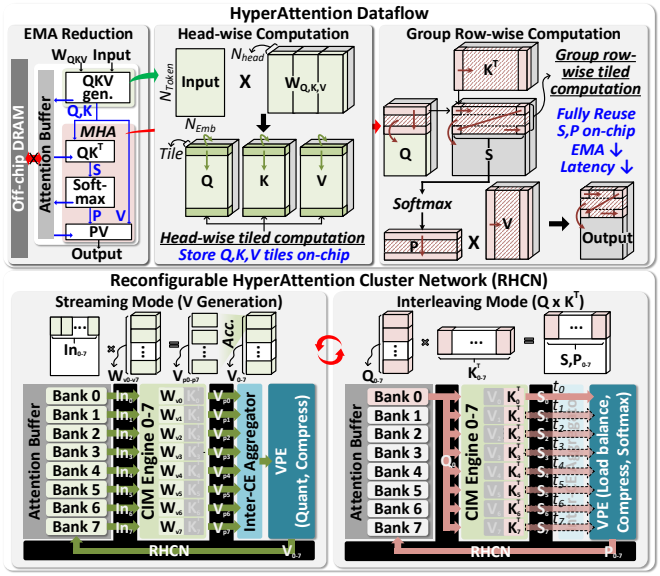


Fig. 3. HyperAttention (HA) dataflow and RHCN to reduce EMA

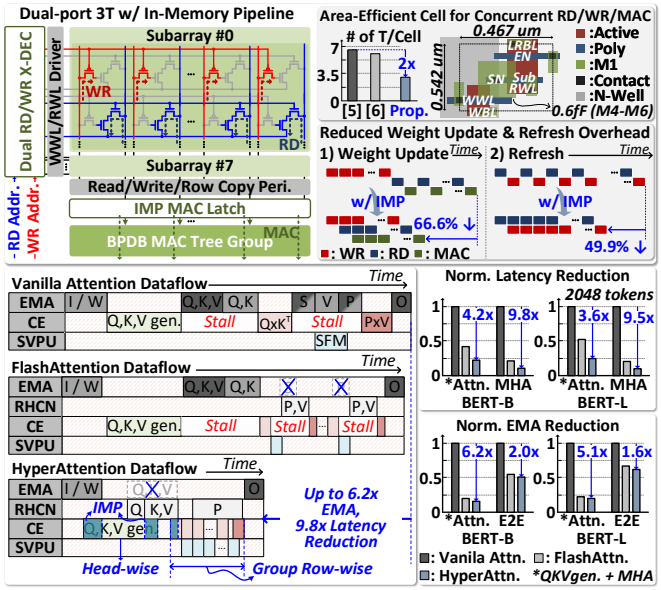


Fig. 4. Dual-port 3T with in-memory pipeline (IMP) and HA dataflow

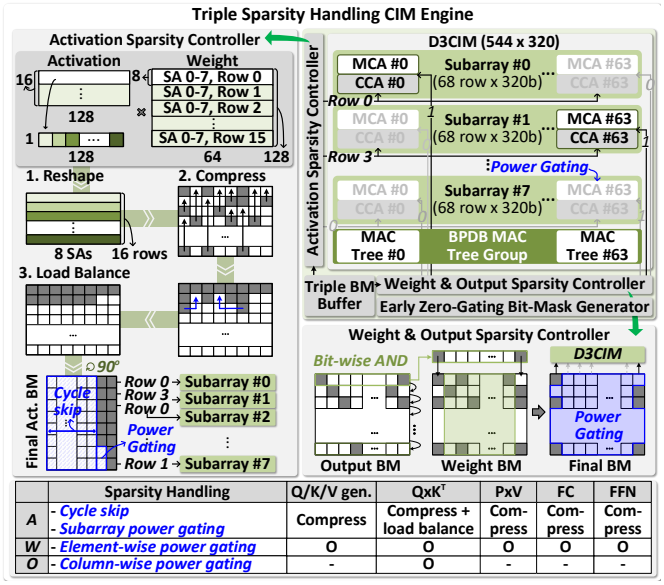


Fig. 5. Triple sparsity handling CIM engine for activation, weight, and output

optimal data movements for HA. RHCN reconfigures the on-chip network for streaming mode and interleaving mode. The streaming mode enables high BW data fetch from the attention buffer to VPE through CEs and ICEA, while the interleaving mode broadcasts data from a single bank of the attention buffer to CEs and interleaves CEs to fetch S to VPE for softmax. The IMP, enabled by the area-efficient dual-port 3T cells, dual X-DEC, and MAC latch, concurrently processes read, write, and MAC, reducing W fetching and refresh latency by 66.6% and 49.9%, respectively, as shown in Fig. 4. This concurrent read/write/MAC is achieved by only 3.33 transistors per cell, 1.9× lower than [5], [6]. Thanks to RHCN and IMP, HA ultimately allows off-chip access only for input and output of attention, achieving up to 6.2× EMA and 9.8× latency reduction over vanilla attention.

B. Triple Sparsity Handling CIM Engine

Fig. 5 demonstrates the TSH for A/W/O sparsity to increase computation efficiency. CE handles sparsity in activations for cycle skip and subarray gating, weights for element-wise gating, and output for column-wise gating. Based on the activation BM, pre-processed in the compressor, the TSC only activates nonzero subarrays of the D3CIM to reduce energy and separately controls the row address of each subarray to skip computation cycles. To skip cycles further, the TSC applies load balancing by dynamically redistributing the W with the row copy. The copied W can be reused for subsequent computations based on the column-wise spatial locality in

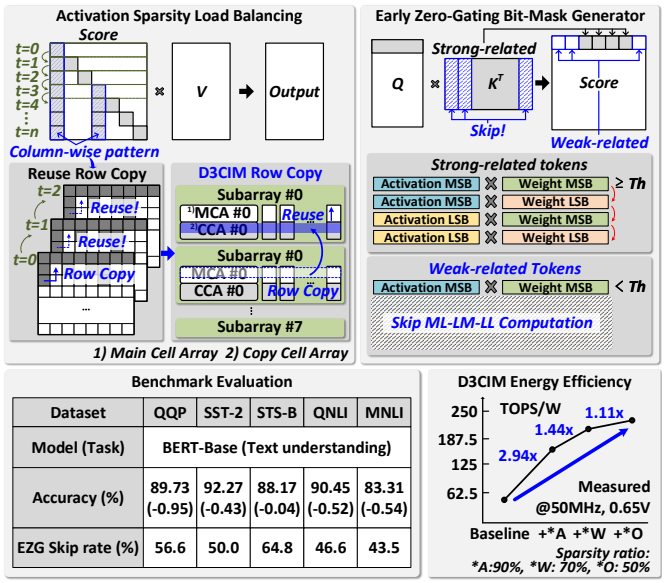


Fig. 6. Activation and output sparsity handling with load balancing and EZG

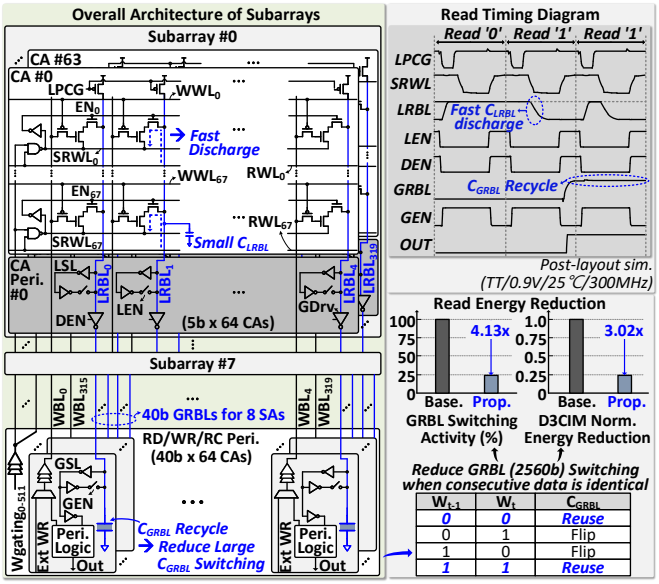


Fig. 8. Charge recycling subarray for high energy efficiency

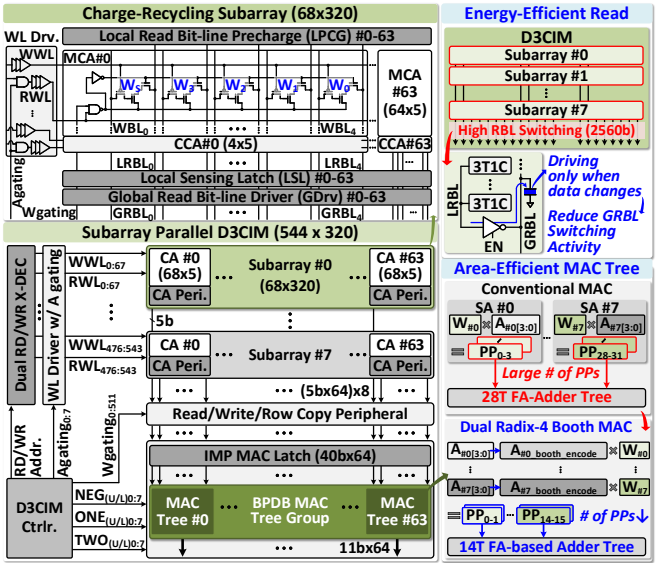


Fig. 7. Overall architecture of energy and area efficient D3CIM Macro

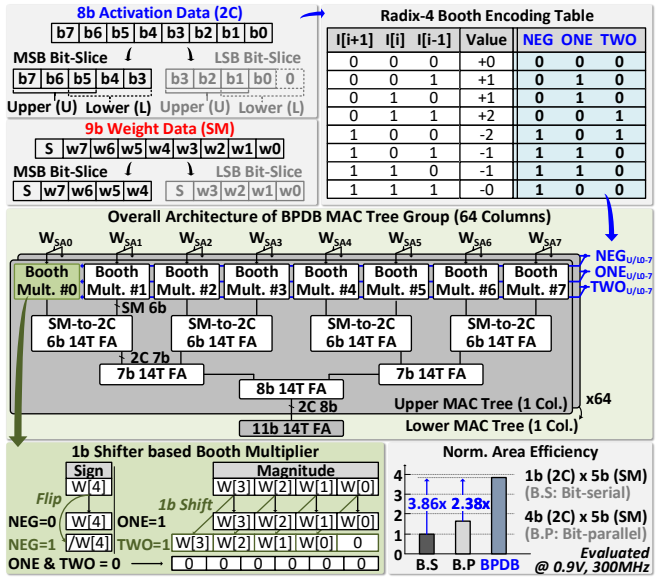


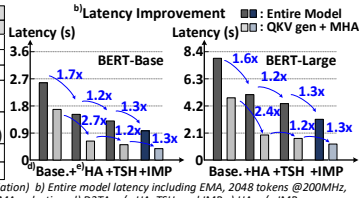
Fig. 9. BPDB MAC tree group for high area efficiency

$Q \times K^T$ [7], eliminating redundant data movement, as shown in Fig. 6. Furthermore, the EZG skips the computations in the weak-related tokens for the O sparsity when the MSB bit-slice computation is smaller than the threshold. The EZG BM generator dynamically generates the output BM and combines it with weight BM through bit-wise AND for the final BM. Thanks to the EZG, $Q \times K^T$ can be skipped by 43.5-64.8% while the accuracy loss is less than 0.95%. Finally, D3CIM achieves 4.7× higher energy-efficiency with sparsity.

C. Energy/Area Efficient D3CIM Macro

Fig. 7 shows D3CIM to reduce energy and area overhead. First, each subarray uses hierarchical read bit-line (RBL) with CR, as shown in Fig. 8. A local sensing latch (LSL) that reads data stored in cells and a global read bit-line driver (GDrv) that drives the global read bit-line (GRBL) to ensure that the global sensing latch (GSL) senses the read data. Thanks to the small capacitance of the local read bit-line (LRBL), the cell discharges the LRBL quickly even with transistors of low driving strength, and data can be sensed accurately with small-sized LSLs. CR minimizes the consecutive GRBL switching to reduce energy, achieving over 4× less GRBL switching and 3× less energy. Second, Fig. 9 shows BPDB MAC tree group with high area-efficiency through booth-encoding-based 4b×5b (2C×SM) MAC and 14T FAs [8]. Each 4b A for 8 subarrays is divided into upper and lower blocks, and the CE controller encodes those blocks to generate $NEG_{U/L}$, $ONE_{U/L}$, and $TWO_{U/L}$ signals for 8 subarrays. The generated

Evaluation on Transformer				
Model	BERT-Base		BERT-Large	
Task	Text Understanding			
Dataset	SST-2	QNLI	SST-2	QNLI
Parameter	110 M		340 M	
D3CIM & SVPU Precision	8b/9b & 32b		8b/9b & 32b	
[a]Accuracy (%)	92.27 (-0.43)	90.45 (-0.52)	92.76 (-0.31)	91.56 (-0.40)
[b]Latency	976.03 ms		3298.87 ms	
[c]EMA Reduction	2.0x / 6.2x		1.6x / 5.1x	

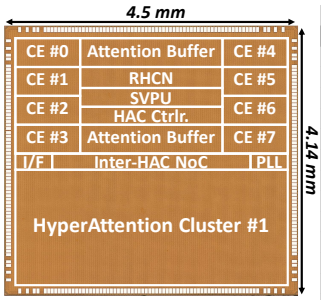


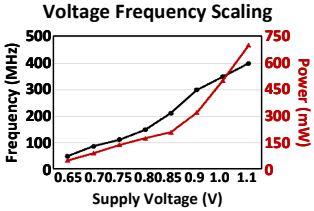
a) Accuracy w/ EZG (accuracy loss due to EZG, not PVT variation) b) Entire model latency including EMA, 2048 tokens @200MHz, 0.9V c) Entire model EMA reduction / (QKV gen. + MHA) EMA reduction d) D3TA w/o HA, TSH, and IMP e) HA w/o IMP

	ISSCC'22 [1]	ISSCC'23 [2]	ISSCC'23 [3]	ISSCC'24 [9]	ISSCC'23 [10]	SoVC'24 [11]	This Work
Technology [nm]	28	28	28	28	28	28	**28**
Transformer Target	O	O	O	O	X	O	**O**
Compute Type	Digital CIM	Digital CIM	Digital CIM	Digital + SNN	Analog CIM	Analog/ Digital CIM	**Digital CIM**
Cell Type	6T SRAM+ 4T NOR	6T SRAM	8T/10T SRAM	[3]-	3T2C eDRAM	1T1C eDRAM	**3T1C eDRAM**
Macro Size [KB]	24	128	144	[3]-	1200	3456	**340**
Supply Voltage [V]	0.6 – 1.0	0.6 – 1.0	0.64–1.03	0.7 – 1.1	1.0	1.0	**0.65–1.1**
Frequency [MHz]	80 – 240	85 – 275	20 – 320	50 – 200	250	250	**25 – 400**
Area [mm^2]	6.83	14.36	3.93	20.25	20.25	6.48	**18.63**
[1]Peak Performance [TOPS]	1.48 (8b/8b)	3.55 (8b/8b)	3.33 (8b/8b)	3.41 (8b/8b)	4.9 (8b/9b)	-	[4][6]**15.53 (8b/9b)**
[1]Peak Energy Efficiency [TOPS/W]	12.5–20.5 (8b/8b)	48.4–101 (8b/8b)	1.96–25.2 (8b/8b)	22.9–47.8 (8b/8b)	7.8–13.8 (8b/9b)	7.4 - 27.2 (8b/8b)	[5][7]**4.1–[4][7]38.9 (8b/9b)**
[1]Peak Area Efficiency [TOPS/mm^2]	0.22 (8b/8b)	0.25 (8b/8b)	0.85 (8b/8b)	0.17 (8b/8b)	0.24 (8b/9b)	-	**0.83 (8b/9b)**
TSH Support	X	X	X	X	X	X	**O**
IMP Support	X	X	X	X	X	X	**O**
[2]System FoM	4.51	25.275	21.437	8.126	3.312	-	**32.287**

1) 1MAC = 2 Operation 2) System peak energy efficiency (TOPS/W) x system peak area efficiency (TOPS/mm^2) based on 8b 3) It is not CIM-based transformer accelerator 4) With sparsity A: 90% w/ compression and load balancing, W: 70%, O: 50% 5) w/o sparsity 6) @400MHz, 1.1V 7) @50MHz, 0.65V

Fig. 10. Measurement results and performance comparison table





Specifications	
Technology	Samsung 28 nm
Chip Area	18.63 mm^2
Macro Area	0.223 mm^2
Bit-Cell Area	0.253 um^2
[1]Retention Time	257 us
Supply Voltage	0.65 – 1.1 V
Clock Frequency	25 – 400 MHz
Macro Performance	
Input Bit Precision	4b (2C)
Weight Bit Precision	5b (SM)
Macro Density	0.74 Mb/mm^2
[2][3][5]Peak Energy Efficiency	241.3 TOPS/W
[2][3][6]Peak Area Efficiency	5.41 TOPS/mm^2
System Performance	
Input Bit Precision	8b (2C)
Weight Bit Precision	9b (SM)
[2][4][6][7]Peak Throughput	15.53 TOPS
[2][4][5][7]Peak Energy Efficiency	38.9 TOPS/W
[2][4][6][7]Peak Area Efficiency	0.83 TOPS/mm^2

1) Measured w/ D3CIM
2) 1MAC = 2 Operation
3) With Sparsity A: 90% w/ compression, W: 70%, O: 50%
4) With Sparsity A: 90% w/ compression and load balancing, W: 70%, O: 50%
5) @50MHz, 0.65V
6) @400MHz, 1.1V
7) BERT-Base w/ 1024 token length

Fig. 11. Chip photograph and performance summary

48b encoded data broadcasts to 64 columns and is multiplied with 40b W data through the 1b-shifter-based booth multiplier in each column. If ONE and TWO are zero, multiplication is skipped. The final outputs are 64 11b, considering the bit position. The BPDB MAC tree group achieves 3.86× and 2.38× higher area-efficiency compared to bit-serial and bit-parallel computing.

III. Measurement Results and Conclusion

The D3TA operates at 25-to-400MHz with 0.65-to-1.1V, and by leveraging key features, it achieves up to 2.6× latency reduction and 2.0× EMA reduction compared to the baseline, as shown in Fig. 10. The system peak throughput is 15.53 TOPS with energy efficiency of 38.9 TOPS/W, and the system FoM is 1.28-to-9.75× higher than [1]–[3], [9], [10]. Fig. 11 shows the chip photograph and performance summary of D3TA fabricated in 28nm CMOS technology. In conclusion, D3TA optimizes transformer acceleration in memory, achieving high energy and area efficiency with reduced EMA and latency.

Acknowledgment

This work was supported by IITP grant funded by the Korea government (MSIT) (No.2022-0-01037, Development of High Performance Processing In Memory Technology based on DRAM and IITP-2025-RS-2023-00256472, Graduate School of Artificial Intelligence Semiconductor).

References

[1] F. Tu *et al*. "A 28nm 15.59 μJ/token full-digital bitline-transpose CIM-based sparse transformer accelerator with pipeline/parallel reconfigurable modes" 2022 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2022.

[2] F. Tu *et al*. "16.1 MuITCIM: A 28nm 2.24μJ/token Attention-Token-Bit Hybrid Sparse Digital CIM-Based Accelerator for Multimodal Transformers." 2023 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2023.

[3] S. Liu *et al*. "16.2 A 28nm 53.8 TOPS/W 8b sparse transformer accelerator with in-memory butterfly zero skipper for unstructured-pruned NN and CIM-based local-attention-reusable engine" 2023 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2023.

[4] 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" 2024 European Solid-State Electronics Research Conference (ESSERC). IEEE, 2024.

[5] W.-S. Khwa *et al*. "34.2 A 16nm 96Kb Integer/Floating-Point Dual-Mode-Gain-Cell-Computing-in-Memory Macro Achieving 73.3-163.3 TOPS/W and 33.2-91.2 TFLOPS/W for AI-Edge Devices" 2024 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2024.

[6] H. Fujiwara *et al*. "34.4 A 3nm, 32.5 TOPS/W, 55.0 TOPS/mm 2 and 3.78 Mb/mm 2 Fully-Digital Compute-in-Memory Macro Supporting INT12× INT12 with a Parallel-MAC Architecture and Foundry 6T-SRAM Bit Cell" 2024 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2024.

[7] A. Yazdanbakhsh *et al*. "Sparse attention acceleration with synergistic in-memory pruning and on-chip recomputation" 2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO). IEEE, 2022.

[8] Y.-D. Chih *et al*. "16.4 An 89TOPS/W and 16.3 TOPS/mm 2 all-digital SRAM-based full-precision compute-in memory macro in 22nm for machine-learning edge applications" 2021 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2021.

[9] S. Kim *et al*. "20.5 C-Transformer: A 2.6-18.1 μJ/Token Homogeneous DNN-Transformer/Spiking-Transformer Processor with Big-Little Network and Implicit Weight Generation for Large Language Models" 2024 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2024.

[10] S. Kim *et al*. "16.5 DynaPlasia: An eDRAM in-memory-computing-based reconfigurable spatial accelerator with triple-mode cell for dynamic resource switching" 2023 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2023

[11] S. Hong *et al*. "Dyamond: A 1T1C DRAM In-memory Computing Accelerator with Compact MAC-SIMD and Adaptive Column Addition Dataflow" 2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits). IEEE, 2024.