CENTAUR: A 38.5-TFLOPS/W 600MHz Floating-Point Digital Compute-In-Memory Engine with 40nm Fusion RRAM-eDRAM Macros Featuring 3D-MAC Operation

Luqi Zheng[1], Amir Massah Bavani[1], Shuting Du[1], Te-Yu Hsin[1], Mufeng Chen[1], Win-San Khwa[2], Ashwin Lele[3], Harry Chuang[2], Yu-Der Chih[2], Meng-Fan Chang[2], Haitong Li[1]

[1]Purdue University, West Lafayette, United States [2]TSMC, Hsinchu, Taiwan [3]TSMC, San Jose, United States

Non-volatile memory (NVM) based compute-in-memory (CIM) accelerators are being actively developed for energy-constrained edge AI applications, where an increasing number of workloads demand high-precision floating-point (FP) data formats [1-2]. Existing NVM-based FP-CIM macro designs face three major challenges: (1) significant area and energy overhead from on-chip integer/FP conversion or large pre-alignment logic, (2) accuracy degradation due to architectural limitations, such as row-wise pre-alignment of weights, and (3) limited operating frequency constrained by slow NVM sensing [3-4]. To address these limitations, we develop and present CENTAUR, a floating-point CIM engine featuring RRAM-eDRAM fusion macros and a novel FP 3D-MAC dataflow. Our new CIM architecture eliminates non-computational (alignment-induced) accuracy loss, reduces area overhead, and enables high-speed, energy-efficient FP computation. Fabricated in 40 nm CMOS with foundry RRAM and validated on a full-stack testing platform, CENTAUR achieves 600 MHz operating frequency, 38.5 TFLOPS/W energy efficiency, and high inference accuracy with Tiny-ViT (Vision Transformer) on CIFAR-10 with only 1.75% accuracy degradation compared to software baseline. CENTAUR marks the first NVM-eDRAM CIM chip.

As shown in Fig. 1, floating-point matrix-vector multiply-accumulate (MAC) involves two distinct components: static, input-agnostic mantissa processing, and dynamic, input-dependent exponent operations. This intrinsic partitioning motivates our reliability-aware, RRAM-eDRAM co-computation flow: static weight mantissas are stored and processed in RRAM, while exponent-related dynamic computations are mapped to refresh-free eDRAM. In the proposed FP 3D-MAC operation, conventional exponent alignment is reformulated as a 3-operand multiplication using a shift vector S_{xy}, effectively forming a 3D dot-product. This 3-operand computation maps efficiently onto a 3T1C gain-cell structure, leveraging eDRAM's storage node and two input ports. The fusion CIM co-computation begins with RRAM-based multiplication of weight mantissa and input mantissa. The resulting partial product is then transferred to eDRAM, where the 3-operand multiplication is performed. Final results are accumulated in a digital 3D-MAC accumulator, which also performs overflow detection and conversion back to floating-point format.

Fig. 2 illustrates CENTAUR's top-level architecture and dataflow. The fusion CIM engine consists of four RRAM-eDRAM fusion CIM macros, a sign-exponent processing core, a top-level engine controller, a global SRAM buffer, and an input/output SRAM buffer. Weight mantissas and exponents are stored in the RRAM macros and the exponent core, respectively. The input vector is split into mantissa and sign-exponent parts, temporarily stored in the global buffer. The exponent core reads input sign and exponent, fetches weight exponent from RRAM, performs sign multiplication, and writes the result to the sign register to steer accumulation. The exponents are bias-adjusted, one-hot encoded into shift vectors, and written into eDRAM via the global buffer. Input mantissas are then injected into both RRAM and eDRAM macros for co-computation. RRAM handles mantissa multiplication, while eDRAM performs shift-vector-based multiplication. Positive/negative MAC results are then separated and sent to the 3D-MAC accumulator. Data transfer between RRAM and eDRAM is managed by the fusion CIM bridge, while local and global flows are controlled by the macro and top engine controllers. Leveraging the decoupled read/write ports of gain-cell eDRAM, our design enables parallel computation and data updates across different rows, reducing latency over conventional alignment-based approaches.

Fig. 3 presents a detailed example of the fusion CIM 3D-MAC operation in FP5 format for simplified illustration, showing how a conventional floating-point 2D-MAC is transformed into a fixed-point 3D-MAC via RRAM-eDRAM co-computation, and cycle-accurate results at each computation stage. The digital 3D-MAC accumulator consists of four stages: data accumulation, partial sum tree, spatial sum tree, and temporal accumulation. After temporal accumulation, leading-one detection and fixed-point position prediction are performed to extract the resulting exponent and mantissa, which are then packed back into FP format with overflow detection. To meet timing requirements, the accumulator is deeply pipelined and implemented using a power-aware synthesis flow with a multi-Vt library. Post-routing simulation results show that the proposed 3D-MAC fusion CIM scheme achieves a 1.76× improvement in energy efficiency over conventional alignment-based FP-CIM, primarily by replacing high-overhead pre-alignment logic with eDRAM-based CIM operations.

Fig. 4 presents the macro-level circuit design, including detailed schematics of the RRAM and eDRAM macros, as well as the fusion CIM bridge. To improve RRAM read speed, a voltage-mode two-stage high-speed sense amplifier is implemented, consisting of a pre-amplifier followed by a strong-arm latch. The simulated offset is less than 5 mV, with an operating speed exceeding 800MHz. The gain-cell eDRAM adopts a 3T1C structure with a metal-stacked MOM capacitor and high-Vt access transistor, enabling refresh-free CIM operations with over 200 µs retention time. The timing diagram in Fig. 4 illustrates macro-level sensing, control signal sequencing, and cross-domain data transfer through the fusion CIM bridge. To enhance throughput, an interleaving scheme is applied across multiple RRAM and eDRAM macros, effectively doubling data bandwidth and improving overall system performance.

Fig. 5 shows a testing setup connecting the CENTAUR test chip to a Xilinx Ultrascale+ FPGA board, where a MicroBlaze CPU executes C-coded initialization and configuration routines, interfacing DDR4 and the test chip through a DDR4 controller and a custom AXI-to-chip interface IP. Internal signals are monitored using Vivado ILA, while external waveforms are captured with an oscilloscope. We evaluate inference accuracy and power consumption running Tiny-ViT on CIFAR-10. CENTAUR achieves 84.5% top-1 inference accuracy, with only 1.75% degradation from the software baseline. Oscilloscope waveforms capture an end-to-end latency of 11.86 µs for a typical fusion CIM task when operating at 600 MHz.

Fig. 6 summarizes the area and power breakdown, figure-of-merit (FoM) comparison, and benchmarking against prior works. The engine-level area breakdown shows that RRAM and eDRAM macros together occupy 69.9% of the total area, while the 3D-MAC accumulator accounts for 22.3%. Measured power profiling reveals that over 70% of the total power is consumed by digital logic—consistent with the characteristics of our fully digital FP-CIM implementation and design specs. These results suggest that the CENTAUR architecture could gain substantially from scaled NVMs on advanced nodes. For benchmarking, we first define quality-of-result (QoR) as the ratio of on-chip inference accuracy to the ideal software baseline accuracy. The overall FoM—defined as energy efficiency × operating frequency × quality-of-results (QoR) / normalized area—provides a more comprehensive metric by incorporating both operating speed and inference quality, better reflecting real-world utility. CENTAUR achieves up to 4.84× FoM improvement over SRAM-CIM [5] and 8.48× over RRAM-CIM [6]. Finally, the testing setup, die photo, and detailed chip summary are provided in Fig. 7.

Acknowledgement: This work was supported in part by NSF FuSe2 (# 2425498) and UPWARDS. The authors thank additional TSMC colleagues for chip fabrication support.

References:

[1] T.-H. Wen *et al*., "A 22nm 16Mb Floating-Point ReRAM Compute-in-Memory Macro with 31.2TFLOPS/W for AI Edge Devices," ISSCC, pp. 580-582, 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," ISSCC, pp. 568-570, 2024.

[3] D. -Q. You et al., "A 22nm Nonvolatile AI-Edge Processor with 21.4TFLOPS/W using 47.25Mb Lossless-Compressed-Computing STT-MRAM Near-Memory-Compute Macro," VLSI 2024.

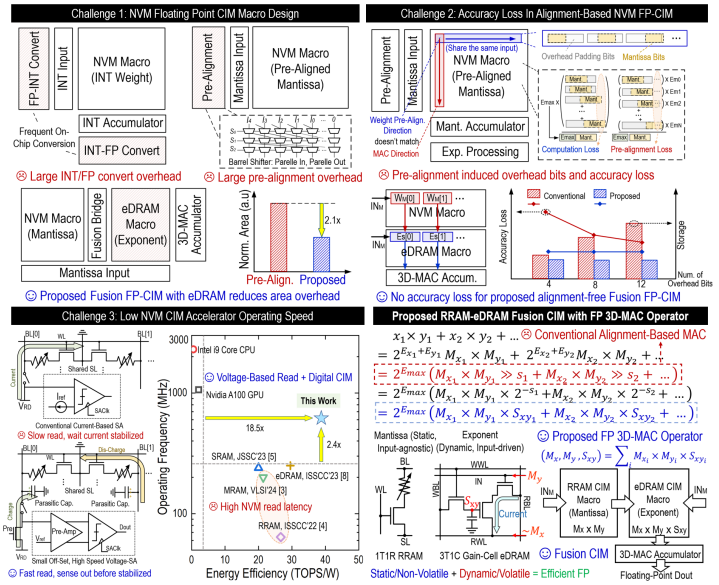


Fig. 1. FP-CIM challenges and our solutions (fusion architecture).

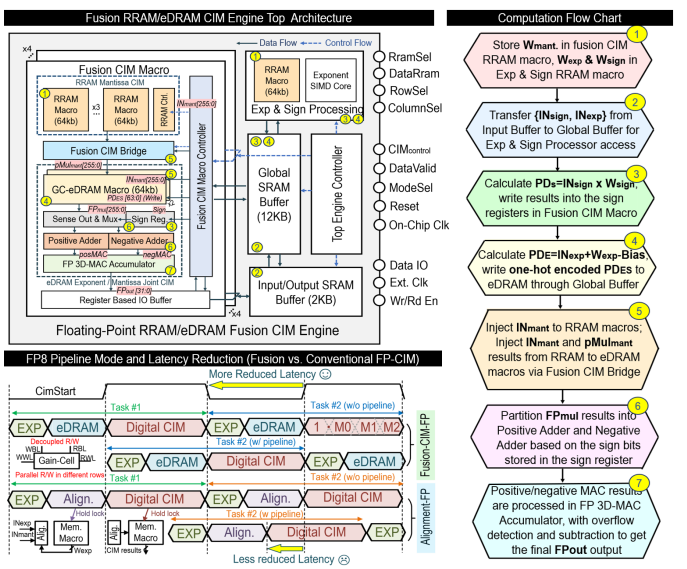


Fig. 2. System architecture, dataflow, and pipeline optimization of our RRAM-eDRAM fusion CIM Engine.

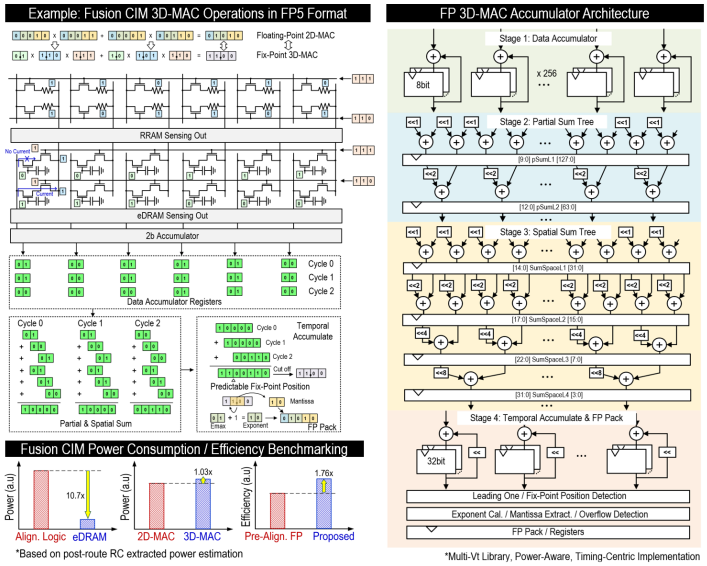


Fig. 3. Fusion CIM 3D-MAC operation, accumulator architecture, and simulated performance benchmarking.

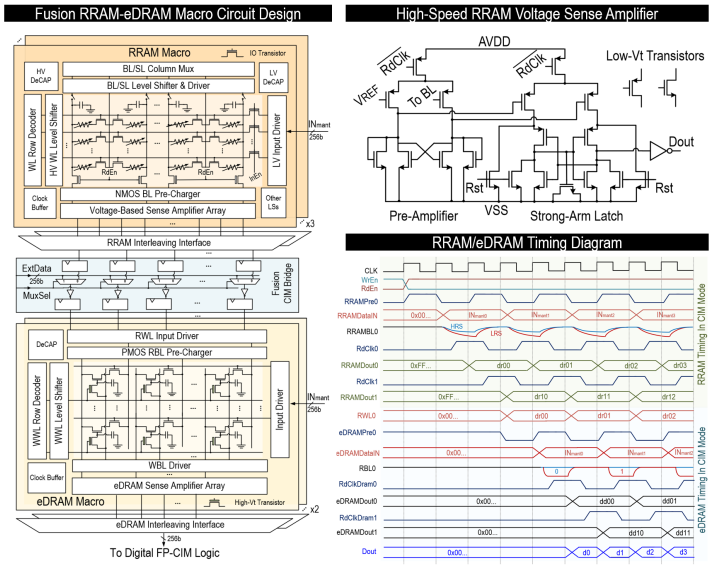


Fig. 4. RRAM-eDRAM macro circuit design, high-speed RRAM sense amplifier, and interface timing for fusion CIM.

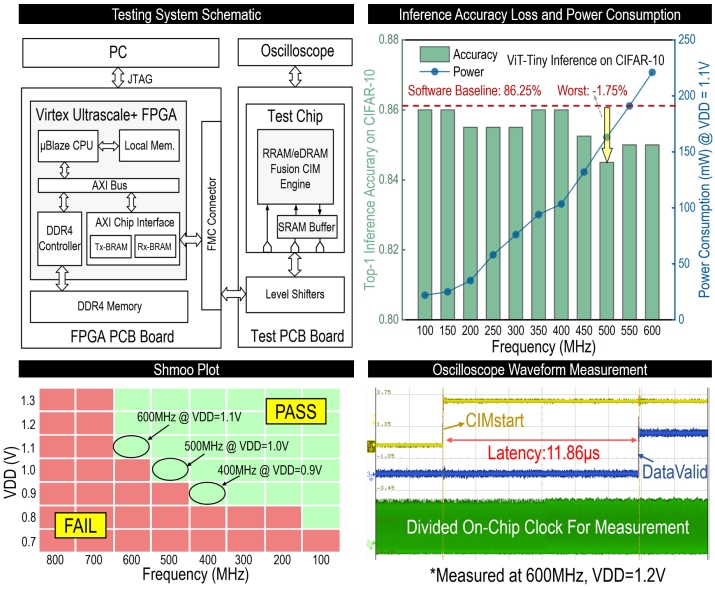
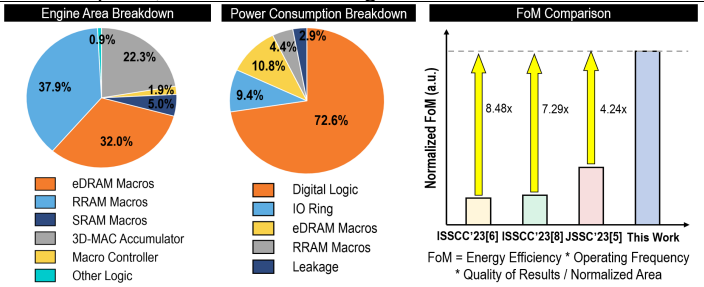


Fig. 5. Testing methods and silicon measurement results.

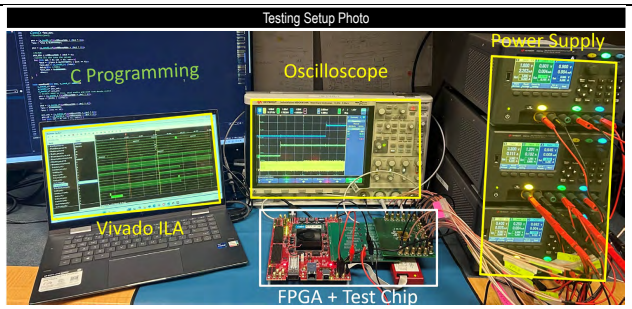


Comparison To Prior Works

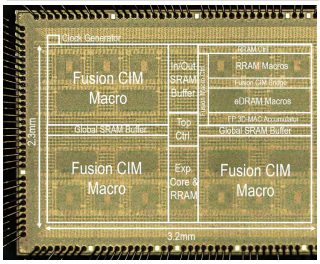
Reference	JSSC'23 [5]	ISSCC'23 [6]	ISSCC'24 [7]	ISSCC'23 [8]	ASSCC'24 [9]	This Work
Technology Node	28 nm	22 nm	40 nm	28 nm	28 nm	**40 nm**
Area (mm²)	6.69	24.48	20.25	20.25	0.05	**7.4**
Memory Type	SRAM	RRAM	RRAM+SRAM	eDRAM	eDRAM	**RRAM+eDRAM**
MAC Unit	Digital CIM	Analog CIM	Digital MAC	Analog CIM	Analog CIM	**Fusion Digital CIM**
CIM Memory Size (Kb)	192	32000	NA	9600	72	**1100***
Cell Density (Kb/mm²)	68.6	1300	2070*	950	1420	**885#**
eDRAM Cell / Refresh	NA	NA	NA	3T2C/Refresh	4T1C/Refresh	**3T1C/Refresh-Free**
Data Precision	INT8/INT16	INT4/INT8	INT8	INT4/INT8	INT4	**FP8**
Max. Frequency (MHz)	240	200	210	250	250	**600**
Peak Energy Efficiency (TOPS/W)	20.5	51.4	1.14	29.5	291.6	**38.5**

*RRAM macros only, without factoring in SRAM area; *Total RRAM + eDRAM capacity; #RRAM-eDRAM Fusion Cell Density

Fig. 6. Engine-level area/power breakdown and FoM comparison.



Die Photo



Chip Summary

Engine Area	2.3 mm x 3.2 mm
Technology Node	40 nm CMOS + Foundry RRAM
RRAM Capacity (Kb)	850
eDRAM Capacity (Kb)	256
SRAM Buffer (KB)	14
Supply Voltage (V)	1.1
Clock Frequency (MHz)	600
Weight/Input/Output Precision	FP8-FP8-FP8
Inference Accuracy Loss (ViT-Tiny on CIFAR-10)	1.75%
Power Consumption (mW)	212
Energy Efficiency (TFLOPS/W) (Peak Performance)	38.5

Fig. 7. CENTAUR's testing setup, die photo and chip summary.

Additional References:

[4] S. -D. Spetalnick *et al.*, "A 40nm 64kb 25.56TOPS/W 2.37Mb/mm2 RRAM Binary/Compute-In-Memory Macro with 4.23x Improvement in Density and >75% Use of Sensing Dyanmic Range," ISSCC, pp. 268-270, 2022.

[5] F. Tu *et al.*, "TranCIM: Full-Digital Bitline-Transpose CIM-based Sparse Transformer Accelerator With Pipeline/Parallel Reconfigurable Modes," JSSC, vol. 58, no. 6, pp. 1798-1809, 2023.

[6] W. -H. Huang *et al.*, "A Nonvolatile AI-Edge Processor with 4MB SLC-MLC Hybrid-Mode ReRAM Compute-in-Memory Macro and 51.4-251TOPS/W," ISSCC, pp. 15-17, 2023.

[7] S. D. Spetalnick *et al.*, "A 40nm VLIW Edge Accelerator with 5MB of 0.256pJ/b RRAM and a Localization Solver for Bristle Robot Surveillance," ISSCC, pp. 482-484, 2024.

[8] S. Kim *et al.*, "DynaPlasia: An eDRAM In-Memory-Computing-Based Reconfigurable Spatial Accelerator with Triple-Mode Cell for Dynamic Resource Switching," ISSCC, pp. 256-258, 2023.

[9] D. Kim *et al.*, "DPe-CIM: A 4T1C Dual-Port eDRAM Compute-in-Memory for Simultaneous Computing and Refresh with Adaptive Refresh and Data Conversion Reduction Scheme," ASSCC 2024.