Heterogeneous Compute-in-Memory Fabrics for Efficient, Scalable Edge Inference and Learning

Luqi Zheng, Zeshu Wang, Shuting Du, Mufeng Chen, Amir Massah Bavani, and Haitong Li
Elmore Family School of Electrical and Computer Engineering, Purdue University
West Lafayette, IN 47907, USA
haitongli@purdue.edu

***Abstract*—Edge deployment of LLMs and neuro-symbolic AI is increasingly constrained by the memory bottleneck and power walls, demanding new co-designed hardware solutions for energy-efficient, scalable inference and learning. We advocate treating memory heterogeneity and 'CMOS+X' integration collectively as a unified, first-class design principle for next-generation cognitive AI hardware. This paper reviews our recent work on heterogeneous compute-in-memory (CIM) fabrics: (i) CENTAUR, a 40 nm floating-point RRAM–eDRAM fusion CIM chip; (ii) an analog MLC eDRAM–RRAM CIM architecture co-designed with zeroth-order fine-tuning; (iii) monolithic 3D co-design methodology with emerging oxide-semiconductor transistors (OSFETs); and (iv) 3D-CIMlet, an open-source modeling framework to explore heterogeneous CIM chiplets with 2.5D/3D integration.**

***Index Terms*—Cognitive Computing, Compute-in-Memory, Memory-Centric Computing, 3D Integration**

I. Introduction

Cognitive computing with large language models [1] and neuro-symbolic AI [2] is catalyzing rapid progress across a wide range of societal and scientific domains. At the same time, translating these cognitive capabilities from cloud to edge exposes two fundamental challenges for the next-generation edge-AI hardware: energy efficiency and scalability. This work outlines a few conceptual and practical answers to address the efficiency and scalability challenges, exploiting heterogeneous compute-in-memory (CIM) fabrics. Towards this vision, we review and summarize our recent efforts spanning algorithm-hardware co-design, cross-layer modeling, chip prototyping, and exploration of emerging semiconductor technologies. First, we discuss our recent CENTAUR chip [3], the first floating-point RRAM-eDRAM fusion CIM demonstration, enabling high-precision, energy-efficient edge inference with heterogeneous computational memories. Second, we revisit zeroth-order fine-tuning and show how an analog heterogeneous CIM architecture can be co-designed to support memory-efficient on-device learning [4]. Third, we explore the vertical dimension and 'CMOS+X' integration, highlighting emerging oxide-semiconductor FETs (OSFETs) as back-end-of-line (BEOL) building blocks for co-designed 3D edge-AI systems that sustain scaling beyond planar CMOS while enriching functionality through heterogeneity. Finally, to systematically navigate the large design space of heterogeneous CIM fabrics in scaled-up/scaled-out 2.5D/3D integrated systems, we provide an overview of our open-source 3D-CIMlet modeling framework [5], and discuss key insights from an edge-LLM inference and continual learning case study.

II. CENTAUR: RRAM-eDRAM Fusion CIM Chip

As edge AI workloads evolve toward Transformer-based and hybrid models, maintaining numerical accuracy while operating under strict energy and area constraints becomes increasingly challenging. Conventional CIM accelerators based on a single memory technology struggle to simultaneously satisfy precision, efficiency, and scalability requirements while supporting floating-point (FP) computation. To address these challenges, our recent CENTAUR chip [3] introduced a floating-point RRAM–eDRAM fusion CIM architecture that leverages functional heterogeneity across memory technologies, enabling high-precision computation with improved energy efficiency and scalability.

Fig. 1 summarizes key challenges faced by prior NVM-based FP-CIM designs, which motivate a heterogeneous CIM design paradigm. Floating-point matrix–vector multiply–accumulate (MAC) operations naturally decompose into two distinct components: static, input-agnostic mantissa computation and dynamic, input-dependent exponent manipulation. Existing designs often force these components into a homogeneous, alignment-based dataflow, resulting in excessive non-computational overhead, accuracy degradation, and limited scalability. In contrast, a heterogeneous CIM fabric explicitly maps different computational roles to different memory substrates, enabling algorithm-aware specialization. In this context, RRAM is well suited for static mantissa storage and multiplication, while gain-cell eDRAM efficiently supports dynamic, refresh-free operations associated with exponent processing. Reformulating exponent alignment as a shift-vector-based multiplication further transforms conventional 2D FP-MAC into a fixed-point 3D-MAC operation, eliminating alignment-induced accuracy loss.

Building upon this principle, Fig. 2(a) illustrates a concrete realization of the heterogeneous CIM fabrics. The CENTAUR architecture integrates multiple RRAM–eDRAM fusion CIM macros, a sign–exponent processing core, distributed SRAM buffers, and a centralized controller. Input activations are decomposed and routed through parallel data paths, allowing mantissa and exponent computations to proceed concurrently across distinct memory domains. A dedicated fusion CIM bridge coordinates cross-domain data movement, enabling

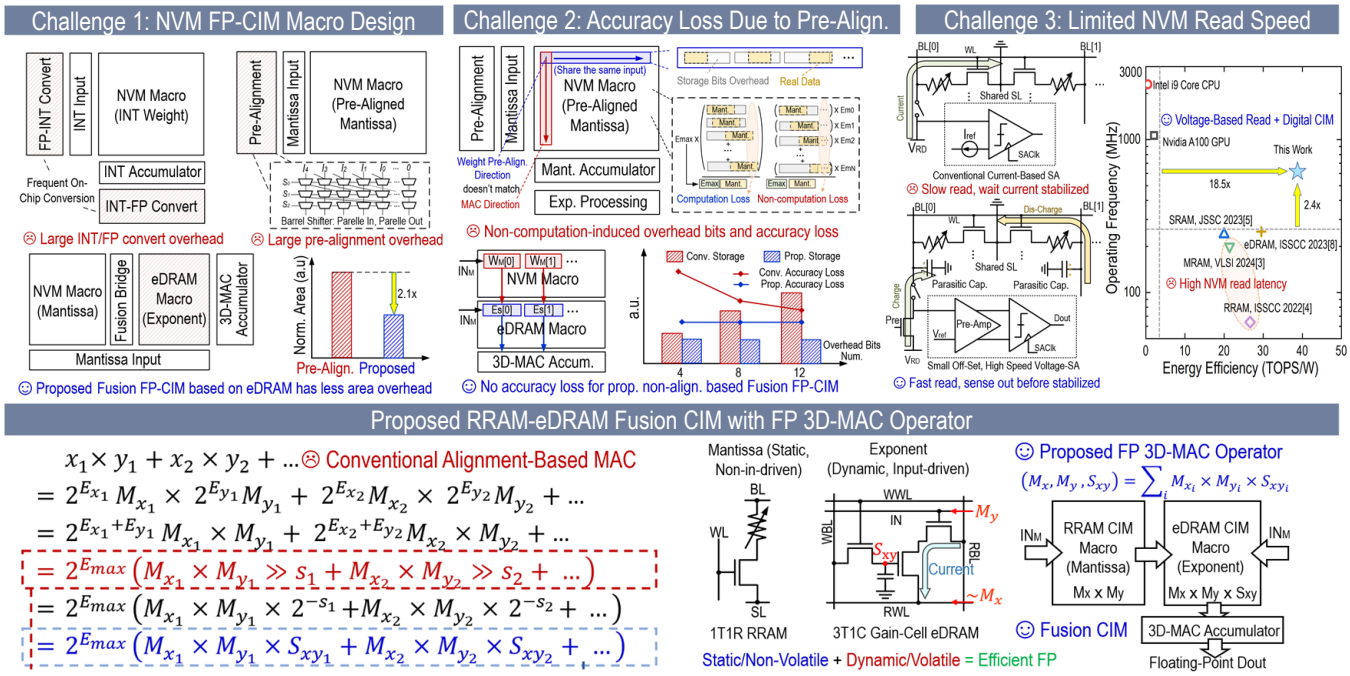


Fig. 1. FP-CIM challenges and our solutions presented in [3]: (1) large INT/FP conversion and pre-alignment overhead in NVM FP-CIM macros, (2) accuracy loss caused by alignment-based FP accumulation, and (3) limited NVM read speed and system throughput. To address these issues, a fusion RRAM–eDRAM CIM architecture with a 3D FP-MAC operator is proposed, which enables non-alignment-based FP accumulation, reduces storage and conversion overhead, and achieves fast, energy-efficient voltage-based reads while preserving computation accuracy.

tightly coupled co-computation without centralized data marshaling. This fabric-level organization allows the system to scale by replicating fusion macros and exploiting locality, rather than relying on monolithic, tightly synchronized CIM blocks. Fig. 2(b) further highlights how scalability is achieved through interleaved RRAM–eDRAM operations. By leveraging the decoupled read and write ports of gain-cell eDRAM, our design supports overlapped sensing, computation, and data update across different rows and macros. This interleaving scheme effectively hides NVM access latency and improves utilization, demonstrating how heterogeneity at the device and circuit levels translates into system-level throughput and scalability benefits.

The CENTAUR silicon prototype [3] is shown in Fig. 3, which exemplifies the heterogeneous CIM design paradigm. Fabricated in 40 nm CMOS with foundry RRAM, the chip integrates RRAM, eDRAM, and SRAM into a unified floating-point CIM fabric operating at 600 MHz and 1.1 V. System-level evaluation on Tiny-ViT (Vision Transformer) inference shows only 1.75% accuracy degradation compared to a software baseline, while achieving a peak energy efficiency of 38.5 TFLOPS/W. Together, these results validate that heterogeneous CIM fabrics can simultaneously deliver high numerical precision, energy efficiency, and robustness, which are key requirements for practical edge inference and learning systems.

III. Analog CIM for Zeroth-Order Fine-tuning

As on-device personalization and domain adaptation become increasingly important for LLM deployment at the edge, fine-tuning efficiency is increasingly dictated by memory overhead and data movement. Zeroth-order (ZO) optimization provides a promising alternative to backpropagation [6], [7], as gradient information is approximated through weight perturbation, avoiding explicit backward passes and reducing the need to store large volumes of intermediate activations and gradients. Existing memory-centric accelerator architectures fail to fully exploit these algorithmic benefits due to architectural inefficiencies in balancing bit density, compute-in-memory capability, and endurance-retention trade-offs.

To address this gap, our recent work [4] introduced a reliability-aware, analog multi-level-cell (MLC) eDRAM-RRAM CIM solution co-designed with zeroth-order forward-gradient optimization for language model fine-tuning, as summarized in Fig. 4. In the ZO optimization scheme, a single forward pass computes both the output and the Jacobian-vector product via weight perturbation, enabling efficient gradient updates without backpropagation. Our heterogeneous analog CIM architecture integrates MLC eDRAM and RRAM at the memory array level, eliminating redundant peripherals and enabling transposed-weight-free computation. The dataflow for forward-gradient computation is illustrated in Fig. 4. During the forward pass, the weights and a perturbation vector regenerated from a fixed random seed are each multiplied by the input and summed to produce both the output and the Jacobian–vector product (JVP). The loss function then uses the regenerated perturbation to form the local weight gradient, which drives the weight update. During testing, the updated weights (i.e., the base weights plus accumulated

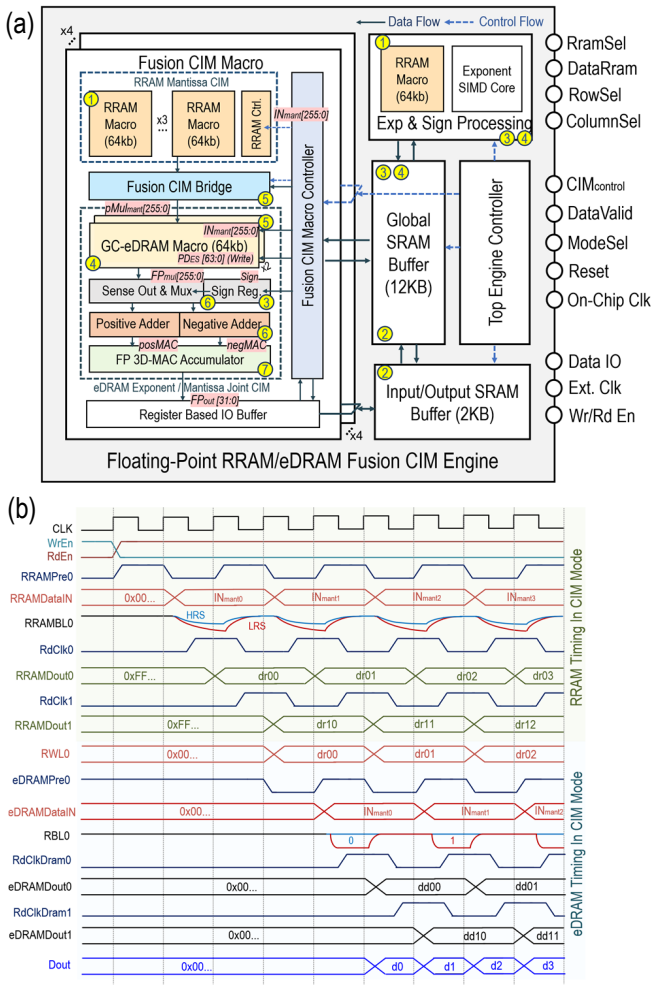
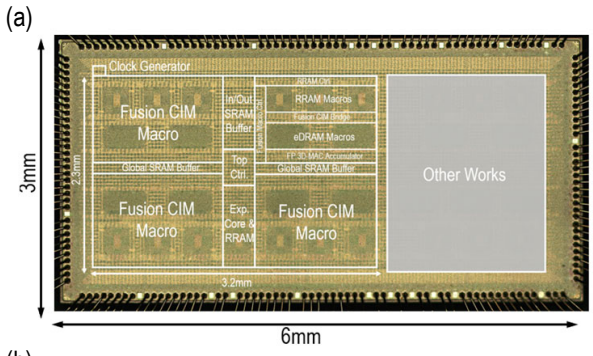


Fig. 2. (a) Top-level architecture of the CENTAUR floating-point RRAM–eDRAM fusion CIM engine. (b) Timing diagram illustrating interleaved RRAM–eDRAM digital CIM operations and control sequencing.



(b)

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. 3. (a) Die photo of the CENTAUR chip [3]. (b) Summary of key chip specifications and performance.

updates) are used for the forward pass. In our reliability-aware eDRAM-RRAM CIM architecture, RRAM cells handle static weights while endurance-unlimited eDRAM cells are tailored for dynamic operations that do not require long data retention.

A key circuit-level contribution is an RRAM-assisted eDRAM MLC programming and accumulation scheme that leverages DC-current-free charge sharing within the integrated array. The RRAM-assisted eDRAM MLC scheme can robustly program 16 levels into eDRAM, paired with a PVT-robust subthreshold time-to-digital converter (TDC). Designed in a 40 nm process, our silicon MLC-eDRAM combining middle-end-of-line (MEOL) capacitance enhancement with the back-end-of-line (BEOL) RRAM-assisted MLC programming provides $12\times$ improvement in bit density over state-of-the-art MLC design [8]. We further extended the BEOL co-design with ultra-thin-channel ALD In_2O_3 FETs to further improve bit density and retention, leveraging the high ON/OFF, low leakage, and CMOS BEOL compatibility of this emerging oxide-semiconductor transistor technology [9]. With the BEOL-stacked In_2O_3 FET serving the read transistor in the gain-cell eDRAM, bit density and retention are further improved by $5\times$ and $2\times$, respectively, compared to our silicon MLC-eDRAM design.

IV. Monolithic 3D Co-Design with OSFETs

Future edge inference and learning systems will likely integrate a diverse set of processing kernels, spanning digital and analog matrix-multiplication engines, convolution and attention accelerators, layer-normalization and activation units, neuromorphic primitives (e.g., event-driven spiking and hyperdimensional computing), probabilistic computing blocks, and dense, energy-efficient on-chip memories. Meeting the resulting demands for both efficiency and flexibility will require highly optimized yet versatile building blocks that can augment and extend beyond silicon CMOS, forming the foundational hardware fabrics for scalable edge inference and on-device learning. Towards such vision, vertically stacking layers of BEOL circuits on front-end-of-line (FEOL) silicon CMOS ('CMOS+X') through fine-grained and dense vertical inter-layer vias (ILVs), monolithic 3D integrated circuits (M3D-ICs) enable higher integration density, greater energy efficiency, higher bandwidth, and enhanced functionality [10]–[15]. Among emerging BEOL-compatible semiconductor device technologies, oxide semiconductor (e.g. InOx, ITO, IWO,

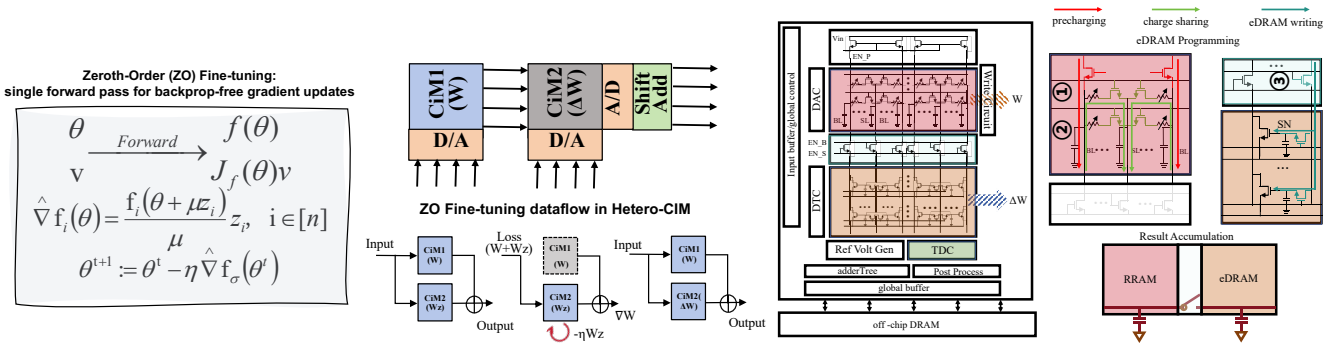


Fig. 4. Overview of the analog MLC eDRAM-RRAM CIM architecture co-designed with the zeroth-order fine-tuning scheme [4], where a single forward pass computes both the output and the Jacobian-vector product via weight perturbation, enabling efficient gradient updates without backpropagation. Heterogeneous analog CIM performs zeroth-order forward-gradient computation by using RRAM for static operations and eDRAM for dynamic updates, alleviating RRAM endurance and eDRAM retention constraints.

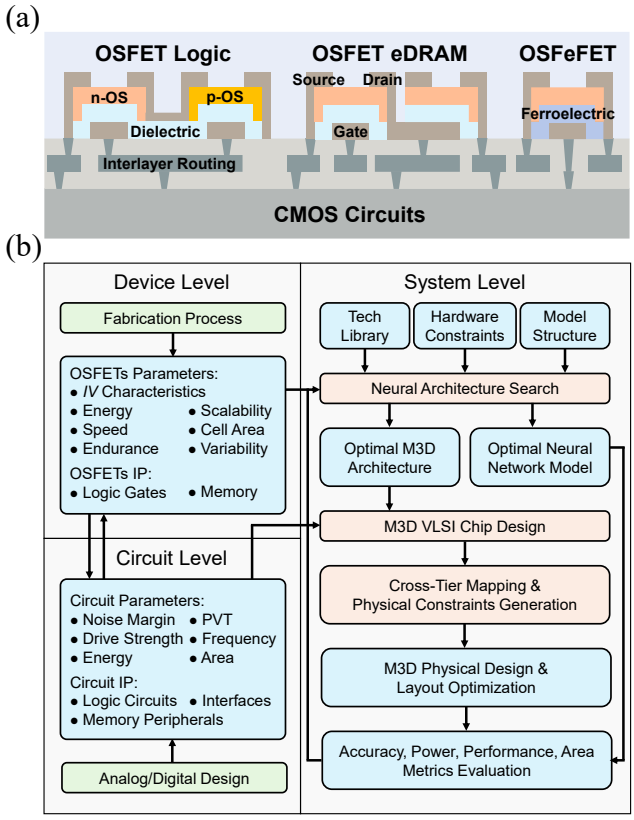


Fig. 5. (a) 'CMOS+X' with BEOL-compatible oxide semiconductor transistors for 3D logic and memory applications. (b) Overview of a device-circuit-system co-design pathway for efficient edge-AI hardware based on CMOS-OSFET integration in 3D.

IGZO, TeOx) field-effect transistors (OSFETs) have shown great potential for building energy-efficient M3D-ICs due to their advantages such as ultra-thin channels [16], low leakage current [17], satisfactory carrier mobility [12], excellent scalability [18], and wafer-scale fabrication process with BEOL-compatible temperature [19]. Here we advocate that the emerging OSFETs are energy-efficient, versatile building blocks for new flavors of heterogeneous CIM fabrics exploiting the third dimension. As illustrated in Fig. 5(a), OSFET-based logic and memory [12], [20] (e.g., digital logic, eDRAM, FeFET) can be integrated on silicon CMOS circuits, interconnected through ILVs in 3D, enabling high-performance 3D computing systems. Realizing the full benefits of 'CMOS+OSFET' M3D-ICs, however, requires more than device stacking: it calls for a device–circuit–system co-design methodology that jointly optimizes the BEOL devices, the interconnected compute/memory fabrics, and the end-to-end architecture for target inference and learning applications. While prior prototypes have demonstrated BEOL OSFET stacking on CMOS substrates [20], VLSI-scale integration strategies that translate these device capabilities into optimized 3D computing systems remain underexplored. To close this gap, we pursue the cross-layer co-design flow depicted in Fig. 5(b).

In the proposed co-design framework, at the device level, process flow and OSFET structures are tuned to meet target circuit requirements through device-circuit co-simulation and technology-aware evaluation. At the circuit level, logic and memory modules for key kernels are tailored to exploit OSFET properties while mitigating non-idealities. At the system level, the M3D architecture and neural models are first co-optimized via neural architecture search (NAS) methods using OSFET device characteristics, FEOL/BEOL technology options, hardware and 3D-IC design constraints, and model requirements as inputs. The resulting M3D architecture and tailored circuit design space then directly guide 3D VLSI implementation: logic and memory tiers are partitioned through cross-tier mapping to optimize inter-tier connectivity, signal integrity, and thermal behaviors, while extracted physical constraints (e.g., routing congestion and via parasitics) inform 3D physical design optimization. Performance-energy-density metrics extracted from the model-hardware co-optimized M3D-IC designs are finally fed back into NAS for closed-loop iterative optimization to continuously provide efficiency and functionality gains over planar silicon CMOS systems.

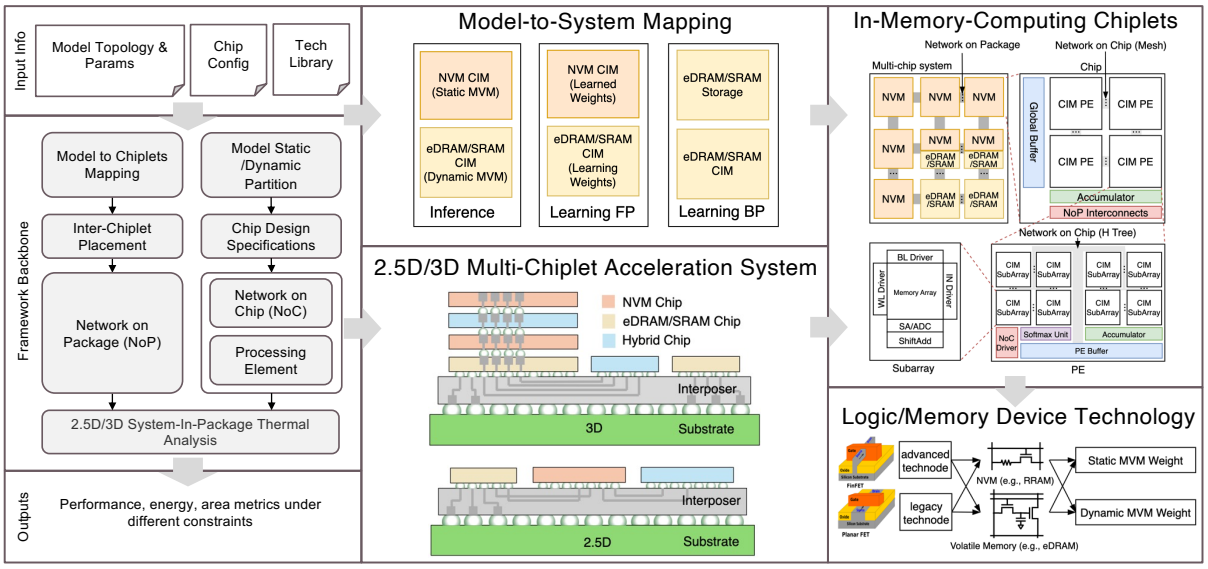


Fig. 6. Overview of 3D-CIMlet [5], a 2.5D/3D CIM chiplet modeling and co-design framework for edge LLM inference and continual learning.

V. 3D-CIMLET MODELING FRAMEWORK

The third dimension indeed opens complementary pathways for both scale-up and scale-out of edge inference and learning systems. In addition to the scale-up through monolithic 3D integration of BEOL device technologies as discussed in Section IV, the scale-out can be realized through heterogeneous chiplet assemblies using through-silicon vias (TSVs) and hybrid bonding with 2.5D/3D packaging, allowing system capability to grow by composing specialized CMOS and 'CMOS+X' dies with mixed technology options. In this context, we recently developed and open-sourced 3D-CIMlet [5], a cross-layer modeling and co-design framework enabling the joint design space exploration of heterogeneous CIM fabrics with 2.5D/3D integration. As illustrated in Fig. 6, the framework bridges the device technology libraries and system-level metrics by exposing a cross-layer design space that incorporates calibrated models for silicon and beyond-silicon computational memories, die-to-die (D2D) interconnects, and 2.5D/3D integration technologies. To demonstrate the capabilities and utilities of the 3D-CIMlet framework, we designed an edge-LLM system for inference and continual learning as a case study, and explored heterogeneous architectures composed of modular CIM primitives, including resistive RAM (RRAM), gain-cell embedded DRAM (eDRAM), and hybrid chiplets. The framework explicitly captures the resulting reliability spectrum: eDRAM offers effectively unlimited endurance but requires refresh management due to limited retention, whereas RRAM provides non-volatility but requires write-minimization and endurance-aware update strategies. Supporting such heterogeneous fabrics, the framework models hierarchical Network-on-Package (NoP) and Network-on-Chip (NoC) structures. By employing customized cycle-accurate simulation, it captures the complexities of diverse interconnect topologies across 2D, 2.5D, and 3D schemes, taking D2D interconnect specifications and intra-chiplet routing overheads into consideration. At the micro-architectural level, CIM chiplets are built from processing engines (PEs) equipped with charge-based or NVM-based subarrays for matrix-vector multiplication (MVM), alongside essential peripheral circuitry like softmax units and accumulators. The explored design space spans heterogeneous technology nodes from cost-effective 40 nm RRAM chiplets to more advanced 14/16 nm eDRAM implementations enabling systematic trade-off analysis for edge-LLM systems.

Leveraging the open-source 3D-CIMlet framework, we distill key design insights for scaling edge LLM systems based on heterogeneous CIM fabrics from the following perspectives:

1) *Architectural Efficiency Gains via Heterogeneity.* Transitioning from 2D monolithic designs to 2.5D/3D chiplet architectures can yield substantial energy efficiency gains, particularly for memory-intensive workloads like continual learning. This improvement comes from the reduced communication overhead by advanced packaging, and from the heterogeneity-aware mapping of workloads to memory technologies (e.g., RRAM, SRAM, eDRAM) that best match device characteristics.
2) *On-Chip Memory Trade-offs in Continual Learning.* For on-chip data storage of backpropagation-related dynamic weights, a central trade-off exists between SRAM leakage and eDRAM refresh energy. For larger-scale workloads, increasing SRAM buffer capacity reduces the reliance on costly eDRAM refreshes, often offsetting the SRAM leakage penalty. Additionally, the choice between RRAM and eDRAM for gradient storage is dictated by the learning schedule: parallel learning modes tend to favor eDRAM due to relaxed retention requirements, whereas sequential modes with large batches favor RRAM to avoid excessive refresh costs when

longer retention intervals are needed.

3) *Scaling and Thermal Trade-offs.* As systems scale, balancing inter-chiplet NoP and intra-chiplet NoC traffic is pivotal for minimizing communication costs. Regarding thermal considerations, while 3D stacking maximizes compute density, it also induces higher thermal stress due to increased thermal resistance. In contrast, 2.5D integration could offer a more uniform thermal profile, though localized hot spots in high-power density chiplets still motivate targeted, thermal-aware partitioning and floorplanning.

VI. Conclusion

Sustaining and scaling the cognitive AI capabilities at the edge call for new heterogeneous, memory-centric fabrics that co-optimize algorithms, architectures, circuits, and devices, while exploiting the third dimension for 'CMOS+X' integration and scaling. Realizing this vision will require continued efforts in developing efficient, heterogeneous CIM prototypes, experimentally-validated co-design approaches, cross-layer modeling tools, and 3D 'CMOS+X' primitives. Together, these directions point to a practical roadmap for energy-efficient, scalable edge inference and learning that can keep pace with rapidly evolving cognitive AI workloads.

Acknowledgment

This work was supported in part by the U.S. National Science Foundation under Award No. 2425498 with industry partners as specified in the Future of Semiconductors (FuSe2) program, and in part by the UPWARDS for the Future Network program. We thank Dr. Win-San Khwa and Prof. Meng-Fan Chang from TSMC Corporate Research along with other colleagues at TSMC for the collaboration on the CENTAUR chip.

References

[1] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi *et al.*, "Deepseek-r1 incentivizes reasoning in llms through reinforcement learning," *Nature*, vol. 645, no. 8081, pp. 633–638, 2025.

[2] T. H. Trinh, Y. Wu, Q. V. Le, H. He, and T. Luong, "Solving olympiad geometry without human demonstrations," *Nature*, vol. 625, no. 7995, pp. 476–482, 2024.

[3] L. Zheng, A. M. Bavani, S. Du, T.-Y. Hsin, M. Chen, W.-S. Khwa, A. Lele, H. Chuang, Y.-D. Chih, M.-F. Chang, and H. Li, "CENTAUR: A 38.5-TFLOPS/W 600MHz Floating-Point Digital Compute-In-Memory Engine with 40nm Fusion RRAM-eDRAM Macros Featuring 3D-MAC Operation," in *IEEE Asian Solid-State Circuits Conference (ASSCC)*, 2025.

[4] M. Chen, L. Zheng, J.-Y. Lin, P. D. Ye, and H. Li, "Analog Multilevel eDRAM-RRAM CIM for Zeroth-Order Fine-tuning of LLMs," in *2025 IEEE International Memory Workshop (IMW)*. IEEE, 2025, pp. 1–4.

[5] S. Du, L. Zheng, A. M. Parvathy, F. Xie, T. Wei, A. Raghunathan, and H. Li, "3D-CIMlet: A Chiplet Co-Design Framework for Heterogeneous In-Memory Acceleration of Edge LLM Inference and Continual Learning," in *2025 62nd ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2025, pp. 1–7, available as open-source: https://github.com/NanoX-Lab/3D-CIMlet.

[6] G. Hinton, "The forward-forward algorithm: Some preliminary investigations," *arXiv preprint arXiv:2212.13345*, vol. 2, no. 3, p. 5, 2022.

[7] Y. Zhang, P. Li, J. Hong, J. Li, Y. Zhang, W. Zheng, P.-Y. Chen, J. D. Lee, W. Yin, M. Hong *et al.*, "Revisiting zeroth-order optimization for memory-efficient llm fine-tuning: A benchmark," *ICML*, 2024.

[8] J. Song, X. Tang, H. Luo, H. Zhang, X. Qiao, Z. Sun, X. Yang, Y. Wang, R. Wang, and R. Huang, "A calibration-free 15-level/cell eDRAM computing-in-memory macro with 3T1C current-programmed dynamic-cascoded MLC achieving 233-to-304-TOPS/W 4b MAC," in *2023 IEEE Custom Integrated Circuits Conference (CICC)*. IEEE, 2023, pp. 1–2.

[9] P. Ye, "Atomic-layer-deposited Atomically Thin In2O3 Channel for BEOL Logic and Memory Applications," in *2022 International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)*. IEEE, 2022, pp. 1–1.

[10] K. Akarvardar and H.-S. P. Wong, "Technology prospects for data-intensive computing," *Proceedings of the IEEE*, vol. 111, no. 1, pp. 92–112, 2023.

[11] H.-S. P. Wong, K. Akarvardar, D. Antoniadis, J. Bokor, C. Hu, T.-J. King-Liu, S. Mitra, J. D. Plummer, and S. Salahuddin, "A density metric for semiconductor technology [point of view]," *Proceedings of the IEEE*, vol. 108, no. 4, pp. 478–482, 2020.

[12] S. Datta, E. Sarkar, K. Aabrar, S. Deng, J. Shin, A. Raychowdhury, S. Yu, and A. Khan, "Amorphous Oxide Semiconductors for Monolithic 3D Integrated Circuits," in *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. IEEE, 2024, pp. 1–2.

[13] D. Jayachandran, R. Pendurthi, M. U. K. Sadaf, N. U. Sakib, A. Pannone, C. Chen, Y. Han, N. Trainor, S. Kumari, T. V. Mc Knight *et al.*, "Three-dimensional integration of two-dimensional field-effect transistors," *Nature*, vol. 625, no. 7994, pp. 276–281, 2024.

[14] L. Zheng and H. Li, "CMOS+ X Technologies for Neuro-Vector-Symbolic Computing," in *Proceedings of Device Research Conference (DRC)*, 2024, pp. 1–2.

[15] S. Liu, R. M. Radway, X. Wang, F. Moro, J.-F. Nodin, K. Jana, L. Yan, S. Du, L. R. Upton, W.-C. Chen *et al.*, "Monolithic 3-d integration of diverse memories: Resistive switching (rram) and gain cell (gc) memory integrated on si cmos," *IEEE Transactions on Electron Devices*, 2025.

[16] Z. Chen, Y. Yan, G. Ding, Y. Zhou, S. Han, and M. Zhang, "Annealing strategy toward achieving high-performance indium tungsten oxide thin-film transistors by equilibrating oxygen vacancy and chemisorbed oxygen," *IEEE Transactions on Electron Devices*, vol. 72, no. 3, pp. 1167–1173, 2025.

[17] S. Fujii, T. F. Lu, K. Ikeda, S. Y. Chang, K. Sakamoto, L. W. Chung, M. Okajima, J.-Y. Tsai, T. Kuroda, C. P. Hao *et al.*, "Oxide-Semiconductor Channel Transistor DRAM (OCTRAM) with 4F2 Architecture," in *2024 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2024, pp. 1–4.

[18] K. Hikake, X. Huang, S.-H. Kim, K. Sakai, Z. Li, T. Mizutani, T. Saraya, T. Hiramoto, T. Takahashi, M. Uenuma *et al.*, "Scaling Potential of Nanosheet Oxide Semiconductor FETs for Monolithic 3D Integration–ALD Material Engineering, High-Field Transport, Statistical Variability," in *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. IEEE, 2024, pp. 1–2.

[19] C. Niu, P. Tan, J.-Y. Lin, L. Long, Z. Lin, Y. Zhang, H. Wang, G. D. Wilk, and P. D. Ye, "First Demonstration of BEOL Wafer-Scale All-ALD Channel CFETs Using IGZO and Te for Monolithic 3D Integration," in *2024 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2024, pp. 1–4.

[20] S. Liu, R. M. Radway, X. Wang, F. Moro, J.-F. Nodin, K. Jana, S. Du, L. R. Upton, W.-C. Chen, J. Chen *et al.*, "Edge Continual Training and Inference with RRAM-Gain Cell Memory Integrated on Si CMOS," in *2024 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2024, pp. 1–4.