3D-CIMlet: A Chiplet Co-Design Framework for Heterogeneous In-Memory Acceleration of Edge LLM Inference and Continual Learning

Shuting Du[1], Luqi Zheng[1], Aradhana Mohan Parvathy[1], Feifan Xie[2], Tiwei Wei[2], Anand Raghunathan[1], Haitong Li[1]
[1]Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA
[2]School of Mechanical Engineering, Purdue University, West Lafayette, IN 47907, USA
{du335, zheng782, amohanpa, xie477, wei427, araghu, haitongli}@purdue.edu

***Abstract*—The design space for edge AI hardware supporting large language model (LLM) inference and continual learning is underexplored. We present 3D-CIMlet, a thermal-aware modeling and co-design framework for 2.5D/3D edge-LLM engines exploiting heterogeneous computing-in-memory (CIM) chiplets, adaptable for both inference and continual learning. We develop memory-reliability-aware chiplet mapping strategies for a case study of edge LLM system integrating RRAM, capacitor-less eDRAM, and hybrid chiplets in mixed technology nodes. Compared to 2D baselines, 2.5D/3D designs improve energy efficiency by up to 9.3x and 12x, with up to 90.2% and 92.5% energy-delay product (EDP) reduction respectively, on edge LLM continual learning.**

***Index Terms*—Transformers, Memory-Centric Computing, Continual Learning, Heterogeneous Integration, Chiplets**

I. Introduction

The rapid growth in large language models (LLMs) has fueled a pressing need for efficient on-device inference and learning solutions. The structural differences between LLM training, fine-tuning, inference, and continual learning introduce unique design challenges that have been overlooked in prior efforts. These challenges, particularly in dataflow, memory access, and computation, represent key gaps in the development of hardware accelerators for on-device training [1]. As depicted in Fig.1, BERT and Adapter-BERT [2] exhibit structural differences that highlight distinct hardware requirements for continual learning and fine-tuning. To meet the demands of high performance, flexibility, and adaptability, a large amount of edge ML accelerator studies focused on co-design integrating architectural and technological innovations. These include circuit designs leveraging memory-centric architectures based on silicon [3]–[5] or non-volatile memory (NVM) [6], [7], as well as advanced modeling and design exploration tools [8]–[11]. Furthermore, scaling and adapting the single-chip designs to multi-die architectures [11]–[13] has proven effective in bridging the gap between resource-constrained edge and high-performance edge systems capable of LLM workloads. However, despite rapid advancements across the technology-to-system stack, significant challenges persist in achieving efficient on-device LLM inference and continual learning, largely due to the absence of comprehensive co-design frameworks.

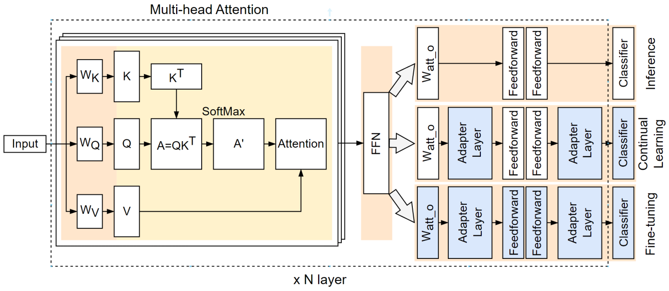

Fig. 1. Transformer-based LLMs create diverse requirements for edge inference and learning, adapter-based continual learning, fine-tuning, and inference layer structures highlighted in our case study.

In this work, a 2.5D/3D thermal-aware co-design framework leveraging heterogeneous in-memory computing is developed to design efficient chiplet-based edge LLM engines supporting inference and continual learning. With the developed co-design and modeling framework, efficient edge LLM inference and continual learning are demonstrated through a case study of reliability-aware heterogeneous eDRAM/RRAM chiplet designs in cost-effective, mixed-technology nodes. The framework is open-sourced.[1] The contribution of this work is summarized as follows:

- We develop 3D-CIMlet, a modeling and co-design framework that allows rapid design space exploration of 2.5D/3D chiplet-based accelerator architectures for transformers, leveraging heterogeneous memory technologies.
- Based upon 3D-CIMlet, we develop a heterogeneous RRAM/eDRAM CIM system with 2.5D and 3D integration schemes and corresponding reliability-aware mapping strategies to support efficient inference and continual learning of edge LLMs.
- Through chiplet-to-package, multi-scale design space explorations (DSE), we provide co-optimization guidelines spanning CIM chiplet designs (intra-chiplet and inter-chiplet), cost-aware and thermal-aware system integration, and runtime optimizations for continual learning.

[1] https://github.com/NanoX-Lab/3D-CIMlet

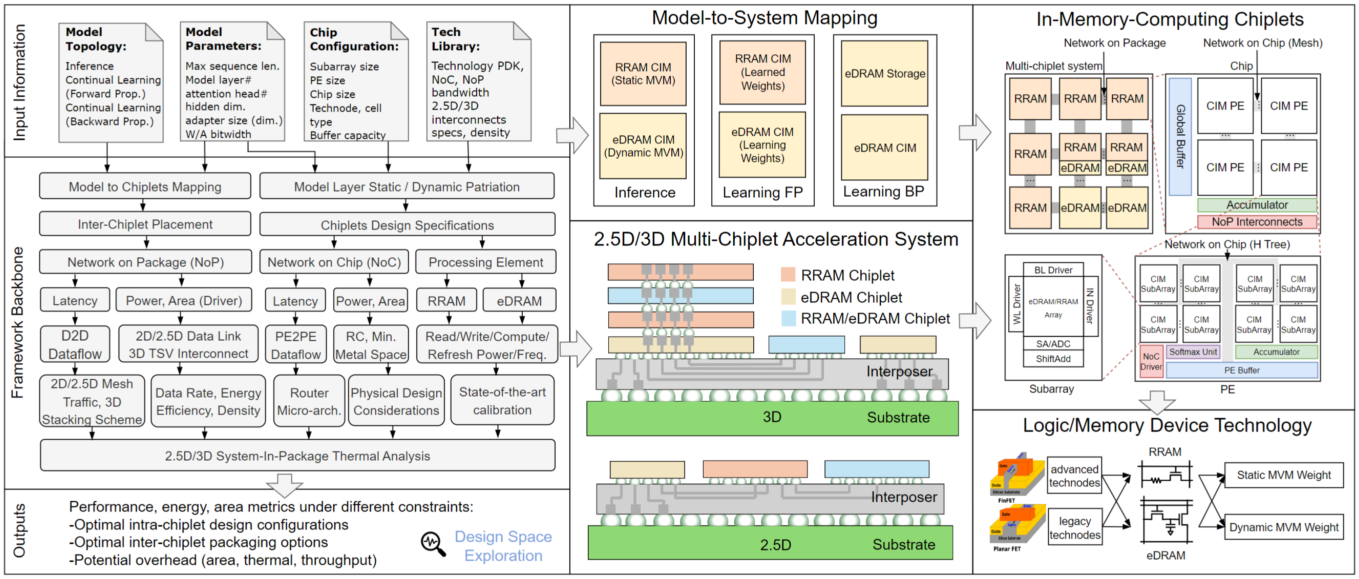

Fig. 2. Overview of 3D-CIMlet, a 2.5D/3D chiplet-based modeling and co-design framework for edge LLM inference and continual learning.

II. Background and Related Works

Single-chip accelerators for transformer-based models, such as SRAM-based CIM accelerators [5] and RRAM-based accelerators capable of storing all model weights [14], have shown promise for LLM edge applications. However, these single-chip solutions face limitations: they can only parallelize a few transformer layers [5] and accommodate very tiny models [14] within the on-chip embedded NVM. Benefiting from the advanced packaging technologies, multi-chiplet systems combine the strengths of mature single-die transformer accelerator designs with high-density edge NVM storage, enabling more robust and scalable edge LLM applications.

Multi-chiplet architectures have improved the performance of deep neural networks (DNNs) inference and learning significantly in the past decades. CIM architecture based prototypes include the CHIMERA Illusion System [12] and the multi-chip module (MCM) TensorCIM [13]. Moreover, [11] proposed a 2.5D chiplet-based RRAM-CIM accelerator simulator and benchmarking tool for DNNs. Despite these advancements, current architectures that rely on single-chiplet design foundations have not been fully expanded to support diverse chiplet configurations across systems to optimize for LLM workloads.

Previous modeling frameworks have primarily focused on single-die accelerators, leaving multi-die designs for edge LLM applications largely unexplored. Tools such as Timeloop/Accelergy [8] and AccelTran [9] targeted DNNs and sparsity-aware inference, respectively, but are limited to single-die architectures and inference scenarios. 3D Neurosim [10] explored advanced packaging with CIM architectures but remains limited in scalability for complex chiplet systems. Similarly, SIAM [11] benchmarked 2.5D RRAM-CIM accelerators but lacks support for continual learning workloads which are essential for edge LLMs.

III. 3D-CIMlet: Co-Design Framework

A. Methodology

3D-CIMlet is a two-prong co-design framework: (1) Exploiting transformer CIM designs with various chiplet architectures through advanced packaging, multi-scale technology-to-system design space is fully exposed. Thermal-aware design space explorations (DSE) can be conducted to probe into design tradeoffs and optimization opportunities. (2) Building upon the distributed memory-centric architectures, the framework explores memory-reliability-aware mapping strategies for both inference and continual learning. This is achieved by connecting inference and continual learning characteristics of edge LLMs to the heterogeneous memory characteristics from the chiplet primitives.

Figure 2 provides an overview of 3D-CIMlet methodology and key backbones. The framework takes transformer model inputs and calibrated technology libraries. The technology libraries include silicon and beyond-silicon computational memories, die-to-die (D2D) interconnects, and 2.5D/3D integration technologies as the foundational enablers. The scope of this work is focused on a case study analyzing heterogeneous computational memories in heterogeneous architectures. Hence, we architect the edge LLM system consisting of resistive RAM (RRAM), capacitor-less embedded DRAM (eDRAM), and hybrid RRAM/eDRAM chiplets designed as modular CIM primitives at both chiplet level and system level. The network-on-package (NoP) and network-on-chip (NoC) hierarchies support D2D interconnection and routing of CIM processing engines (PEs). Each CIM chiplet contains CIM PEs for matrix-vector multiplication (MVM), global buffers, accumulators, and NoP drivers. Each CIM PE includes CIM subarrays exploiting charge-based or non-charge-based computations, softmax units, accumulators, buffers and NoC drivers. Since

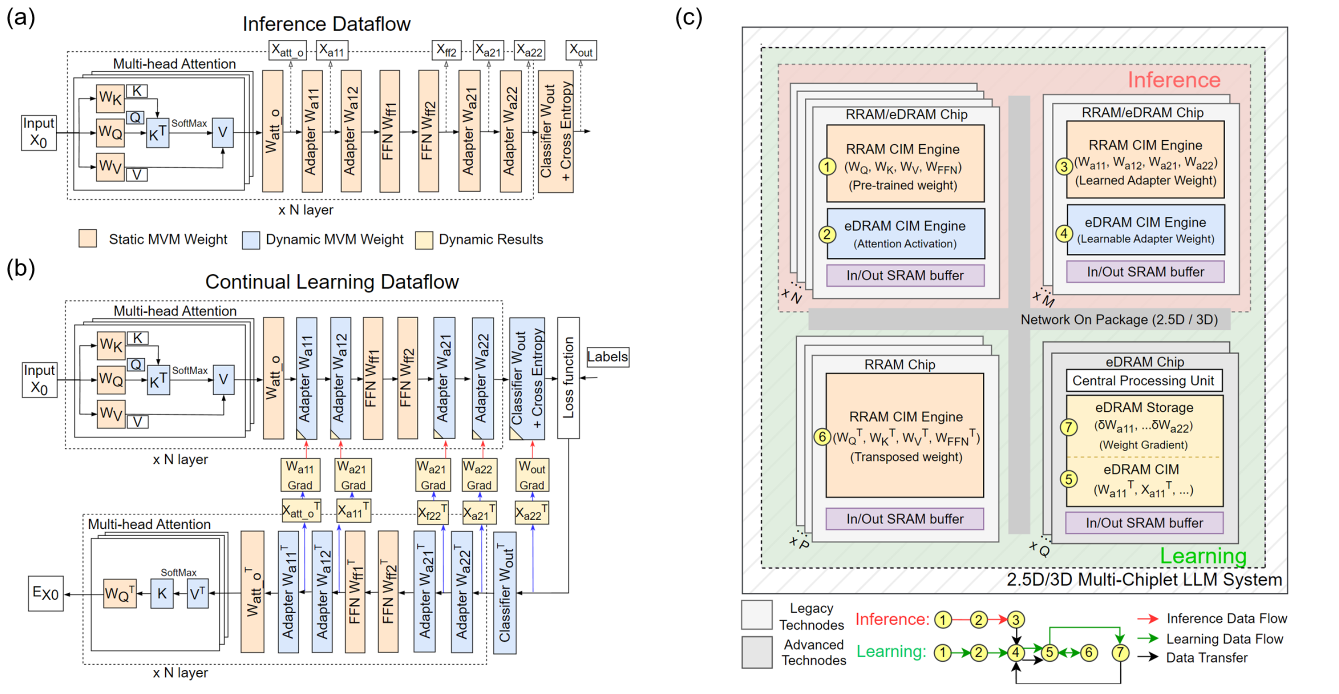

Fig. 3. The dataflow and static/dynamic operation partitioning in adapter-based (a) inference and (b) continual learning, together with (c) a reliability-aware mapping strategy for the heterogeneous multi-chiplet edge LLM system. With mixed technology nodes supported, RRAM/eDRAM chiplets are leveraged for inference and forward propagation during learning, while eDRAM chiplets are utilized for backpropagation.

this work focuses on edge LLM use cases, our case study takes a relatively conservative, cost-driven assumption to land on a sweet spot between cost [15], [16], embodied carbon footprint [17], and system efficiency: RRAMs are based on 40 nm foundry RRAM technology in this cost-effective legacy node [6], [12], [18], whereas capacitor-less eDRAMs being used are mature in both planar [3], [4] and FinFET (14/16 nm) nodes [19], [20]. For intra-chiplet NoC and inter-chiplet NoP modeling in the backbone of the framework, BookSim [21] is customized incorporating the 2D [22], 2.5D [23], [24], and 3D [25]–[27] system integration technologies.

First, for intra-chiplet design explorations, the framework categorizes and compiles input model layers into static and dynamic operations, and optimizations around those operations are guided by heterogeneous memory characteristics seen in silicon charge-based memories as well as beyond-silicon non-volatile memories (NVMs) across chiplets. In our case study, RRAM and eDRAM technologies form a wide ”reliability spectrum”: capacitor-less eDRAMs with unlimited endurance would suffer from limited retention leading to frequent refreshes, whereas excessive writes to RRAMs with long retention must be minimized. NoC behaviors and performance are simulated in a cycle-accurate fashion, focusing on PE-to-PE dataflow and router architecture, with power and area estimations based on RC considerations and minimum metal spacing. Second, inter-chiplet design exploration is associated with model-to-chiplet mapping strategies (Section IV). NoP behaviors and performance are analyzed considering D2D dataflow and traffic patterns, including 2D/2.5D mesh and 3D stacking schemes. Power and area are further modeled based on 2D/2.5D data links and 3D TSV interconnects [23]–[27].

B. Chiplet-to-Package Thermal Modeling

Finite Element Method (FEM) is employed to thoroughly analyze the thermal characteristics and performance of both 2.5D and 3D integration. The front end of line (FEOL) is modeled as a surface heat source, with all chiplets mounted on a 100 μm thick silicon interposer substrate. The chips are mechanically reinforced and encapsulated with an adiabatic molding compound. Heat transfer coefficients of 1000 W/(m²·K) on the top of silicon dies and 20 W/(m²·K) on the bottom of the organic substrate are used, reflecting moderate forced convection cooling on the package surface.

IV. Reliability-Aware Chiplet Mapping Strategies

The 3D-CIMlet framework aims to enable efficient chiplet design and system integration by combining inference and continual learning acceleration within a single chiplet-based system. Hardware resources are allocated dynamically to inference and continual learning tasks as needed. The 40 nm RRAM, 40 nm RRAM/eDRAM, and 14 nm eDRAM chiplets serve as modular, reusable components that facilitate the mapping of MVM operations and on-chip data storage within an integrated edge LLM engine (Fig. 3). Additionally, for inference-only edge use cases, the chiplet-based approach allows for a cost-effective, scaled-back version with chiplets in low-cost technology node only.

The dataflow of inference with learnable adapter layers is shown in Fig. 3(a). Adapter layers are lightweight modules

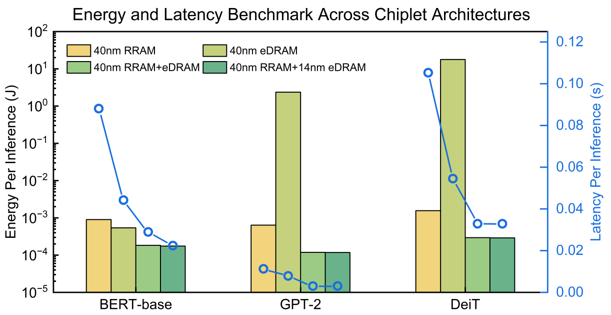

Fig. 4. The energy consumption and latency benchmark across different chiplet architectures for inference workloads. Heterogeneous RRAM/eDRAM designs improve the total inference energy by up to 1.4x and the inference latency by up to 3.9x from RRAM-CIM chiplet baseline.

added into pre-trained backbone networks, adapting the models to new tasks without modifying the original pre-trained weights. During inference, apart from the dynamic MVM operations between K, Q, and V in the attention layers of transformers, all other weights remain static since they belong to the pre-trained backbone and adapter layers. As shown in Fig. 3(c), inference workloads are mapped onto 40 nm RRAM/eDRAM chiplets with the wide retention spectrum: static MVM weights are handled by RRAM CIM engines, including the weights of K, Q, V projection layers, feedforward layers ①, and the trained adapter layers ③, which remain the same with varying input activations. The dynamic attention activations ② varying with input activations are allocated to eDRAM CIM engines without endurance limitations while incurring the trade-off of short retention times.

In addition to the RRAM/eDRAM chiplets accelerating inference, 14 nm capacitor-less eDRAM and 40 nm RRAM chiplets are leveraged to expand the system for continual learning, with the dataflow and mapping shown in Fig. 3(b) and Fig. 3(c). Forward propagation (FP) during learning is executed on the same chiplets allocated for inference, with the key distinction that adapter weights are learnable during FP as the signature of continual learning. To avoid endurance issues with RRAM during weight updates in backpropagation (BP), these learnable weights are initially loaded from the RRAM CIM engines ③ to the eDRAM CIM engines ④. BP operations are distributed across RRAM chiplets and 14 nm eDRAM chiplets. Errors propagated backward are multiplied by the transposed adapter weights and the transposed FP output activations to compute weight gradients. These computations occur in the endurance-optimal eDRAM CIM engines ⑤, with transposed static weights in the retention-optimal RRAM CIM engines ⑥. In contrast to high-bandwidth memory (HBM) based designs, the 14 nm eDRAM chiplets allow on-die integration with a processor unit, which orchestrates the process of storing the generated weight gradients in eDRAM chiplets ⑦. These gradients are retained until needed for weight updates during the subsequent FP iteration in the eDRAM CIM engines ④.

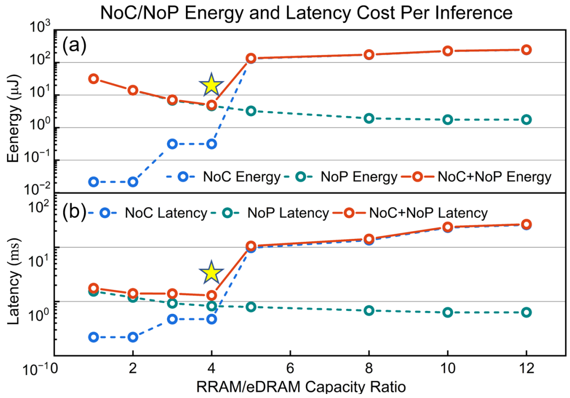

Fig. 5. The NoC and NoP energy consumption and latency associated with 40 nm RRAM/eDRAM chiplets with varying RRAM-to-eDRAM capacity ratios. (a) NoC and NoP energy cost per inference, (b) NoC and NoP latency per inference. At a ratio of 4, the balance of NoC and NoP is reached with minimized communication overheads.

V. Design Explorations and Analysis

A. Edge LLM Inference

For edge LLM inference, we evaluated models across two key aspects: (1) different input modalities (text and image) which require distinct embedding strategies, and (2) diverse transformer architectures, including encoder-only and decoder-only models. Language and vision models differ in embedding approaches; language models use embedding tables, whereas vision models often rely on convolutional layers. Our experiments included a convolution-based embedding layer for vision transformers to reflect this difference. Due to their autoregressive nature, decoder-only models process one query at a time and perform vector-matrix computations, in contrast to encoder-only models, which predominantly have matrix-matrix multiplications. Based on these distinctions, we evaluated a diverse suite of models: (1) BERT-base, an encoder-only text model, on the GLUE dataset (average of 128 tokens); (2) GPT-2, a decoder-only text generation model (128 tokens and 1024 tokens) on the Wikitext-2 dataset; (3) DeiT-base, an encoder-only image model, on ImageNet (196 tokens).

We illustrate 3D-CIMlet for inference through two sets of DSE analyses. First, we analyze energy consumption and performance across various chiplet configurations for inference workloads, including BERT, GPT-2, and DeiT. As shown in Fig. 4, across transformer model architectures, heterogeneous chiplet designs with RRAM and eDRAM available in the system demonstrate up to 3.9x and 2.6x performance and up to 1.4x and more than 10^{4}x energy efficiency improvement compared to those with NVMs or charge-based memories only.

Second, RRAM-to-eDRAM capacity ratio in 40 nm RRAM/eDRAM chiplets is a key design knob in our DSE experiments. To better understand the overall impact of NoP and NoC, their energy and latency overheads are evaluated. As shown in Fig. 5, the NoC energy and latency costs per inference gradually increase, while the NoP energy and latency decrease as the RRAM-to-eDRAM capacity ratio scales. This analysis reveals that an optimal balance in communication

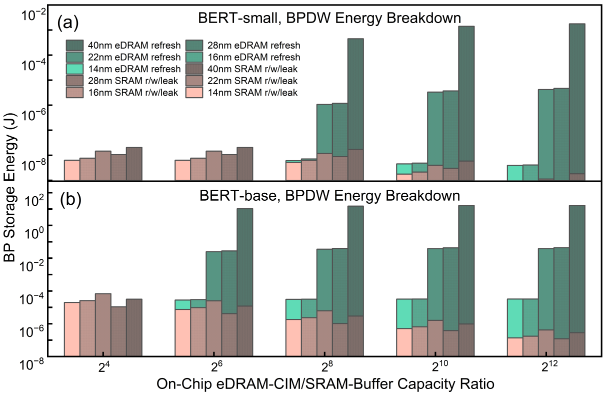

Fig. 6. Energy costs of backpropagation dynamic weight (BPDW) storage across various eDRAM-to-SRAM capacity ratios on eDRAM chiplets, in multiple technology nodes, analyzed for continual learning in (a) BERT-small and (b) BERT-base models. SRAM leakage and eDRAM refresh play key roles in BPDW memory optimization for continual learning workloads of different scales.

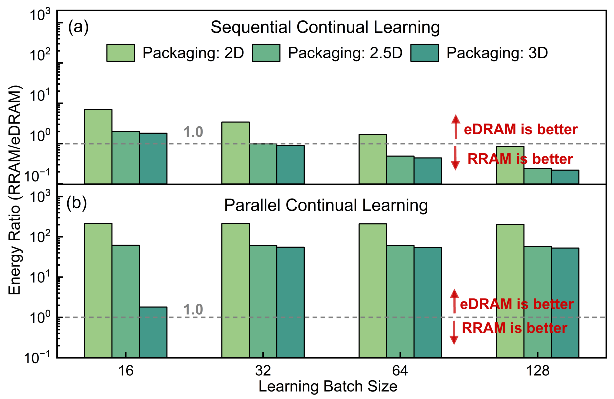

Fig. 7. Relative energy costs for weight gradient storage between 40 nm RRAM (due to write-back) and 14 nm eDRAM (due to refresh) in (a) sequential-mode continual learning and (b) parallel-mode continual learning, analyzed across various integration schemes.

overhead between NoC and NoP can be obtained at a ratio of 4 for RRAM/eDRAM capacity allocation across chiplets.

B. Edge LLM Continual Learning

For continual learning with 3D-CIMlet, Adapter-BERT is evaluated on two configurations: (1) BERT-base (12 layers, 12 attention heads) on the GLUE dataset (average 128 tokens), and (2) BERT-small (4 layers, 4 attention heads, average 32 tokens). During continual learning tasks, layer output activations during FP act as dynamic weights during BP, referred to as backpropagation dynamic weights (BPDW), alongside learnable adapter weights. The required data retention time for BPDW, in the context of eDRAM refreshes, varies by layer position within or across encoder/decoder modules. Similarly, the retention time for weight gradients before the next FP iteration depends on the learning mode (sequential or parallel) and batch size. Sequential learning mode processes input samples without overlapping layer computations within a single iteration. This approach trades off longer retention times for weight gradients with potentially lower system power. In contrast, parallel learning mode pipelines input samples across multiple layers simultaneously. 3D-CIMlet orchestrates on-chip BP data storage through DSE on two fronts: (1) exploring and optimizing the on-die eDRAM-to-SRAM buffer capacity ratio in eDRAM chiplets of multiple technology nodes for BPDW storage, and (2) optimizing the on-chip storage of weight gradients between RRAM and eDRAM chiplets.

For continual learning, each eDRAM chiplet is configured with storage and CIM partitions to store and process BPDW. With on-die SRAM buffers, we first evaluate the impact of hybrid SRAM/eDRAM storage. A mapping strategy is taken such that early-generated BPDW are stored in an SRAM buffer until full, with subsequent weights stored directly in eDRAM on the same die. SRAM-stored weights are transferred to the eDRAM-CIM for computation as needed, while eDRAM-stored weights are refreshed until the next computation. For the BERT-small model with shorter iteration durations, reduced energy costs are observed at 14 nm and 16 nm nodes with decreasing SRAM buffer capacities (Fig.6(a)). In contrast, for the BERT-base model, increasing the eDRAM/SRAM capacity ratio leads to higher energy costs for BPDW storage across planar and FinFET silicon nodes (from 40 nm to 14 nm), due to low SRAM leakage offset by high eDRAM refresh energy over time (Fig.6(b)). These results suggest that advanced-node eDRAM, with low refresh power, is better suited for BPDW storage in small-scale continual learning scenarios. For larger-scale continual learning, increasing SRAM buffer sizes reduces energy costs despite SRAM leakage, while the density benefit of capacitor-less eDRAMs mitigates the area overhead of larger SRAM buffers within a given BPDW energy budget.

Second, the weight gradient storage options under sequential and parallel learning modes are analyzed to identify runtime optimization opportunities for continual learning workloads (Fig. 7). In sequential mode (less thermal stress), the required retention time for weight gradients decreases with smaller learning batch sizes, which increases the likelihood of using eDRAM. For larger batch sizes, high refresh energy costs of eDRAM tend to push the storage needs towards RRAM, especially with low energy overhead of 2.5D/3D packaging. In parallel mode (higher performance), the inherently relaxed retention requirement results in eDRAM being the preferred storage medium across all batch sizes and integration schemes.

In broader use cases, continual learning and inference tasks are periodically interleaved. 3D-CIMlet's DSE provides insights into the impact of different chiplet stacking architectures on energy efficiency and energy-delay product (EDP) across various model architectures and application scenarios (Fig. 8).

For inference (Fig. 8(a)), 2.5D/3D architectures significantly improve energy efficiency and latency of traditional 2D packages, with TOPS/W increased by 9.9× for BERT-base and 4.5× for DeiT-base, compared to 1.1× for GPT-2. The primary driver of these energy efficiency and latency gains with 3D stacking is reduced communication overhead from enhanced D2D connectivity. The D2D communication energy in BERT-base and DeiT-base constitutes 90.8% and 78.3% of total

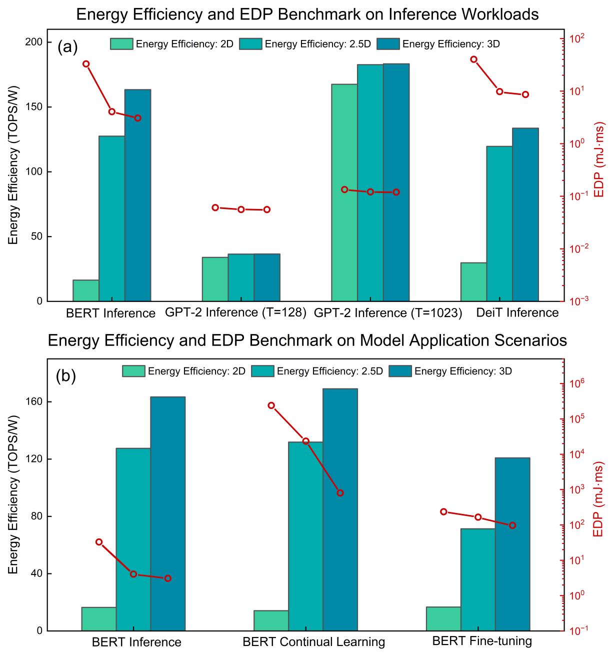

Fig. 8. The energy efficiency and energy-delay product (EDP) of 3D-CIMlet designs for (a) inference with 40 nm RRAM/eDRAM chiplets, and (b) Adapter-BERT inference, continual learning, and fine-tuning with integrated 14 nm eDRAM and 40 nm RRAM/eDRAM chiplets.

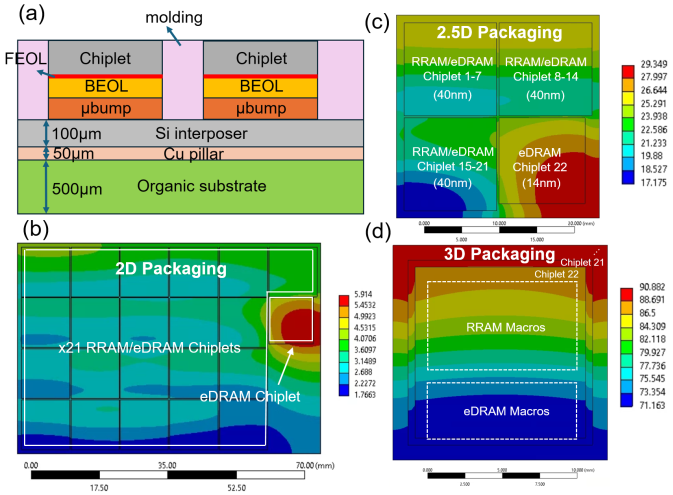

Fig. 9. Chiplet-to-package thermal analysis for edge LLM continual learning across various packaging configurations. (a) Schematic of the 2.5D/3D packages for thermal modeling and analysis. Peak temperature increase contours are generated across three different packaging designs: (b) traditional 2D packaging layout with a large form factor, (c) 2.5D chiplet integration with limited die stacking, and (d) full 3D stacking.

energy in the 2D architecture, significantly benefiting from chiplet integration. In GPT-2 architecture, the autoregressive design leads to sequential token generation, resulting in negligible D2D communication costs compared to computation. NoP accounts for only 8.7% of total energy during 1024-token inference with modest TOPS/W improvements.

For continual learning (Fig. 8(b)), 9.3× and 12.0× improvements in energy efficiency, along with a 90.2% and 92.5% reduction in EDP, respectively, are achieved by the 2.5D and 3D designs through the 3D-CIMlet framework compared to 2D baselines. The inference exhibits fewer gains with only 40 nm RRAM/eDRAM chiplets. For fine-tuning, the rise in on-chip computations outweighs the increase in NoP communication costs, resulting in diminished improvements overall.

C. Thermal Analysis

Chiplet-to-package thermal analysis plays a pivotal role in edge LLM continual learning by highlighting key thermal challenges and trade-offs in managing temperatures for advanced 2.5D/3D designs. Figure 9 illustrates the packaging-level thermal modeling setup (a), and the peak temperature increase contours with various packaging schemes (b)–(d). A non-uniform temperature distribution is observed due to varying power characteristics across chiplets. Thermal modeling reveals distinct temperature profiles for 2.5D/3D packaging schemes. In the 2.5D setup, the maximum temperature increase ($T_{\max}$) reaches 29.3 K, with a temperature difference (ΔT) of 12.2 K, indicating moderate thermal stress. However, persistent hotspots are found in the 14 nm eDRAM chiplet due to high power density and limited cooling pathways. Adopting a worst-case scenario, full 3D stacking shows highest thermal stress driven by increased thermal resistance and inefficient heat dissipation, calling for additional cooling to mitigate thermal buildup. It is worth noting that the on-die memory area allocation between RRAM and eDRAM macros in the 3D chiplets and the package design jointly lead to non-uniform peak temperature distribution with cooler eDRAMs, which helps mitigate the impact on eDRAM retention. Comparatively, 2.5D designs offer better temperature uniformity compared to 3D designs but still face challenges with localized thermal hotspots, trading off system-level compute density.

VI. Conclusion

We present 3D-CIMlet, a co-design framework that harnesses the unique capabilities of heterogeneous CIM chiplets within 2.5D/3D multi-die architectures, enabling efficient inference and continual learning for edge LLMs. 2.5D and 3D designs with heterogeneous RRAM/eDRAM chiplets lead to significant energy efficiency and EDP benefits compared to 2D architecture for continual learning. At the core of 3D-CIMlet's co-design capabilities are diverse embedded computational memories, in-memory compute-storage allocation strategies, NoP/NoC interplays with intra-chiplet designs, flexible model-to-architecture mapping space, and chiplet-to-package thermal analysis. These features, extending beyond the case study presented, will enable memory-reliability-aware and thermal-aware system designs for scalable and energy-efficient deployment of future LLM workloads at the edge and beyond.

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, in part by the UPWARDS for the Future Network program (with funding from NSF Award No. 2329784, Micron Technology, and TEL), and also in part by CoCoSys, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) Program sponsored by DARPA.

References

[1] J. Lee and H.-J. Yoo, "An overview of energy-efficient hardware accelerators for on-device deep-neural-network training," *IEEE Open Journal of the Solid-State Circuits Society*, vol. 1, pp. 115–128, 2021.

[2] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. D. Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, May 2019, pp. 2790–2799.

[3] Z. Chen, X. Chen, and J. Gu, "15.3 a 65nm 3t dynamic analog ram-based computing-in-memory macro and cnn accelerator with retention enhancement, adaptive analog sparsity and 44tops/w system energy efficiency," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64, 2021, pp. 240–242.

[4] Y. He, S. Fan, X. Li, L. Lei, W. Jia, C. Tang, Y. Li, Z. Huang, Z. Du, J. Yue, X. Li, H. Yang, H. Jia, and Y. Liu, "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 *2024 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 67, 2024, pp. 578–580.

[5] F. Tu, Z. Wu, Y. Wang, L. Liang, L. Liu, Y. Ding, L. Liu, S. Wei, Y. Xie, and S. Yin, "Trancim: Full-digital bitline-transpose cim-based sparse transformer accelerator with pipeline/parallel reconfigurable modes," *IEEE Journal of Solid-State Circuits*, vol. 58, no. 6, pp. 1798–1809, Jun. 2023.

[6] S. D. Spetalnick, A. S. Lele, B. Crafton, M. Chang, S. Ryu, J.-H. Yoon, Z. Hao, A. Ansari, W.-S. Khwa, Y.-D. Chih, M.-F. Chang, and A. Raychowdhury, "30.1 a 40nm vliw edge accelerator with 5mb of 0.256pj/b rram and a localization solver for bristle robot surveillance," in *2024 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 67, 2024, pp. 482–484.

[7] C.-X. Xue, Y.-C. Chiu, T.-W. Liu, T.-Y. Huang, J.-S. Liu, T.-W. Chang, H.-Y. Kao, J.-H. Wang, S.-Y. Wei, C.-Y. Lee *et al.*, "A cmos-integrated compute-in-memory macro based on resistive random-access memory for ai edge devices," *Nature Electronics*, vol. 4, no. 1, pp. 81–90, 2021.

[8] Y. N. Wu, J. S. Emer, and V. Sze, "Accelergy: An Architecture-Level Energy Estimation Methodology for Accelerator Designs," in *IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, 2019.

[9] S. Tuli and N. K. Jha, "Acceltran: A sparsity-aware accelerator for dynamic inference with transformers," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 42, no. 11, pp. 4038–4051, 2023.

[10] X. Peng, W. Chakraborty, A. Kaul, W. Shim, M. S. Bakir, S. Datta, and S. Yu, "Benchmarking monolithic 3d integration for compute-in-memory accelerators: overcoming adc bottlenecks and maintaining scalability to 7nm or beyond," in *2020 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2020, pp. 30–4.

[11] G. Krishnan, S. K. Mandal, M. Pannala, C. Chakrabarti, J.-S. Seo, U. Y. Ogras, and Y. Cao, "Siam: Chiplet-based scalable in-memory acceleration with mesh for deep neural networks," *ACM Transactions on Embedded Computing Systems*, vol. 20, no. 5s, pp. 1–24, Oct. 2021.

[12] K. Prabhu, A. Gural, Z. F. Khan, R. M. Radway, M. Giordano, K. Koul, R. Doshi, J. W. Kustin, T. Liu, G. B. Lopes, V. Turbiner, W.-S. Khwa, Y.-D. Chih, M.-F. Chang, G. Lallement, B. Murmann, S. Mitra, and P. Raina, "Chimera: A 0.92-tops, 2.2-tops/w edge ai accelerator with 2-mbyte on-chip foundry resistive ram for efficient training and inference," *IEEE Journal of Solid-State Circuits*, vol. 57, no. 4, pp. 1013–1026, 2022.

[13] F. Tu, Y. Wang, Z. Wu, W. Wu, L. Liu, Y. Hu, S. Wei, and S. Yin, "16.4 tensorcim: A 28nm 3.7 nj/gather and 8.3 tflops/w fp32 digital-cim tensor processor for mcm-cim-based beyond-nn acceleration," in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2023, pp. 254–256.

[14] K. Prabhu, R. M. Radway, Y. Jeffrey, K. Bartolone, M. Giordano, F. Peddinghaus, Y. Urman, W.-S. Khwa, Y.-D. Chih, M.-F. Chang *et al.*, "Minotaur: An edge transformer inference and training accelerator with 12 mbytes on-chip resistive ram and fine-grained spatiotemporal power gating," in *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. IEEE, 2024, pp. 1–2.

[15] A. Graening, S. Pal, and P. Gupta, "Chiplets: How small is too small?" in *2023 60th ACM/IEEE Design Automation Conference (DAC)*, 2023, pp. 1–6.

[16] Y. Feng and K. Ma, "Chiplet actuary: a quantitative cost model and multi-chiplet architecture exploration," in *Proceedings of the 59th ACM/IEEE Design Automation Conference*, ser. DAC '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 121–126. [Online]. Available: https://doi.org/10.1145/3489517.3530428

[17] M. Garcia Bardon, P. Wuytens, L.- Ragnarsson, G. Mirabelli, D. Jang, G. Willems, A. Mallik, A. Spessot, J. Ryckaert, and B. Parvais, "Dtco including sustainability: Power-performance-area-cost-environmental score (ppace) analysis for logic technologies," in *2020 IEEE International Electron Devices Meeting (IEDM)*, 2020, pp. 41.4.1–41.4.4.

[18] C.-C. Chou, Z.-J. Lin, P.-L. Tseng, C.-F. Li, C.-Y. Chang, W.-C. Chen, Y.-D. Chih, and T.-Y. J. Chang, "An n40 256k× 44 embedded rram macro with sl-precharge sa and low-voltage current limiter to improve read and write performance," in *2018 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE, 2018, pp. 478–480.

[19] R. Giterman, A. Shalom, A. Burg, A. Fish, and A. Teman, "A 1-mbit fully logic-compatible 3t gain-cell embedded dram in 16-nm finfet," *IEEE Solid-State Circuits Letters*, vol. 3, pp. 110–113, 2020.

[20] C.-H. Lin, B. Greene, S. Narasimha, and e. a. Cai, J., "High performance 14nm soi finfet cmos technology with 0.0174μm2 embedded dram and 15 levels of cu metallization," in *2014 IEEE International Electron Devices Meeting*, 2014, pp. 3.8.1–3.8.3.

[21] N. Jiang, J. Balfour, D. U. Becker, B. Towles, W. J. Dally, G. Michelogiannakis, and J. Kim, "A detailed and flexible cycle-accurate network-on-chip simulator," in *2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. Austin, TX, USA: IEEE, Apr. 2013, pp. 86–96.

[22] X. Bai, J. Zhao, S. Zuo, and Y. Zhou, "A 2.5 Gbps, 10-lane, low-power, LVDS transceiver in 28 nm CMOS technology," *Electronics*, vol. 8, no. 3, p. 350, 2019.

[23] Y. Zhang, X. Zhang, and M. S. Bakir, "Benchmarking digital die-to-die channels in 2.5-d and 3-d heterogeneous integration platforms," *IEEE Transactions on Electron Devices*, vol. 65, no. 12, pp. 5460–5467, 2018.

[24] M.-S. Lin, T.-C. Huang, C.-C. Tsai, K.-H. Tam, K. C.-H. Hsieh, C.-F. Chen, W.-H. Huang, C.-W. Hu, Y.-C. Chen, S. K. Goel, C.-M. Fu, S. Rusu, C.-C. Li, S.-Y. Yang, M. Wong, S.-C. Yang, and F. Lee, "A 7-nm 4-ghz arm^1-core-based cowos^1 chiplet design for high-performance computing," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 4, pp. 956–966, 2020.

[25] X. Sun, X. Peng, S. Q. Zhang, J. Gomez, W.-S. Khwa, S. S. Sarwar, Z. Li, W. Cao, Z. Wang, C. Liu *et al.*, "Estimating power, performance, and area for on-sensor deployment of ar/vr workloads using an analytical framework," *ACM Transactions on Design Automation of Electronic Systems*, vol. 29, no. 6, pp. 1–27, 2024.

[26] S. W. Liang, G. C. Y. Wu, K. C. Yee, C. T. Wang, J. J. Cui, and D. C. H. Yu, "High performance and energy efficient computing with advanced soic™ scaling," in *2022 IEEE 72nd Electronic Components and Technology Conference (ECTC)*, 2022, pp. 1090–1094.

[27] C. Hu, M. Chen, W. Chiou, and C. Doug, "3d multi-chip integration with system on integrated chips (soic™)," in *2019 Symposium on VLSI Technology*. IEEE, 2019, pp. T20–T21.