2025 IEEE 33rd Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM) | 979-8-3315-0281-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/FCCM62733.2025.00033

C2OPU: Hybrid Compute-in-Memory and Coarse-Grained Reconfigurable Architecture for Overlay Processing of Transformers

Siyuan Miao[1], Lingkang Zhu[2], Chen Wu[3,4], Shaoqiang Lu[3], Jinming Lyu[3,4], Lei He[1]
[1]University of California, Los Angeles, United States [2]University of Nottingham Ningbo China
[3]Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[4]Engineering Research Center of Chiplet Design and Manufacturing of Zhejiang Province, Ningbo, Zhejiang, China
Email: cwu@idt.eitech.edu.cn lei.hexun@gmail.com

I. C2OPU Architecture

Transformers become increasingly memory-bound due to the computational patterns of feed-forward networks and the attention mechanism. Analog Computing-In-Memory (ACIM) has emerged as a promising approach for its high performance and area efficiency. However, challenges remain in accelerating Transformers on ACIM, including mismatched computational patterns and lower computing accuracy.

Fig. 1 shows the overview of C2OPU, a hybrid ACIM-CGRA (Coarse-Grained Reconfigurable Architecture) processor for the overlay processing of Transformers. The ACIM acts as a specialized accelerator for high-performance weight-stationary matrix multiplications (MMs), while the CGRA serves as a general-purpose parallel computing unit handling diverse workloads and delivering high-precision computation.

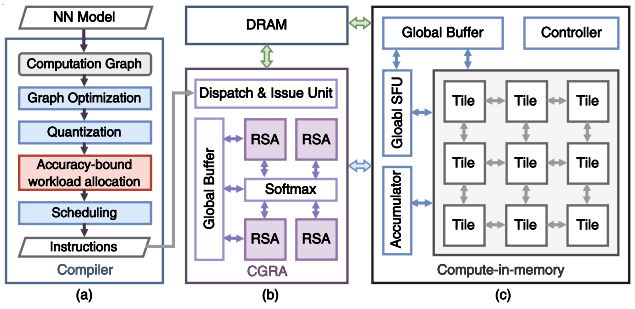

Fig. 1. The C2OPU Architecture.

The CGRA targets dynamic MMs within attention heads and certain accuracy-sensitive weight-stationary MMs. The reconfigurable systolic arrays (RSAs) are adopted to improve processing element (PE) utilization when handling variable-length inputs. The PEs within the ACIM tiles perform the majority of computations, delivering the highest computing power. 2T2R RRAM [1] cells are employed in the PEs to mitigate the IR drop issue in memristors. Each ACIM PE is restricted to compute at most one neural network (NN) layer, and a large NN layer can be partitioned across multiple PEs.

The accuracy degradation of ACIM can significantly undermine the feasibility of NN applications. We develop an efficient algorithm to simulate ACIM computations that considers nonidealities. The C2OPU compiler encompasses an effective strategy for accuracy-bound workload allocation. The C2OPU is designed for high scalability, where configurable RSA sizes that can be optimized for the target model.

[1]This work was supported by Ningbo Key Technologies of "Science and Technology Innovation in Yongjiang 2035" (2024Z283).

II. Architecture Evaluations

We evaluate C2OPU using a variety of Transformer-based models on GLUE benchmark, and compare C2OPU with an Intel Xeon Gold 6348 CPU, an NVIDIA A100 GPU, and SOTA CIM processors including Science23 [2], Nature23 [3], and VLSI24 [4], summarized in Table I.

We present a case study of optimizing C2OPU for BERT-Small. With optimizations of workload allocation, CIM parallelism, and RSA size. Table II summarizes the optimized C2OPU configuration along with its key metrics.

TABLE I
Speedup of C2OPU over SOTA processors.

Model		BERT		GPT-2		T5	DistilBERT	Average
Variant		Small	Large	Small	XL	3B		
C2OPU speedup over	CPU	133.41×	143.50×	158.75×	143.14×	131.21×	159.22×	145.41×
	GPU	4.34×	4.67×	5.16×	4.65×	4.27×	5.18×	4.73×
	Science23	4.25×	4.57×	5.06×	–	–	5.08×	4.70×
	Nature23	3.48×	3.74×	4.14×	–	–	4.15×	3.85×
	VLSI24	1.24×	1.33×	1.47×	1.33×	–	1.48×	1.37×

TABLE II
Optimized C2OPU configurations and key metrics

CIM (28nm)		**CGRA** (14nm)	
Tiles per CIM	12	RSA row size	16
# PEs per tile	4	RSA column size	64
Crossbar size	512 × 512	Number of RSA	8
# ADCs per PE	128	**Inter-die Connection**	
# DACs per PE	256	PCIe5 x8	
CIM area (mm^2)	140.59	CGRA area (mm^2)	52.55
C2OPU TP* ($sentences/s$)	11.56	C2OPU power (W)	2.66

* TP for throughput

References

[1] Q. Liu *et al.*, "33.2 A Fully Integrated Analog ReRAM Based 78.4TOPS/W Compute-In-Memory Chip with Fully Parallel MAC Computing," in *2020 IEEE International Solid-State Circuits Conference - (ISSCC)*, 2020, pp. 500–502.
[2] W. Zhang *et al.*, "Edge learning using a fully integrated neuro-inspired memristor chip," *Science*, vol. 381, pp. 1205–1211, 2023.
[3] S. Ambrogio *et al.*, "An analog-AI chip for energy-efficient speech recognition and transcription," *Nature*, vol. 620, pp. 768–775, Aug 2023.
[4] S. Liu *et al.*, "HARDSEA: Hybrid Analog-ReRAM Clustering and Digital-SRAM In-Memory Computing Accelerator for Dynamic Sparse Self-Attention in Transformer," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 32, no. 2, pp. 269–282, 2024.

2576-2621/25/$31.00 ©2025 IEEE
DOI 10.1109/FCCM62733.2025.00033