15.3 COMB-MCM: Computing-on-Memory-Boundary NN Processor with Bipolar Bitwise Sparsity Optimization for Scalable Multi-Chiplet-Module Edge Machine Learning

Haozhe Zhu*[1], Bo Jiao*[1], Jinshan Zhang*[1], Xinru Jia[1], Yunzhengmao Wang[1], Tianchan Guan[2], Shengcheng Wang[2], Dimin Niu[2], Hongzhong Zheng[2], Chixiao Chen[1], Mingyu Wang[1], Lihua Zhang[1], Xiaoyang Zeng[1], Qi Liu[1], Yuan Xie[2], Ming Liu[1]

[1]Fudan University, Shanghai, China
[2]Alibaba DAMO Academy, Shanghai, China

*Equally Credited Authors (ECAs)

Recently, computing-in-memory (CIM) macros, originally designed to reduce the intensive memory accesses of AI tasks, have been employed in low-power machine learning SoCs due to their ultra-high computing efficiency [1, 2, 3]. These CIM macros still access weight data through on/off-chip memories, similar to processing elements in near-memory-computing architectures. The implementation poses challenges when counting the overall SoC energy efficiency (Fig. 15.3.1). First, the memory wall issue is unsolved. The weight updates affect overall system performance when large networks are deployed and massive off-chip weight data transfer occurs. Even for tiny machine learning tasks, power consumption and latency of constant weight updates cannot be neglected, because MAC computing efficiency is optimized and closely matches the efficiency of on-chip memory access (2pJ/b vs. 1pJ/b). Second, the viability of structured and coarse-grained sparsity optimization is highly algorithm dependent and requires explicit zero-detection blocks. Power optimization schemes for fine-grained or even arbitrary-sparsity patterns are lacking. Third, edge machine learning chips are cost sensitive. The conventional monolithic SoC design strategy, fabricating one specific SoC for each application, is not affordable in terms of NRE costs.

To overcome these challenges, this work proposes: 1) computing-on-memory-boundary (COMB), a compromise between in-memory and near-memory designs featuring both high macro computing energy efficiency and low system power overhead by static weight placement; 2) a bipolar power-optimized COMB macro, enabling simultaneous power reduction when zeros or multiple ones occur and compatibility with arbitrary sparsity patterns without explicit zero-detection modules; 3) a scalable system architecture supporting different edge AI tasks of varying complexity by assembling multi-chiplet modules (MCM) of varying chip numbers. For validation, two COMB NN processors are fabricated in 65nm and 28nm, respectively. Four 65nm chiplets are assembled into an MCM system by a 2.5D package, while the 28nm design features better performance.

Figure 15.3.2 compares CIM and COMB architectures and demonstrates the overall COMB processor architecture. CIM processors have separate weight buffers and computing macros, between which weight data keep moving back and forth. In contrast, the COMB architecture eliminates the back-and-forth transfers by distributing MACs along the boundary of weight memory banks. Through this approach, the intensive weight data movement is replaced by one-shot memory access. Specifically, the proposed architecture includes a COMB Macro Array (CMA), a Vector Quantization Pipeline (VQP), a 256KB SRAM-based Activation Buffer (ABUF), a Chip-to-Chip Link (C2CLink), and a system controller. The CMA consists of 326KB COMB macros, providing a total of 192KB of weight storage and 4608 MAC operations per cycle. The weights in the COMB are uploaded from the host, and remain static during inference. The macros in the CMA share 144×4b activations fetched from the ABUF, and deliver 32×16b MAC results to the VQP. The VQP performs accumulation, pooling, normalization, and ReLU in sequence, and finally quantizes to 4b. To deploy large networks on the MCM, the C2CLink transfers activations among connected chips. The system controller manages the workload allocation and communicates with a host.

Figure 15.3.3 shows the COMB macro implementation details, including 432 COMB rows and two bitwise adder trees. Each COMB row has multiple 8T-SRAM cells and multipliers with embedded dynamic latches on the memory boundary (MELOM). In every MAC operation, one cell of each COMB row is selected to be multiplied by input activations (IA), generating bitwise partial products. All bitwise products are accumulated by a digital adder tree, thus lossless as digital CIM [5]. Computing two adjacent layers inside one COMB processor is exemplified in Fig.15.3.3. Before inference, weights of the same layer are initially distributed to different COMB macros with the same address index, whereas weights of different layers are mapped onto different address sets. During inference of the second layer, one group of activation SRAM banks (#0-3) serves as the input source of COMB macros, while the other group (#4-5) collects the computing results. To enable the inter-chip layer-wise pipeline, the C2CLink, including Rx and Tx, connects the activation SRAM banks simultaneously, receiving the next image and transmitting the computing result to the next chip, respectively.

Figure 15.3.4 shows the circuit implementation of MELOM: a custom-designed gate enabling bipolar sparsity optimization. The 65nm MELOM v1 gate incorporates an IA-controlled NMOS connected read bitline (RBL), a pre-charge PMOS, and a TSPC second stage. It fuses readout circuits, multiplier (AND logic), and a dynamic latch into one gate. The RBL discharges only if both weight and activation are logic-1. In other words, the dynamic switching power on the RBL is saved if either activation or weight is zero. However, the NMOS-only MELOM causes the pre-charge voltage on the RBL to be (V_{DD} – V_{thN}), introducing errors during low voltage operation. The 28nm MELOM v2 replaces the NMOS with a CMOS transmission gate. Another modification of MELOM v2 is RBL sharing, wherein four MELOMs share one RBL performing 4b MAC in one shot. The sharing scheme not only increases the throughput from bit-serial to bit-parallel, but also saves (K-1)× RBL switching power when K MELOMs are on. The parallel MELOM topology realizes a power-reduction when multiple logic-1 IA values are present.

Figure 15.3.5 demonstrates scalable MCM systems using the proposed COMB NN processors (COMB-MCM). Layer-wise pipeline mapping schemes are adopted to deploy NNs of different sizes on COMB-MCM, where the chiplet number is a dimension to scale. For large NNs, the overall workload is partitioned into multiple sub-workloads, each containing one or a few layers. The number of chiplets equals the number of sub-workloads. Activations flow among different chiplets sequentially, while weights are all statically placed. Thus, the memory wall issue is alleviated without external DRAM access. Synchronization is controlled by the host, and die-to-die communication circuits are implemented by single-ended all-digital links. Three common edge machine learning tasks, including keyword spotting, CIFAR-10 image classification, and tiny-YOLO-based object detection, are deployed on COMB-MCM using one, two, and four chiplets, respectively.

Figure 15.3.7 shows the die and 2.5D package photos of 65nm and 28nm COMB processor prototypes. Four 65nm prototype chiplets are assembled clockwise in an integrated fanout (InFO) package, minimizing the inter-chip distance between a Tx-Rx pair to 0.8mm. The C2CLink rate is maximized by the ultra-short range. The measurement results are demonstrated in Fig. 15.3.6. The 65nm and 28nm chip can work under a maximum rate of 120MHz and 105MHz with 0.71/1.31V and 0.63/0.95V COMB/digital voltage, respectively. Peak macro/overall system energy efficiency of 14.1/8.6TOPS/W and 45.7/32.9TOPS/W is realized on the 65nm and 28nm prototypes, respectively. All three machine-learning testcases are mapped and measured; utilization details are also shown. In contrast with SOTA in/near-memory-computing designs, the COMB NN processor maintains high computing efficiency and low system power overhead. The overhead ratio between the computing macro and system energy efficiency is 1.39, 2.25-to-3.5× lower than other SOTA CIM SoCs. The low system overhead facilitates COMB-MCM's memory-wall-friendly scalability when mapping various edge AI tasks.

Acknowledgement:
This work was supported by the National Key R&D Program Grant # 2018YFA0701500, NSFC Grant # 61974033, 61732020, 61821091, and the Strategic Priority Research Program of CAS under Grant # XDB44000000, Shanghai Qi Zhi Institute, Biren Technology, and MOE innovation platform. Corresponding author: Chixiao Chen.

References:
[1] J. Yue et al., "A 2.75-to-75.9TOPS/W CIM NN Processor Supporting Set-Associate Block-Wise Zero Skipping and Ping-Pong CIM," *ISSCC*, pp. 238-240, 2021.
[2] R. Guo et al., "A 6.54-to-26.03 TOPS/W CIM RNN Processor using Input Similarity Optimization and Attention-based Context-breaking with Output Speculation," *IEEE Symp. VLSI Circuits*, 2021.
[3] H. Jia et al., "A Programmable Neural Network Inference Accelerator Based on Scalable In-Memory Computing," *ISSCC*, pp. 236-238, 2021.
[4] B. Zimmer et al., "A 0.32–128 TOPS, Scalable MCMBased Deep Neural Network Inference Accelerator with Ground-Referenced Signaling in 16 nm," *IEEE JSSC*, vol. 55, no. 4, pp. 920-932, 2020.
[5] Y.D. Chih et al.," An 89TOPS/W and 16.3TOPS/mm2 AllDigital SRAM-Based Full-Precision CIM Macro in 22nm for Machine-Learning Edge Applications," *ISSCC*, pp. 252-254, 2021.
[6] Z. Chen et al.," A 65nm 3T Dynamic Analog RAM Based CIM Macro and CNN Accelerator with Retention Enhancement, Adaptive Analog Sparsity and 44TOPS/W System Energy Efficiency," *ISSCC*, pp. 240-241, 2021.

2022 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-6654-2800-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISSCC42614.2022.9731657

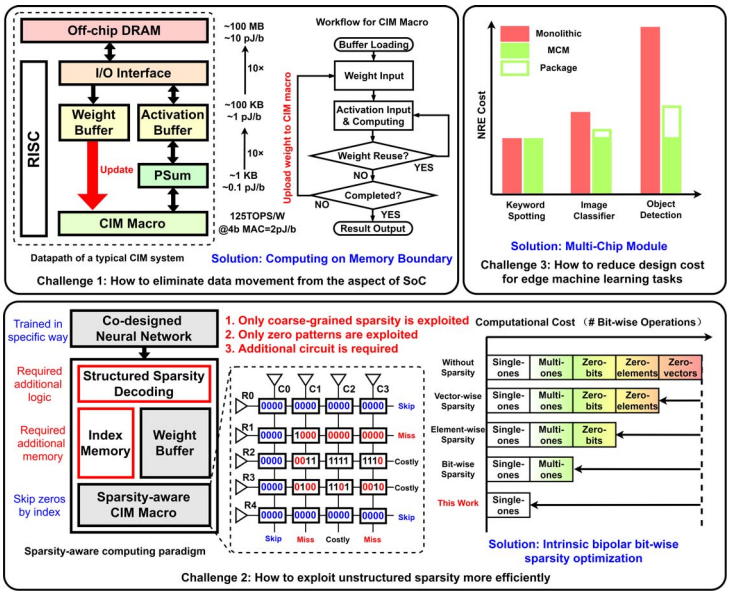


Figure 15.3.1: Challenges of existing computing-in-memory based SoC designs.

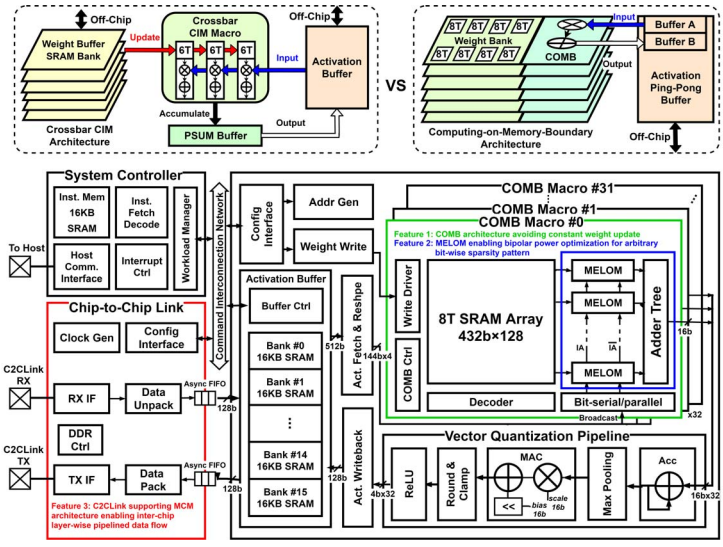


Figure 15.3.2: Proposed COMB based NN processor architecture supporting MCM scalable systems.

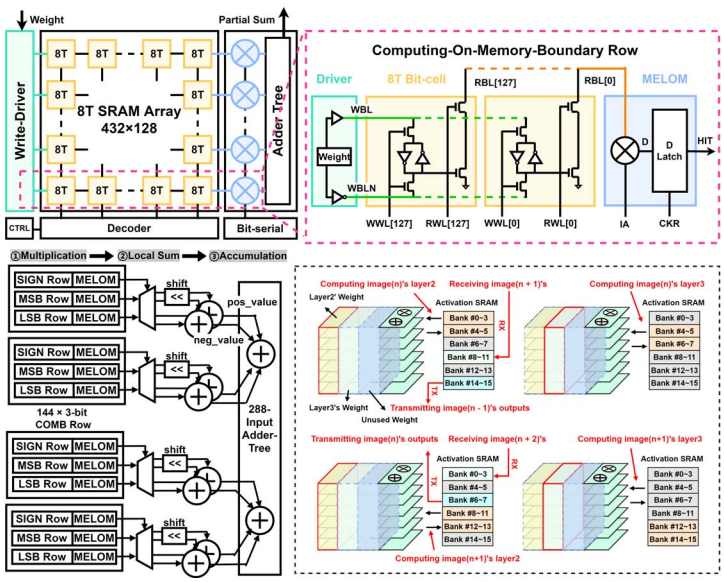


Figure 15.3.3: COMB macro organization and dataflow mapping.

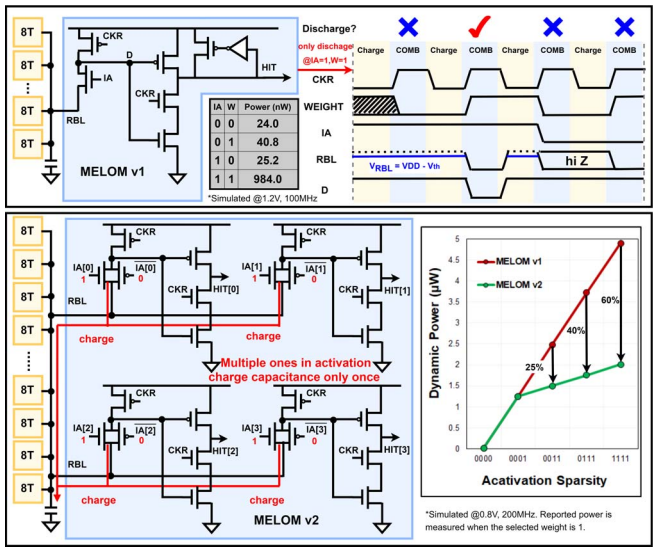


Figure 15.3.4: The circuit implementation of multiplication with embedded dynamic latch on the memory boundary (MELOM) and the bipolar sparsity optimization schemes.

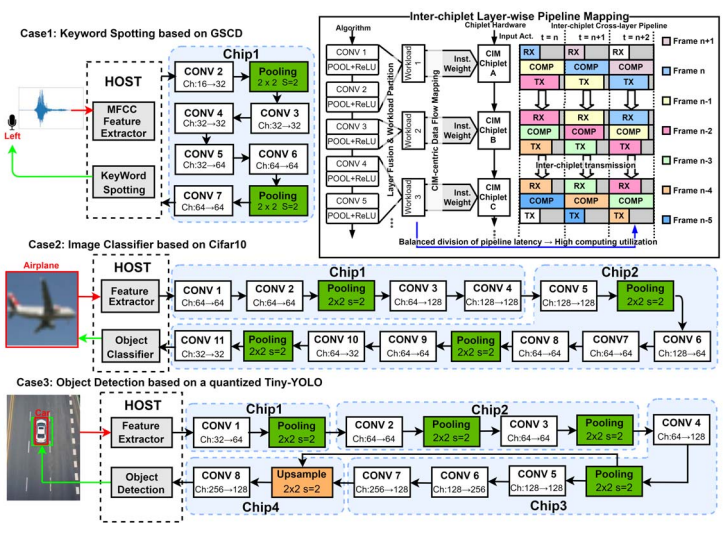
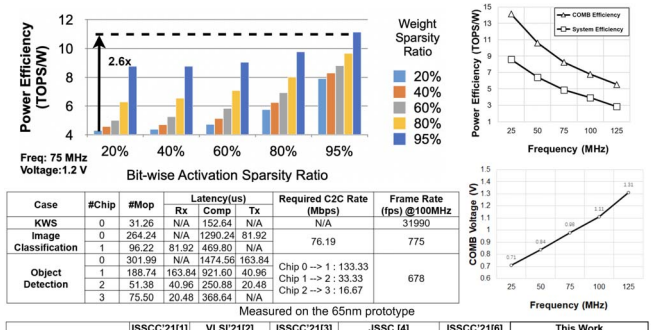


Figure 15.3.5: Layer-wise pipeline-based scalable MCM systems targeting varying edge machine-learning tasks.



Case	#Chip	#Mop	Latency(us)			Required C2C Rate (Mbps)	Frame Rate (fps) @100MHz
			Rx	Comp	Tx		
KWS	0	31.26	N/A	152.64	N/A	N/A	31990
Image Classification	0	264.24	N/A	1290.24	81.92	76.19	775
	1	96.22	81.92	469.80	N/A		
Object Detection	0	301.99	N/A	1474.56	163.84	Chip 0 --> 1 : 133.33 Chip 1 --> 2 : 33.33 Chip 2 --> 3 : 16.67	678
	1	188.74	163.84	921.60	40.96		
	2	51.38	40.96	250.88	20.48		
	3	75.50	20.48	368.64	N/A		

Measured on the 65nm prototype

	ISSCC'21[1]	VLSI'21[2]	ISSCC'21[3]	JSSC [4]	ISSCC'21[6]	This Work	
Technology	65nm	65nm	16nm	16nm	65nm	65nm	28nm
Voltage (V)	0.62-1	0.75-1.1	0.8	0.52-1.1	0.85-1.1	0.71-1.31	0.63-0.95
Bit Precision	A: 2/4/6/8 bit W: 1-8 bit	A: 2/4 bit W: 2/4/8 bit	1 - 8 bit	8 bit	A: 4/8 bit W: 2/4/8 bit	A: 3 bit W: 3 bit	A: 4 bit W: 3 bit
Memory Cell Structure	6T SRAM	6T SRAM	8T SRAM	-	3T eDRAM	8T SRAM	8T SRAM
Array Size	64 Kbit	64 Kbit	4608 Kbit	-	8 Kbit	1152 Kbit	1728 Kbit
Total Digital SRAM	294 KB	32 KB	-	0.625 MB	172 KB	170 KB	256KB
Frequency (MHz)	25-100	160-302	200	161-2001	105	25-125	30-105
Peak Performance (TOPS)	0.10-3.16 @2A,1W	0.307-0.579 @4b 1.221-2.304 @2b	11.8 @4b	0.32-4.01 (1 chip) 3.93-15.7 (4 chip) 32.5-127.8 (36 chip)	0.43 @4b	0.25 (1 chip) 1.00 (4 chip)	0.97 (1 chip)
Macro Energy Efficiency (TOPS/W)	7.32-370	-	121 @1b	-	102.2 @4b	14.1*	45.7*
System Energy Efficiency (TOPS/W)	2.75-75.9	6.54 @4b 26.03 @2b	-	0.64-6.24 (4 chip)	32.5 @4b	8.6*	32.9*
Macro/System Energy Efficiency Ratio	2.66-4.87	-	-	-	3.14	1.64	1.39
Scalability	NO	NO	Intra-chip	MCM	NO	MCM	MCM

*The results are measured on a single chip constellation when running case 3 (TinyYOLO). Sparsity rate is 68%.

Figure 15.3.6: Measurement results and comparison table.

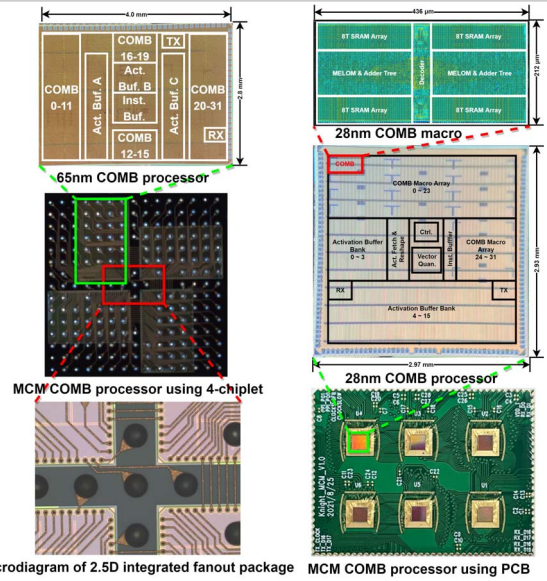


Figure 15.3.7: Die micrographs and MCM package photo of both 65nm and 28nm COMB processors.

978-1-6654-2800-2/22/$31.00 ©2022 IEEE