15.5 A 28nm 29.2TFLOPS/W BF16 and 36.5TOPS/W INT8 Reconfigurable Digital CIM Processor with Unified FP/INT Pipeline and Bitwise In-Memory Booth Multiplication for Cloud Deep Learning Acceleration

Fengbin Tu[1,2], Yiqi Wang[1], Zihan Wu[1], Ling Liang[2], Yufei Ding[2], Bongjin Kim[2], Leibo Liu[1], Shaojun Wei[1], Yuan Xie[2], Shouyi Yin[1]

[1]Tsinghua University, Beijing, China
[2]University of California, Santa Barbara, CA

Many computing-in-memory (CIM) processors have been proposed for edge deep learning (DL) acceleration. They usually rely on analog CIM techniques to achieve high-efficiency NN inference with low-precision INT multiply-accumulation (MAC) support [1]. Different from edge DL, cloud DL has higher accuracy requirements for NN inference and training, which demands extra support for high-precision floating-point (FP) MAC. As shown in Fig. 15.5.1, applying CIM techniques to cloud DL has three main limitations: 1) FP MAC has tightly coupled exponent alignment and INT mantissa MAC. Implementing complex exponent alignment in memory will harm CIM's direct accumulation structure and reduce efficiency. 2) FP MAC's energy is dominated by INT mantissa MAC. Further acceleration on CIM-based INT MAC is critical for processor efficiency. 3) Previous cloud DL processors usually have separate FP and INT engines, but only activate one engine at once [2], which causes high area overhead and low resource utilization.

Targeting the above limitations, this work proposes a reconfigurable FP/INT CIM processor ReCIM, with three features from top to bottom: 1) ReCIM is designed based on an in-memory alignment-free FP MAC pipeline that interleaves exponent alignment and INT mantissa MAC. Both inputs and weights are pre-aligned to their local maximum exponents, so CIM focuses on only MAC acceleration without complex alignment logic. 2) A Bitwise in-Memory Booth Multiplication (BM^2) architecture is designed with bitwise input Booth encoding in the BM^2 controller and partial-product recoding in the CIM macro. Compared with conventional bit-serial CIM, BM^2 reduces cycle count and bitwise multiplications by nearly 50%. 3) The proposed pipeline offers new opportunities to reuse CIM for constructing a unified FP/INT MAC dataflow. ReCIM implements hierarchical and reconfigurable in-memory accumulators to enable flexible support of BFloat16 (BF16)/FP32 and INT8/16 in the same CIM macro.

Figure 15.5.2 shows the overall architecture of the ReCIM processor. It consists of 16 CIM cores, a 32KB global buffer, a SIMD core, and a top controller. Each core has a 2KB Input-Exponent Memory (IEMEM), a 4KB Input-Mantissa Memory (IMMEM), an Input Alignment Unit (IAU), a BM^2 controller, and 4 BM^2-CIM macros. In FP mode, IEMEM stores input exponent bits, and IMMEM stores input mantissa bits in a new format called Mantissa+, which is the 2's complement representation of {sign, 1, mantissa}. IAU conducts exponent pre-alignment for inputs and sends them to the BM^2 controller for Booth encoding. The encoded inputs are fed to the BM^2-CIM macros and perform INT mantissa MAC with the weights, which are pre-aligned offline and stored in the Mantissa+ format. The outputs are sent to the SIMD core for operations like activation, normalization and rounding. In the INT mode, IEMEM and IMMEM are combined for input storage. Inputs are directly sent to the BM^2 controller and CIM macros for INT MAC computation.

Figure 15.5.3 shows the in-memory alignment-free FP MAC pipeline. We interleave the tightly coupled exponent alignment and INT mantissa MAC, to avoid alignment in CIM. IAU's exponent normalizer loads the exponents of 32 inputs from IEMEM and finds their maximum value E_{Imax} with a comparison tree. Each input's exponent offset is computed by E_{Imax}-E_I. IAU's mantissa shifter loads the corresponding mantissas from IMMEM, and right-shifts them by the offset bits. As for weights, only the mantissas are stored in CIM. E_{Wmax} of each weight column is previously obtained from the DL model. The weight mantissas are stored with our Mantissa+ format after right-shifting by E_{Wmax}-E_W bits. Value locality of DL models makes the exponent offset usually fewer than 4b, leading to no accuracy loss produced by shifting. After exponent pre-alignment to E_{Imax} and E_{Wmax}, the shifted input and weight mantissas can directly perform INT MAC operations in CIM. As shown in the example of Fig. 15.5.3, the output exponent equals E_{Imax}+E_{Wmax}-127. The final outputs are then transformed to the standard FP format in the SIMD core. Compared to the conventional FP MAC unit, our pre-alignment technique achieves 56.7% smaller area and 44.1% lower power. The CIM-based MAC implementation further reduces area and power by 71.9% and 78.3%.

Figure 15.5.4 shows the BM^2 architecture from a CIM subarray view with 8b weights. Every 8 encoders in the BM^2 controller are assigned to one 8×48 SRAM-CIM subarray. The BM^2 encoder is implemented by replacing the traditional radix-4 Booth's signal ONE with signal ZERO (represents partial product 0). Each input is fed into a BM^2 encoder to generate a set of encoding signals NEG, TWO, ZERO per cycle, which recode the partial product as {0, ±W,±2W}. NEG and TWO are sent to one row of the subarray through the bitline pair (BL/BLB). The BM^2-CIM macro is designed with full-digital logic to realize BM^2 partial-product recoding. A 4T XOR gate based on transmission gate logic is connected to the cell and BL for bitwise negation. Two transistors are added to the XOR output for 1b left shift, under the control of signal TWO. Thus, the intermediate partial product P[8:0] can represent {W,2W,$\overline{W}$,2$\overline{W}$}. The 8 rows of P[8:0] are sent to the subarray adder for zero setting, sign extension, and 8-row addition. Signal ZERO of each row decides whether to directly set P[8:0] as 0. Owing to the 2's complement representation, an additional 1 and 2 should be added to $\overline{W}$ and 2$\overline{W}$, respectively, to transform them to -W and -2W. Therefore, a 5b compensation sum is generated by the 8 BM^2 encoders and added to the 8-row sum. We use a digital CIM solution instead of analog CIM, in order to support higher precision (>8b) with no accuracy loss. Meanwhile, partial-product recoding can be easily implemented by customizing the digital CIM with only 6.40% area and 3.06% power overhead. Compared to conventional bit-serial CIM design [3], the BM^2 architecture achieves 1.6-to-2.0× speedup and 1.47-to-1.84× energy savings under different numerical precisions.

Figure 15.5.5 shows the hierarchical and reconfigurable in-memory accumulator design that supports BF16/FP32 and INT8/16 in one CIM macro. The macro's 48 columns are divided into 6 groups, so the subarray level has 6 corresponding adders. The macro level has 3 stages: 1) 6 adders that perform vertical accumulation for the 4 subarrays' partial sums. 2) 6 BM^2 accumulators that perform shift-accumulation for the bitwise encoded inputs. 3) Precision fusion that merge results from different columns. The subarray- and macro-level reconfigurations are defined by the operation sign and precision. For example, we use 24b to store an FP32 weight's mantissa, so every 3 adders of the subarray are combined to construct an FP32 MAC: The 1st adder is configured as signed 8b, and the remaining two are configured as unsigned 8b. The reconfigurable extension unit controls the sign mode of a subarray adder, according to the sign flag and Q[8:7]. In the macro level, the 24b fusion mode is enabled to merge the results of every 32 columns and generate 2 outputs per macro. The bottom left of Fig. 15.5.5 shows the detailed configurations for all supported precisions. Compared to a conventional architecture with separate FP and INT engines like [2], our design reuses CIM macros for the core INT MAC operations with the combination of exponent pre-alignment and reconfigurable CIM, achieving a reduction of 78.1% in area and 84.8% in power.

Figure 15.5.6 shows the measurement results of the ReCIM processor, fabricated in 28nm CMOS technology. The chip can work at 0.6-to-1.0V supply, 50-to-220MHz. For fair comparison, we implement a 28nm digital architecture baseline similar to the state-of-the-art cloud DL processor [2], with a separate FP engine for BF16/FP32 and INT engine for INT8/INT16. Evaluation is conducted on 3 different tasks with the ImageNet Dataset. ReCIM achieves 11.61× energy savings on ResNet-50 inference (INT8) over the baseline. Our digital CIM design has only 0.26% accuracy loss, much lower than the 2.24-to-3.45% loss reported in [1]. ReCIM runs EfficientNet-B0 inference in BF16 to maintain accuracy, and the total energy savings reaches 8.92×. Considering all on-chip components, ReCIM's peak system-level FP energy efficiency is 29.2TFLOPS/W at BF16, 0.65V, 95MHz, which is 20.42× higher than a recent FP CIM processor [4]. This is because [4] still uses digital MAC units, and designs the CIM just for exponent alignment. Figure 15.5.7 shows ReCIM's die photo, voltage-frequency scaling curves and summary table.

Acknowledgement:
This work was supported in part by National Key R&D Program 2018YFB2202600, NSFC Grant U19B2041 and 61774094, Beijing S&T Project Z191100007519016 and Beijing Innovation Center for Future Chip. The corresponding author of this paper is Shouyi Yin (yinsy@tsinghua.edu.cn).

References:

[1] J. Yue et al., "A 2.75-to-75.9TOPS/W Computing-in-Memory NN Processor Supporting Set-Associate Block-Wise Zero Skipping and Ping-Pong CIM with Simultaneous Computation and Weight Updating," *ISSCC*, pp. 238-239, 2021.
[2] A. Agrawal et al., "A 7nm 4-Core AI Chip with 25.6 TFLOPS Hybrid FP8 Training, 102.4 TOPS INT4 Inference and Workload-Aware Throttling," *ISSCC*, pp. 144-145, 2021.
[3] Y. Chih et al., "An 89TOPS/W and 16.3TOPS/mm^2 All-Digital SRAM-Based Full-Precision Compute-In-Memory Macro in 22nm for Machine-Learning Edge Applications," *ISSCC*, pp. 252-253, 2021.
[4] J. Lee et al., "A 13.7 TFLOPS/W Floating-point DNN Processor using Heterogeneous Computing Architecture with Exponent-Computing-in-Memory," *IEEE Symp. VLSI Circuits*, 2021.
[5] J. Park et al., "A 40nm 4.81 TFLOPS/W 8b Floating-Point Training Processor for Non-Sparse Neural Networks Using Shared Exponent Bias and 24-Way Fused Multiply-Add Tree," *ISSCC*, pp. 148-149, 2021.

2022 IEEE International Solid-State Circuits Conference (ISSCC) | 978-1-6654-2800-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISSCC42614.2022.9731762

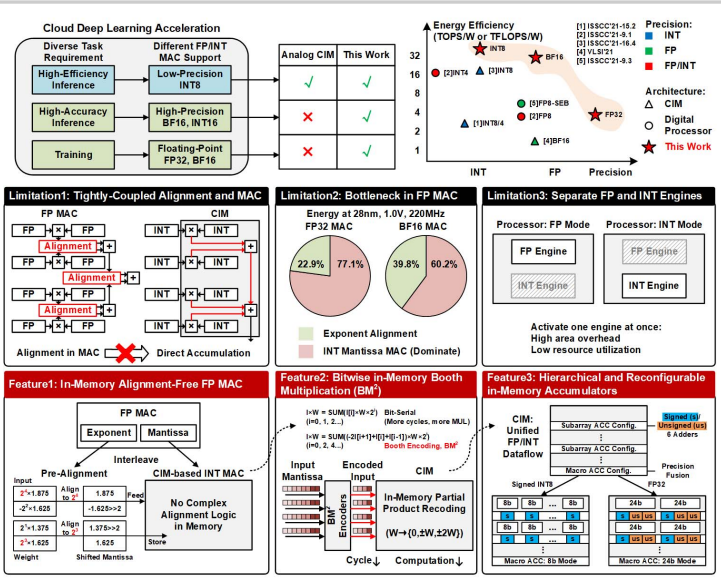


Figure 15.5.1: Limitations of designing a CIM processor for cloud deep learning acceleration and ReCIM's three features.

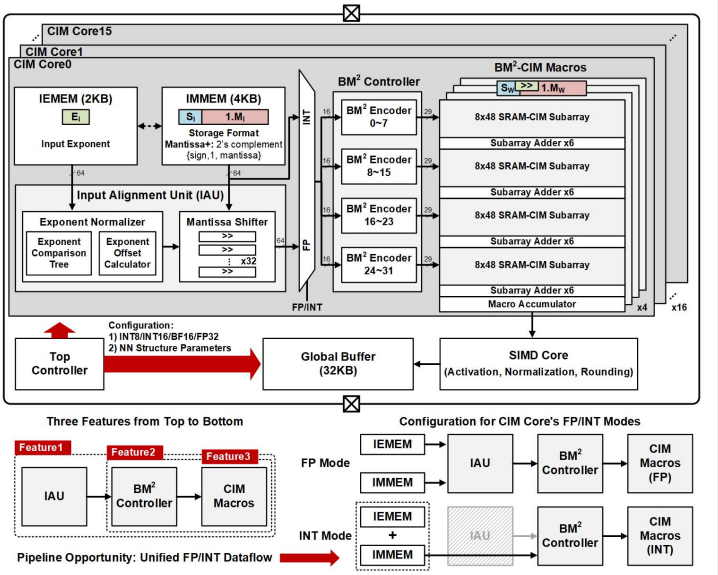


Figure 15.5.2: ReCIM's overall architecture, multi-level design features, and uniform FP/INT pipeline.

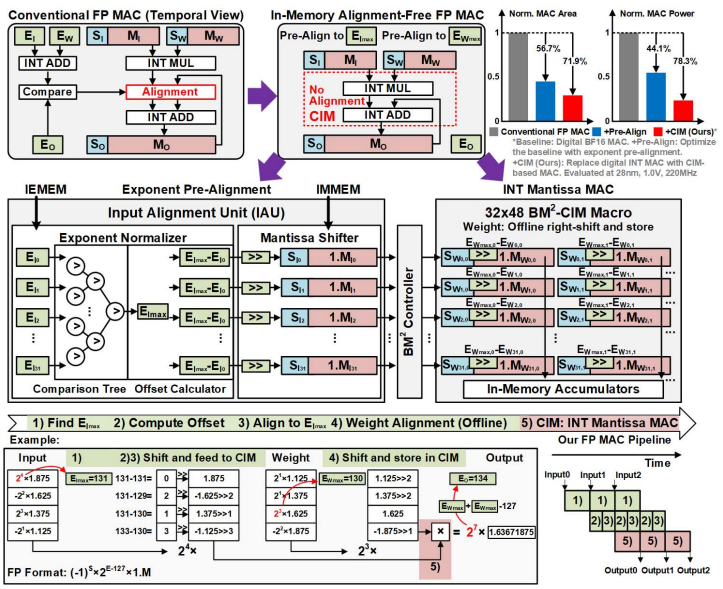


Figure 15.5.3: In-memory alignment-free FP MAC pipeline.

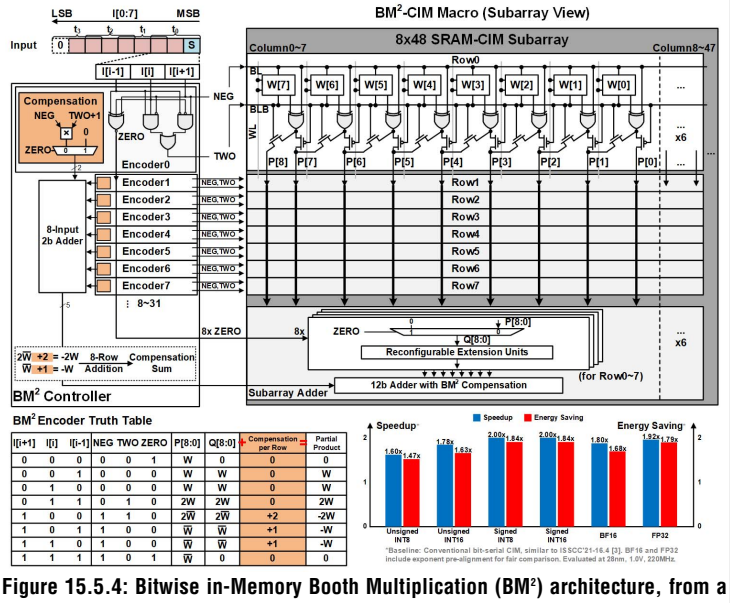


Figure 15.5.4: Bitwise in-Memory Booth Multiplication (BM²) architecture, from a CIM subarray view.

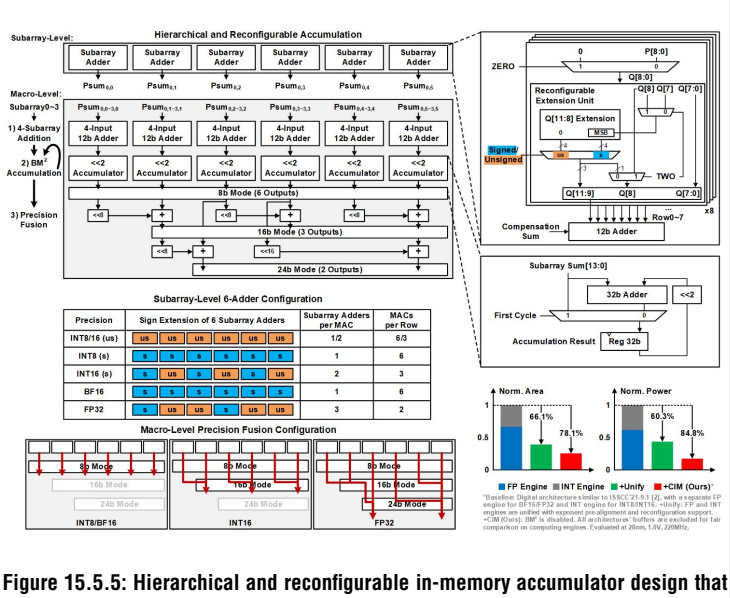


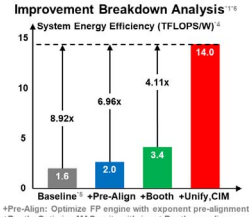
Figure 15.5.5: Hierarchical and reconfigurable in-memory accumulator design that supports BF16/FP32 and INT8/16 in one CIM macro.

Evaluation on Different Cloud Deep Learning Tasks[1]

Model	ResNet-50	EfficientNet-B0	EfficientNet-B0
Dataset	ImageNet	ImageNet	ImageNet
Task	Inference	Inference	Training
Data Precision	INT8	BF16	FP32
Top-1 Accuracy	75.87%	76.23%	76.28%
Accuracy Loss[2]	-0.26%	-0.05%	0
Execution Time	6.85ms	0.94ms	20.8ms[3]
CIM Macro Energy Efficiency	30.4TOPS/W	23.2TFLOPS/W	3.0TFLOPS/W
System Energy Efficiency[4]	21.7TOPS/W	14.0TFLOPS/W	1.8TFLOPS/W
Energy Saving[5]	11.61x	8.92x	9.86x

*1: Measured at 1.0V, 220MHz for high-performance evaluation. Off-chip memory is not included.
*2: Compared with Top-1 accuracy at FP32.
*3: An image's one round of feedforward, backpropagation, and gradient update passes.
*4: Include all on-chip components.
*5: We implement a 28nm digital architecture baseline like ISSCC'21-9.1 [2], with a separate FP engine for BF16/FP32 and INT engine for INT8/INT16.
*6: Evaluated on EfficientNet-B0 inference at BF16.

Improvement Breakdown Analysis[1][6]



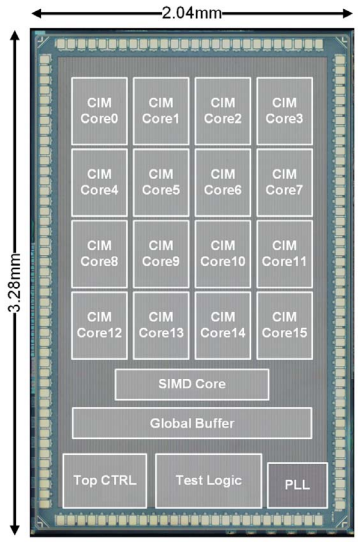
+Pre-Align: Optimize FP engine with exponent pre-alignment.
+Booth: Optimize MAC units with input Booth encoding.
+Unify,CIM (Ours): Combine FP and INT engines. Replace MAC units with BM²-CIM macros.

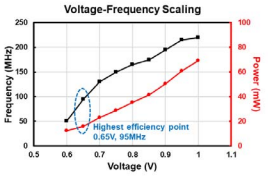
Comparison with State-of-the-art Deep Learning Processors

	ISSCC'21-9.1 [2]	ISSCC'21-9.3 [5]	ISSCC'21-15.2 [1]	ISSCC'21-16.4 [3]	VLSI'21 [4]	This Work
MAC Implementation	Digital MAC	Digital MAC	Analog CIM	Digital CIM	Digital MAC, Exponent-in-Memory	Digital CIM
INT Precision	INT2,INT4	×	INT2/4/6/8	INT4/8/12/16	×	√ INT8,INT16
FP Precision	HFP8,FP16,FP32	FP8-SEB	×	×	BF16	√ BF16,FP32
Technology	7nm	40nm	65nm	22nm	28nm	28nm
Supply Voltage (V)	0.55-0.75	0.75-1.1	0.62-1.0	0.72	0.76-1.1	0.6-1.0
Frequency (MHz)	1000-1600	20-180	25-100	500	~250	50-220
Die Area (mm²)	36	7.29	12	0.202 (macro)	5.832	6.69
Power (mW)	-	13.1-230	18.6-84.1	-	1.2-156.1	12.5-69.4
Peak Performance[1] (TOPS or TFLOPS) INT	102.4 (INT4)	N/A	0.013 (INT8)	0.917 (INT8)	N/A	1.35 (INT8)[4]
Peak Performance[1] (TOPS or TFLOPS) FP	25.6 (FP8)	0.567 (FP8-SEB)	N/A	N/A	0.119 (BF16)	1.08 (BF16)[4] 0.14 (FP32)[4]
Energy Efficiency[1,2] (TOPS/W or TFLOPS/W) INT	16.5 (INT4)	N/A	2.75 (INT8/4)	17.7 (INT8)[3]	N/A	36.5 (INT8)[5]
Energy Efficiency[1,2] (TOPS/W or TFLOPS/W) FP	3.5 (FP8)	4.81 (FP8-SEB)	N/A	N/A	1.43 (BF16)	29.2 (BF16)[5] 3.7 (FP32)[5]

*1: Evaluated in the dense mode. One operation (OP) represents one multiplication or one addition.
*2: System energy efficiency includes all on-chip components. Off-chip memory is not included.
*3: On-chip buffers are included for fair comparison. Buffer capacity is 128KB, same as ReCIM.
*4: Measured at the highest performance point, 1.0V, 220MHz.
*5: Measured at the highest efficiency point, 0.65V, 95MHz.

Figure 15.5.6: Measurement results and comparison with state-of-the-art deep learning processors.





Technology	TSMC 28nm CMOS	
Die Area	3.28mm×2.04mm	
Supply Voltage	0.6–1.0V	
Frequency	50–220MHz	
Buffer Size	128KB	
CIM Size	12KB	
CIM Macro Area	0.0147mm²×64	
Data Precision	INT8, INT16, BF16, FP32	
Power	12.47mW @0.6V, 50MHz	
	69.35mW @1.0V, 220MHz	
Peak Performance*1	INT8	1.35TOPS
	BF16	1.08TFLOPS
	FP32	0.14TFLOPS
Area Efficiency*1*3	INT8	1.43TOPS/mm² (macro) 0.20TOPS/mm² (system)
	BF16	1.15TFLOPS/mm² (macro) 0.16TFLOPS/mm² (system)
	FP32	0.15TFLOPS/mm² (macro) 0.02TFLOPS/mm² (system)
Energy Efficiency*1*2*3	INT8	30.8–57.8TOPS/W (macro) 19.5–36.5TOPS/W (system)
	BF16	24.6–46.2TFLOPS/W (macro) 15.6–29.2TFLOPS/W (system)
	FP32	3.2–5.9TFLOPS/W (macro) 2.0–3.7TFLOPS/W (system)

*1: Highest performance (lowest efficiency) point, 1.0V, 220MHz.
*2: Highest efficiency point, 0.65V, 95MHz.
*3: Evaluated in the dense mode. Macro evaluation includes only CIM macros. System evaluation includes all on-chip components.

Figure 15.5.7: Die micrograph, voltage-frequency scaling curves, and summary table.

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