A 28nm 244.45TOPS/W Winograd-Standard Fusion accelerator with Symmetric Hybrid Domain CIM Groups for Edge AI devices

An Guo[1#], Zhichao Liu[1#], Wentao Zheng[1], Yutong Zhang[1], Tianhui Jiao[1], Fangyuan Dong[1], Mingzi Wang[2], Shaochen Li[1], Zhican Zhang[1], Yuhui Shi[1], Xing Wang[1], Xin Si[1], Xin Wang[2*], Wenwu Zhu[2*]

[1]Southeast University, Nanjing, China, [2]Tsinghua University, Beijing, China

Email: xin_wang@tsinghua.edu.cn/, wwzhu@ tsinghua.edu.cn/

[#]Equally contributed authors *Corresponding authors

The rapid advancement of AI technologies has garnered widespread attention, but the substantial power consumption of sophisticated AI models presents a significant challenge. Various AI accelerators [1-10] have been developed to optimize energy efficiency for AI tasks. Winograd Convolution, as a CNN accelerating method, can save more than 2.25 times operations, as illustrated in Fig.1. This method involves four steps: (1) Winograd-domain (WD) weights transformation to W' using the G matrix; (2) WD features transformation to F' using the B matrix; (3) Hadamard production between W' and F' to obtain WD results Y'; and (4) Inverse WD transformation into final neural network results Y. Winograd Convolution typically employs multiple transform sizes, such as (2,3), (4,3), (6,3), where the numbers represent output and input tile sizes (OTS and ITS) respectively. Larger OTS yields more operation savings but increases accuracy loss, with OTS >4 leading to task failure in YOLO inference tests. Given the additional memory requirements of Winograd convolution, CIM emerges as a promising solution to efficiently implement this algorithm. This work proposes a Winograd AI accelerator with symmetric digital-analog hybrid domain CIM groups addressing: (1) Winograd's limitation to stride-1 layers, as decomposition algorithms for stride-2 CONV lead to large redundant calculations; (2) Significant storage and access overhead in WD weights; and (3) Complex Winograd operations, including GWGT, BTFB, Hadamard production and ATY'A. A 28nm fabricated Winograd/standard fusion AI accelerator achieves an INT8 energy efficiency of 79.9TOPS/W and a mean average precision (mAP) of 57.88% with retraining for YOLOv8 with Winograd CONV.

Fig. 2 illustrates the architecture of the proposed Winograd and standard convolution fusion (WSF) hybrid CIM chip, featuring: (1) a WSF CNN CIM architecture for edge image and video processing; (2) a Left Feature Parallel, Right Feature Serial Adder (LFp-RFsA) and WSF-digital-to-macro buffer (WSF-D2MB) based pre-macro processor for input stationary flow; (3) a bit-configurable digital-analog hybrid domain CIM with symmetric structure (S-hybrid CIM); and (4) a Winograd-advanced-calculation standard-ordinary-accumulation (WaC-SoA) based post-macro processor. The WSF chip comprises a main control, IO, Input preprocessor, 568Kb global buffer, SIMD core for nonlinear functions, WSF mode controller, ReLU early termination, and WSF core. Winograd calculations are divided into four parts: (1) F' = BTFB, computed in the LFp-RFsA pre-macro processor, where input data is split into LFp- and RFs-groups, processed in parallel and bit-serial over 128 cycles respectively, with WSF-D2MB using paired DFFs for pipeline input scheduling; (2) W'=GTWG, calculated offline for multi-time reuse; (3) Y' = F'⊙W', executed in 4 symmetric-hybrid CIMs capable of INT2/4/8 mode operations, solving 128-ichs MAC operations; and (4) Y = ATY'A, processed in the WaC-SoA aggregator, computing ATY' and half of Y'A in advance to accelerate post-macro processing. Final results are sent to an output buffer for subsequent ReLU and other nonlinear functions.

Fig. 3 illustrates the proposed LFp-RFsA and WSF-D2MB pre-macro processor, which operates in two modes: Winograd and standard convolution. In Winograd mode, input data undergoes BTFB transformation, allowing multi-position input sharing. The process is divided into two parts, calculated alternately. Each part shares 4 of 12 inputs between left- and right-groups, with different input rows directed to 4 S-hybrid CIM macros. The LFp-RFs scheduler manages left-groups using 2 8b individual and 2 8b shared inputs for one LFp-adder per macro, while right-groups use 8×2 1b individual and 8×2 1b shared inputs for 8 RFs-adders per macro. LFp-adders perform 4 8b accumulations or subtractions per cycle, producing 8b macro input data stored in a temporary storage (TS) buffer. RFs-adders calculate 4 1b operations per cycle, with 8 adders introduced to maximize utilization of CIM macros' 1b feature and 8b weight MAC capabilities for 16-ochs in 16 cycles. RFs-adder results are stored in TS buffers with 3b carry bit storage DFF and 1b extra DFF for the last 1b PSUM. To address carry-in issues, RFsAs operate two cycles earlier in the Winograd BTFB pipeline pre-macro flow. In standard convolution mode, inputs from different ochs are divided into left- and right-groups, with left features scanned from high to low bits and right features from low to high. These inputs bypass LFp-RFsA and WSF-D2MB, proceeding directly to S-hybrid CIM macros.

Fig. 4 illustrates the proposed symmetric hybrid CIM macro, comprising an INDRV, WL driver, key switch driver, IO, multi-level accumulator, main and CIM controllers, and 8 128×32 banks. Each bank processes 128 1b and 8b MAC operations for 2 output channels (ochs), one in analog and one in digital domain. The macro employs an input-stationary inner-macro data flow, maximizing input utilization through local weight switching. The left side of the figure details the computing units, where upcells and downcells connect to digital computing units (DCU) and analog computing units (ACU), controlled by K0. Different banks are managed by distinct Kns, as shown in the top right. The 8b vertical cut example operates in 8 steps: (1) With K0-6=0 and K7=1, left features input highest bits and right features input lowest bits. Left features and upcells compute digitally, except for the lowest bit in analog, while right features and downcells compute in analog, except for the highest bit in digital. This process repeats for 16 cycles, switching local weights for 16 ochs. (2) K0-5=0 and K6-7=1, inputting MSB-1 for left features and LSB+1 for right features. Left features and upcells compute digitally except for the lowest 2 bits in analog, and right features and downcells compute in analog except for the highest two bits in digital. (3) Similar operations as (1-2) continue with Kn and left-right feature switches, completing the 8-step process for full 8b computation.

Fig.5 top illustrates the proposed WaC-SoA post-macro processor, utilizing S-hybrid CIM macro in INT4/8 mode. The S-hybrid CIM macro incorporates two multi-level accumulators, generating two 128 1b feature and 4b weight PSUMs, which are processed by distinct Winograd advanced ATY'A processors. Each WaC processor divides calculations into top- and bottom- A matrix multiplications, computing ATY'& top of Y'A initially and ATY'& bottom of Y'A subsequently, with shared input data. In INT8 mode, post-WaC processor PSUMs undergo shift-and-add operations to produce final INT8 results. Fig.5 bottom showcases performance improvements: the LFp-RFs pre-macro processor achieves 64x adders, 1.79x power, and 4.59x area savings, while the WaC-SoA Aggregator post-macro processor realizes 1.50x power reduction and 3.04x speed increase. The CIM macro achieves an area efficiency of 840.9GOPS/mm^2@0.9V and an energy efficiency of 79.88TOPS/W@0.6V. Three hybrid domain CIM structures - lightning-like (LLS) [5], vertical-cut (VCS) [6], and symmetric - were evaluated using 1000 random vectors with identical digital and analog circuits, resulting in 2-8x RC overhead with maximum 4.01% accuracy loss. In the WSF chip, pre-, inner-, and post-macro components consume 7.3%, 20.5%, and 72.2% of power and 10.7%, 17.6%, and 71.7% of area, respectively, while conducting 4.2%, 8.2%, and 87.6% of operations. The additional operations in pre- and post-macros are attributed to BTFB and ATY'A computations.

Fig.6 displays the measured shmoo plot of the proposed WSF-chip, along with measured results on VGG16, ResNet18, and Yolov8, and a comparison table with previous work. The chip achieves 620MHz@0.9V and 165MHz@0.6V, with an energy efficiency of 45.46-243.62TOPS/W. Fig.7 presents the die photo and chip summary of the fabricated 28nm CMOS technology Winograd-standard CNN fusion Hybrid CIM accelerator. Proposed chip features a core area of 0.9mm2 with four 0.11mm2 hybrid CIM macros, achieving an energy efficiency of 62.93-244.45 TOPS/W @INT4 precision and 14.85-57.68 TOPS/W @INT8 precision. Additionally, it demonstrates an inference accuracy of 70.72% on ResNet-18@ImageNet and a mean average precision (mAP) of 57.88% on Yolov8@COCO.

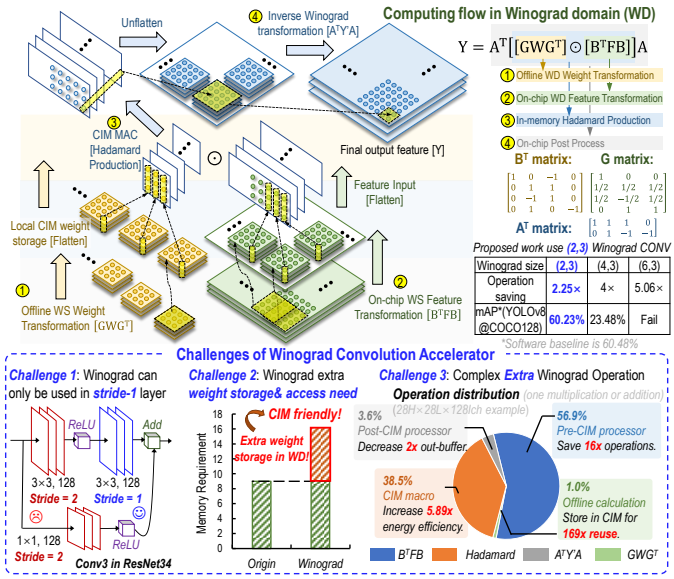


Fig. 1. Design challenges of Winograd CIM accelerator.

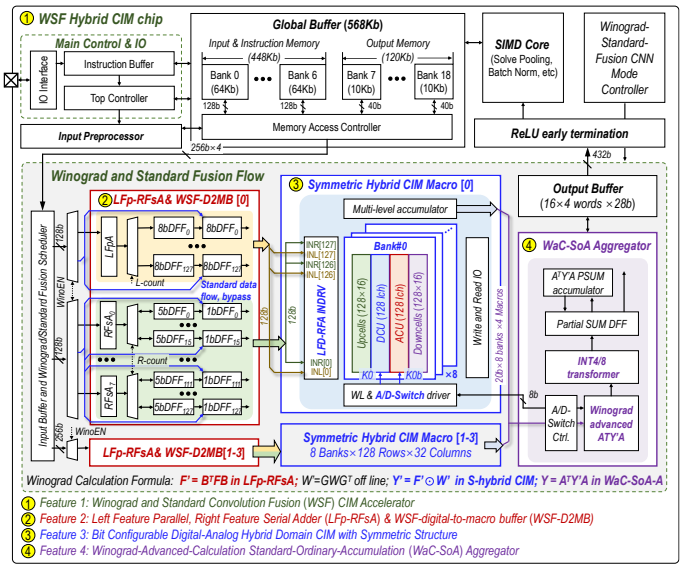


Fig. 2. Overall system architecture of proposed Winograd-Standard convolution fusion CIM accelerator.

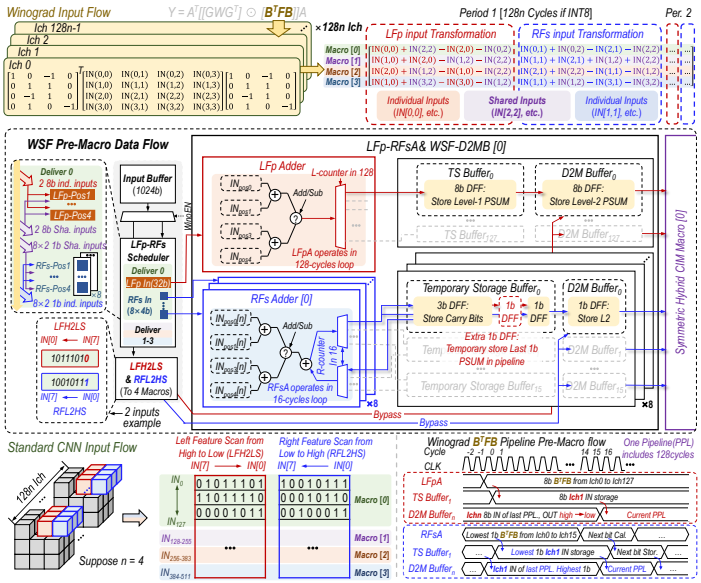


Fig. 3. Proposed LFp-RFsA and WSF-D2MB pre-macro processor.

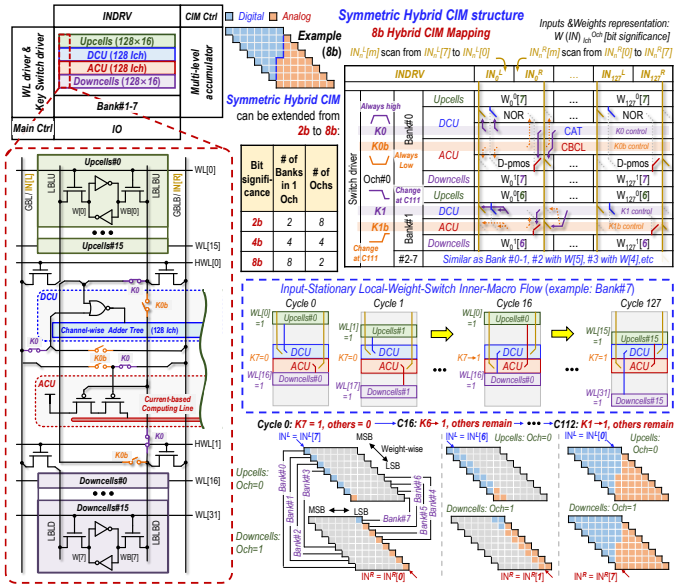


Fig. 4. Proposed symmetric hybrid CIM macro.

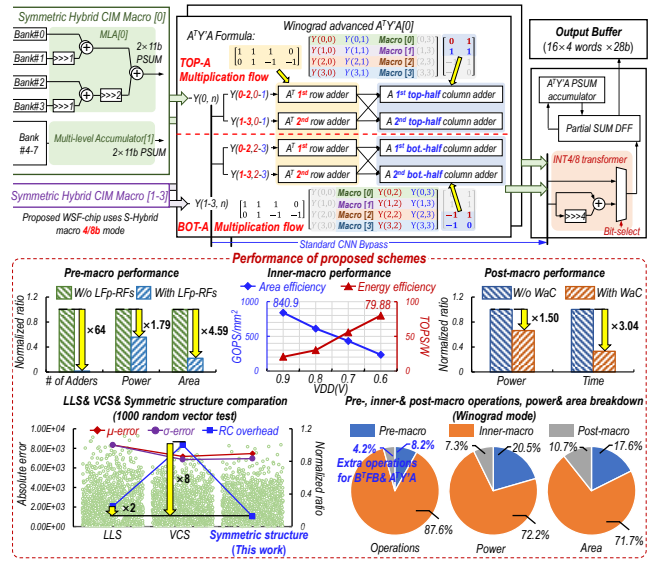
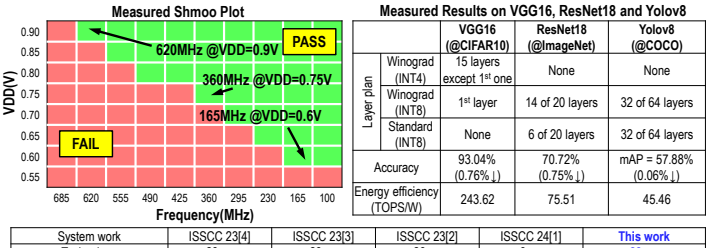


Fig. 5. Proposed WaC-SoA post-macro processor and performance improvements.

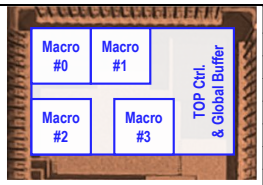


Measured Results on VGG16, ResNet18 and Yolov8

		VGG16 (@CIFAR10)	ResNet18 (@ImageNet)	Yolov8 (@COCO)
Layer plan	Winograd (INT4)	15 layers except 1st one	None	None
	Winograd (INT8)	1st layer	14 of 20 layers	32 of 64 layers
	Standard (INT8)	None	6 of 20 layers	32 of 64 layers
Accuracy		93.04% (0.76% ↓)	70.72% (0.75% ↓)	mAP = 57.88% (0.06% ↓)
Energy efficiency (TOPS/W)		243.62	75.51	45.46

System work	ISSCC 23[4]	ISSCC 23[3]	ISSCC 23[2]	ISSCC 24[1]	This work
Technology	28nm	28nm	28nm	3nm	28nm
MAC implementation	Digital CIM	Digital CIM	Digital MAC	Digital CIM	Hybrid CIM
Model	ResNet50	Transformer	Mobile/ Res-Net	Super Resolution	ResNet50, Yolov8
CIM capacity(Kb)	4×16	16×64	-	1248	4×32
On-chip SRAM(KB)	512	192	1120	917	71
System size(mm^2)	2.52×1.8=4.54	4.46×3.22=14.36	7.81	1.37	1.2×0.75=0.9
Macro size(mm^2)	0.56×0.12 =0.067	-	-	-	0.330×0.353 =0.116
Supply voltage(V)	0.397-0.9	0.6-1.0	0.66-1.3	0.46-1.0	0.6-0.9
Frequency (MHz)	10-400	85-275	100-500	400-1300	165-620
Input precision (bit)	INT4/8, FP/BF16	INT8/16	INT8	INT12	INT4/ INT8
Weight precision (bit)	INT4/8, FP/BF16	INT8/INT16	INT8	INT12	INT4/ INT8
Peak performance(TOPS)	9.63(INT4)	0.89/3.55	0.46	16.5	1.152
System energy efficiency (TOPS/W)	12.8-75	12.1-101.1	3.33-4.98	16.3-23.2	14.85-244.45
System area efficiency (TOPS/mm^2)	0.361-2.12	0.062/0.247	0.059	12.0	0.129-1.94
Macro energy efficiency (TOPS/W)	68.7-403	-	-	-	20.57-341.01
Macro area efficiency (TOPS/mm^2)	-	0.06/0.25	-	-	0.235-3.364

Fig. 6. Measurement results and performance comparison table.



[1]High point: test using Yolov8@COCO under 0.6V at 165MHz; low point: test using Yolov8@COCO under 0.9V at 650MHz.
[2]High point: test under 0.9V; low point: test under 0.6V.
[3]Using ResNet-18 model and the software baseline is 71.27%.
[4]Using Yolov8 model and the software baseline is 57.94%.

CHIP SUMMARY			
Technology		28nm CMOS	
Core area		1.2×0.75=0.9 mm^2	
CIM Macro Area		4×0.330×0.353 mm^2	
Supply voltage(V)		0.6-0.9	
Frequency		165-620MHz	
CIM size		4×32Kb	
SRAM size		568Kb	
CIM memory density(Kb/mm^2)		90.65	
Peak performance(TOPS)		1.152	
Precision(bit)		INT 4	INT 8
System energy efficiency[1] (TOPS/W)		62.93-244.45	14.85-57.68
Macro energy efficiency[1] (TOPS/W)		93.18-341.01	20.57-79.9
System area efficiency[2] (TOPS/mm^2)		0.516-1.94	0.129-0.485
Macro area efficiency[2] (TOPS/mm^2)		0.94-3.364	0.235-0.841
Neural network index of different AI tasks	Inference accuracy[3] (ResNet-18 @ImageNet)	70.72% (16, 4 of total 20 layers using Winograd INT8 and Standard INT8 convolution)	
	mAP[4] (Yolov8@COCO)	57.88% (32, 32 of total 64 CONV layers using Winograd INT8 and Standard INT8 convolution)	

Fig. 7. Die photo and chip summary table.

Acknowledge

This work is funded with NSTMP under Grant 2022ZD0118901, NSFC under Grant 62522403, 62204036, 92464302, Jiangsu Provincial RDP under Grant BE20230201and the Fundamental Research Funds for the Central Universities.

References:

[1] M. Shih et al., " NVE: A 3nm 23.2TOPS/W 12b-Digital-CIM-Based Neural Engine for High-Resolution Visual-Quality Enhancement on Smart Devices," ISSCC, pp. 128-130, 2024.

[2] C. Du et al., "A 28nm 11.2TOPS/W Hardware-Utilization-Aware Neural-Network Accelerator with Dynamic Dataflow," ISSCC, pp. 332-333, 2023.

[3] F. Tu et al., "MulTCIM: A 28nm 2.24μJ/Token Attention-Token-Bit Hybrid Sparse Digital CIM-Based Accelerator for Multimodal Transformers," ISSCC, pp. 248-249, 2023.

[4] J. Yue et al., "A 28nm 16.9-300TOPS/W Computing-in-Memory Processor Supporting Floating-Point NN Inference/Training with Intensive-CIM Sparse-Digital Architecture," ISSCC, pp. 252-254, 2023.

[5] A. Guo et al., "A 22nm 64kb Lightning-Like Hybrid Computing-in-Memory Macro with a Compressed Adder Tree and Analog-Storage Quantizers for Transformer and CNNs," ISSCC, pp. 570-571, 2024.

[6] P. -C. Wu et al., " A 22nm 832Kb Hybrid-Domain Floating-Point SRAM In-Memory-Compute Macro with 16.2-70.2TFLOPS/W for High-Accuracy AI-Edge Devices," ISSCC, pp. 126-128, 2023.

[7] A. Guo et al., " A 28nm 64-kb 31.6-TFLOPS/W Digital-Domain Floating-Point-Computing-Unit and Double-Bit 6T-SRAM Computing-in-Memory Macro for Floating-Point CNNs," JSSC, vol.59, no.9, 2024.

[8] V. Chikin and V. Kryzhanovskiy,"Channel Balancing for Accurate Quantization of Winograd Convolutions," CVPR, pp. 12497-12506,2022.

[9] C. Mu et al., "A 28nm 76.25TOPS/W RRAM/SRAM-Collaborative CIM Fine-Tuning Accelerator with RRAM-Endurance/Latency-Aware Weight Allocation for CNN and Transformer," 2024 IEEE Asian Solid-State Circuits Conference (A-SSCC), Hiroshima, Japan, 2024, pp. 1-3.

[10] Z. Wei, Y. Su, T. T. -H. Kim and Y. Zheng, "A 280.580 μm2/Grid, 30.7pJ TMRA-PIM (Time-Multiplexed Random-Access Processing-in-Memory) Programmable Accelerator for Partial Differential Equations Solving and Edge Processing," 2024 IEEE Asian Solid-State Circuits Conference (A-SSCC), Hiroshima, Japan, 2024, pp. 1-3.