34.4 A 3nm, 32.5TOPS/W, 55.0TOPS/mm^2 and 3.78Mb/mm^2 Fully-Digital Compute-in-Memory Macro Supporting INT12 × INT12 with a Parallel-MAC Architecture and Foundry 6T-SRAM Bit Cell

Hidehiro Fujiwara[1], Haruki Mori[1], Wei-Chang Zhao[1], Kinshuk Khare[1], Cheng-En Lee[1], Xiaochen Peng[2], Vineet Joshi[3], Chao-Kai Chuang[1], Shu-Huan Hsu[1], Takeshi Hashizume[4], Toshiaki Naganuma[4], Chen-Hung Tien[1], Yao-Yi Liu[1], Yen-Chien Lai[1], Chia-Fu Lee[1], Tan-Li Chou[1], Kerem Akarvardar[2], Saman Adham[3], Yih Wang[1], Yu-Der Chih[1], Yen-Huei Chen[1], Hung-Jen Liao[1], Tsung-Yung Jonathan Chang[1]

[1]TSMC, Hsinchu, Taiwan
[2]TSMC, San Jose, CA
[3]TSMC, Ottawa, Canada
[4]TSMC, Yokohama, Japan

2024 IEEE International Solid-State Circuits Conference (ISSCC) | 979-8-3503-0620-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISSCC49657.2024.10454556

Compute-in-memory (CIM) is being widely explored to minimize power consumption related to data movement and multiply-and-accumulate (MAC) operations for AI edge devices. Compared to analog based CIMs, digital-based CIMs (DCIM), which include small, distributed SRAM banks and a customized MAC unit, realize massively parallel computation with no accuracy loss and better power-performance-area (PPA) scaling with advanced technologies. However, balancing operating efficiency per area (TOPS/mm^2) and bit density (Mb/mm^2) is one of the challenges in prior DCIMs because of low operating throughput caused by bit-serial input and a small number of rows. In this paper, we introduce a 3nm SRAM-based DCIM macro, which is based on foundry 6T SRAM bit cells and a parallel-MAC scheme to improve bit cell density and operating throughput. The DCIM macro is implemented with CIM BIST in a test chip, and confirms ultra-low voltage MAC operation down to 360mV, and 1.5GHz operation at 0.9V. The DCIM achieves 32.5TOPS/W (assuming a 25% input toggle rate and a 50% weight = 1 distribution), 55.0TOPS/mm^2 and 3.78 Mb/mm^2.

Figure 34.4.1 summarizes the architecture of the 3nm DCIM macro. The DCIM macro is configured with 72 input channels, 4 output channels and 18 weight sets. The DCIM macro supports MAC operations with signed INT12 format [1], offering a good tradeoff between accuracy and PPA (compared to INT8 and FP16). To balance TOPS/mm^2 vs Mb/mm^2, we use a foundry provided 6T-SRAM bit cell, with a 0.026μm^2 cell size, and a parallel-MAC scheme based on look-up-tables (LUTs) [2]. Weight memory storage is partitioned into 18 segments for 18 input channels. Each segment consists of 18 rows and 192 columns. 192 columns include 12b for INT12, 4 input channels and 4 output channels (i.e. 12 × 4 × 4). The MAC unit processes the 12b × 12b parallel MAC operation and it receives 3456 bits weight data (12b × 72 input channels ×4 output channels) from the memory and 864 bits input data (12b × 72 input channels). Because there is a gap of minimum operating voltage (V_{MIN}) between the SRAM bit cell and the MAC logic, there are two power domains (V_{DD} and V_{DDM}) in the macro: $V_{DDM} \geq V_{DD}$ and V_{DD} can be reduced below 0.4V. V_{DDM} is assigned to the SRAM array to ensure SRAM functionality and V_{DD} is assigned to interface and MAC unit for power saving. Low-V_t devices are mainly used in the V_{DD} domain to support high-speed operation even in the ultra-low voltage range. High-V_t devices are used in the V_{DDM} domain for leakage reduction. Level shifters (LS) are implemented before the address decoder to transfer data between the V_{DD} and V_{DDM} domains. On the other hand, no level shifters are required between SRAM output and MAC unit since the signal propagation from V_{DDM} (high) to V_{DD} (low voltage) presents no issues in terms of functionality.

Figure 34.4.2 shows the schematic for the SRAM array and local IO (LIO) circuitry for one output channels (48 columns). To increase cell efficiency, we use the flying-BL scheme [3] and arrays of two segments are directly abutted, eliminating any empty space. Because of the cell efficiency improvement due to the flying-BL scheme, the total macro area is around 5% smaller. Unlike conventional SRAM operations, all segments are simultaneously accessed and in total 18 WLs/macro are asserted to enable high bandwidth readout. The readout data from all columns in all segments are then propagated to the MAC unit. We eliminated sense amplifiers because there are only 18 rows in each segment and the BL discharge period is short. This simplifies the read path in LIO circuitry and minimizes the area and read energy. Moreover, only one WL in a macro is selected and 4:1 multiplexer is used in the LIO write circuitry to mitigate routing congestion. For the write operation the weight data is updated for only one input channel (i.e., 12 bit × 4 output channels) in each cycle. The parallel MAC scheme not only achieves a higher throughput but also a lower energy consumption. Figure 34.4.3 explains the data pattern dependency in real workloads and the data toggle rate difference between serial and parallel MAC operations. We analyzed the data pattern and toggle rate difference on AlexNet, ResNet-50, MobileNetV2 and Inception-v1 running inference on the ImageNet dataset. As shown on Fig. 34.4.3(left), the input data toggle rate for the MSBs is lower than that for the LSBs, for all CNNs we analyzed because the MSBs show a stronger data correlation compared to the LSBs. Due to the bit-serial input, from MSB to LSB, the serial MAC misses the MSB data correlation. The toggle rate comparison summary shown in Fig 34.4.2(right) shows a reduced input toggle rate in the parallel MAC compared to the serial MAC, which lowers power consumption. In addition to parallel operation, we have also leveraged LUTs in the MAC unit [2]. Because the LUT approach can share summation results f two weights (weight0, weight1, weight1 + weight0 or 0) with multiple bits of input, the device count and delay are reduced. As an example, compared to a Booth multiplier, a 7% smaller device count and a 5% faster speed can be achieved by using the LUT approach.

Figure 34.4.4 shows the three-sage pipelined operation in the DCIM macro. The first stage is the SRAM array access, and the MAC operation is divided into two subsequent stages. Because of the weight stationary dataflow, SRAM array accesses occur only when the weight update for a MAC is required. In addition, SRAM array accesses, MAC stage 1 and 2, can be executed simultaneously without incurring any penalty in timing (see cycle #5 in Fig. 34.4.4). The cycle time for each stage when the V_{DDM} voltage is set to 0.675V is shown on the bottom left of Fig. 34.4.4. Because we fine tune the gate and RC delays in MAC stages using a device-level STA tool, the cycle time for MAC stage1 and stage2 remain essentially the same across voltage. Operating speed is determined by the MAC unit when V_{DD} is lower than V_{DDM}, while operating speed is nearly the same among SRAM array access, MAC stage1 and stage2 when V_{DD} and V_{DDM} are similar. When the input toggle rate is 25% and weight=1 probability is 50%, the energy efficiency at 0.55V is 32.5TOPS/W, 22.5TOPS/W and 12.1TOPS/W for no SRAM array access, an SRAM array access every four cycles, and an SRAM array access every cycle, respectively. When the SRAM array is accessed, additional power is consumed, in both the SRAM array and the MAC unit, due to signal toggles from both input and weight.

Figure 34.4.7 shows the 3nm FinFET test chip micrograph. The DCIM macro area is 0.0157mm^2. The bit capacity per macro is 60.75kb; hence, the bit density per area is 3.78Mb/mm^2. Since BIST and DFT schemes are key components for mass production, we also implemented CIM BIST based on a commercial MBIST engine. The CIM BIST includes a custom algorithm for the DCIM macro and realizes >99% fault coverage in a gate-level fault simulator. The area overhead of CIM BIST is around 4% for one CIM BIST per macro and testing time for each macro is <2ms. The area overhead can be smaller if one CIM BIST is shared across multiple macros, although that would increase the testing time. Figure 34.4.5 summarizes measurement results for the SS corner wafer. The chart on the left is V_{MIN} measurement results at –40°C. The plots using a red color are for when both V_{DD} and V_{DDM} are swept at the same time (i.e. $V_{DD} = V_{DDM}$). We confirmed V_{MIN} becomes below 0.6V with 95% yield. On the other hand, the plots using a blue color are for when V_{DD} is swept and V_{DDM} is fixed. We observed V_{MIN} then becomes 0.36V with a 95% yield. Since the MAC unit, which is the most critical component for energy consumption, can operate at a ultralow voltage, the DCIM can realize better TOPS/W compared to the single rail design (i.e. V_{DD}=V_{DDM}). We observed a good match between the simulated and measured power. We also measured the frequency vs. voltage Shmoo at —40°C. At a fixed V_{DDM} voltage the measured operating frequency is 300MHz, 650MHz and 1.5GHz at 0.4, 0.5 and 0.9V. Figure 34.4.6 compares this work with prior work: compared to [1], TOPS/W, Mb/mm^2 and TOPS/mm^2 are 47, 23 and 65% better.

References:

[1] H. Mori et al., "A 4nm 6163-TOPS/W/b 4790–TOPS/mm^2/b SRAM Based Digital-Computing-in-Memory Macro Supporting Bit-Width Flexibility and Simultaneous MAC and Weight Update," *ISSCC*, pp. 132-133, 2023.
[2] C.-F. Lee et al., "A 12nm 121-TOPS/W 41.6-TOPS/mm^2 All Digital Full Precision SRAM-based Compute-in-Memory with Configurable Bit-width For AI Edge Applications," *IEEE Symp. on VLSI Tech. and Circ.*, pp. 24-25, June 2022.
[3] J. Chang et al., "A 7nm 256Mb SRAM in high-k metal-gate FinFET technology with write-assist circuitry for low-VMIN applications," *ISSCC*, pp. 206-207, 2017.
[4] G. Jedhe et al., "A 12nm 137 TOPS/W Digital Compute-In-Memory using Foundry 8T SRAM Bitcell supporting 16 Kernel Weight Sets for AI Edge Applications," *IEEE Symp. on VLSI Tech. and Circ.*, June 2023.
[5] G. Desoli et al., "A 40-310TOPS/W SRAM-Based All-Digital Up to 4b In-Memory Computing Multi-Tiled NN Accelerator in FD-SOI 18nm for Deep-Learning Edge Applications," *ISSCC*, pp. 260-261, 2023.

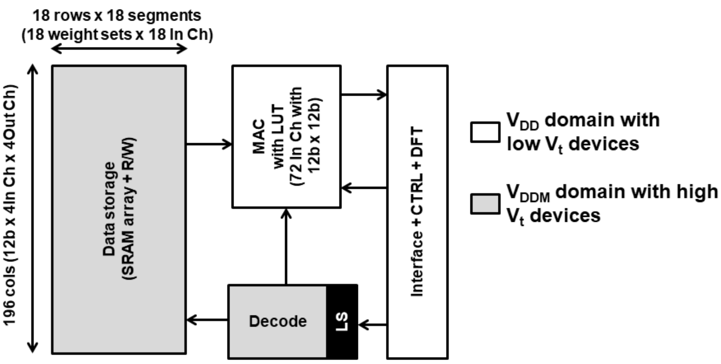

18 rows x 18 segments
(18 weight sets x 18 In Ch)
196 cols (12b x 4In Ch x 4Out Ch)
Data storage
(SRAM array + R/W)
MAC
with LUT
(72 In Ch with
12b x 12b)
Interface + CTRL + DFT
Decode
LS
V_{DD} domain with low V_t devices
V_{DDM} domain with high V_t devices

Figure 34.4.1: CIM architecture and dual-rail power assignment.

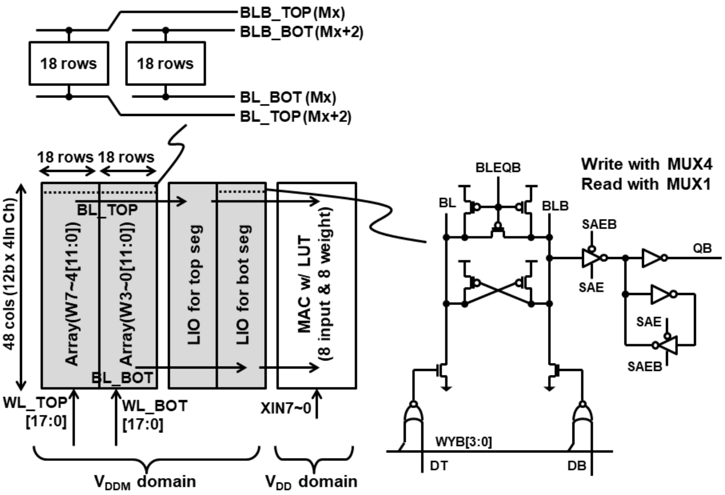


Figure 34.4.2: Array with flying-BL and local-IO circuitry.

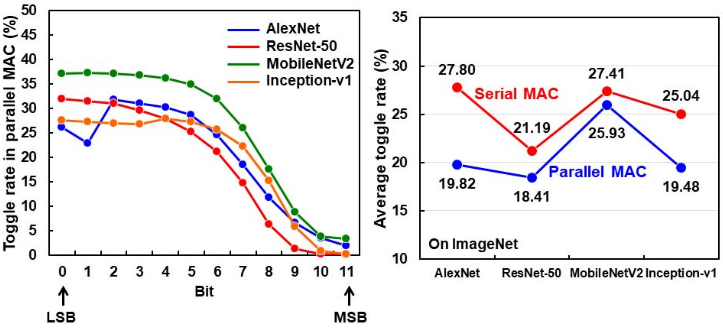


Figure 34.4.3: Parallel vs serial MAC comparison.

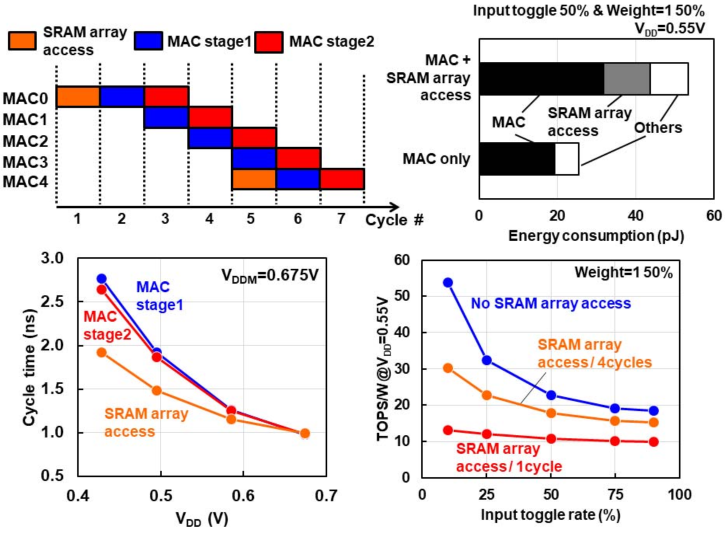


Figure 34.4.4: SRAM array access and pipeline operation.

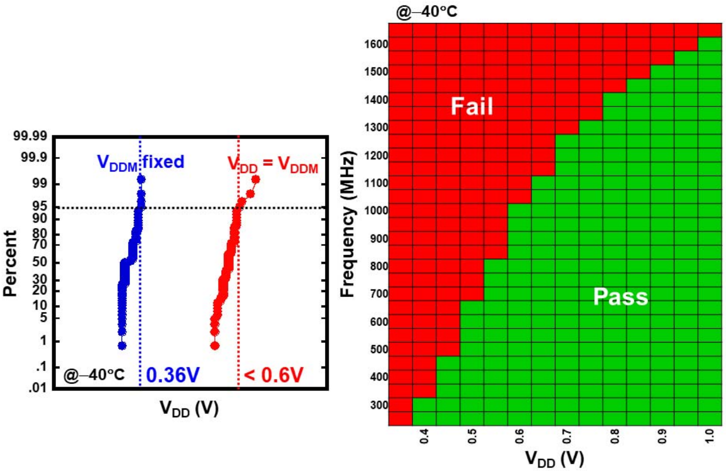
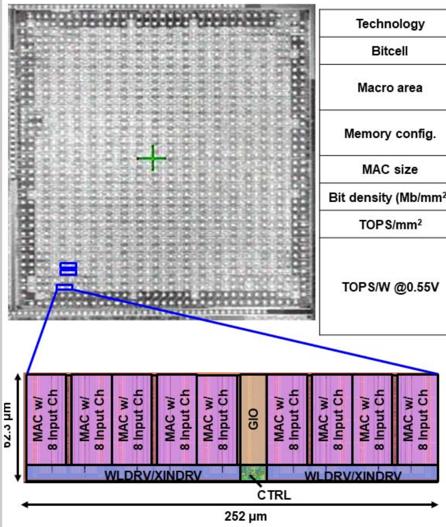


Figure 34.4.5: Chip measurement results with CIM BIST.

	ISSCC'23 [1]	VLSI'22 [2]	VLSI'23 [4]	ISSCC'23 [5]	This work
Technology	4nm	12nm	12nm	18nm	3nm
Bitcell	8Tx2 +OAI (2bit)	NA	8T2P	8T2P	6T
VDD	0.32 ~ 1.1V	0.72V	0.55 ~ 0.99V	0.525 ~ 1.0V	0.36 ~ 1.1V
Array size (Kb)	54	8	64	256	60.75
weight sets	2	1	16	32	18
Macro area (mm^2)	0.0172	0.0323	0.0455	0.526	0.0157
Simultaneous MAC + Write	Yes	No	Yes	Yes	Yes
Mb/mm^2	3.07	0.24	1.37	0.48	3.78
Input Ch	64	64	64	32	72
Output Ch	64	16	16	64	4
Format	INT12	INT4	INT4	INT4	INT12
TOPS/W (*) (weight=1 50%)	319 (input=1 10%) 199 (input=1 25%) @0.55V	121 (input=1 10%) @0.72V	137 @0.55V	58 (input=1 25%) @0.525V	484 (input 10% toggle) 293 (input 25% toggle) @0.55V
TOPS/mm^2 (*)	299.7 @0.9V	41.6 @0.72V	11.3 @0.99V	13.6 @1.0V	495.3 @0.9V

(*) Normalized to INT4

Figure 34.4.6: PPA summary and comparison to prior work.



Technology	3nm HKMG FinFET
Bitcell	6T cell (0.026 µm^2)
Macro area	0.0157 mm^2
Memory config.	60.75 Kb (1296 words x 48 IOs w/ MUX4)
MAC size	72 Input Ch x 4 Output Ch x 18 weight set
Bit density (Mb/mm^2)	3.78
TOPS/mm^2	10.6 @0.4V ~ 55.0 @0.9V
TOPS/W @0.55V	Input toggle rate = 10% - 53.8 w/o row switch - 30.3 w/ row switch every 4cycles Input toggle rate = 25% - 32.5 w/o row switch - 22.9 w/ row switch every 4cycles

Figure 34.4.7: Micrograph of test chip and design summary table.

979-8-3503-0620-0/24/$31.00 ©2024 IEEE