11.6 A 5-nm 254-TOPS/W 221-TOPS/mm² Fully-Digital Computing-in-Memory Macro Supporting Wide-Range Dynamic-Voltage-Frequency Scaling and Simultaneous MAC and Write Operations

Hidehiro Fujiwara[1], Haruki Mori[1], Wei-Chang Zhao[1], Mei-Chen Chuang[1], Rawan Naous[2], Chao-Kai Chuang[1], Takeshi Hashizume[3], Dar Sun[1], Chia-Fu Lee[1], Kerem Akarvardar[2], Saman Adham[4], Tan-Li Chou[1], Mahmut Ersin Sinangil[2], 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, Yokohama, Japan
[4]TSMC, Austin, TX

2022 IEEE International Solid-State Circuits Conference (ISSCC) | 978-1-6654-2800-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISSCC42614.2022.9731754

Computing-in-memory (CIM) is being widely explored to minimize power consumption in data movement and multiply-and-accumulate (MAC) for edge-AI devices. Although most prior work focuses on analog-based CIM (ACIM) to leverage the BL charge/discharge operation, the lack of accuracy caused by transistor variation and the ADC is an issue [1-3]. In contrast, a digital-based CIM (DCIM) approach realizes enough accuracy and flexibility for various input and weight bit widths [4, while also benefiting from technology scaling. This paper proposes a 64kb DCIM macro using a one-read and one-write (1R1W) 12T bitcell. The DCIM macro can realize simultaneous MAC + write operations and wide range dynamic voltage-frequency scaling (DVFS) due to the 12T cell's 1R1W functionality and low-voltage operation. Further improvements in power-performance-area (PPA) are obtained by optimizing the circuit architecture and layout topology.

Figure 11.6.1 summarizes DCIM advantages when technology scaling is considered. Digital circuit design, with technology scaling from 7 to 5nm, achieves 1.5-2× logic density, ~1.2× power efficiency and performance improvement using automatic place and route (APR). In contrast, there are a lot of challenges for ACIM designs in the advanced technology nodes: attention must be paid to the ADC's dynamic range due to voltage scaling, a more non-linear drain current (I_D) due to drain-induced-barrier-lowering (DIBL) for current-based ACIM [1], and smaller capacitances due to the technology scaling for capacitance-based ACIM [2, 3]. Before starting CIM macro design, a comparison was made between DCIM and a charge-injection type ACIM in the 5nm technology node: the DCIM achieved a better PPA than the ACIM in the design benchmark result. The DCIM realizes >1.3× better power efficiency and >1.4× better performance/area efficiency than ACIM. Furthermore, DCIM realizes no accuracy loss typically due to truncation or gain errors in the ADC, while also achieving flexibility in input and weight bit width. DCIM also has the advantage with respect to design for testability (DFT).

Figure 11.6.2 shows the overall architecture of DCIM in the 5nm technology node. The DCIM macro includes 64 MAC arrays with 64kb of 12T bitcells for signed or unsigned weight storage. XIN drivers (XINLB drivers) for the 64 4b signed or unsigned input (XIN) and read and write address decoders are shared by the 64 MAC arrays. A 4:1 multiplexer in the XINLB driver outputs 1b of XIN per cycle for bitwise multiplication. Each MAC array comprises of 1kb SRAM bitcells, a bitwise multiplier and adder tree, and a bit-shifter and final accumulator. To support simultaneous MAC operation with 64 inputs, the 1kb SRAM bitcells are separated into 64 banks: each bank includes four rows and four columns. Every bitwise multiplier (NOR) receives a signal from read BL (RBL) and 1b input information per cycle. The adder tree accumulates the results of NOR (XINLB, RBLB) as a partial sum result. The bit-shifter compensates for the bit significance of XIN and the final MAC result can be obtained as NOUT from the final accumulator through FFs. The bit-shifters and accumulators can optionally use a 2's compliment of the partial sum result to handle signed input calculation. The DCIM macro also supports read operations: read out data from the bitcell array go through the adder tree, that is the hierarchical readout circuitry in a regular 1R1W register file is replaced by MAC circuitry.

Figure 11.6.3 shows the layout floorplan and circuit optimizations for the DCIM macro. The macro, including 12T bitcell array and MAC circuitry, is drawn using standard-cell design rules. Each four banks of the bitcell array are combined with the first and second stages of the adder tree. Reversed-polarity full-adder (FA) cells are introduced in the adder tree and achieves a 12.5% smaller area and a 15% total MAC power reduction compared to a FA cell from the standard-cell library. The 12T bitcell's size, based on standard-cell design rules, is 0.075 μm². Since empty spaces and dummy regions are not required at the transition region between logic and bitcell array, the overall area can be smaller than when using foundry 6T or 8T SRAM bitcells [5]. The location of the cell boundary in the bitcell array region is the source and drain, unlike standard cell placement, to minimize the macro area. The interleaved-write-WL (WWL) scheme mitigates routing congestion in the adder tree (*x*-direction) while the read WL (RWL) is shared between two banks to avoid routing congestion in *y*-direction.

Figure 11.6.4 shows the MAC operation timing diagram. The 12T bitcell, which supports asynchronous 1R1W operation using independent clocks (CLKR for read and MAC and CLKW for write), allows for concurrent MAC operations and weight updates; so long as the access location for MAC operations and the address for write operations are different. The cycle time of MAC operations takes longer than the write access, as such write operation can be sped up if a shorter clock period is used for CLKW. The bitwise operation for XIN causes the MAC operation to require four CLKR cycles; NOUT comes out one clock cycle later. Therefore, to get the MAC result for 64 4b inputs and 4b weights, a total of four and five CLKR cycles are needed for the MAC operation and output latency, respectively. Because the CMOS-type read port (inverter and transmission gate) in the 12T bitcell does not require RBL precharge before/after the read operation, the RWLs maintain the same state during the same weight data are reused. This reduces the power consumption in the array because no RBL charge or discharge occurs. Although the MAC operation is based on a 4b input and a 4b weight, the DCIM macro can support longer bit width for both input and weight: to store a signed 8b weight, the weight is stored using a 4b signed format in the array2 and using a 4b unsigned format array1. Appling an 8b signed input requires two operations: a 4b signed XIN for the first MAC operation and a 4b unsigned XIN for the second MAC operation. This requires additional control logic outside of the DCIM macro, but then one DCIM macro can support various input and weight bit widths using a signed or unsigned format.

Figure 11.6.7 shows a micrograph of the test chip fabricated in a 5nm high-κ metal gate (HKMG) FinFET technology node: eight DCIM macros are implemented on a chip. The area of DCIM macro is 0.0133 mm² (109 × 122μm²). The macro area has been confirmed to be 3× smaller than using a similar architecture implemented using a digital flow: RTL → synthesis → APR. Figure 11.6.5 summarizes the chip measurement results. Since the digital-based 12T bitcell is used the minimum operating voltage (V_{MIN}) can be the same as the standard cell logic; forgoing SRAM dual-power rails or its read/write assist techniques. V_{MIN} is measured to be below 0.5V at –40°C with a 95% yield. The maximum operating frequency (f_{MAX}) is 0.36, 0.96 and 1.44GHz at 0.5, 0.7 and 0.9V; which correlates well with simulation results. Dynamic current is measured with a weight bit sparsity is 50%, input bit sparsity is 10%, 25% and 50%. As shown by the dynamic current graph, when the input sparsity is low the dynamic power is small. For an input bit sparsity of 10%, the dynamic current is 16.2, 23.4 and 32.5μA/MHz at 0.5, 0.7 and 0.9V. The PPA is summarized as a TOPS/W vs TOPS/mm² chart: 221.2 and 55.3TOPS/mm² are measured at 0.9 and 0.5V, using a 4b input and a 4b weight. Note that by reducing the memory capacity for weight storage (i.e. the number of rows) increases TOPS/mm². Efficiencies of 253.5, 205.0 and 155.2TOPS/W are achieved at 0.5V with 10, 25 and 50% input bit sparsity, and 41.9, 55.8 and 69.9TOPS/W at 0.9V. Figure 11.6.6 shows a comparison summary to prior work. The 1R1W 12T bitcell and digital based design allows the DCIM macro to support a wide range DVFS without any accuracy loss due to truncation or gain loss in the ADC, while also supporting simultaneous MAC and weight update operations.

References:

[1] Q. Dong et al., "A 351TOPS/W and 372.4GOPS Compute-in-Memory SRAM Macro in 7nm FinFET CMOS for Machine-Learning Applications," *ISSCC*, pp. 242-243, 2020.
[2] X. Si et al., "A 28nm 64Kb 6T SRAM Computing-in-Memory Macro with 8b MAC Operation for AI Edge Chips," *ISSCC*, pp. 246-247, 2020.
[3] S. Yin et al., "PIMCA: A 3.4-Mb Programmable In-Memory Computing Accelerator in 28nm for On-Chip DNN Inference," *IEEE VLSI*, 2021.
[4] Y.-D. Chih et al., "An 89TOPS/W and 16.3TOPS/mm2 All-Digital SRAM-Based Full-Precision Compute-In Memory Macro in 22nm for Machine-Learning Edge Applications," *ISSCC*, pp. 252-253, 2021.
[5] H. Fujiwara et al., "A 5nm 5.7GHz@1.0V and 1.3GHz@0.5V 4kb Standard-Cell- Based Two-Port Register File with a 16T Bitcell with No Half-Selection Issue," *ISSCC*, pp. 340-341, 2021.
[6] J.-S. Park et al., "A 6K-MAC Feature-Map-Sparsity-Aware Neural Processing Unit in 5nm Flagship Mobile SoC," *ISSCC*, pp. 152-153, 2021.

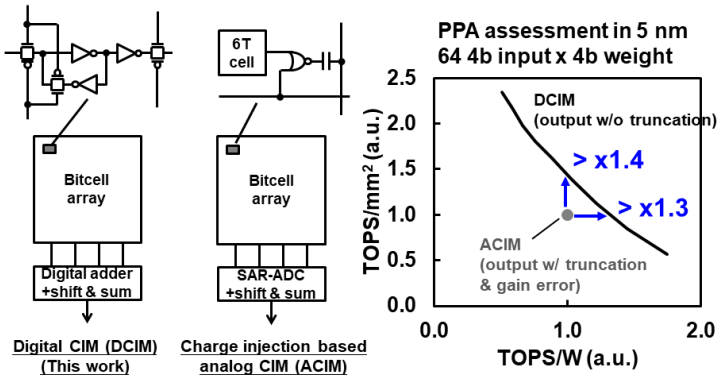


Figure 11.6.1: PPA comparison between digital CIM and analog CIM in a 5nm technology node.

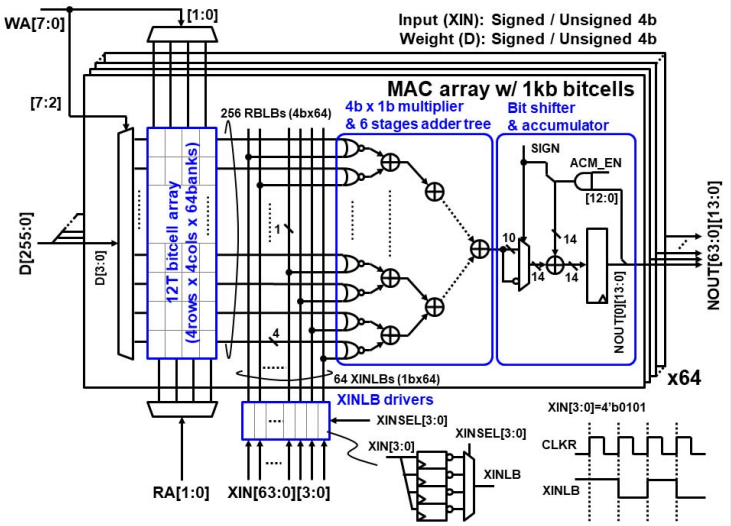


Figure 11.6.2: Overall architecture of DCIM macro.

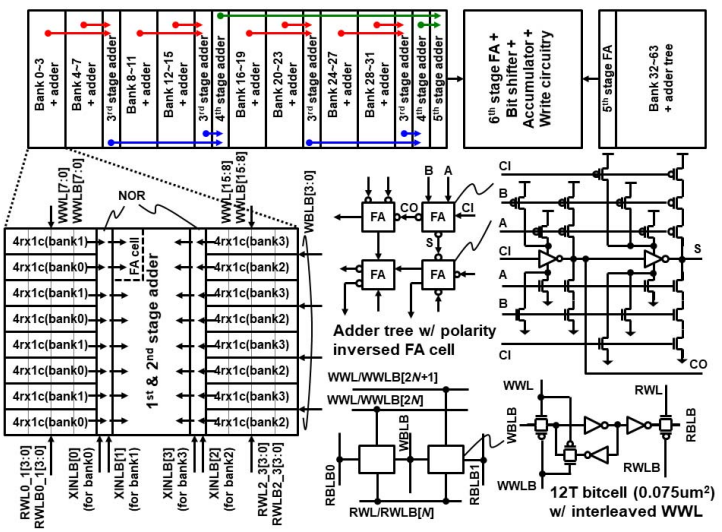


Figure 11.6.3: MAC array floor plan with polarity inversed FA cell and 12T bitcell.

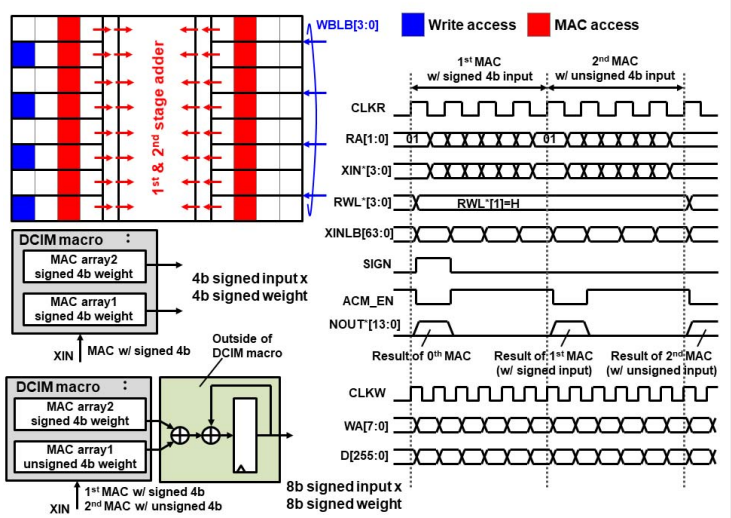


Figure 11.6.4: Timing diagram with simultaneous MAC and write accesses.

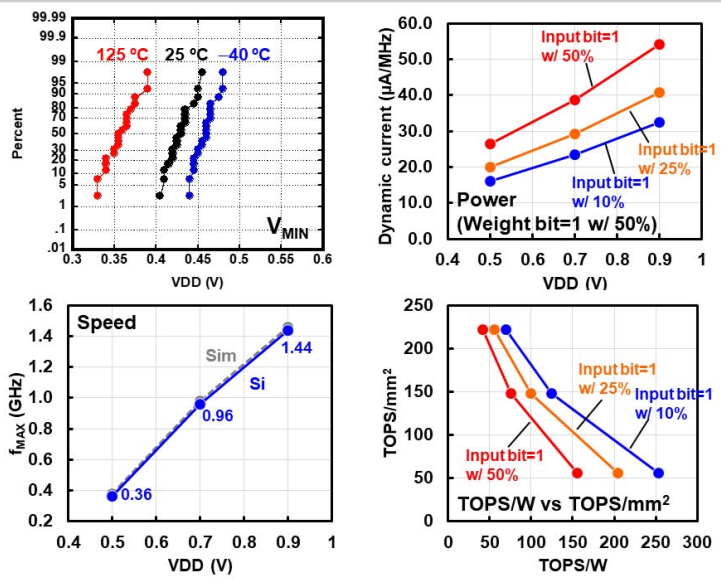
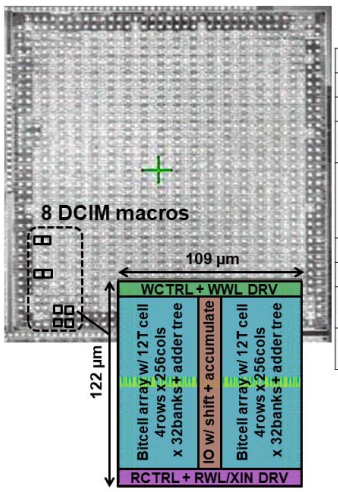


Figure 11.6.5: Measurement results: V_{MIN}, f_{MAX}, dynamic power and TOPS/mm² vs TOPS/W.

	ISSCC'21 [6]	ISSCC'20 [1]	ISSCC'20 [2]	VLSIC'21 [3]	ISSCC'21 [4]	This work
Technology	5 nm	7 nm	28 nm	28 nm	22 nm	5 nm
MAC operation	MAC array in NPU core	Analog CIM (current base)	Analog CIM (charge share)	Analog CIM (charge inject)	Digital CIM	Digital CIM
Array size	2 kb	4 kb	64 kb	32 kb	64 kb	64 kb
Bitcell type	N/A	1R1W 8T	1RW 6T	10T1C	1RW 6T	1R1W 12T
Bitcell area	N/A	0.053 μm²	0.25 μm²	N/A	0.379 μm²	0.075 μm²
Macro area	0.022 mm²	0.0032 mm²	N/A	N/A	0.202 mm²	0.0133 mm²
Voltage	0.55 V ~ 0.9 V	0.8 V	0.7 V ~ 0.9 V	1 V	0.72 V	0.5 V ~ 0.9 V
# of Input Ch	16	64	16	256	256	64
# of Output Ch	16	16	16	128	64	64
Input bit	8/16	4	4~8	1/2	1~8	4
Weight bit	8/16	4	4/8	1/2	4	4
Output bit	N/A	4 (truncation)	12	4 (truncation)	16	14
Simultaneous MAC + Write	No	No	No	No	No	Yes
TOPS/mm²	28 (8b)	116	N/A	N/A	16	221 (4b) 55 (8b*)
TOPS/W	13.6 (system level)	262~610	68.4	588	89	254 (4b) 63 (8b*)

(*) Estimated value

Figure 11.6.6: PPA summary table and comparison with prior works.



Technology	5 nm HKMG FinFET
Bitcell	Low voltage 12T (0.075µm^2)
Macro size	109 µm x 122 µm
Memory config.	64 kb 256 addresses x 256 IOs w/ MUX1
MAC size	1024 MAC / CLK cycle 64 4b signed/unsigned Input 4b signed/unsigned weight w/ 4rows 64 14b singed/unsigned Output
TOPS	0.74 @0.5V ~ 2.95 @0.9V
TOPS/mm^2	55.3 @0.5V ~ 221.2 @0.9V
Energy (µW/MHz)	16.2 ~ 32.5 (Input bit=1 w/ 10%) 20.0 ~ 40.7 (Input bit=1 w/ 25%)
TOPS/W	69.9 ~ 253.5 (Input bit=1 w/ 10%) 55.8 ~ 205.0 (Input bit=1 w/ 25%)

Figure 11.6.7: Test chip micrograph.

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