A FeFET-based Digital Compute-in-Memory Macro Design

Matthew Chen[1], Shimeng Yu[1*]

[1]*Georgia Institute of Technology, 791 Atlantic Drive, Atlanta, Georgia, USA*

For AI/ML-related computations, energy consumption is often dominated by repetitive memory access. Traditional von Neumann architectures may struggle to meet these energy demands – this has led to rising interest in compute-in-memory (CIM), a computing paradigm with the potential for higher energy efficiency [1]. Although traditionally done in the analog domain for maximal power savings, analog CIM can suffer from reduced accuracy due to low signal-to-noise ratio. All-digital CIM has been proposed – the energy savings may be lower compared to analog CIM, but there is also no accuracy loss and better power-performance-area (PPA) scaling at modern tech nodes [2]. This work demonstrates a digital compute-in-memory (DCIM) macro utilizing a ferroelectric-transistor (FeFET)-based 2F bitcell for weight storage. The proposed bitcell utilizes the voltage divider effect to store single-bit data with a direct voltage output, non-destructive read, and non-volatility. A program and erase scheme is also proposed. Apart from the bitcell, the rest of the design features a lookup-table (LUT)-based multiplier and bit-parallel operation to further improve computing efficiency [3][4]. When simulated against a comparable 28nm SRAM-based DCIM macro with the same synthesized INT4 multiplier/adder tree, the 28nm FeFET-based DCIM macro demonstrates 0.6% lower compute energy, while the bitcell itself consumes 39% less static power vs. a non-push rule 28nm 6T SRAM cell. Compared to other DCIM implementations, it also has competitive PPA of 110.6 TOPS/W and 0.65 TOPS/mm^2, supporting the 2F cell's potential where low standby power and non-volatility for instant on/off operations are attractive features for edge AI applications.

References

[1] S. Yu, H. Jiang, S. Huang, X. Peng and A. Lu, "Compute-in-Memory Chips for Deep Learning: Recent Trends and Prospects," in *IEEE Circuits and Systems Magazine*, vol. 21, no. 3, pp. 31-56, thirdquarter 2021, doi: 10.1109/MCAS.2021.3092533.

[2] Y.-D. Chih *et al.*, "16.4 An 89TOPS/W and 16.3TOPS/mm2 All-Digital SRAM-Based Full-Precision Compute-In Memory Macro in 22nm for Machine-Learning Edge Applications," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2021, pp. 252–254. doi: 10.1109/ISSCC42613.2021.9365766.

[3] C.-F. Lee *et al.*, "A 12nm 121-TOPS/W 41.6-TOPS/mm2 All Digital Full Precision SRAM-based Compute-in-Memory with Configurable Bit-width For AI Edge Applications," in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, Jun. 2022, pp. 24–25. doi: 10.1109/VLSITechnologyandCir46769.2022.9830438.

[4] H. Fujiwara *et al.*, "34.4 A 3nm, 32.5TOPS/W, 55.0TOPS/mm2 and 3.78Mb/mm2 Fully-Digital Compute-in-Memory Macro Supporting INT12 × INT12 with a Parallel-MAC Architecture and Foundry 6T-SRAM Bit Cell," in *2024 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2024, pp. 572–574. doi: 10.1109/ISSCC49657.2024.10454556.

* Corresponding author email: shimeng.yu@ece.gatech.edu

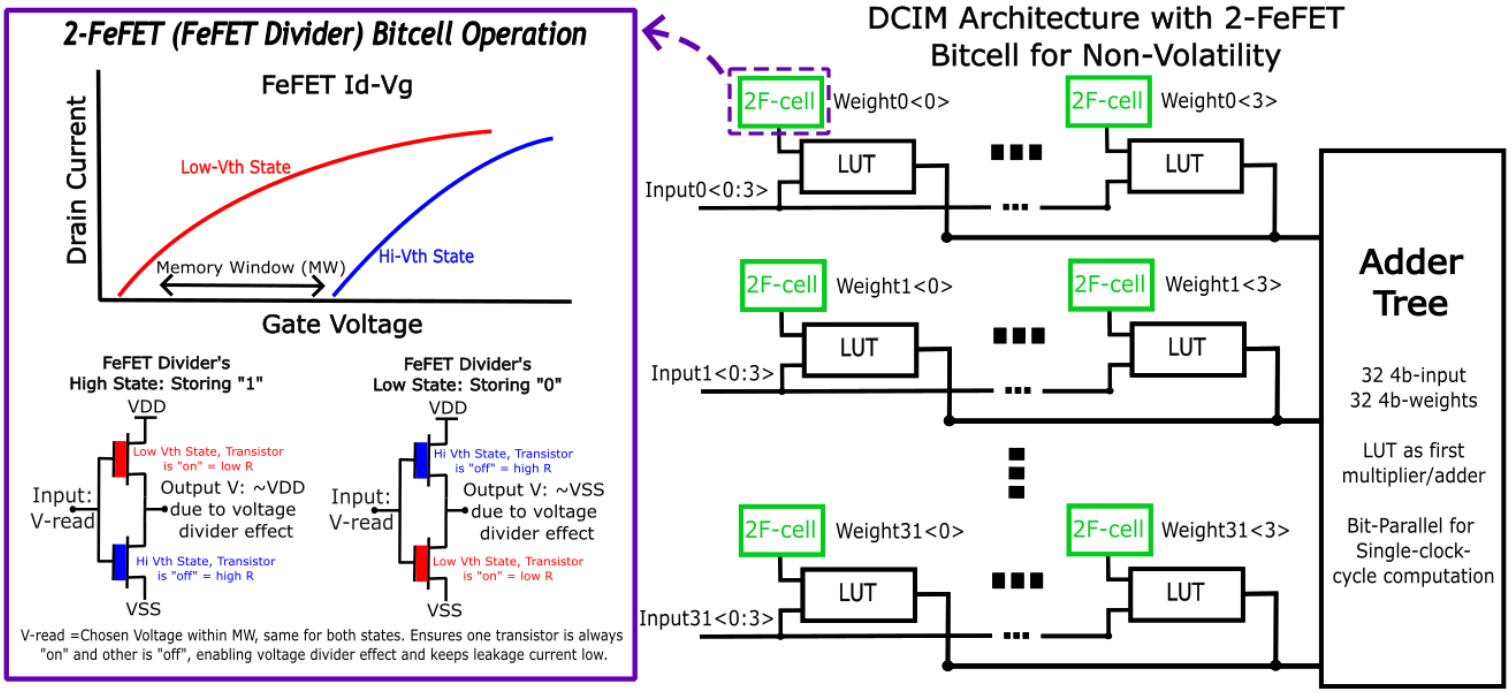
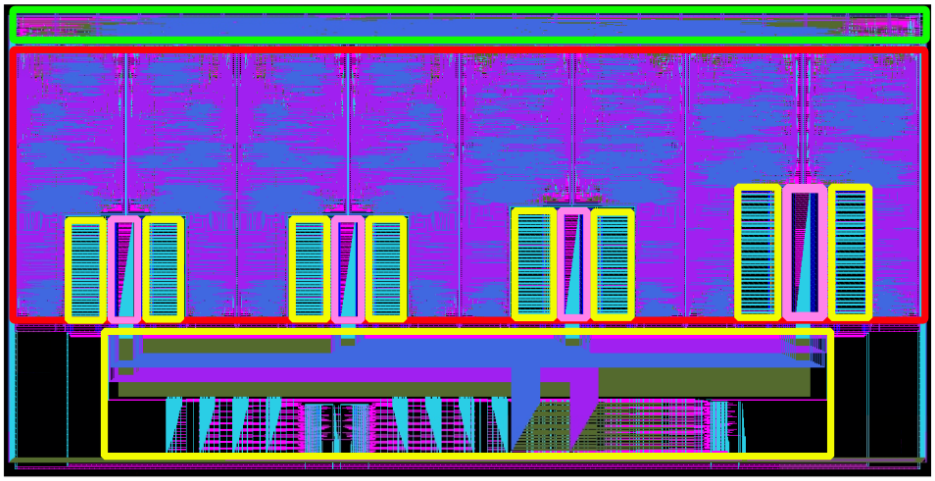


Figure 1: Architecture of the DCIM macro and 2-FeFET bitcell operation.

DCIM Macro Floorplan and PPA Table



	ISSCC '21 [2]	VLSI '22 [3]	ISSCC '24 [4]	**FeFET DCIM**	Comparable SRAM DCIM
Technology	22nm	12nm	3nm	28nm	28nm
Bitcell	6T SRAM	SRAM	6T SRAM	2F Voltage Divider	6T SRAM
Power Supply (V)	0.72	0.72	0.36-1V	0.72-1V	0.72-1V
Array size	64Kb	8Kb	60.75Kb	4Kb	4Kb
Bitcell Area (μm^2)	0.379	N/A	N/A	0.517	0.575
Macro Area (mm^2)	0.202	0.0323	0.0157	0.158	0.119
Input Channels	256	64	72	256	256
Output Channels	64	16	4	32	32
TOPS/W (Weight = 1 50%, INT4)	89	121 (input = 1 10%)	484 (input 10% toggle) @0.55V	110.6 (input 10% toggle) @0.8V	109.9 (input 10% toggle) @0.8V
TOPS/mm^2	16.3	41.6 @ 0.72V	495.3 @0.9V	0.65 @0.8V	0.86 @0.8V

Figure 2: Macro floorplan and PPA comparison.