MIND-MAC: Multi-Level In-memory Quasi Non-Destructive MAC Operation in Compact 2T-nC FeRAM for Efficient DNN Accelerator

Rudra Biswas[1*], Sadik Yasir Tauki[1], Varun Darshana Parekh[1], Prapti Panigrahi[1], Jiahui Duan[2], Shan Deng[2], Kai Ni[2], Vijaykrishnan Narayanan[1]

[1]*The Pennsylvania State University, University Park, PA 16802, USA*
[2]*University of Notre Dame, Holy Cross Dr, Notre Dame, IN 46556, USA*

As DNN models become deeper and more complex, the growing gap between processor speed and external memory bandwidth dominates both energy use and latency. We propose MIND-MAC, a novel multiply-accumulate (MAC) architecture using a quasi-nondestructive read-out (QNRO) enabled 2T-nC ferroelectric memory cell [1, 2]. By integrating multi-level Metal-Ferroelectric-Metal (MFM) capacitors with drain-driven readout in a low energy, compact, stackable structure, our design enables parallel in-memory MAC computation. We employ remanent polarization to encode multiple discrete weight levels (Fig. 1(b)) and apply input vectors as drain voltages on read transistors that produce output currents that inherently represent the MAC result (Fig. 1(c-d)).

Circuit-level functionality was evaluated using SPICE simulations incorporating parasitics (C_{FE}, C_X, C_Z) extracted via TCAD modeling of vertically stacked 2T-4C arrays (Fig. 1e–f). Simulations demonstrated reliable 2-bit weight programming and successful analog MAC operation using 1-bit drain inputs (Fig. 2a). Importantly, the read bit line (RBL) current showed distinct levels corresponding to the programmed polarization states (Fig. 2e), validating the system's multi-level sensing capability and enabling current-based accumulation for in-memory MAC. The 2T-nC FeRAM architecture was experimentally validated using a fabricated MFM capacitor integrated into a 2T-3C cell (Fig. 2b). Measured P–V and I–V characteristics of the MFM (Fig. 2c–d) demonstrated stable remanent polarization and distinct switching behavior. Readout measurements revealed four clearly separated RBL current levels corresponding to different programmed states (Fig. 2e), confirming reliable multi-level sensing.

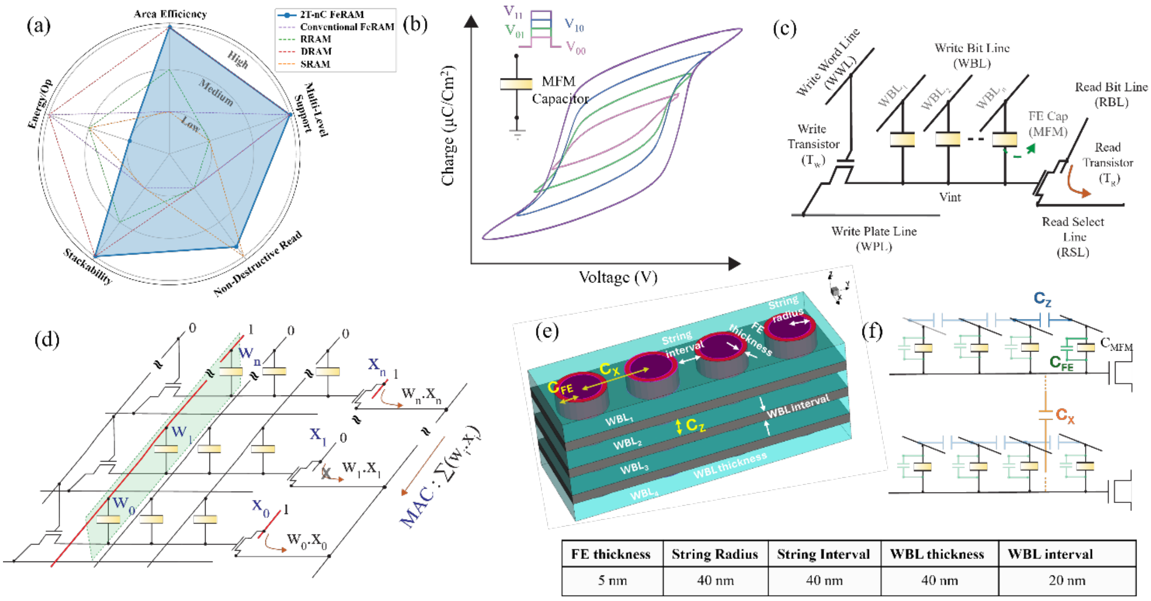
Building on these promising initial results, we will carry out (i) thermal analysis using a HotSpot model (ii) workload mapping of representative DNN inference tasks onto the proposed memory array, to evaluate energy per MAC operation, total energy per inference, and end to end latency. These results will be compared with CiM DRAM, SRAM, and RRAM baselines (Fig. 1(a)) to quantify the benefits of our approach.

References

[1] S. Deng *et al.*, "First Demonstration of Vertical 2T-nC FeRAM Hybrid Cell and its Scalability for High-Density 3D Ferroelectric Capacitor Memory," *2024 IEEE International Electron Devices Meeting (IEDM)*

[2] Y. Xiao *et al.*, "Demonstration of Vertical 2T-nC FeRAM Hybrid Cell and Its Scalability for High-Density 3-D Ferroelectric Capacitor Memory," in *IEEE Transactions on Electron Devices*,

* Corresponding author: email: rfb5659@psu.edu



FE thickness	String Radius	String Interval	WBL thickness	WBL interval
5 nm	40 nm	40 nm	40 nm	20 nm

Figure 1: (a) Comparative overview of 2T-nC FeRAM with conventional compute-in-memory (CiM) technologies, highlighting key advantages. (b) Conceptual representation of a Metal-Ferroelectric-Metal (MFM) capacitor illustrating multi-level states enabled by remanent polarization under varying input biases. (c) Schematic of a 2T-nC FeRAM unit cell, where shared MFM capacitors enable multi-level weight storage. (d) Illustration of Multiply-Accumulate (MAC) operation utilizing QNRO-enabled 2T-nC architecture for in-memory computation. (e) TCAD modeling of a vertically stacked 2T-nC FeRAM multi-string array used for parasitic extraction of C_{FE}, C_X, and C_Z, (f) SPICE-compatible schematic of the 2T-nC FeRAM cell incorporating extracted parasitics for circuit-level evaluation.

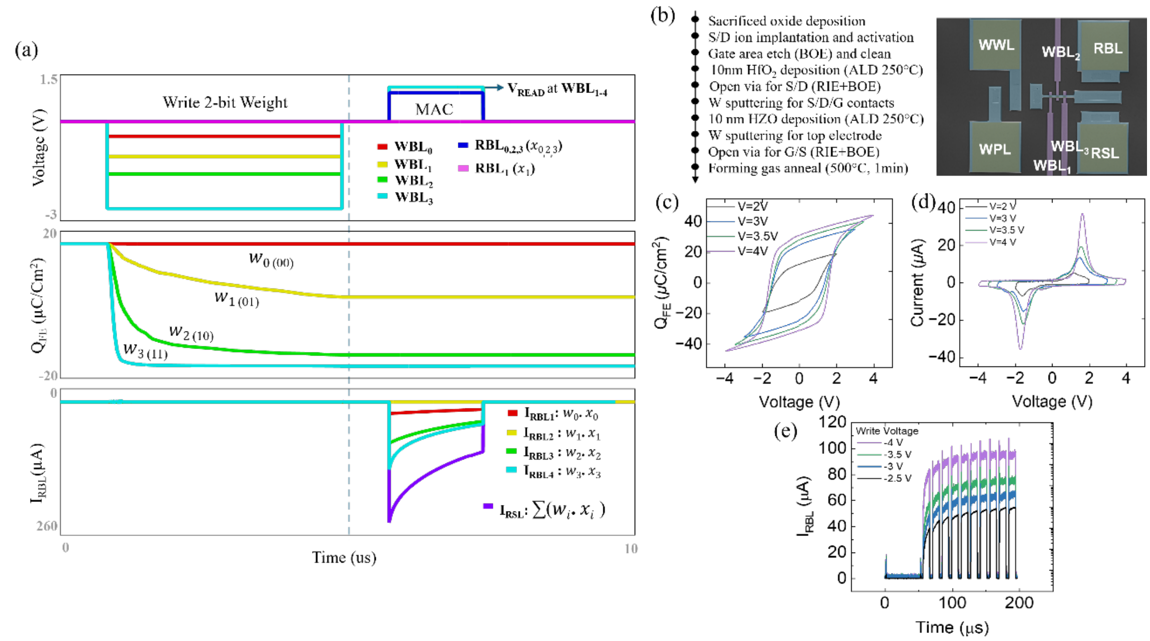


Figure 2: (a) SPICE simulation results demonstrating 2-bit weight ($w_{00\text{-}11}$) programming in MFM capacitors of 2T-nC FeRAM cells, followed by a read pulse and 1-bit input ($x_{0\text{-}1}$) on the read bit line (RBL) to enable MAC operation. (b) Fabrication process flow and SEM image of a 2T-nC FeRAM cell. (c) Measured polarization-voltage (P-V) characteristics of a fabricated MFM capacitor across different voltage sweep conditions, and (d) Corresponding current-voltage response of the same MFM device. (e) Measured RBL current showing distinct levels for four programmed weight states, validating multi-level sensing capability.