Nanoelectronic Artificial Intelligence Processors

CNF Project Number: 316924
Principal Investigator(s): Peter McMahon
User(s): Guilherme Marega, Ruomin Zhu, Yongqi Zhang

Affiliation(s): Applied & Engineering Physics department, Cornell University

Primary Source(s) of Research Funding: the David & Lucile Packard Foundation, Cornell AEP

Contact: pmcmahon@cornell.edu

Website: https://mcmahon.aep.cornell.edu/index.html

Primary CNF Tools Used: Westbond 7400A Ultrasonic Wire Bonder, KLA P7 Profilometer, YES EcoClean Asher, Oxford 81 RIE, Suss MA6|BA6 Aligner, Hamatech Hot Piranha, Oxford PECVD, Oxford 100 ICP Dielectric, Woollam RC2 Spectroscopic Ellipsometer, Zeiss Ultra SEM, Oxford 82 RIE

Abstract:

Resistive crossbar arrays co-locate memory and analog computation to overcome the von Neumann "memory wall," where data movement can dominate energy costs. By encoding weights in multi-level resistive cells and performing parallel dot-product operations directly within a 16 × 16 array fabricated at the Cornell Nanoscale Facility, we demonstrate up to 60.9 TOPS/W and four decades of linear dynamic range [1]. This in-memory computing platform offers a compelling path toward dramatically lower-power AI inference, as well as efficient signal and image processing.

Summary of Research:

The explosive growth of AI services is driving data-center electricity demand toward unsustainable levels. The International Energy Agency projects global data-center energy use to exceed 945 TWh by 2030—more than double current levels—and U.S. data centers, already consuming 4.4 % of national electricity in 2023, may account for up to 12 % by 2028 [4][5]. Even a single Artificial Intelligence query carries a measurable footprint: a typical ChatGPT interaction consumes roughly 0.3 Wh, equivalent to running an LED bulb for several minutes [6].

Conventional von Neumann architectures exacerbate this burden via the "memory wall," in which moving a 64-bit word from Dynamic Random Access Memory (DRAM) to the Central processing unit (CPU) costs on the order of 1,000 pJ—about 50× the energy of a 64-bit floating-point add [7]. Across real workloads, data transfers can account for 60–70 % of total system energy, severely limiting both performance and efficiency.

Resistive crossbar arrays address this challenge by performing matrix-vector multiplications in situ: voltages applied to row lines induce column currents proportional to conductance-encoded weights, realizing massively parallel dot products in one step. Hardware demonstrations include 60.9 TOPS/W for binary neural inference in oxide-based devices [1], 405 TOPS/W in magnetoresistive prototypes at 0.8 V [2], and 3.6 TOPS/W in designs with nonlinear Analog Digital Converters (ADCs) for specialized preprocessing [3].

Since the first memristor crossbar proposal in 2008, the field has advanced rapidly: writes as low as 6 fJ per cell for sparse coding [8], 24 TOPS/W in XNOR-RRAM arrays monolithically integrated with 90 nm CMOS [9], and area efficiencies exceeding 130 TOPS/mm² alongside the aforementioned TOPS/W milestones [2].

In our work, we aim to fabricate resistive crossbar arrays with AI model encoded on them to save energy for AI computing. As a first step, we have fabricated 16 × 16 arrays using a CMOS-compatible process and measured their analog performance at Cornell Nanoscale Facility. By encoding image-processing kernels as conductance matrices and feeding input voltages corresponding to grayscale images, we obtained output currents that reproduce digital convolution outputs with high fidelity (Fig. 4), validating hardware-in-the-loop processing as a viable digital alternative.

Conclusions and Future Steps:

We have shown that resistor-based crossbar arrays can achieve state-of-the-art energy efficiency for low-precision AI and image-processing tasks. To scale this approach, we plan to:

- Increase array dimensions from 16 × 16 to 128 × 128, enabling higher-resolution kernels and larger neural-network layers.

- Expand application domains to voice-signal processing, discrete Fourier transforms and AI computing, leveraging the same in-memory dot-product primitive.

These developments will bring in-memory computing closer to deployment in edge AI accelerators and high-

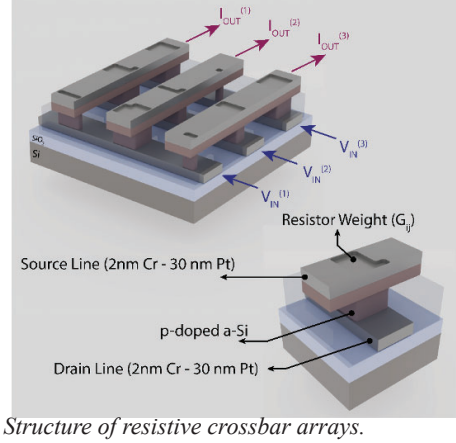

Figure 1: Structure of resistive crossbar arrays.

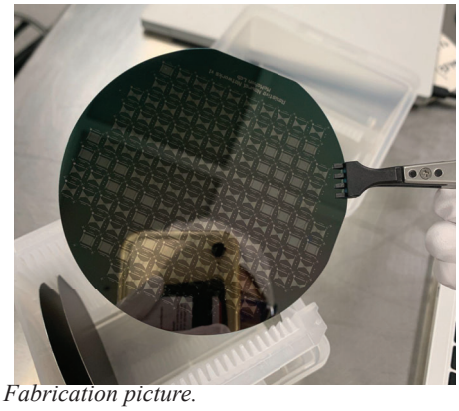

Figure 2: Fabrication picture.

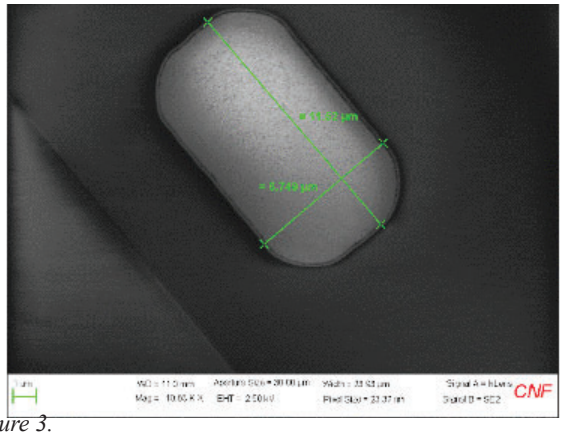

Figure 3.

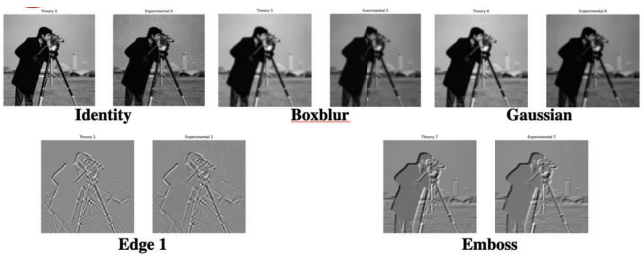

Figure 4: Digital and Hardware processing results.

throughput signal-processing hardware.

References:

[1] Li H, Wang S, Zhang X, Wang W, Yang R, Sun Z, Feng W, Lin P, Wang Z, Sun L, Yao Y. Memristive crossbar arrays for storage and computing applications. Advanced Intelligent Systems. 2021;3(9):n/a. doi:10.1002/aisy.202100017.

[2] Jung S, Lee H, Myung S, Kim H, Yoon S-K, Kwon S-W, Ju Y, Kim M, Yi W, Han S, Kwon B, Seo B, Lee K, Koh G-H, Lee K, Song Y, Choi C, Ham D, Kim S-J. A crossbar array of magnetoresistive memory devices for in-memory computing. Nature. 2022;601:211–216. doi:10.1038/s41586-021-04196-6.

[3] Yang J, Mao R, Jiang M, Cheng Y, Sun P-S V, Dong S, Pedretti G, Sheng X, Ignowski J, Li H, et al. Efficient nonlinear function approximation in analog resistive crossbars for recurrent neural networks. Nat Commun. 2025;16:1136. doi:10.1038/s41467-025-56254-6.

[4] International Energy Agency. AI is set to drive surging electricity demand from data centres while offering the potential to transform how the energy sector works. IEA; 10 April 2025.

[5] Pomerleau S, Luna A. AI Data Centers: Why Are They So Energy Hungry? American Action Forum. July 15, 2025.

[6] You J. How much energy does ChatGPT use? Gradient Updates, Epoch AI; 07 Feb 2025.

[7] Leland RL, Murphy R, Hendrickson B. Large-Scale Data Analytics and Its Relationship to Simulation. Executive Office of the President; January 2014.

[8] Agarwal S, Quach T-T, Parekh O, Hsia AH, DeBenedictis EP, James CD, Marinella MJ, Aimone JB. Energy scaling advantages of resistive memory crossbar based computation and its application to sparse coding. Frontiers in Neuroscience. 2016;9:484. doi:10.3389/fnins.2015.00484.

[9] Yin S, Sun X, Yu S, Seo J-S. High-Throughput In-Memory Computing for Binary Deep Neural Networks with Monolithically Integrated RRAM and 90 nm CMOS. arXiv preprint arXiv:1909.07514; 2019. doi:10.48550/arXiv.1909.07514.