

ReRAM Memristor Crossbars & Analog In-Memory AI

Physical O(1) Matrix-Vector Multiplication: Harnessing Ohm's and Kirchhoff's Laws at 500 TOPS/Watt

EXECUTIVE ABSTRACT: In digital GPUs, 85-90% of energy is lost moving weight matrices across the Von Neumann bus. Resistive RAM (ReRAM) crossbars compute matrix-vector products physically in memory. Guided by Ohm's Law ($I = V \cdot G$) and Kirchhoff's Current Law ($I_{total} = \sum I$), an entire $M \times N$ matrix evaluates in $O(1)$ time in a single clock cycle at 100-500+ TOPS/W. Stacking 1S1R threshold switching selectors overcomes sneak-path leakage, unlocking high-density analog edge AI inference.

Metric	Digital GPU Tensor Core (H100)	Optical Photonic Matrix (MZI)	Analog ReRAM Crossbar (CIM)
Basis	Digital FP8 / FP16 Systolic Array	Mach-Zehnder Interference	Ohm's Law & Kirchhoff's Current Law
Complexity (MVM)	$O(N)$ to $O(N^2)$ Tiled Cycles	$O(1)$ Speed of Light	$O(1)$ Physical Evaluation (1 Cycle)
Energy Efficiency	15 to 30 TOPS / Watt (HBM bound)	200 to 800 TOPS / Watt	100 to 500+ TOPS / Watt (In-Memory)
Accessing Cost	Continuous High-Bandwidth Fetch	Zero (Laser phase modulated)	Zero (Weights stored in conductance)
Area Overhead	Register Files & Cache Hierarchy	Laser Couplers & Modulators	ADCs & DACs (60-70% Die Area)

1. The Von Neumann Wall and the In-Memory Computing Paradigm

Modern artificial intelligence workloads are fundamentally memory-bound: transformer models spend up to 90% of their operational energy moving multi-gigabyte weight matrices from High Bandwidth Memory (HBM) into processing registers. The actual arithmetic floating-point multiply-accumulate (MAC) instructions account for a small fraction of the thermal budget.

Analog Compute-in-Memory (CIM) shatters this barrier by computing directly within storage elements. Non-volatile Resistive RAM (ReRAM) memristors store synaptic weights as analog conductances ($G = 1/R$), eliminating external memory transfers entirely.

2. Physical Evaluation via Fundamental Electromagnetic Laws

In a memristive crossbar array, input activations are converted into analog voltage pulses and driven simultaneously across horizontal wordline rows. At every junction, current conducts through the ReRAM device according to Ohm's Law: $I_{ij} = V_i * G_{ij}$.

Because vertical bitlines are continuous metallic conductors, Kirchhoff's Current Law sums the individual junction currents along each column: $I_{out,j} = \text{Sum}(V_i * G_{ij})$. The dot-product multiplication and accumulation evaluate instantaneously in $O(1)$ time across all matrix dimensions, bounded only by nanosecond RC interconnect delays.

3. Sneak-Path Suppression via 1S1R Selectors and Mixed-Signal Walls

In unselected passive arrays, current bleeds through neighboring low-resistance cells, forming parasitic sneak paths that destroy computation accuracy in arrays exceeding 32x32. Foundries solve this via One-Selector One-Resistor (1S1R) integration, stacking volatile Threshold Switching (TS) amorphous chalcogenide diodes that provide 10^5 On/Off non-linearity.

The remaining commercial hurdle is peripheral mixed-signal overhead: high-speed Analog-to-Digital Converters (ADCs) and DACs consume over 60% of total chip area and power, necessitating multi-bit SAR ADC optimizations.

CRITICAL MICROARCHITECTURAL TAKEAWAYS & PRODUCTION GUIDELINES

- Ohm's Law and Kirchhoff's Current Law collapse matrix multiplication time complexity to $O(1)$ in a single physical clock cycle.
- Analog ReRAM crossbars eliminate weight movement, achieving 100 to 500+ TOPS/Watt energy efficiency.
- 1S1R threshold switching selectors provide 10^5 non-linearity, choking sneak-path leakage in large arrays.
- Quantization-Aware Training (QAT) with noise injection preserves LLM perplexity on 4-bit to 6-bit analog weight cells.