Monolithic 4K Electrochemical RAM-Based Analog AI Chip for Energy-Efficient On-Chip Training

Hyunjeong Kwak

Dept. of Materials Science and Engineering,
Pohang University of Science and Technology, Pohang, 37673, South Korea

Abstract—The rapid advancement of artificial intelligence (AI) has led to an exponential increase in computational demand. However, Moore's Law, which has guided performance improvements in conventional computing for decades, is no longer sufficient to sustain this growth. There is a pressing need for high-performance, energy-efficient AI hardware to overcome the limitations of traditional systems.

Analog AI hardware based on resistive memory arrays has emerged as a promising candidate, offering in-memory vector-matrix multiplication (VMM) with enhanced parallelism and power efficiency. Yet, real-world implementations have faced gaps between theoretical expectations and experimental results due to device nonidealities, variability, and noise.

Electrochemical random-access memory (ECRAM) devices are well suited for analog AI hardware due to their ion-driven, stable conductance modulation and inherently low cycle-to-cycle and device-to-device variation. In this talk, we present a monolithically integrated analog AI chip comprising 4,096 ECRAM devices fabricated atop CMOS peripheral circuits using a back-end-of-line (BEOL)-compatible low-temperature process (< 200°C). This architecture enables scalable and energy-efficient in-memory computing.

The fabricated chip achieves a low total power consumption of 11.08 mW and a peak system-level energy efficiency of 6.17 TOPS/W. Our ECRAM devices demonstrate symmetric and linear conductance updates, robust endurance over 80,000 cycles, and stable selector-free operation across the entire 4K array.

Most notably, we experimentally demonstrate the world's largest in-situ neural network training using a selector-free resistive memory cross-point array. A total of 441 functional devices are used to perform VMM operations and stochastic outer-product updates without the need for external computation. This result demonstrates meaningful progress toward realizing practical neuromorphic computing hardware.

By integrating material engineering, circuit design, and algorithm co-optimization, this work establishes a scalable and energy-efficient hardware platform for analog AI and neuromorphic computing. Furthermore, the monolithic integration scheme is readily extendable to sensor-interfaced systems, paving the way for future on-chip intelligence in neuromorphic sensing applications.