ULTRARAM™: A III–V Non-Volatile Memory Enabling a New Energy–Latency Regime for In-Memory Computing

A. Kumar(1), A. Dasgupta(2), P. D. Hodgson(3), M. Hayne(3)

(1) University of California, Berkeley, USA; (2) IIT Roorkee, India; (3) Lancaster University, United Kingdom

Emerging artificial intelligence (AI) and edge computing workloads are increasingly constrained by memory energy consumption and data movement overheads associated with conventional von Neumann architectures. Existing memory technologies, including DRAM, SRAM, and flash, exhibit intrinsic trade-offs between speed, energy efficiency, and non-volatility, preventing their effective deployment in next-generation in-memory computing systems.

In this work, we present ULTRARAM™, a novel III–V compound semiconductor non-volatile memory device based on InAs/AlSb heterostructures and triple-barrier resonant tunnelling (TBRT). The device employs a quantum well-engineered floating gate architecture that enables ultra-low energy switching while maintaining high endurance (>10^7 cycles) and ultra-long data retention (>1000 years). The TBRT mechanism allows rapid charge transfer at low operating voltages (±2.5 V), achieving switching energies orders of magnitude lower than conventional charge-based memory technologies, positioning ULTRARAM™ within a new energy–latency regime (Fig. 1).

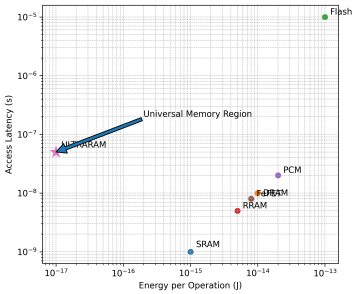

Fig 1: Energy–latency comparison of ULTRARAM™ with conventional and emerging memory technologies, illustrating its position within a new operating regime.

A physics-based compact model captures resonant tunnelling and floating-gate charge evolution, enabling device-to-system evaluation. Using this framework, ULTRARAM™ is assessed within a compute-in-memory architecture for neuromorphic and AI workloads. System-level simulations (VGG-8, CIFAR-10) show comparable accuracy (~91%) to SRAM while improving energy and area efficiency.

Experimentally validated devices show up to 1.8× area and 1.52× energy efficiency improvements over CMOS SRAM, with further gains projected for scaling. At advanced nodes (32 nm), modelling indicates additional improvements in latency, throughput, and energy efficiency relative to SRAM and other emerging memory technologies.

ULTRARAM™ is compatible with heterogeneous integration as a standalone die or chiplet, enabling deployment alongside CMOS logic. These results establish ULTRARAM™ as a strong candidate for next-generation memory architectures, particularly for energy-constrained edge AI, neuromorphic computing, and in-memory processing applications.