INNOVATION PARTNERSHIPS
UNIVERSITY OF MICHIGAN

RecPIM: A PIM–Enabled DRAM–RRAM Hybrid Memory System For Recommendation Models

TECHNOLOGY NUMBER: 2024-344

OVERVIEW

Optimized hybrid memory system for deep learning recommendation models

- Enhances memory performance by in-memory processing and intelligent data mapping
- E-commerce recommendations, social media feeds, and targeted advertisements

BACKGROUND

Deep Learning Recommendation Models (DLRM) have become indispensable in today's data-driven environments, especially for personalizing user experiences in platforms such as e-commerce and social networks. Historically, these models have relied on DRAM for memory needs, but DRAM's performance is constrained by bandwidth limitations, particularly the memory-intensive embedding layers that constitute the majority of DLRM operations. Prior approaches like DRAM-based Near Memory Processing offered incremental improvements but are inadequate for the evolving complexities of DLRM, particularly with their expanding embedding tables. Emerging memory technologies like RRAM promise lower costs and better scaling but require innovation for practical use in such computationally heavy tasks. Therefore, there's a critical need for a hybrid solution that leverages both traditional and novel memory technologies to address these bottlenecks efficiently.

INNOVATION

Researchers at the University of Michigan have developed, RecPIM, that stands at the forefront of memory technology innovation by integrating DRAM and RRAM in its architecture, exploiting RRAM's in-memory processing capabilities. Through Memristor Aided Logic (MAGIC), RecPIM performs complex arithmetic operations directly within memory, significantly enhancing throughput. The system incorporates Access-Pattern-Aware Data Mapping, optimizing DRAM and RRAM data placement according to access frequency, and employs Selective PIM Reduction to strategically deploy in-memory computations only when beneficial. This comprehensive approach results in significant speed improvements, reducing off-chip memory traffic by 49%, and extending the system’s lifetime to over 12 years. Real-world applications abound, from web services deploying recommendation algorithms to data centers optimizing AI workload processing, showcasing the potential of RecPIM to transform data-heavy industries with its resourceful design.

Technology ID

2024-344

Category

Hardware

Engineering & Physical Sciences

Semiconductor, MEMS, and Electronics

Inventor

Nishil Rakeshkumar Talati
Hee Woo Kim
Haojie Ye
Trevor Mudge
Ronald Dreslinski Jr.

Further information

Joohee Kim
jooheek@umich.edu

View online

ADDITIONAL INFORMATION

REFERENCES

H. Kim, H. Ye, T. Mudge, R. Dreslinski and N. Talati, "RecPIM: A PIM-Enabled DRAM-RRAM Hybrid Memory System For Recommendation Models," 2023 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), Vienna, Austria, 2023, pp. 1-6, doi: 10.1109/ISLPED58423.2023.10244420

INTELLECTUAL PROPERTY

Pending