

Processing-in-Storage Architecture for Machine Learning and Bioinformatics

Roman Kaplan, Leonid Yavits and Ran Ginosar

Technion, Israel Institute of Technology

romankap@gmail.com

Extended Abstract

Most contemporary accelerators are von Neumann machines. With the increasing sizes of gathered and then processed data, memory bandwidth is the main limiting of performance. One approach to mitigate the bandwidth constraint is to bring the processing units closer to the data. This approach is known as *near-data processing* (NDP) [1].

The premise of NDP is reducing memory transfer time by cutting the physical distance and increasing the bandwidth between the processing units and memory. However, NDP architectures such as 3D DRAM and CPU stacks, or SSD with embedded CPU, are still inherently limited because they are based on replicating the von Neumann architecture in memory or storage. Since its inception, NDP has mainly meant processing-in-memory (PiM). More recent approach is to exploit processing elements closer to storage, i.e., near-storage processing. However, both PiM and near-storage processing still suffer from the von Neumann bandwidth bottleneck. PiM has a bottleneck between storage and main memory. Near-storage processing suffers from the bottleneck between the storage chips and processing units. Both approaches are inherently limited because they are largely based on the von Neumann architecture model.

This work presents PRINS, a novel PProcessing-in-Storage architecture that employs a Resistive Content Addressable Memory (ReCAM), and applies it to acceleration of bioinformatics and machine learning tasks. PRINS simultaneously functions as a data storage and a massively parallel SIMD accelerator that performs the computations *in-situ*, resulting in increased performance through more complete utilization of the internal storage bandwidth, and reduced energy consumption. The scalability of the system makes it suitable for storing and *in-situ* processing of high volume data intensive applications such as machine learning.

Poster Outline

The first part of the poster presents the bandwidth bottleneck problem of von Neumann architecture (top-left). The second part details the architecture of ReCAM, provides some background on memristors and presents the entire PRINS architecture, composed of multiple ReCAM integrated circuits (top-middle and top-right). Center-left part of the poster describes the Smith-Waterman algorithm, shows how it is mapped to ReCAM in-memory processing and provides performance comparisons of PRINS with other large-scale published solutions. This work was published in IEEE Micro [2]. Center-right describes the K-means algorithm, shows its parallel sections where massive parallelism of ReCAM has advantage over other architectures, demonstrate the K-means mapping to ReCAM and provide performance comparisons with other published solutions. This work was published in IEEE Transactions on Nanotechnology [3]. For brevity and due to its

similarity, only the K-means algorithm is presented. The published work also contains the implementation, evaluation and comparison to other works of the K-Nearest Neighbors (KNN) algorithm.

The last part of the poster (bottom) presents how data deduplication, a method for increasing effective storage size by storing only one copy of each data block. Traditional systems require large compute and memory requirements to manage a key-value data-structure to detect duplicate data blocks. This work presents how deduplication can be performed on ReCAM without requiring to compute a key (hash) for every stored data block. Evaluations show that ReCAM can provide orders of magnitude improvement in bandwidth over a traditional deduplication system, with similar energy requirement. This work was published in the International Symposium on Power And Timing Modeling, Optimization and Simulation [4].

REFERENCES

- [1] R. Balasubramonian, J. Chang, and T. Manning, "Near-data processing: Insights from a MICRO-46 workshop," *IEEE Micro*, 2014.
- [2] R. Kaplan, L. Yavits, R. Ginosar, and U. Weiser, "A Resistive CAM Processing-in-Storage Architecture for DNA Sequence Alignment," *IEEE Micro*, vol. 37, no. 4, pp. 20-28, 2017.
- [3] Kaplan, Roman, Leonid Yavits, and Ran Ginosar. "PRINS: Processing-in-Storage Acceleration of Machine Learning." *IEEE Transactions on Nanotechnology* (2018).
- [4] R. Kaplan, L. Yavits and R. Ginosar, "Deduplication in Resistive Content Addressable Memory Based Solid State Drive," in *PATMOS*, pp. 100-106, 2016.