# PhD Showcase: Accelerator Architectures 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. More recent approach is to exploit processing elements closer to storage, i.e., near-storage processing. However, both NDP and near-storage processing still suffer from the von Neumann bandwidth bottleneck. NDP 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.

My research proposes two main in-storage processing architectures. The following sections shortly describe each architecture and my research contribution .

## 1. PRINS: PRocessing-in-Storage Architecture

The first architecture is called PRINS, a novel PRocessing-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.

The first publication related to PRINS introduced a simple version of ReCAM and focused on *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 [2].

The second publication presented the PRINS architecture, using a modified version of ReCAM. This work focused on the Smith-Waterman algorithm. An in-storage implementation of the algorithm was described, including performance comparisons of PRINS with other large-scale published solutions. This work was published in IEEE Micro [3]. Third publication with the PRINS architecture focused on K-means and K-nearest neighbors algorithm. The work showed how to map the algorithms to ReCAM; presented in implementation of the algorithms in a form of pseudo-code; introduced an addition to ReCAM in order to support several reduction operations required by the algorithms; and provide performance comparisons with other published solutions. This work was published in IEEE Transactions on Nanotechnology [4].

## 2. RASSA: Resistive Approximate Similarity Search Accelerator

Constructing human DNA sequence in real time is paramount to development of precision medicine [5] and on-site pathogen detection of disease outbreaks [6]. Single-molecule, real-time sequencing from Pacific Biosciensces [7] (PacBio) and Oxford Nanopore Technologies [8] (ONT) are new technologies that can produce long reads within minutes, potentially enabling real time genomic analysis. However, long read DNA sequencing poses new challenges. First, long reads contain many thousands of base pairs (bps). Second, long reads tend to exhibit about 15-20% insertion, deletion and substitution errors [7][8].

To construct a complete host sequence, in case a reference sequence exists (from a previously sequenced organism), long reads are mapped to high-similarity locations of the reference sequence. Determining the optimal mapping location of every read onto the reference sequence requires a computationally intensive local alignment procedure (e.g., Smith-Waterman [9]). Its computational time complexity is typically $O(nm)$ for two sequences with lengths $n$ and $m$. Reference sequences vary from several millions to billions of bps. It is therefore computationally prohibitive to perform optimal alignment of every long read with the entire reference sequence.

Read mappers (e.g., minimap [10], minimap2 [11]) find regions of high similarity (overlaps) between reads or between a read and a reference sequence. The overlap locations are then used for correction-free genome assembly or for consensus sequence polishing [12]. Both assembly and polishing require an alignment step to determine the exact edit distance. Once a mapping exists, the alignment can be performed on a specific region of the reference, reducing its duration and resource requirements [12]. Therefore, read mapping can be viewed as a pre-alignment step that reduces the problem size for aligners by narrowing the regions to ones with potentially high-scoring alignment.

Existing pre-alignment hardware solutions [13][14] target short reads (up to several hundreds bps). Such reads contain a small number of errors (less than 5%) and have a different error profile than that of PacBio or ONT long reads [7][8]. High edit distance threshold is required for mapping long but error-prone reads. However, current solutions [13] have high false positive rates when the edit distance is high (i.e., greater than 15). Thus, the current solutions for short reads are not applicable for long reads.

Approximate computing techniques are known to trade accuracy for speed or energy efficiency. In case of long reads, multiple errors are a natural part of the sequencing output. Therefore, long read DNA mapping inherently tolerates the imprecision.

With the end of Dennard scaling and the slowdown of Moore's law, novel hardware solutions for data intensive problems are researched. Emerging technologies such as resistive memories enable new architectures with better performance and energy efficiency. Resistive approximate Hamming distance solutions exist [15]. However, these do not provide the parallelism required to support a high throughput applications such as DNA read mapping.

In this work, a Resistive Approximate Similarity Search Accelerator architecture for long read DNA mapping is presented. RASSA is a massively parallel in-memory processor, facilitating simultaneous compare and mapping of a long read onto a reference sequence. The key performance breakthrough of RASSA is achieved by applying the similarity search in parallel to the entire reference. While the complexity of alignment is O(mn), RASSA employs in-memory parallel computing on O(m) memory cells to reduce computation time to O(n).

RASSA employs resistive elements, memristors, serving at the same time as single bit storage elements and comparators. It allows storing (typically, a data element per memory row) and in-situ processing of large datasets. RASSA enables comparing a key pattern with the entire dataset in parallel. Every number of mismatches (of the key pattern vs. each data element that is in each memory row) causes a specific voltage drop, allowing quantifying the number of mismatching locations (called a *mismatch score*). Additional evaluation transistors translate mismatch scores into voltage levels, which are converted to digital values using Analog to Digital Converters (ADC). The mismatch score is compared with a predefined threshold value to indicate the locations which have the desired degree of similarity with the compared pattern.

This work makes the following contributions:

1. RASSA, an in-memory processing resistive approximate similarity search accelerator, is introduced. The parallel processing architecture is presented bottom-up, from the memristor-based bitcell to base pair encoding and up to a complete RASSA system;
2. RASSA based implementation of long read mapping is developed;
3. Evaluation of RASSA's mapping accuracy and comparative analysis of its execution time is conducted.

To the best of our knowledge, this is the first work to accelerate the problem of DNA long read mapping. This work is currently under review in IEEE Micro.

The accompanied poster presents the following:

- Visual illustration of the read mapping problem
- Hierarchical breakdown of RASSA architecture
- Evaluation results: functionality, chip area, timing and power
- Accuracy and performance figures compared with a state of-art long read mapping tool, minimap2 [11].
- Future work: use RASSA to accelerate overlap finding between pairs of long reads, a fundamental operation in de novo sequence assembly (a newly sequenced organism that).

## 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 and R. Ginosar, "Deduplication in Resistive Content Addressable Memory Based Solid State Drive," in PATMOS, pp. 100–106, 2016.

[3] 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.

[4] Kaplan, Roman, Leonid Yavits, and Ran Ginosar. "PRINS: Processing-in-Storage Acceleration of Machine Learning." IEEE Transactions on Nanotechnology (2018).

[5] Jameson, J.L. and Longo, D.L. "Precision medicine—personalized, problematic, and promising." Obstetrical & gynecological survey, vol. 70, no. 10, pp. 612-614, 2015.

[6] Quick, J., Loman, N.J., Duraffour, S., et al, "Real-time, portable genome sequencing for Ebola surveillance." Nature, 530(7589), pp. 228-232.

[7] Rhoads, Anthony, and Kin Fai Auc. "PacBio sequencing and its applications." Genomics, proteomics & bioinformatics 13.5, pp. 278-289, 2015.

[8] Laver, T., Harrison, J., O'neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. "Assessing the performance of the oxford nanopore technologies minion." Biomolecular detection and quantification, vol. 3, pp. 1-8, 2015.

[9] Kaplan, R., Yavits, L., Ginosar, R. and Weiser, U. "A Resistive CAM Processing-in-Storage Architecture for DNA Sequence Alignment." IEEE Micro, vol. 37, no. 4, pp. 20-28, 2017.

[10] Li, Heng. "Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences." Bioinformatics 32.14, pp. 2103-2110, 2016.

[11] Li, H., Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, vol. 1, p. 7, 2018.

[12] Berlin, K., Koren, S., Chin, C. S., Drake, J. P., Landolin, J. M., & Phillippy, A. M. "Assembling large genomes with single-molecule sequencing and locality-sensitive hashing." Nature biotechnology, 33(6), p. 623, 2015.

[13] Alser, M., Hassan, H., Xin, H., Ergin, O., Mutlu, O. and Alkan, C. "GateKeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping." Bioinformatics, vol. 33, no. 21, pp. 3355-3363, 2017.

[14] Khatamifard, S. K., Chowdhury, Z., Pande, N., Razaviyayn, M., Kim, C., & Karpuzcu, U. R. "A Non-volatile Near-Memory Read Mapping Accelerator." arXiv preprint arXiv:1709.02381.

[15] Imani, M., Rahimi, A., Kong, D., Rosing, T., & Rabaey, J. M. "Exploring hyperdimensional associative memory". In IEEE International Symposium on High Performance Computer Architecture (HPCA), pp. 445-456, 2017.