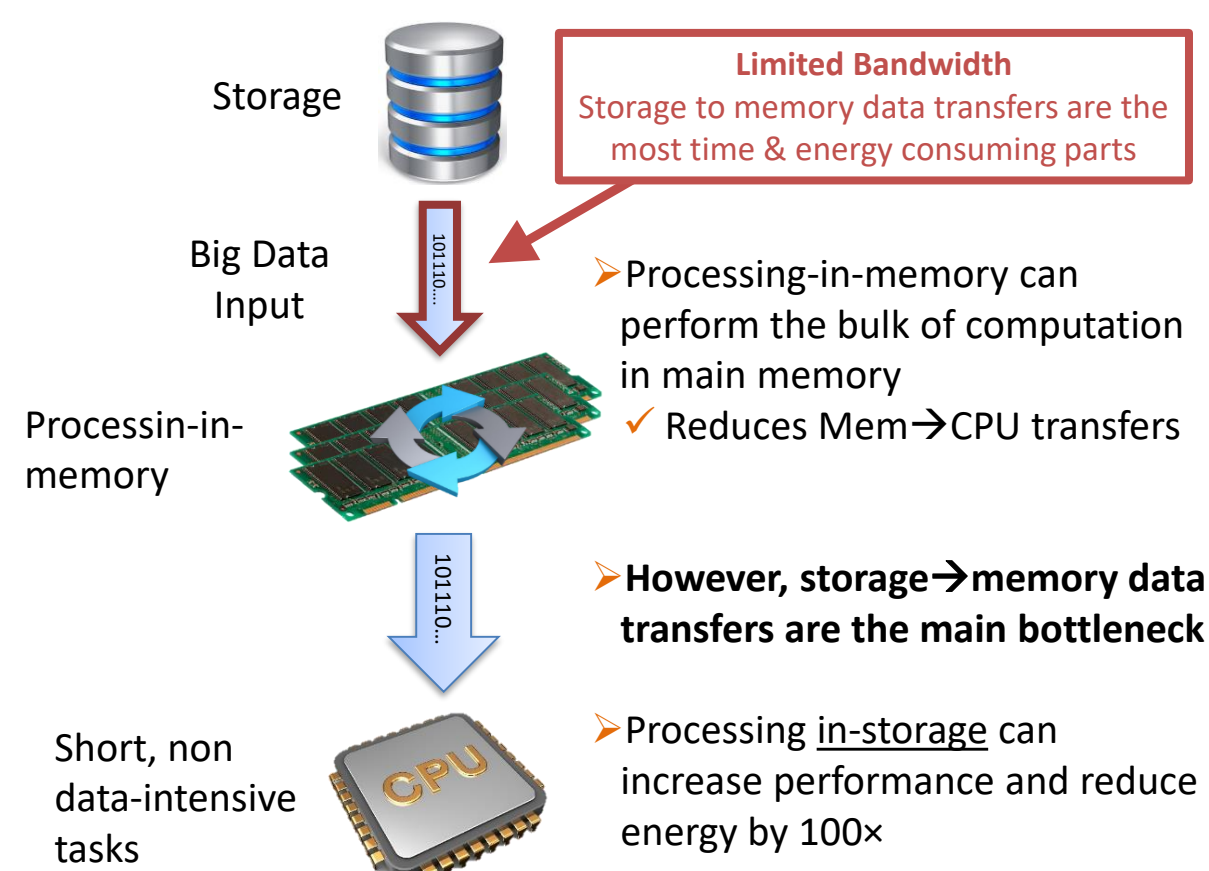


Processing-in-Storage Architecture for Machine Learning and Bioinformatics

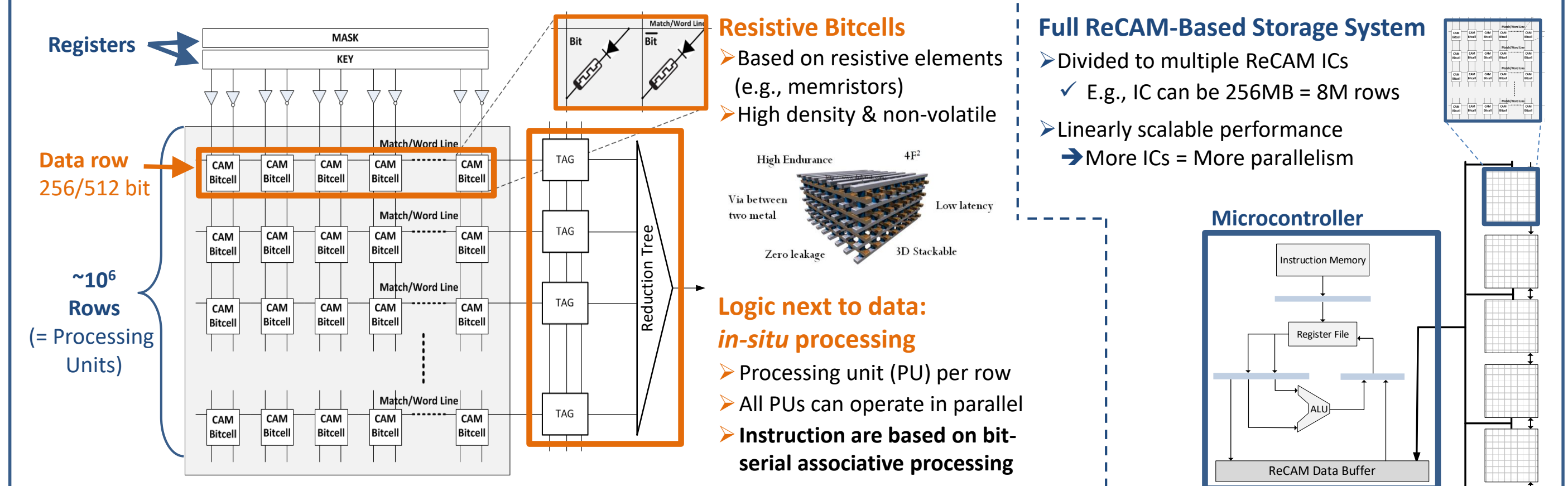
Roman Kaplan, Leonid Yavits and Ran Ginosar

Problem and Solution Overview

Storage ↔ Memory Bottleneck Problem



Resistive Content Addressable Memory (ReCAM): A PRinS Device



PRinS Application DNA Local Sequence Alignment*

Smith-Waterman Algorithm

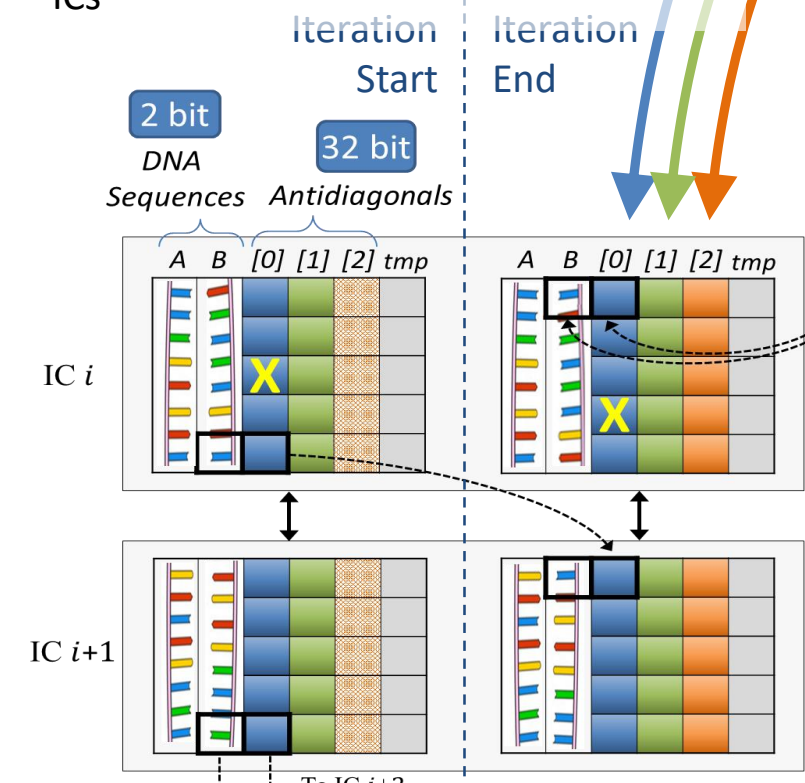
- Finds regions with “highest similarity” between two sequences (DNA/protein)
- Proven to be optimal
- Every “match”, “mismatch” & “gap” has a score
- S-W is based on dynamic programming
- Fills a $m \times n$ matrix
- Has a quadratic computational complexity $O(m \cdot n)$
- We focus on the computationally-heavy scoring step

* This work was done with prof. Uri Weiser

PRinS Implementation & Performance Comparison

In-Memory Computation

- IC = Integrated Circuit
- Shift operations move data between ICs



Performance Comparison

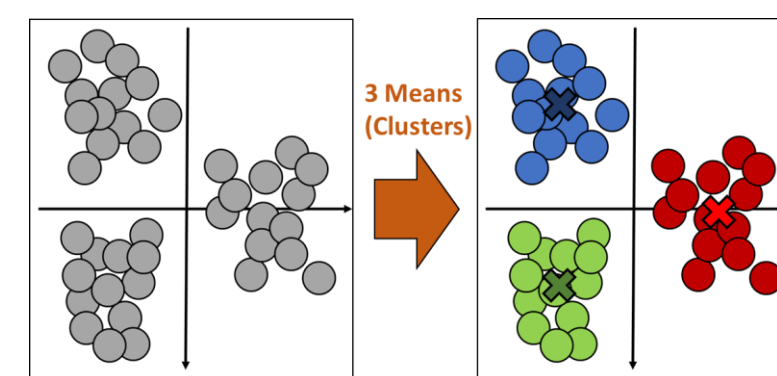
- Cycle-accurate simulator: 8GB of storage running at 500MHz
- Compared to multi-accelerator state-of-the-art solutions: FPGA, Xeon Phi and GPU
- ReCAM shows 4.7x higher performance than a 384-GPU solution

Accelerator	FPGA	Xeon Phi	GPU	ReCAM
Performance (TCUPS)	6.02	0.23	11.08	52.68
# of ICs	128	4	384	32
Publication	L. Weiser, et al. "The PRinS-based High-Performance Consistent Workflow for Applications in Genomics" (2015)	Y. Li and B. Schmidt, 2016, "Accelerating Smith-Waterman protein database search on Xeon Phi coprocessors" in IEEE ASP, 184-185	S. J. Kim, et al. "TCDMP: A Scalable and Efficient Accelerator for Smith-Waterman Alignment in GPU Clusters" (2015)	

PRinS Application K-Means Clustering

K-Means Clustering Algorithm

- Common Machine Learning algorithm
- For each sample, finds its group among possible K
- Widely used in many fields, including:
 - Image processing
 - Anomaly detection
- Data intensive
 - Multiple iterations over entire dataset



Two Main Loops

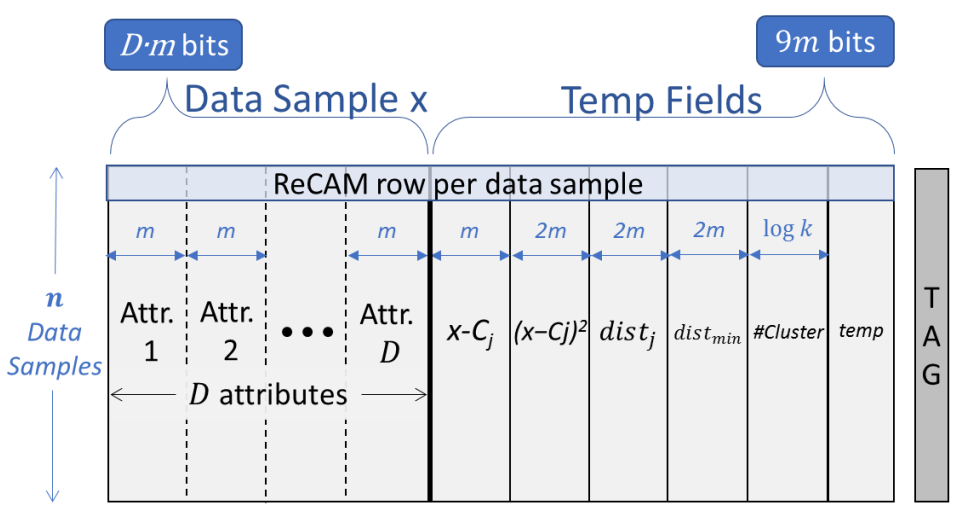
- For each Sample:
For each Center:
Calculate |Sample - Center|
- For each Center:
For each Sample:
Calculate Sum(Sample)

$O(n)$ in a von Neumann machine
 $O(1)$ in ReCAM

PRinS Implementation & Performance Comparison

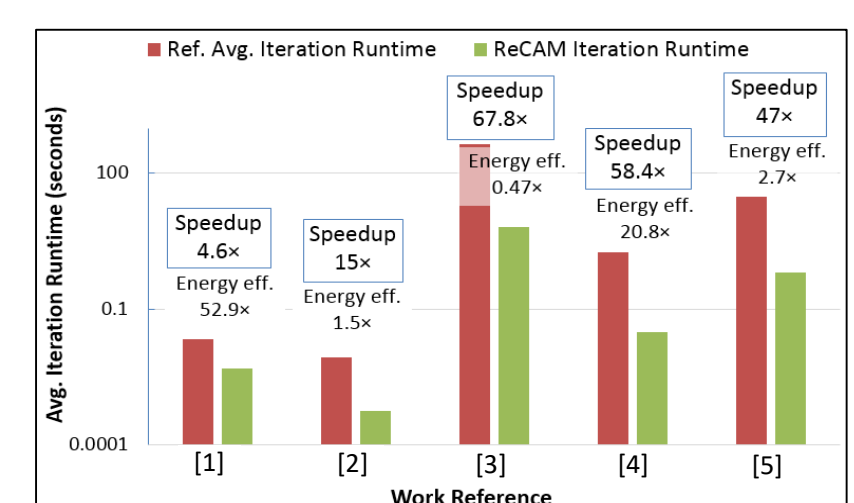
In-Storage Computation

- In large datasets: #Clusters #Data Samples
- Computation over ALL samples in one instruction
- Each sample attributes requires multiple temp fields
 - Difference between center's attribute
 - Squared difference
 - Total distance between sample and center



Comparison to other Works

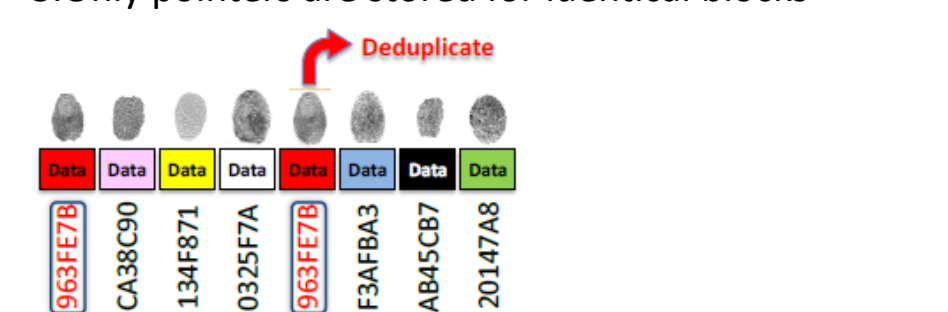
Work Reference	Platform	Samples	Attributes	Size on disk	Clusters
[1]	FPGA	1M	1	4MB	128
[2]	FPGA	2M	4	31.6MB	4
[3]	Intel i7	2.5M	68	318.8MB	10000
[4]	GPU	1.4M	5	21.3MB	240
[5]	10-GPU Cluster	1B	40	152.7GB	120



PRinS Application: In-Storage Deduplication

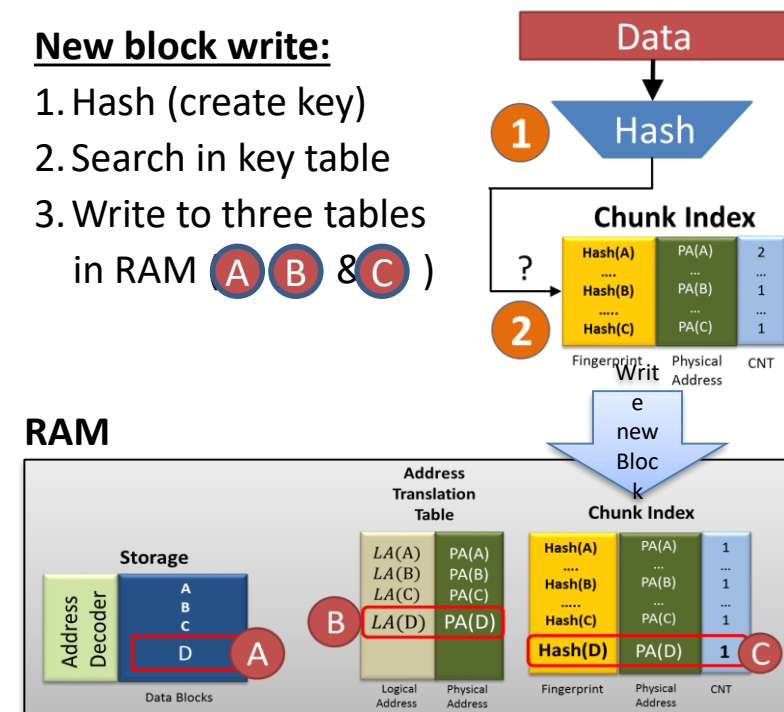
What is Deduplication?

- Deduplication is a technique for storing a single copy of each data block in storage
- Can reach 10x reduction in data volume
- How it works:
 - Data is broken into fixed blocks
 - A fingerprint (FP) is calculated for each block
 - Only pointers are stored for identical blocks

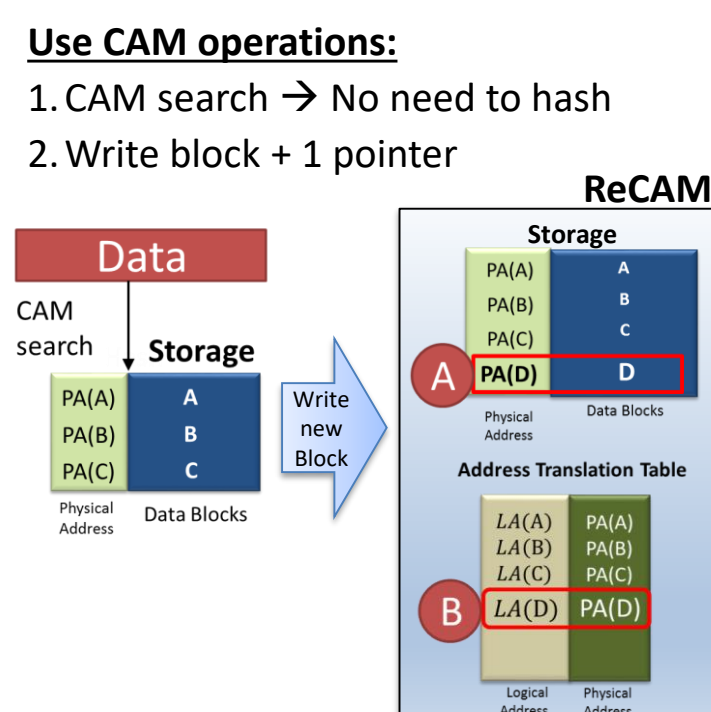


Traditional Systems vs. ReCAM

Traditional (RAM+CPU) Systems



In-ReCAM Deduplication



Performance Evaluations

- ReCAM was simulated with a cycle-accurate simulator
- ReCAM Parameters: 256GB @1GHz
- Openedup executed on high-end server: 4x8 octa-core CPU, 64GB RAM, 800GB SSD drive
- ReCAM has 100x higher throughput than deduplication with RAM+CPU
- Energy consumption is similar or lower for the common block sizes (4 & 8KB)

