

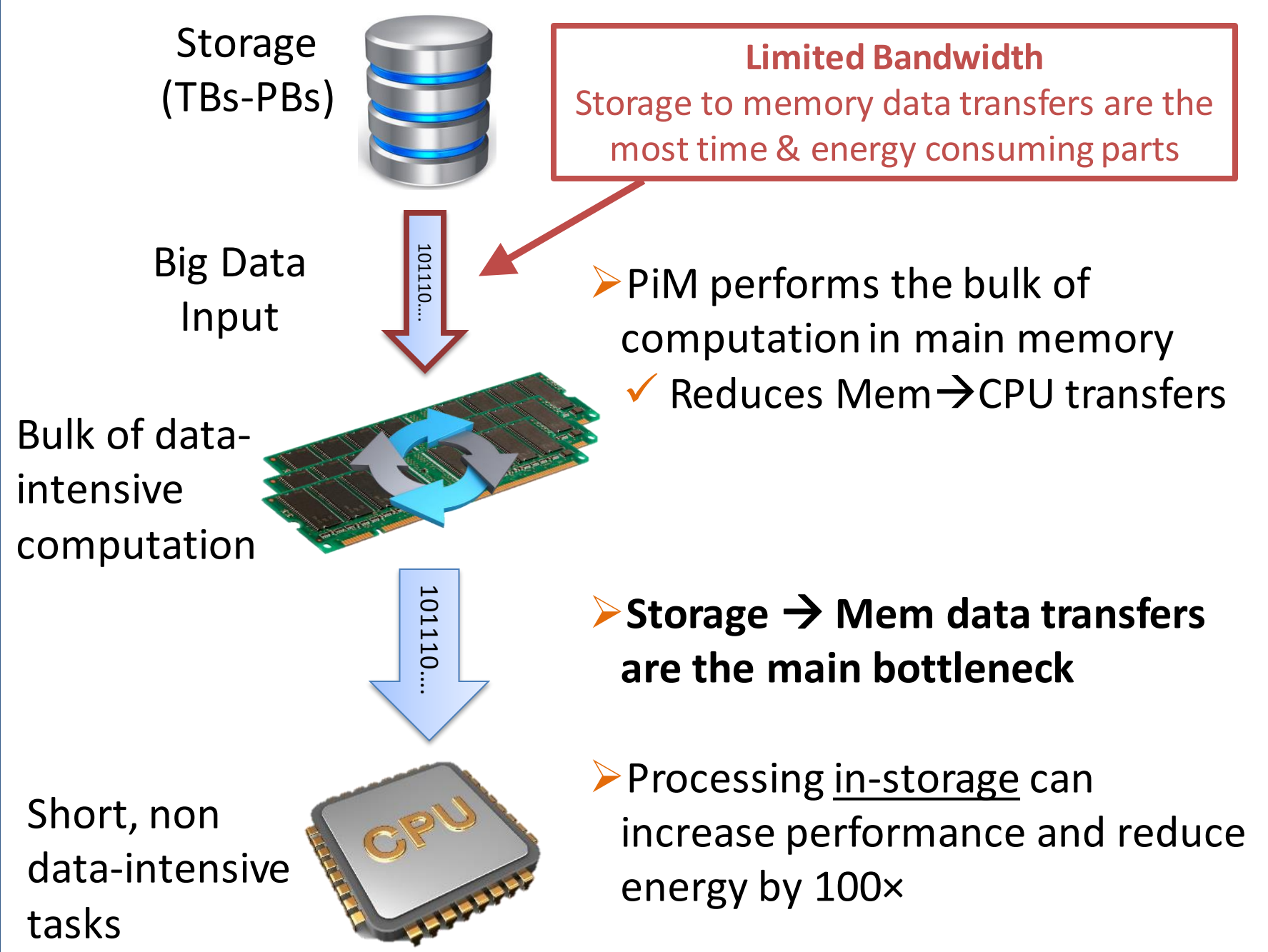
From Processing-in-Memory to Processing-in-Storage (PRinS)

Roman Kaplan, Leonid Yavits and Ran Ginosar

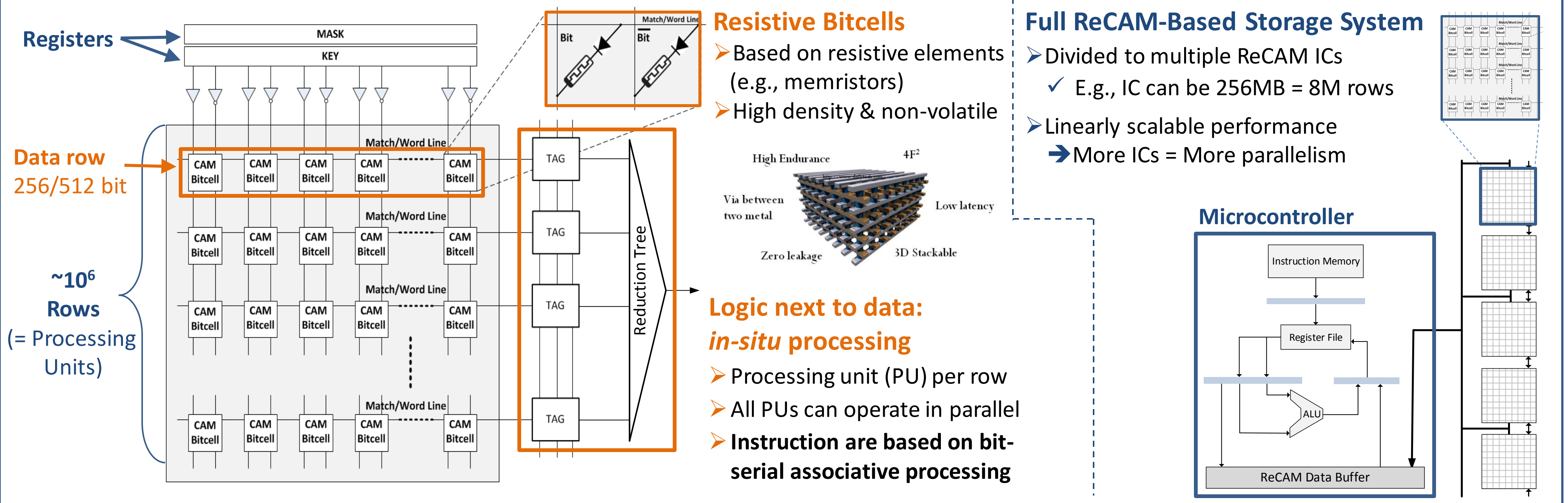


Problem and Solution Overview

Storage ↔ Memory Bottleneck Problem

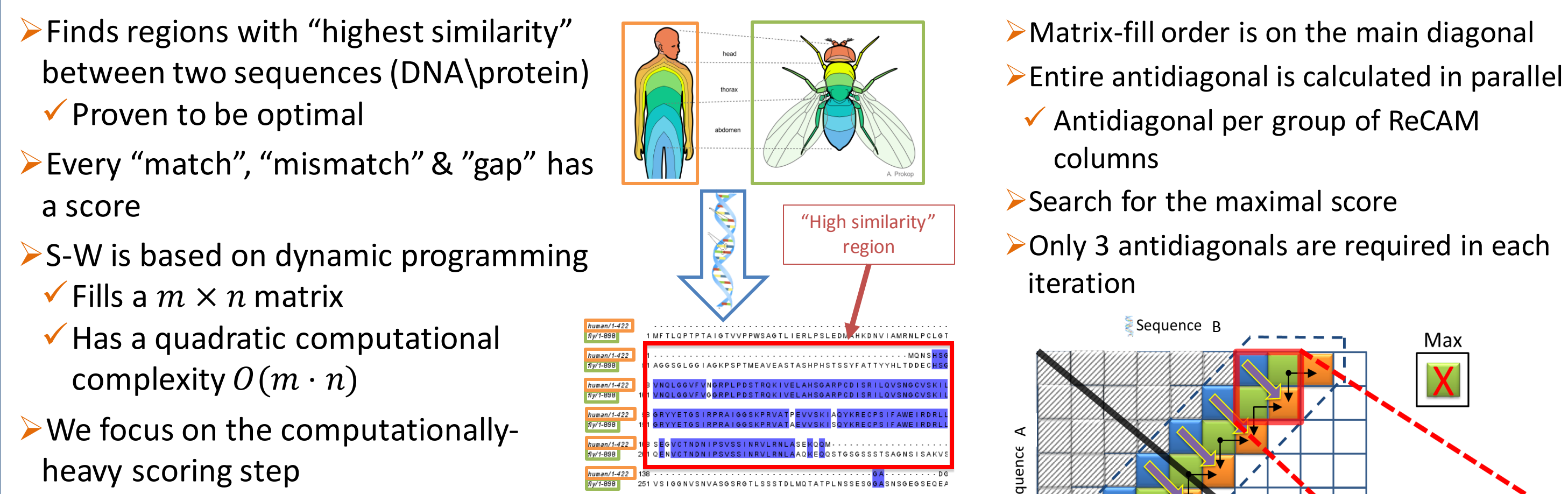


Resistive Content Addressable Memory (ReCAM): A PRinS Device



PRinS Application: DNA Local Sequence Alignment*

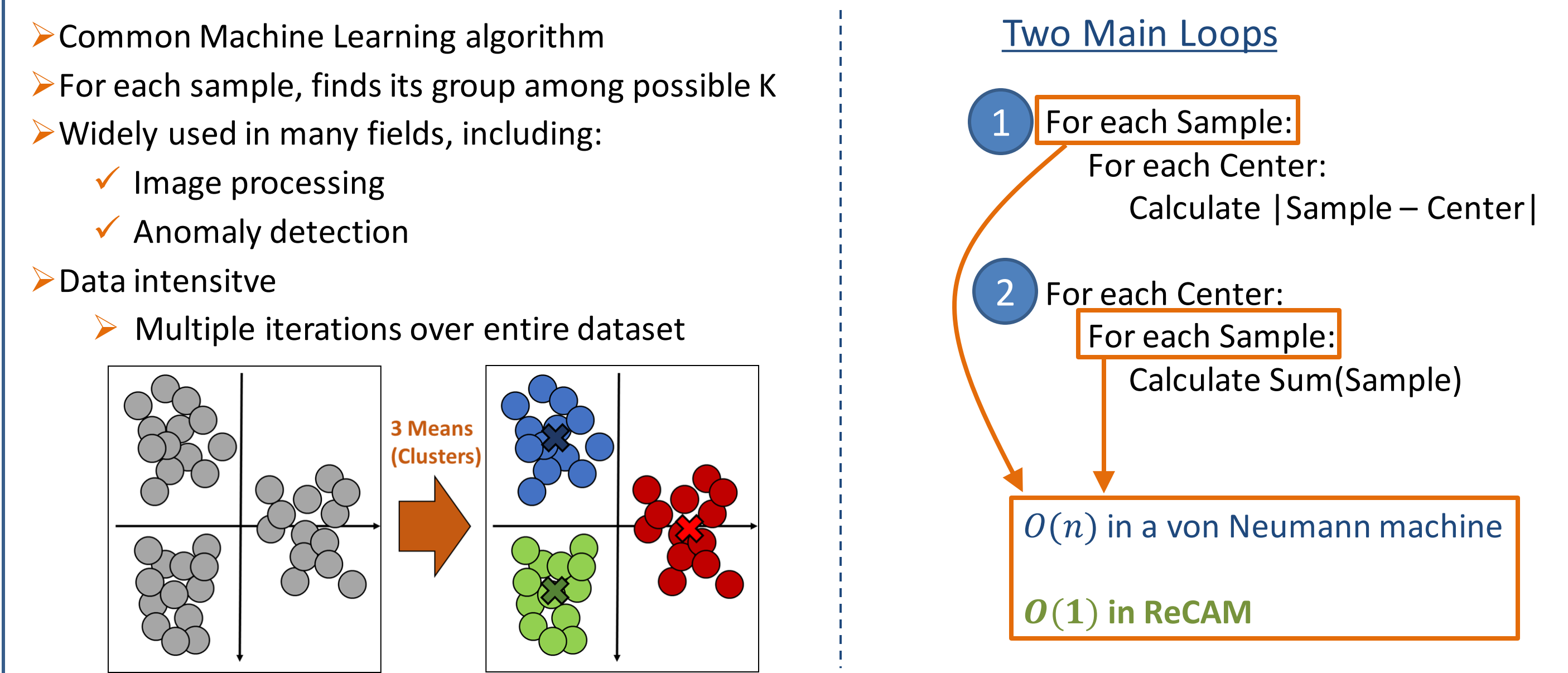
Smith-Waterman Algorithm



* This work was done with prof. Uri Weiser

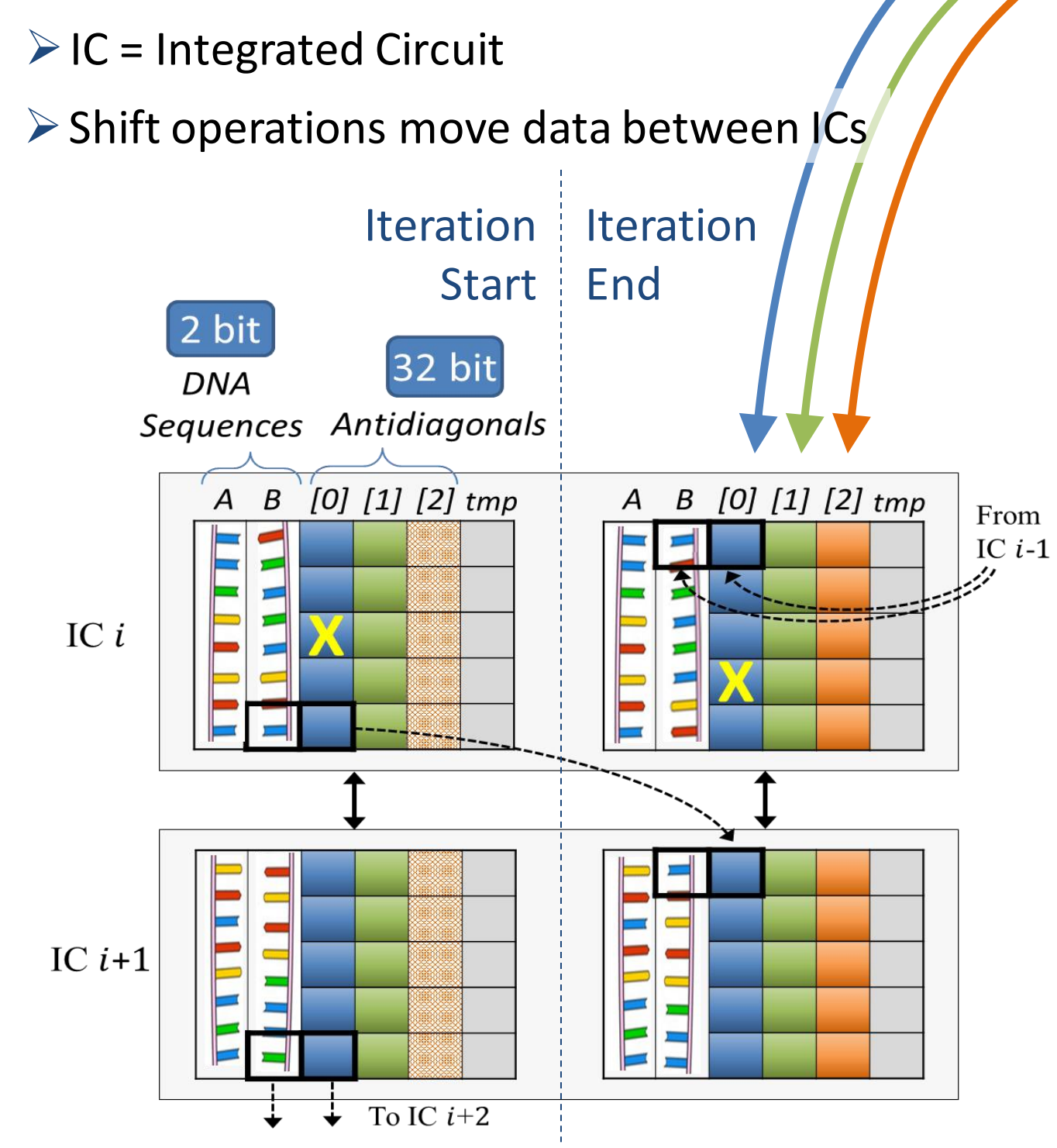
PRinS Application: K-Means Clustering

K-Means Clustering Algorithm



PRinS Implementation & Performance Comparison

In-Storage Computation



Performance Comparison

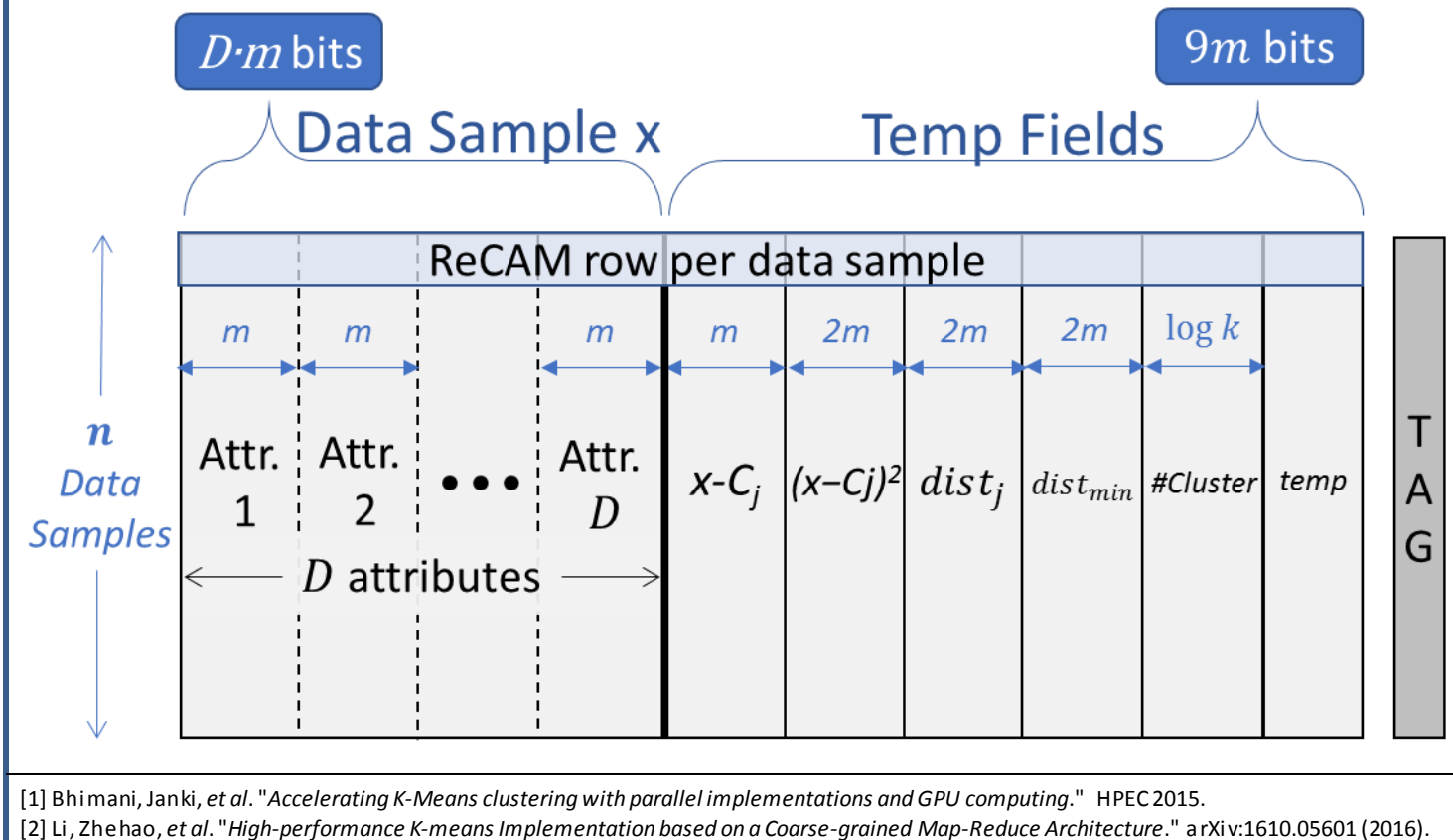
- Cycle-accurate simulator: 8GB of storage running at 500MHz
- Compared to multi-accelerator state-of-the-art solutions: FPGA, Xeon Phi and GPU
- ReCAM shows 4.7x higher performance than a 384-GPU solution

Accelerator	FPGA	Xeon Phi	GPU	ReCAM
Performance (TCUPS)	6.02	0.23	11.08	52.68
# of ICs	128	4	384	32

PRinS Implementation & Performance Comparison

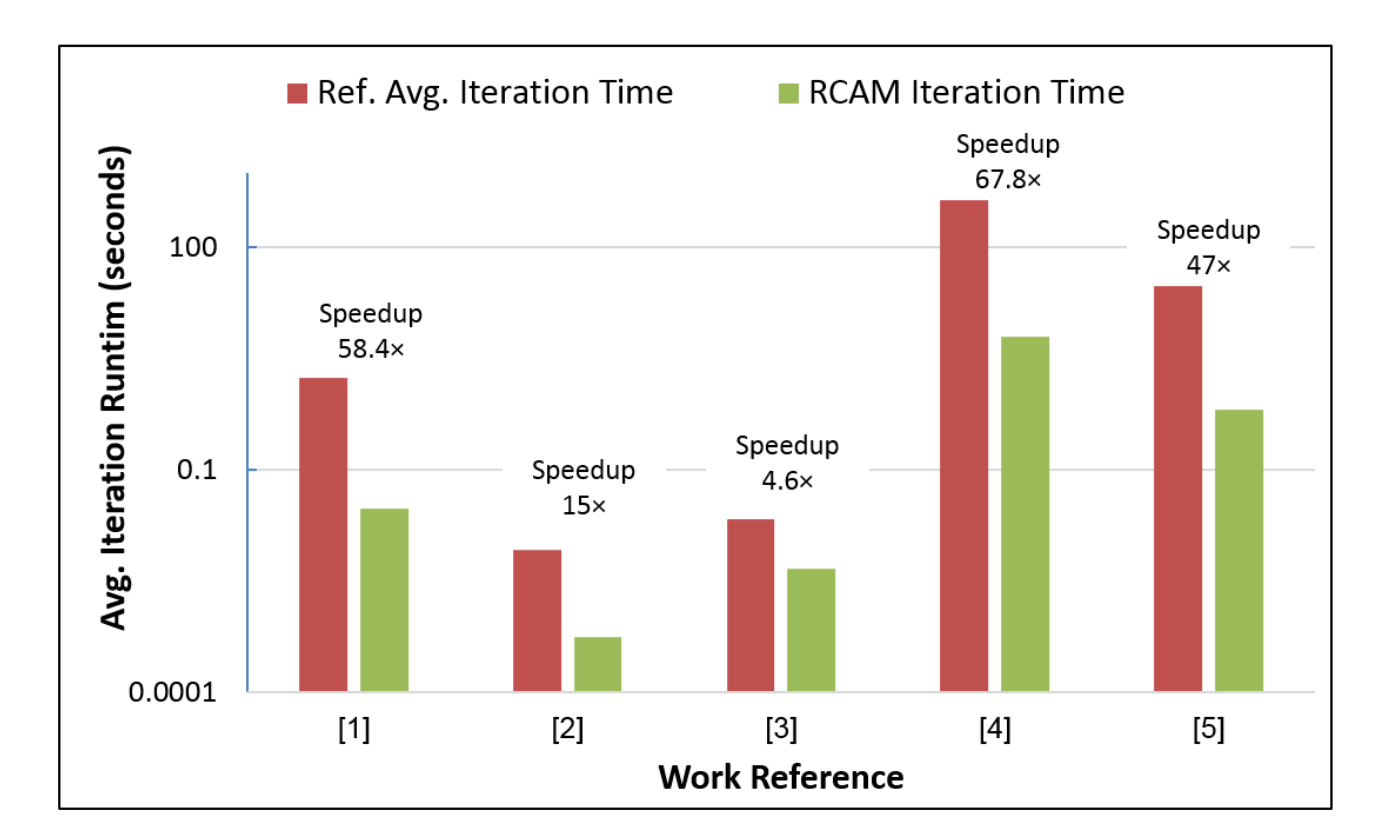
In-Storage Computation

- 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



Performance Comparison to other Works

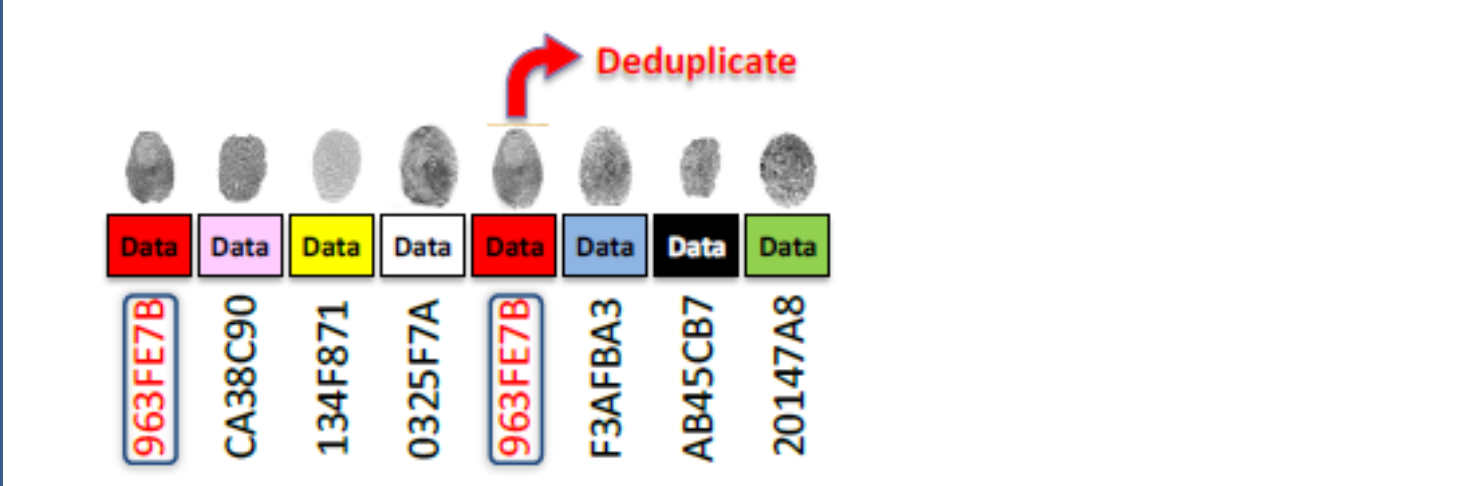
Work Reference	Platform	Dataset			Clusters
		Samples	Attributes	Size on disk	
[1]	GPU	1.4M	5	21.3MB	240
[2]	FPGA	2M	4	31.6MB	4
[3]	FPGA	1M	1	4MB	128
[4]	Intel i7	2.5M	68	318.8MB	10000
[5]	10-GPU Cluster	1B	40	152.7GB	120



PRinS Application: In-Storage Deduplication

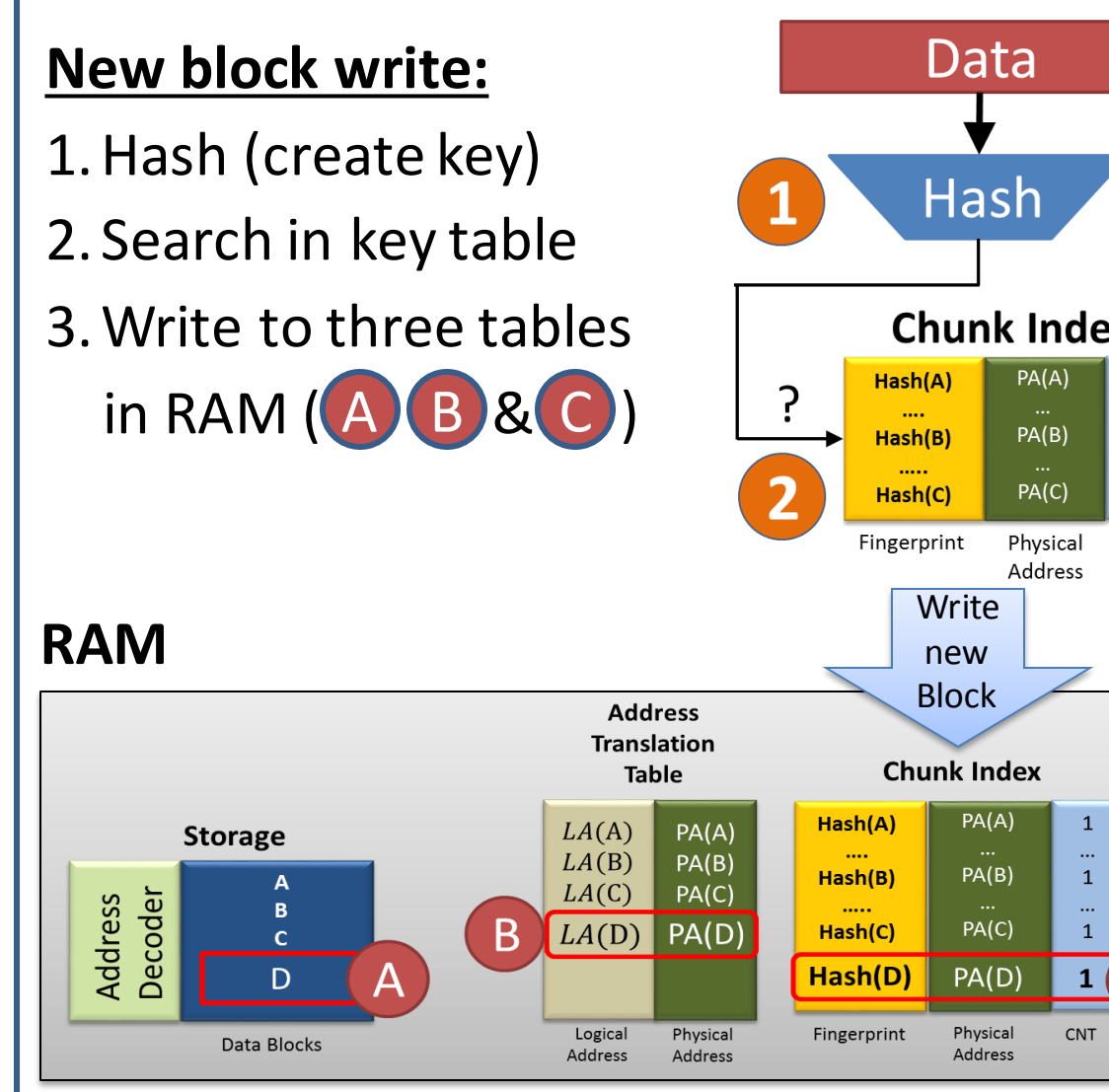
What is Deduplication?

- 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

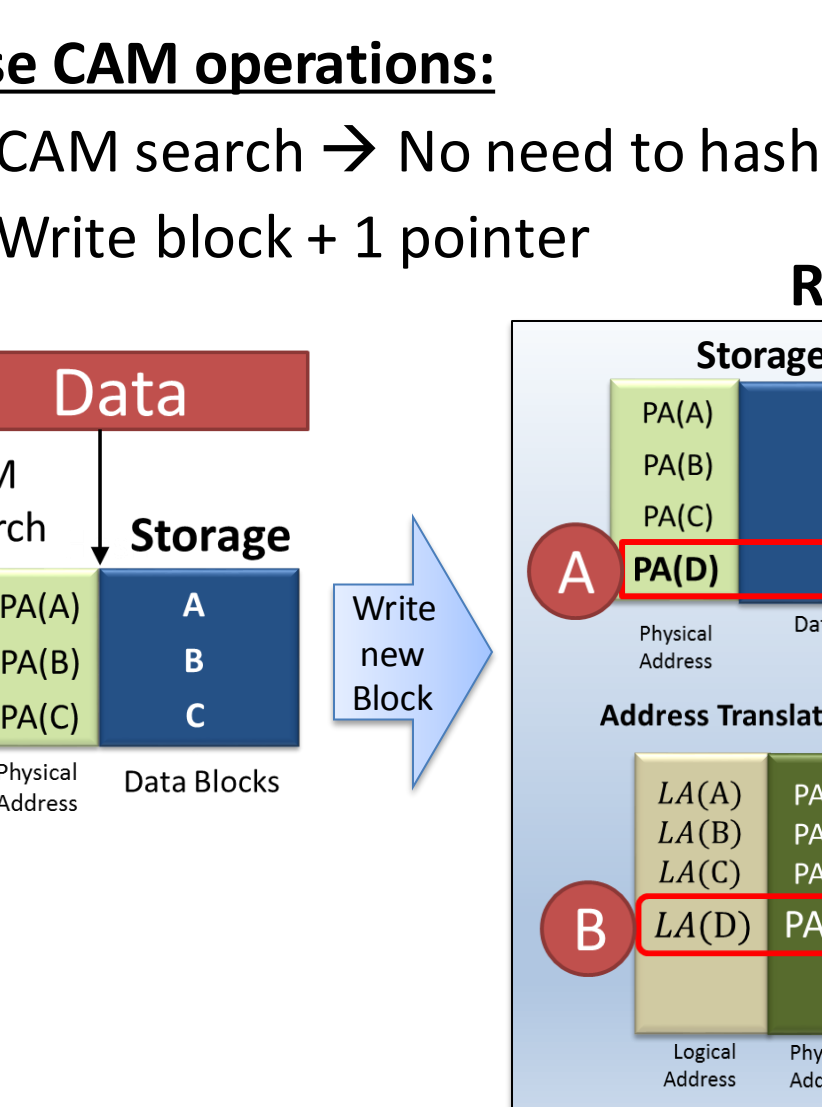


Traditional Systems vs. ReCAM

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)

