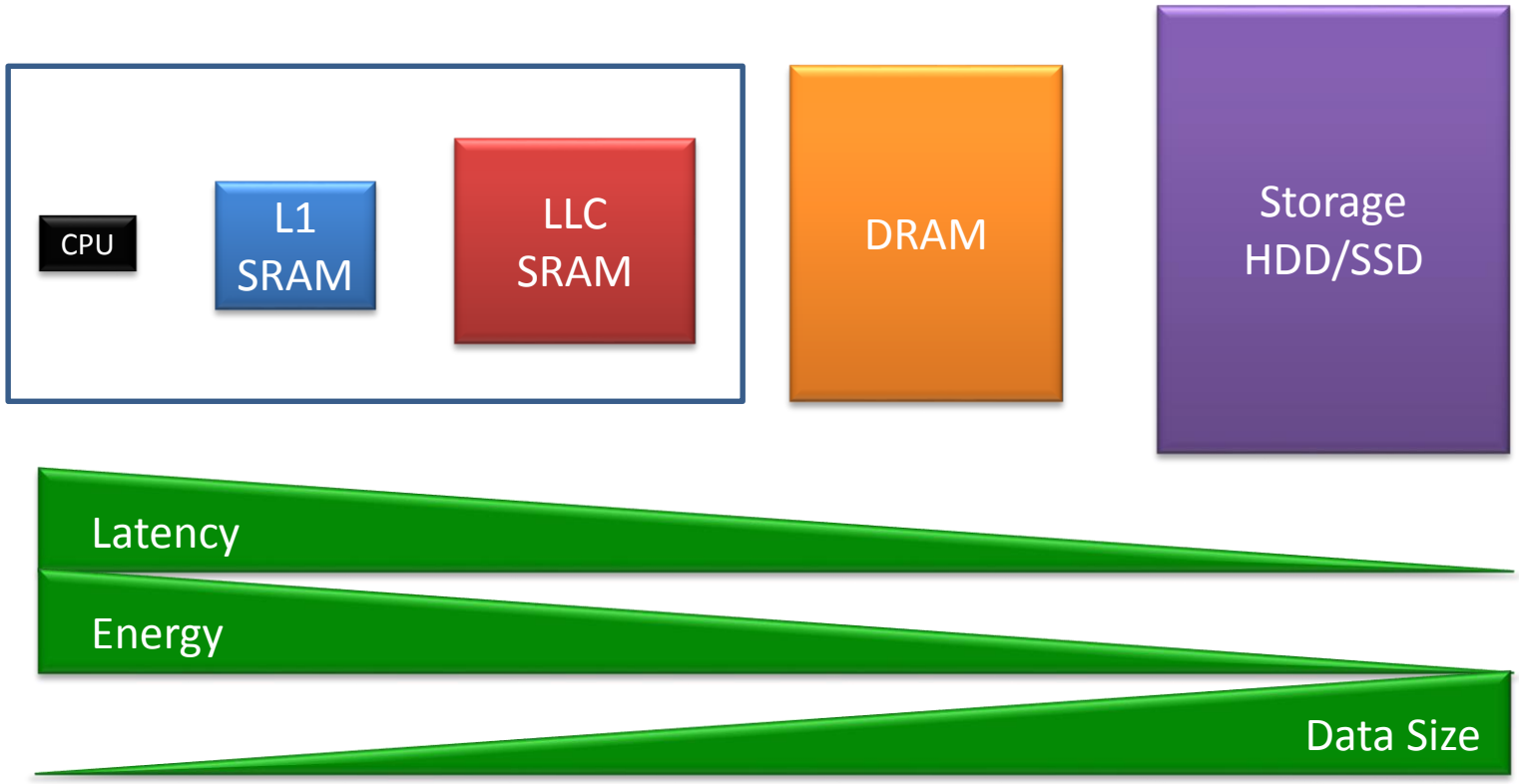
From Processing-in-Memory to Processing-in-Storage

Ran Ginosar
Leonid Yavits, Roman Kaplan

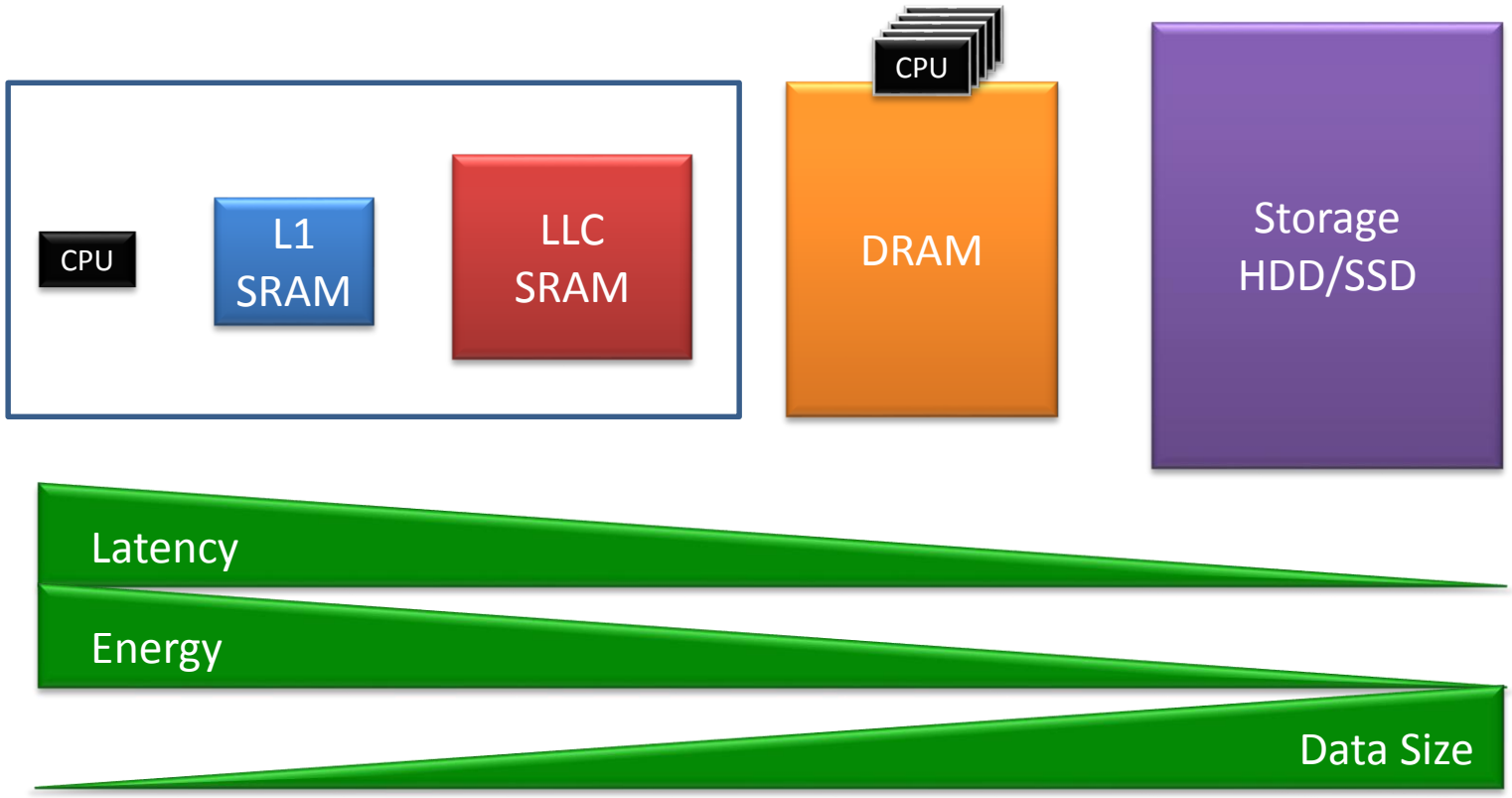
ICRI-CI 2016 Retreat, May 24, 2016

CPU-Mem-Storage Hierarchy



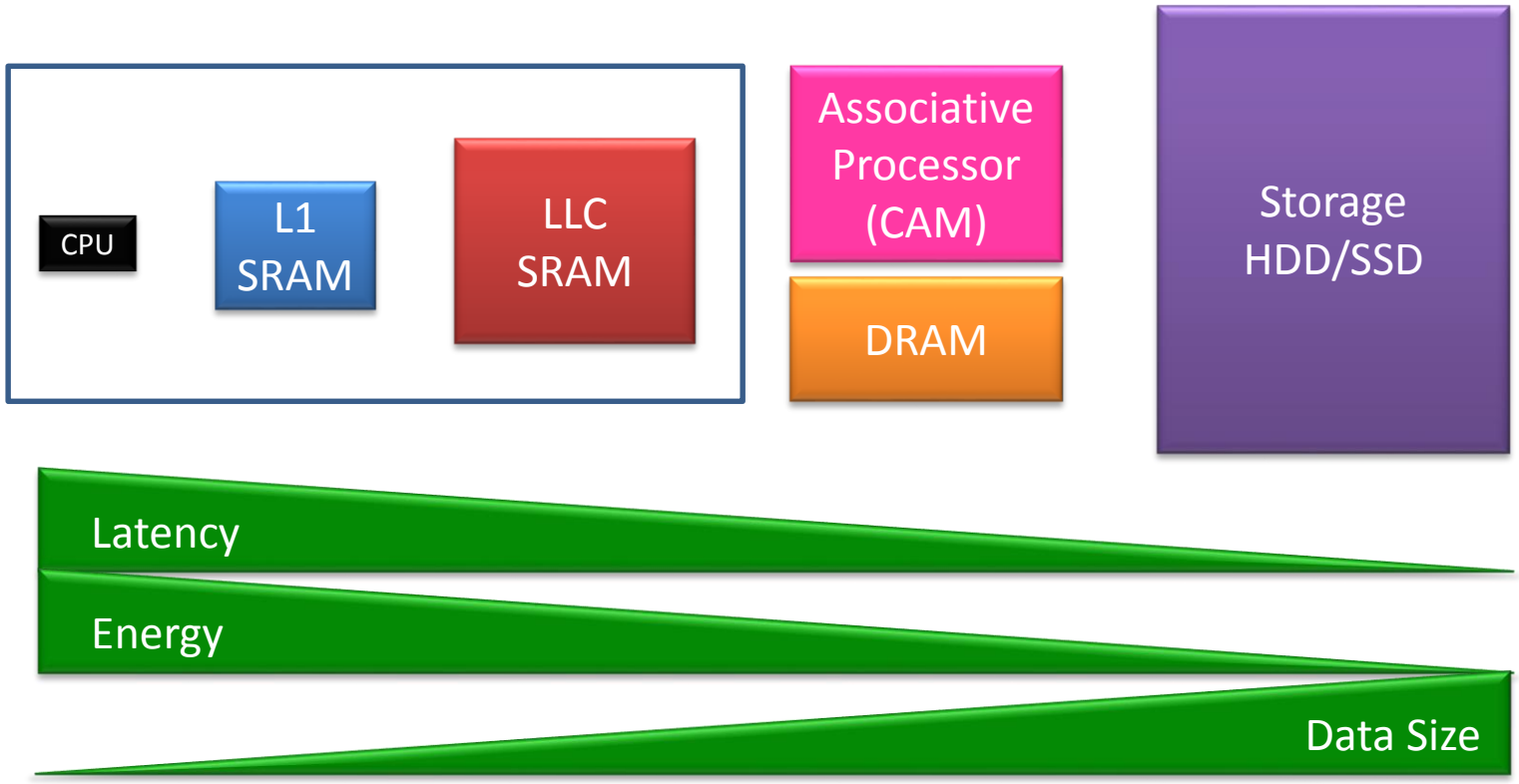
Processing In Memory

The conventional way: processing NEAR memory



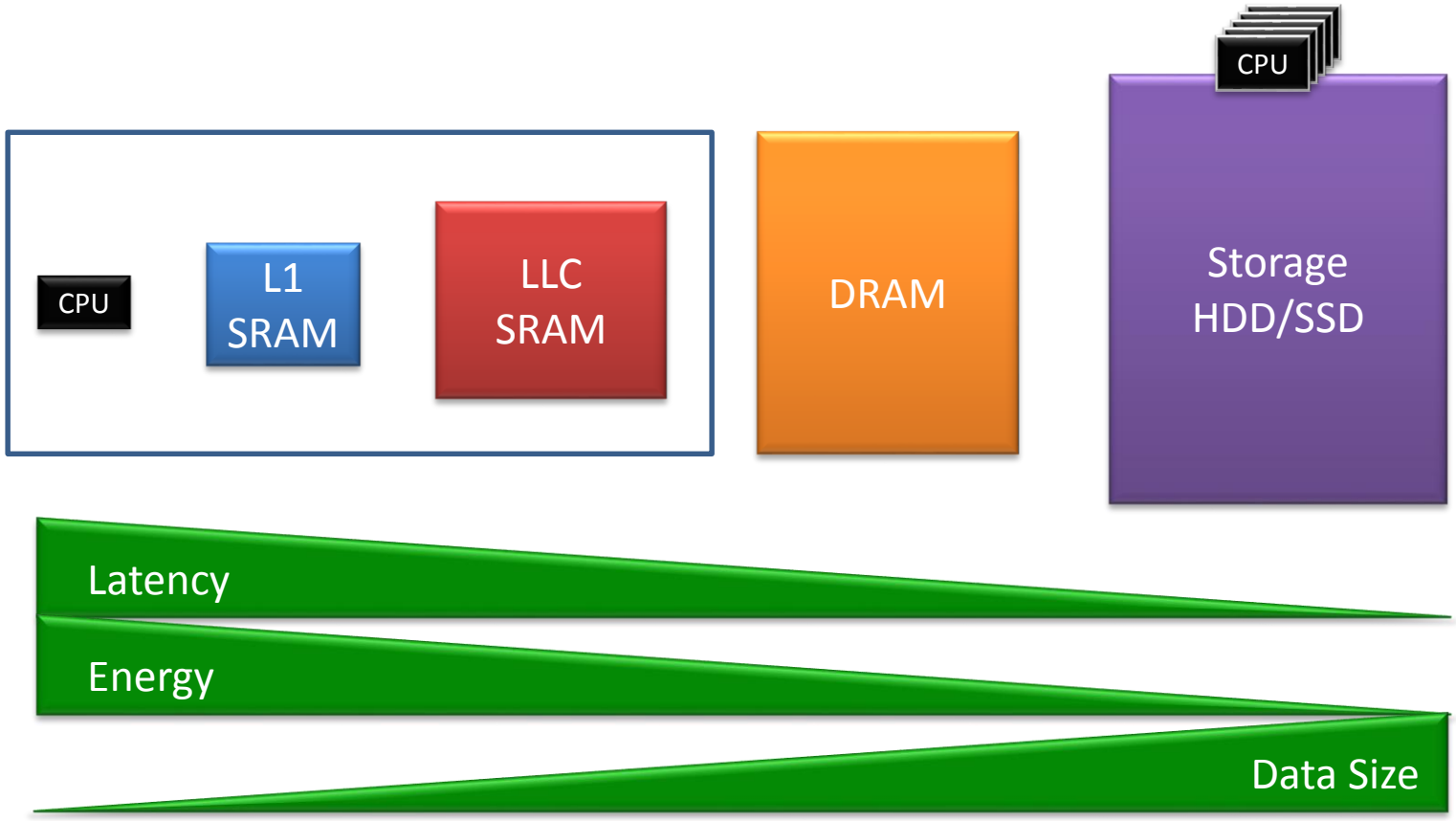
Processing In Memory

Our way: really “in memory”



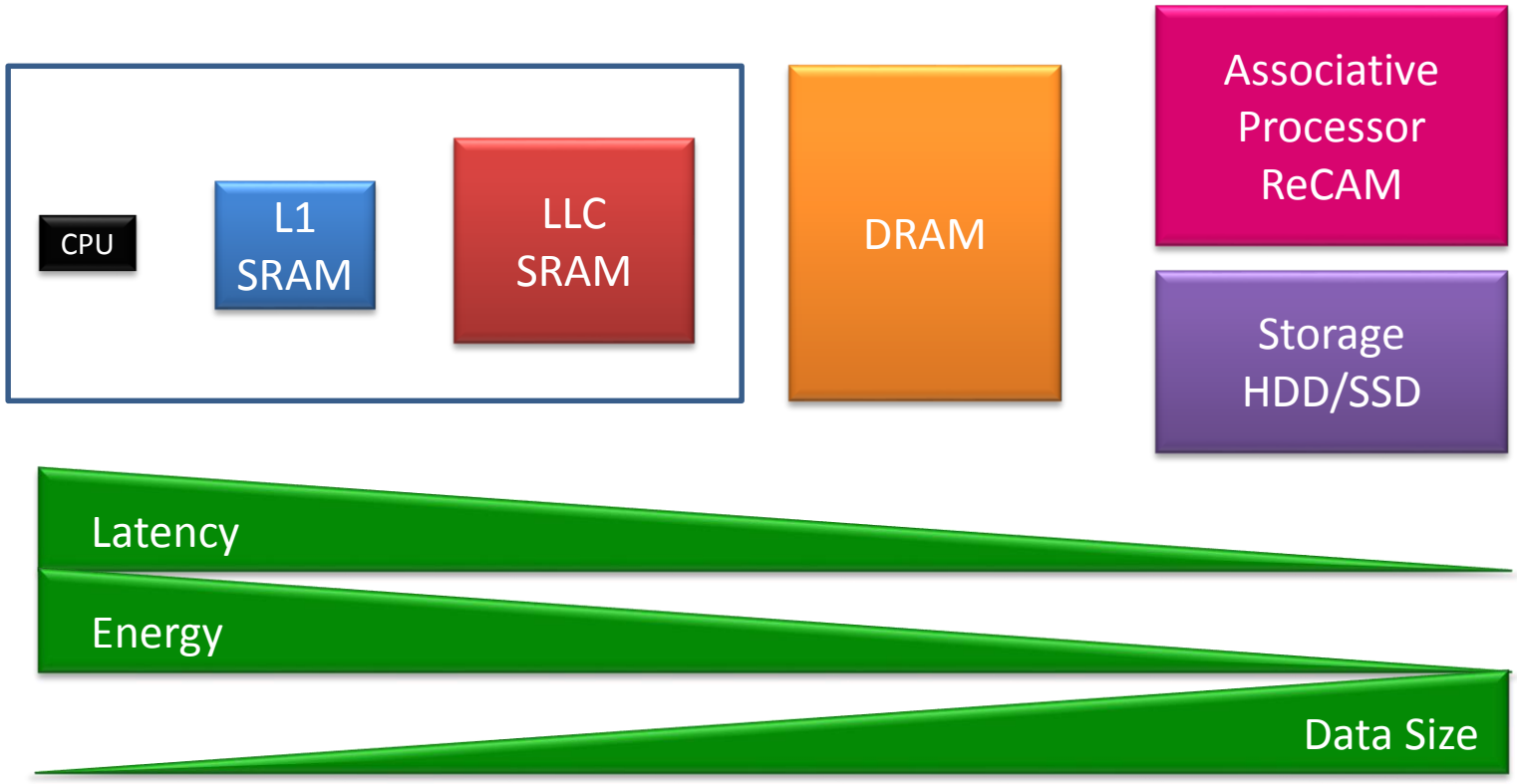
Processing In Storage

The conventional way: processing NEAR storage

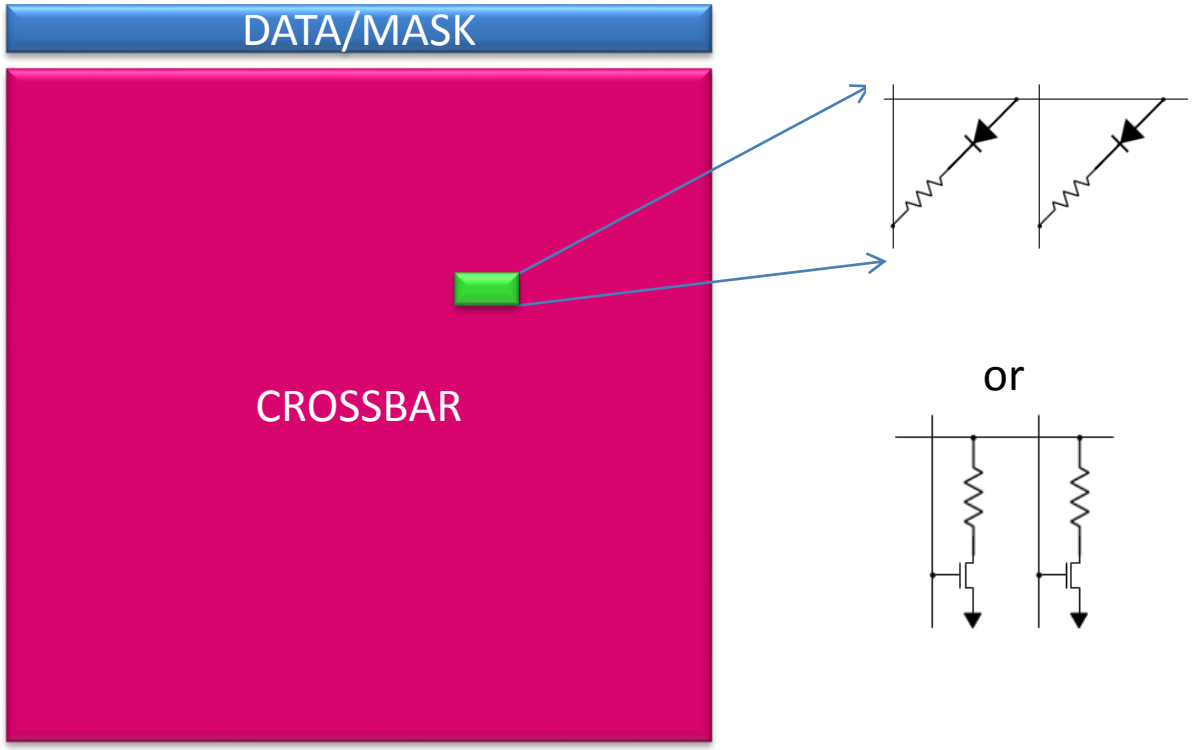


Processing In Storage

Our way: really “in storage”



Introducing ReCAM Storage

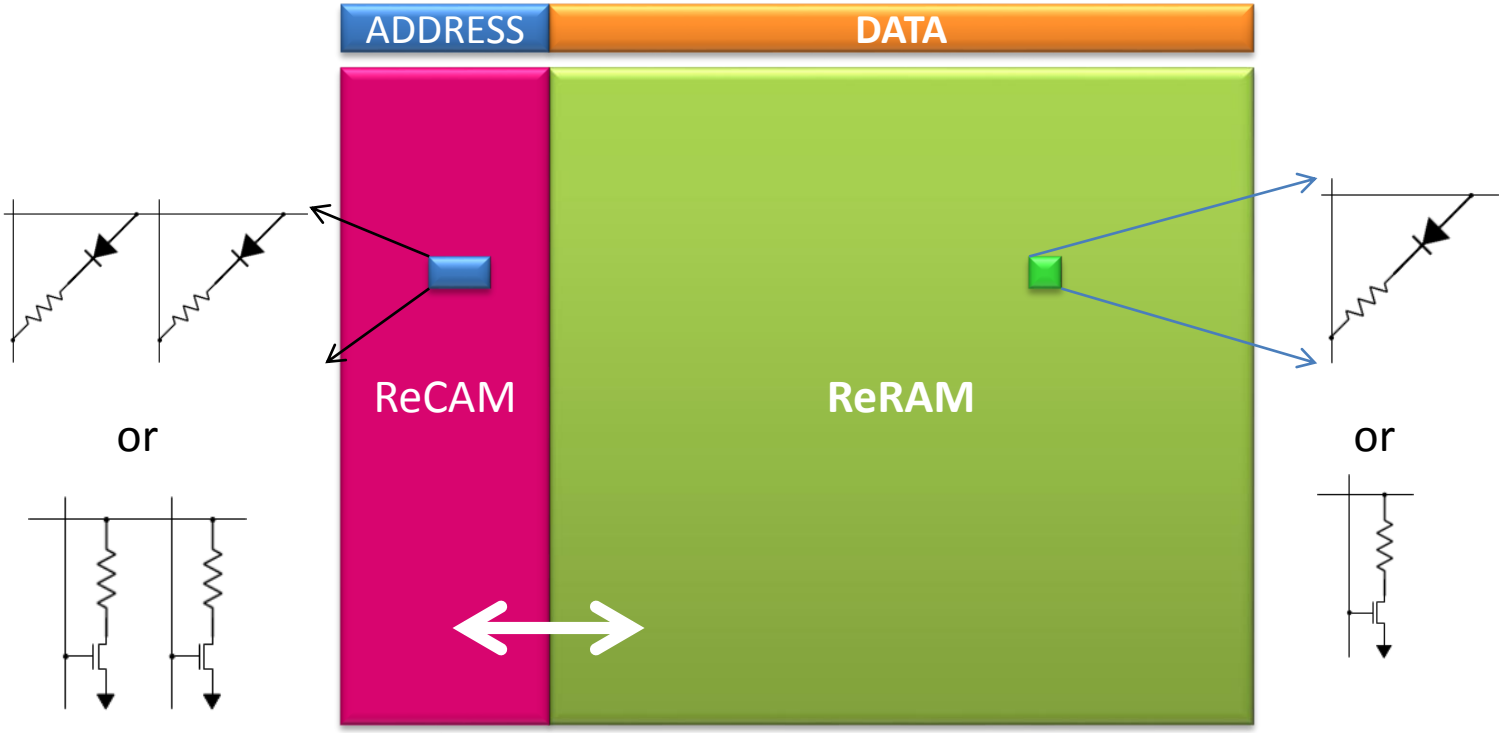


Resistive element

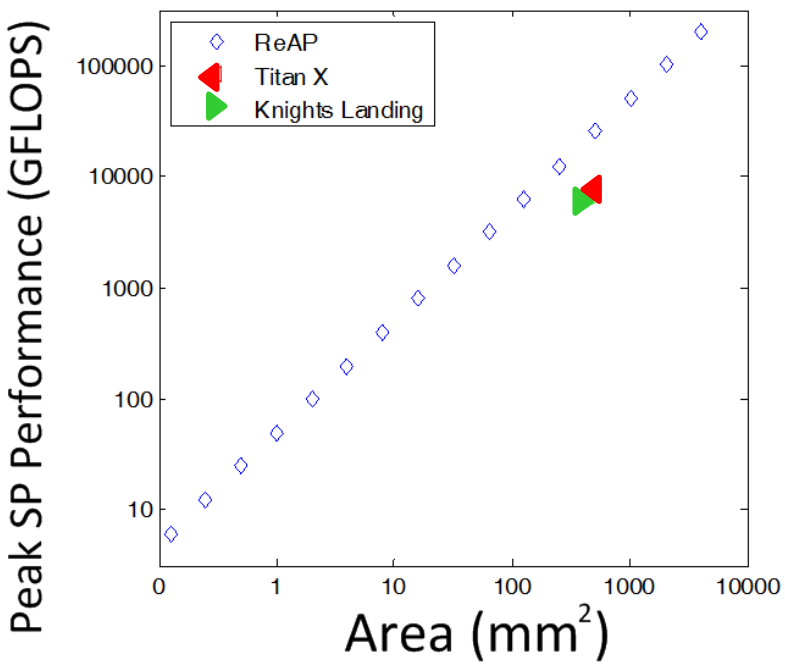
Dual Use

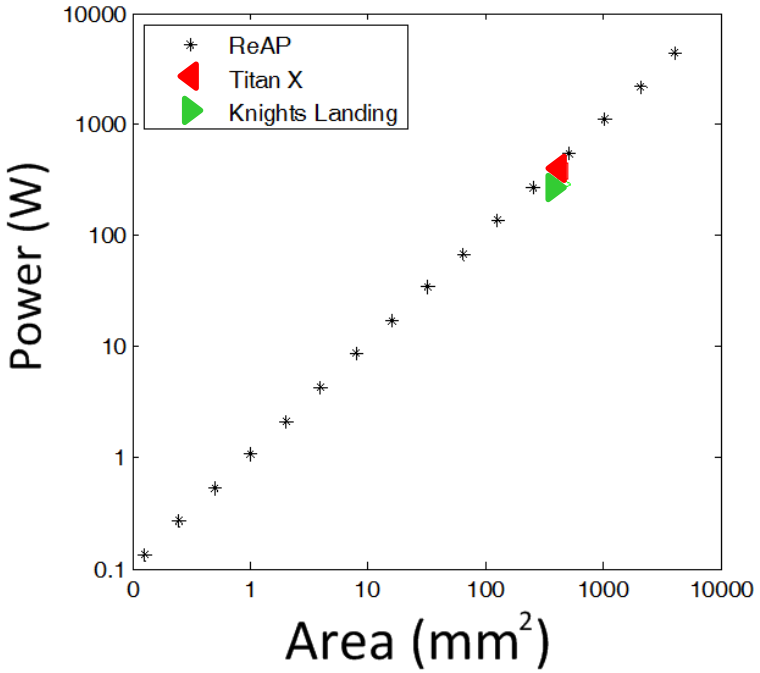
ReCAM is 2x ReRAM => Flexible dual use

- Storage + Processing in-Storage
- Storage only



Processing In Storage Performance and Power





"Resistive Associative Processor", L. Yavits, S. Kvatinsky, A. Morad, R. Ginosar, IEEE CAL 2015

ICRI-CI Intel Collaborative Research Institute for Computational Intelligence

Technion Israel Institute of Technology

Applications: Machine Learning

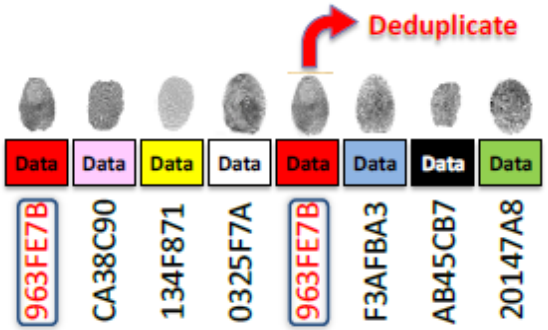
Workload	ReCAM performance	CPU Performance
Pattern Match / grep	$O(1)$	$O(N)$
Max / Min	$O(1)$	$O(N)$
Sort	$O(N)$	$O(N \log N)$
DMM	$O(N^2)$	$O(N^3)$
SpMM	$O(nnz)$	$O(nnz \cdot N)$
Convolution / Correlation	$O(N)$	$O(N \log N)$
Graph Processing (Dijkstra / SSSP)	$O(N)$	$O(N^2)$

Application: Storage

- Online deduplication
- Online compression
- Online security

Example: Online Deduplication

- Goal: store only one instance of data
- Today
  - Complex data structures
  - Computationally expensive hash function
- Processing in Storage
  - Smaller data structures
  - No hash
- Result: 100x higher throughput



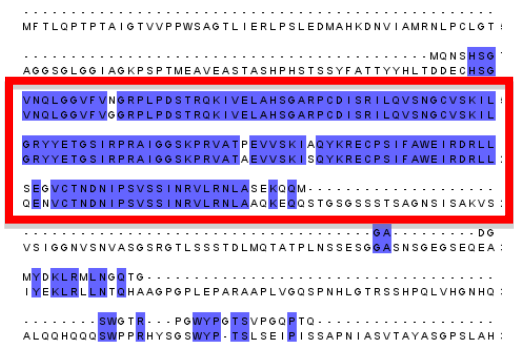
ICRI-CI Intel Collaborative Research Institute for Computational Intelligence

Technion Israel Institute of Technology

Example: DNA Sequence Alignment

- Find regions of similarity in long DNA strands
- $O(N^2)$ matrix
- Processing in Storage: 3x higher performance than a 384-GPU cluster
  - Align human and chimpanzee Chromosome 1
  - 57 Peta score cells (200 PetaBytes)

[Sandes EF, Miranda G, Martorell X, Ayguade E, Teodoro G, Melo AC,
CUDAlign 4.0: Incremental Speculative Traceback for Exact Chromosome-Wide Alignment in GPU Clusters,
IEEE Trans. PDS 2016]



ICRI-CI Intel Collaborative Research Institute for Computational Intelligence

Technion Israel Institute of Technology

Big Data Machine Learning

- Needs MANY nodes
  - Each node needs entire data
- Needs effective network
  - Interconnecting storage, rather than CPUs
- The larger the node,
  - The fewer the nodes
  - The faster the computation
  - The lower the energy

Summary
Processing In Memory and Storage