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

Roman Kaplan, Leonid Yavits and Ran Ginosar

Technion, Israel Institute of Technology

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 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

PRinS Implementation & Performance Comparison

Traditional (RAM+CPU) Systems vs. ReCAM

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:
  1. Data is broken into fixed blocks
  2. A fingerprint (FP) is calculated for each block
  3. Only pointers are stored for identical blocks

Performance Comparison

PRinS Application: In-Storage Deduplication

Traditional (RAM+CPU) Systems

New block write:
1. Hash (create key)
2. Search in key table
3. Write to three tables in RAM

In-ReCAM Deduplication

Use CAM operations:
1. CAM search  No need to hash
2. Write block + 1 pointer

Performance Evaluations

» ReCAM was simulated with a cycle-accurate simulator
» ReCAM Parameters: 256GB @1GHz
» OpenEndup executed on high-end server: 4x 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)

Graphs and diagrams illustrating the performance and implementation of PRinS, including comparisons with traditional systems for deduplication, DNA local sequence alignment, and K-Means clustering. Diagrams depict the architecture and functionality of ReCAM, a hardware component that accelerates machine learning algorithms by integrating processing-in-memory and processing-in-storage techniques.