AC-DIMM: Associative Computing with STT-MRAM

Qing Guo, Xiaochen Guo, Ravi Patel
Engin Ipek, Eby G. Friedman
University of Rochester

Published In: ISCA-2013

Motivation

❑ Prevalent Trends in Modern Computing:

1. **Technology Scaling** > Creates Power and Bandwidth Challenges
   - Transistor density doubles every two years, but power efficiency does not scale proportionally
   - Number of pins grows approximately at 16% / year only
2. **Data-Intensive Work Load**

❑ Resultant Bottlenecks:

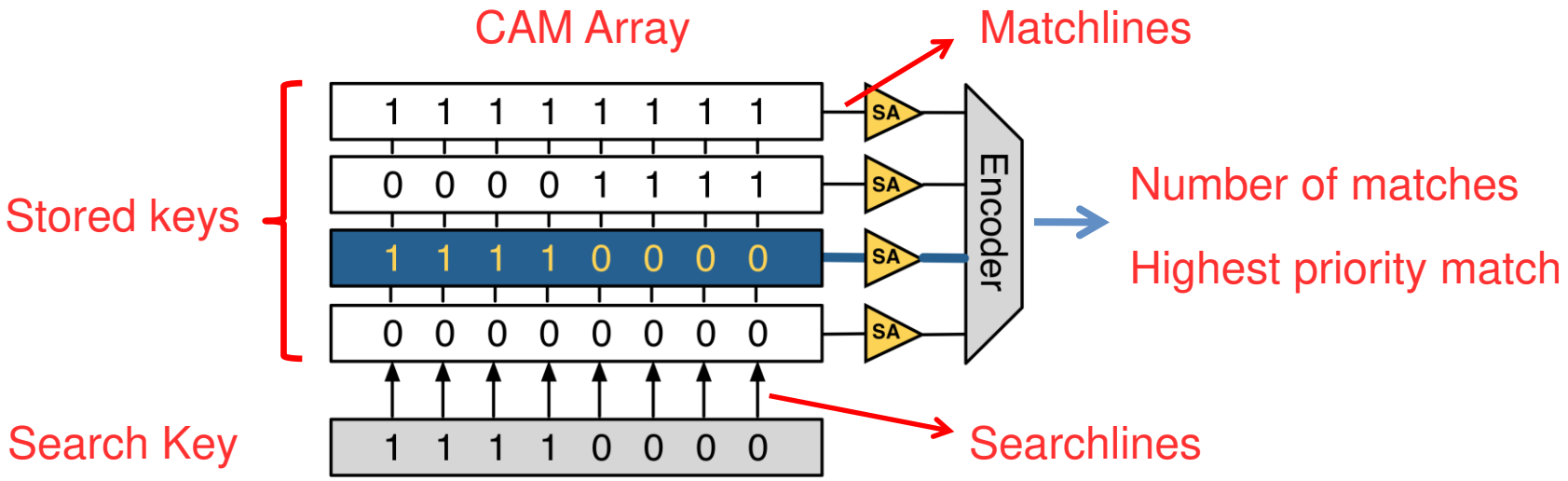
- **On-Chip Power Dissipation**
- **Off-Chip Memory Bandwidth**

❑ One Promising Solution:

Associative Computing Using Content-Addressable Memories (CAM)

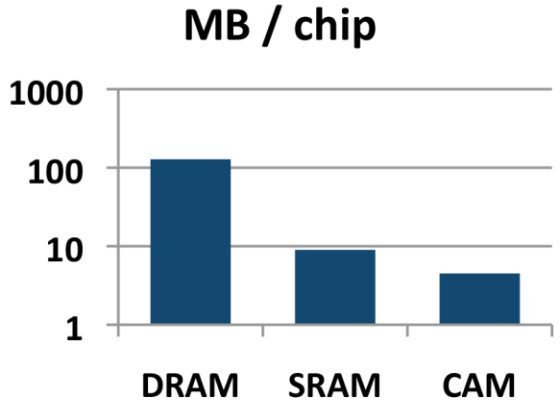
Content Addressable Memory

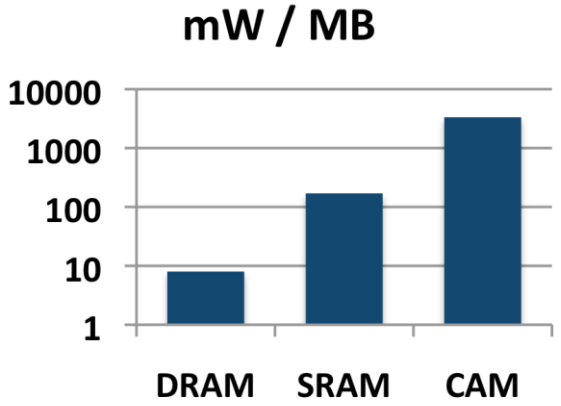
- Simultaneously compares all stored keys against a search key
- Energy- and bandwidth-efficient on an important subset of data-intensive applications



Current Challenges With CAMs

- CMOS-based CAMs are large, costly, and power-hungry



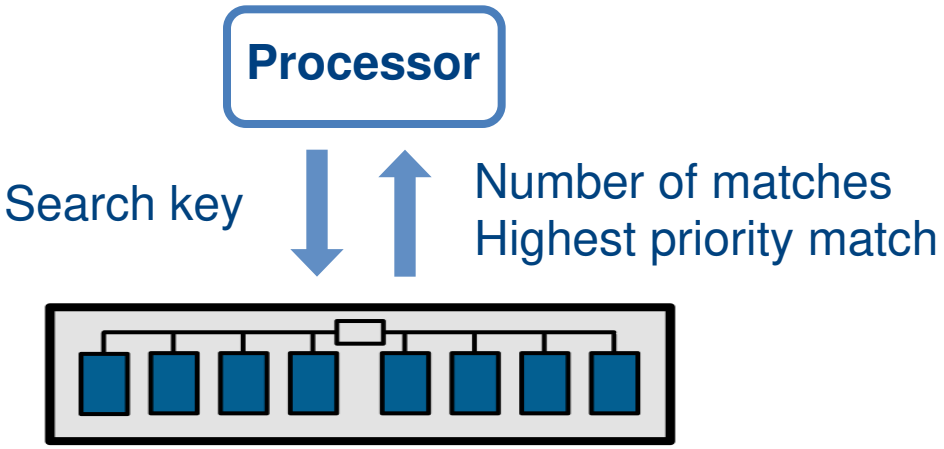


[Goel and Gupta, 2010]

- Commercial uses of CAMs are limited
  - Highly associative caches, TLBs
  - Microarchitectural queues
  - Networking routers

Resistive CAMs

- Resistive memories (e.g., PCM and STT-MRAM) offer high density and very low leakage power
- Previously proposed PCM-based TCAM accelerator [MICRO'11]
  - A gigabyte, DDR3-compatible DIMM
  - TCAM caters to a wide range of search-intensive applications

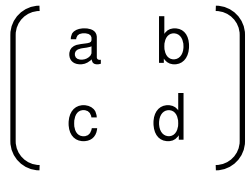


+ Optimized for density and search throughput

− limited functionality

Associative Computing Paradigm

- Broadens the use of CAMs to a more general programming framework
- Data organized by key-value pairs
  - Linked list, array, stack, queue
  - Matrix, tree, graph

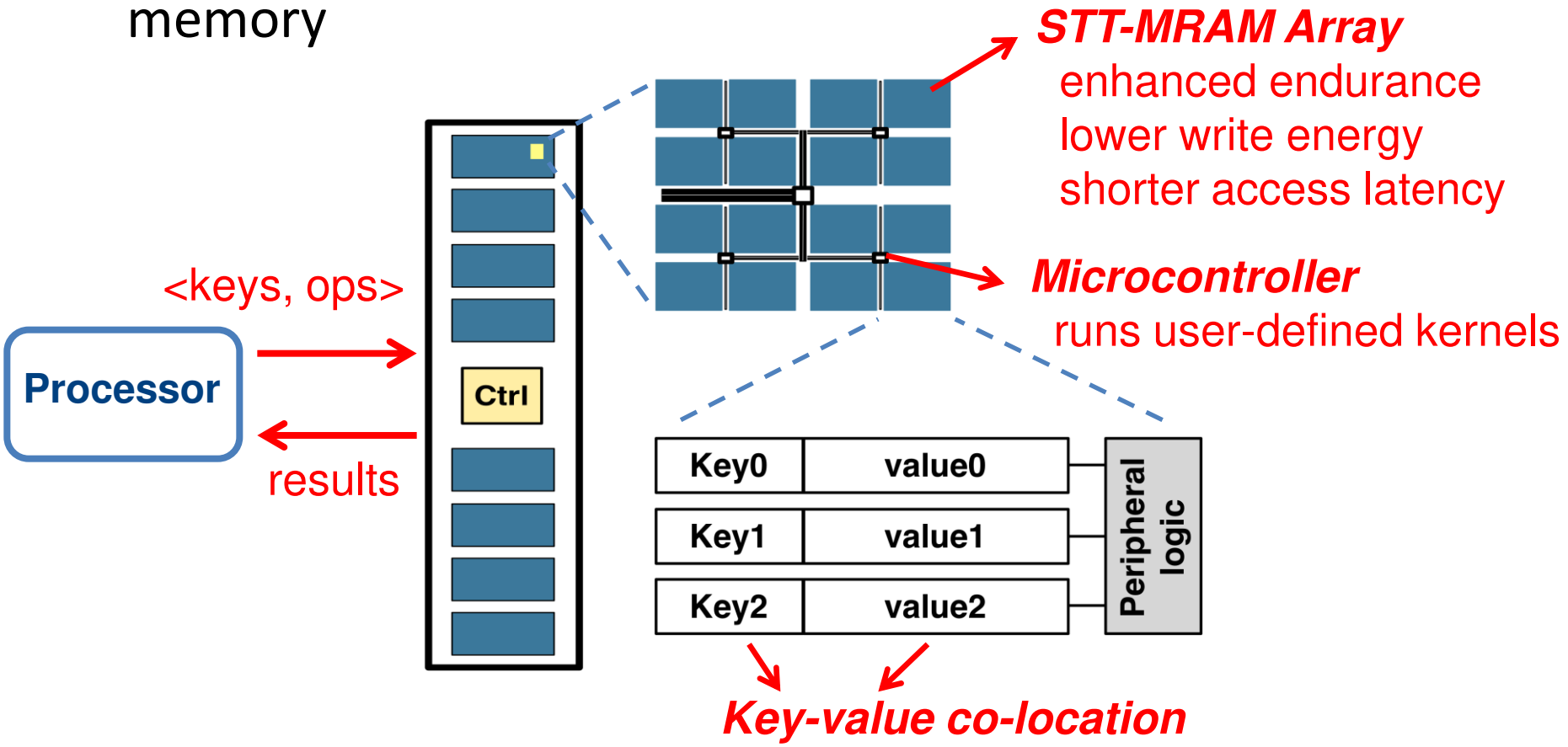


Key (row, col)	Value
(0, 0)	a
(0, 1)	b
(1, 0)	c
(1, 1)	d

L. Potter, "ASC: an associative-computing paradigm", 1994

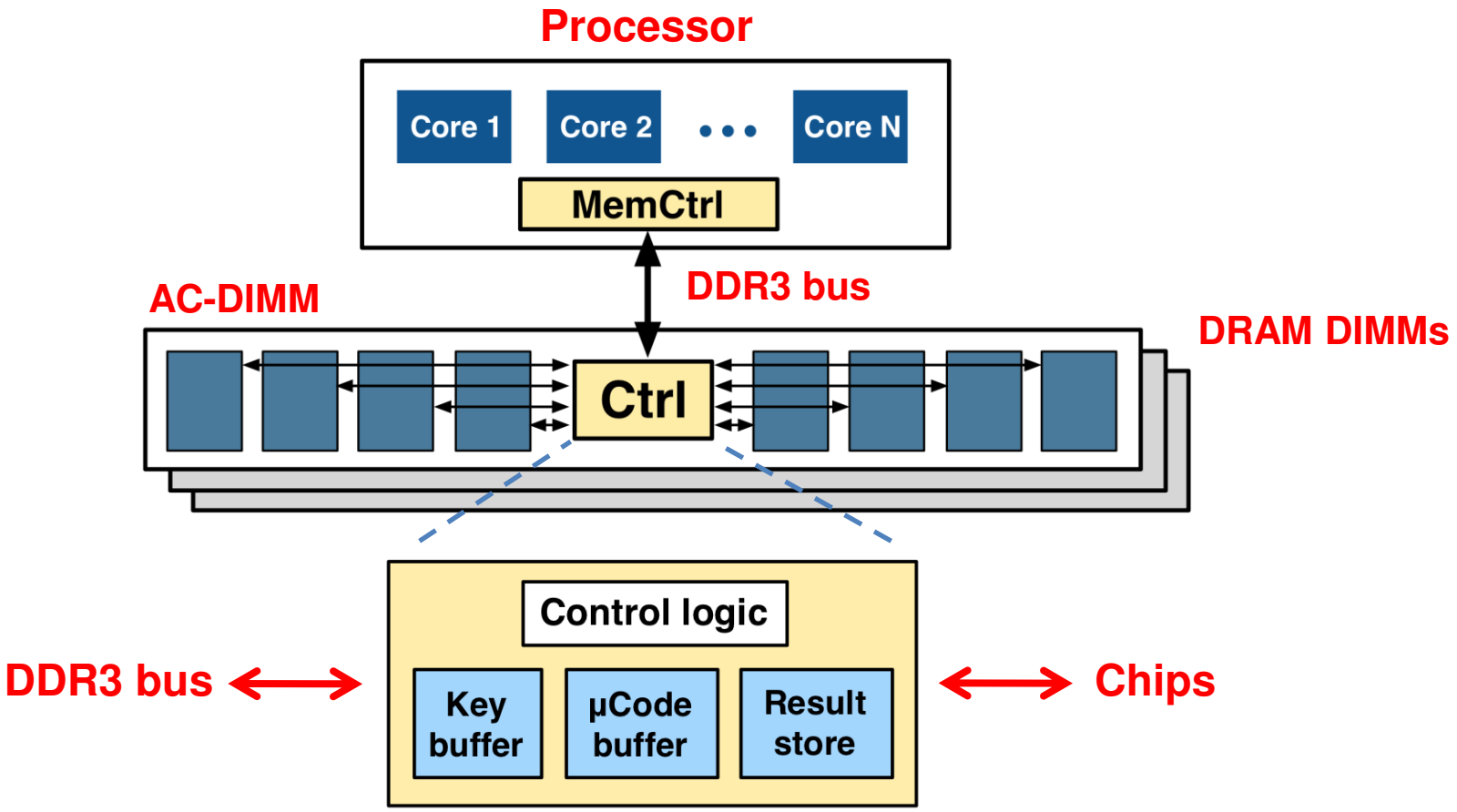
AC-DIMM

- AC-DIMM combines associative lookup and processing in memory



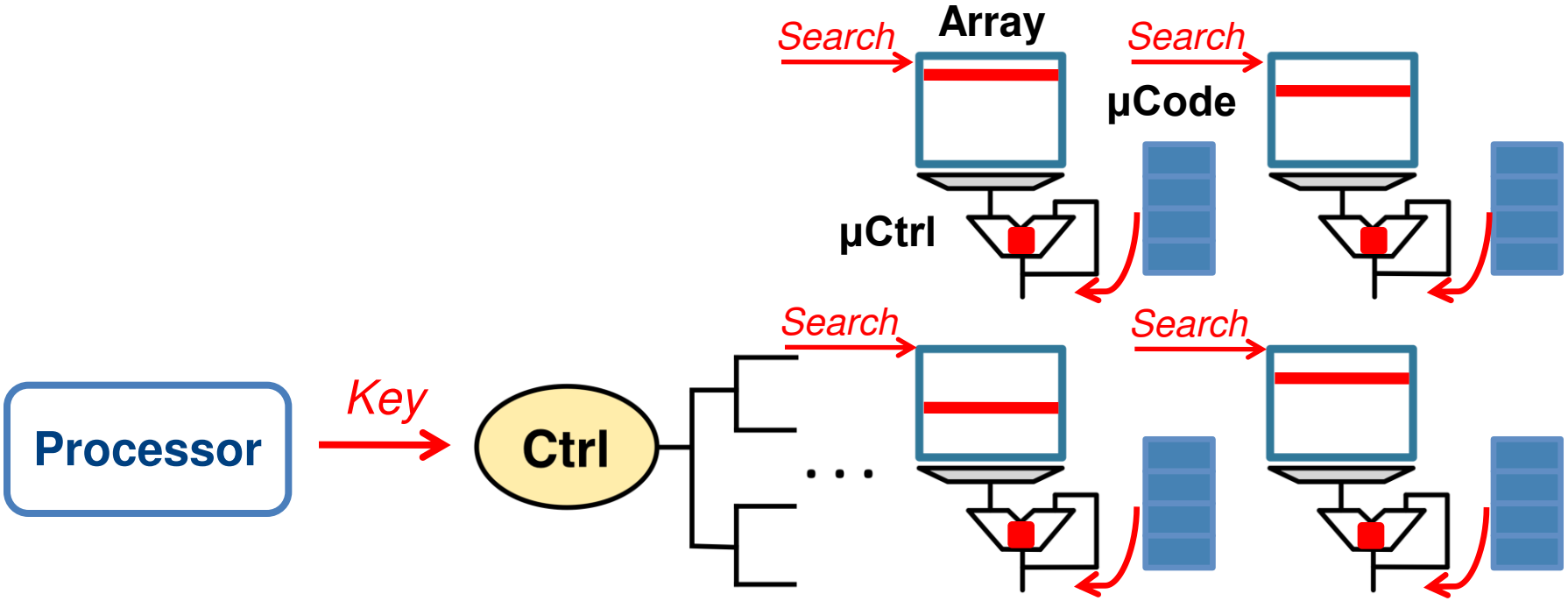
System Interface

- AC-DIMM is a DDR3 compatible module



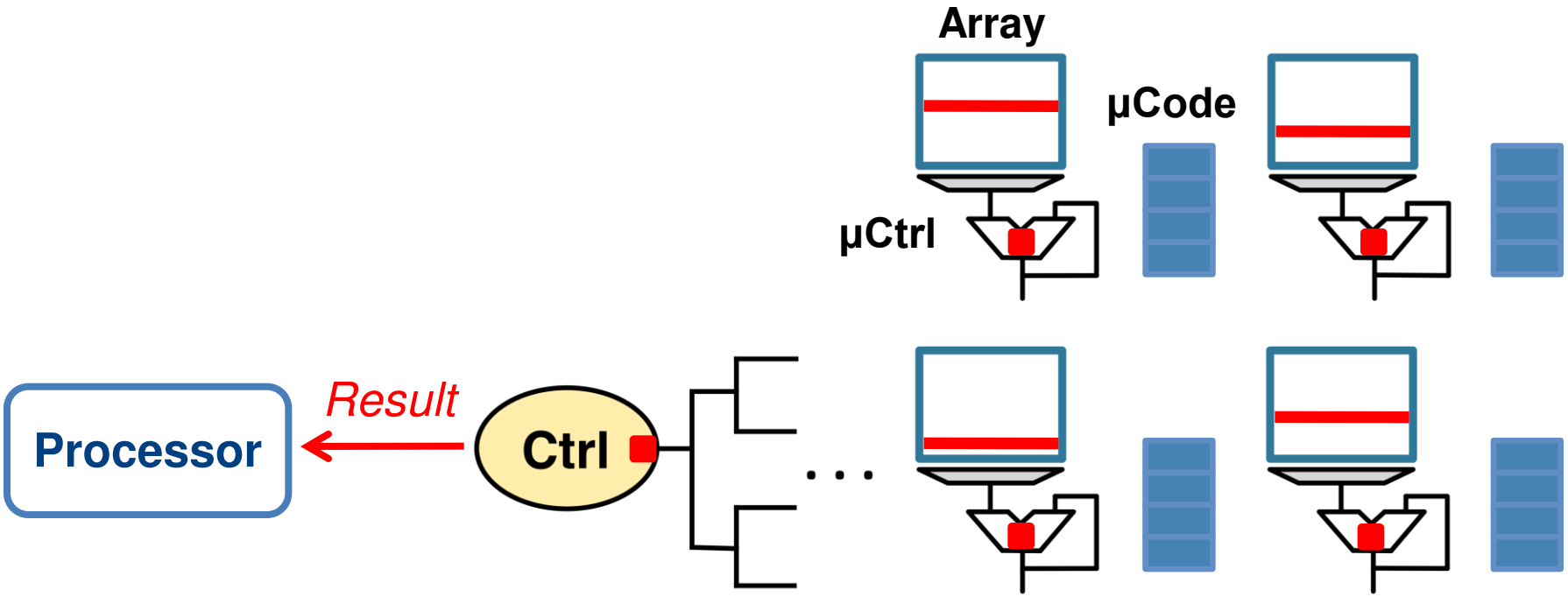
Programming Model

- Program accesses AC-DIMM via a user-level library



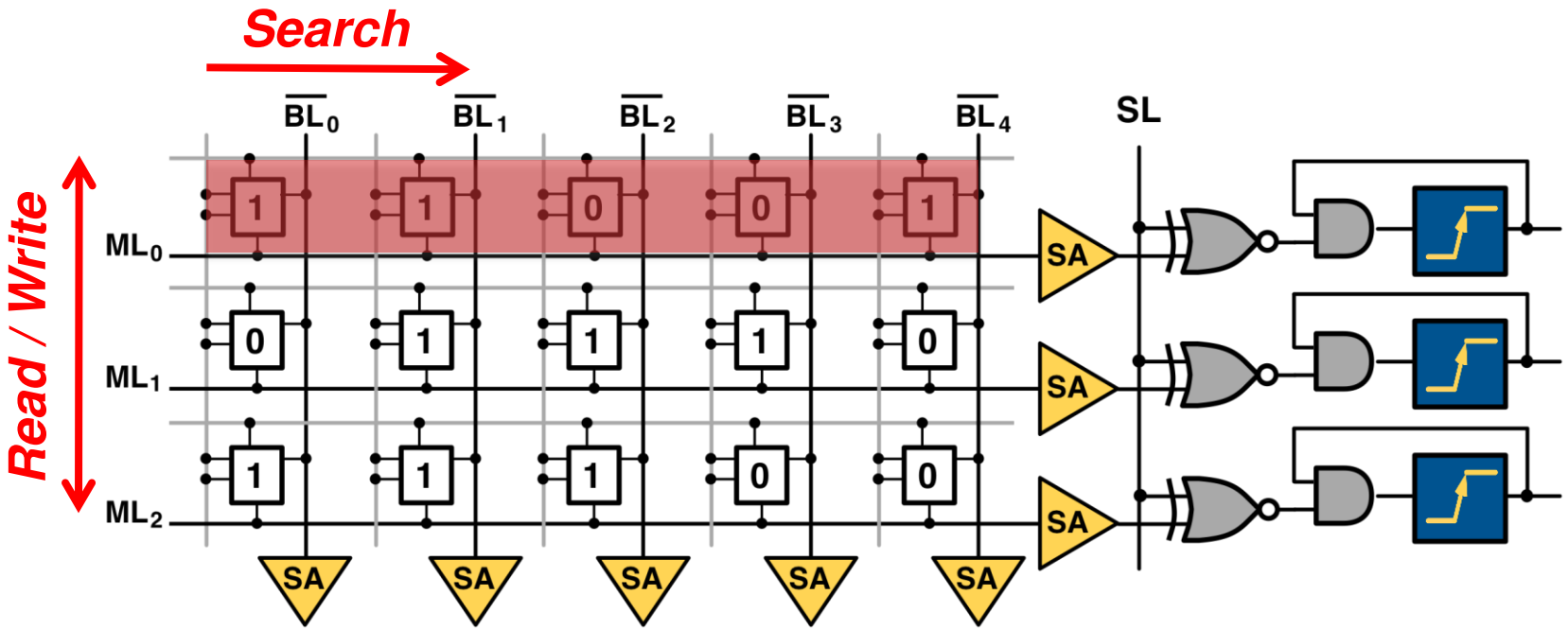
Programming Model

- Program accesses AC-DIMM via a user-level library



Array Organization

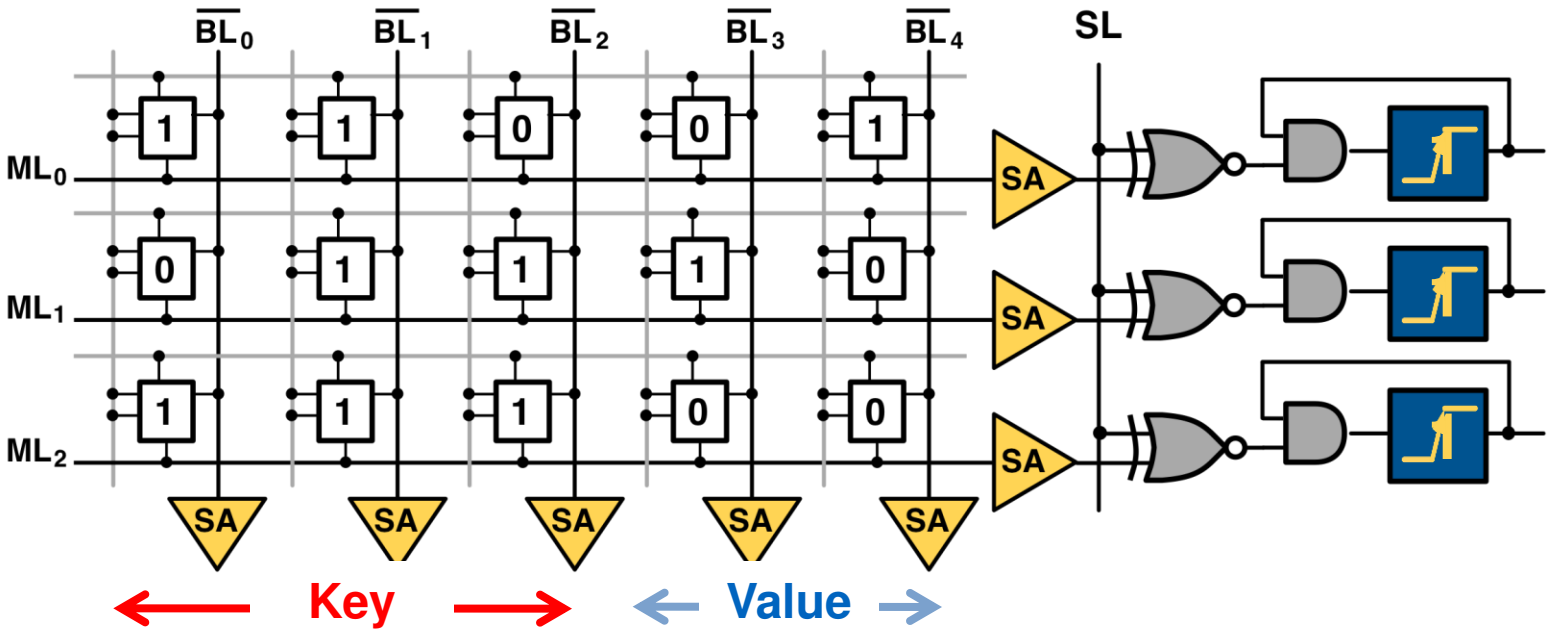
- Memory row can be searched, read, and written
- Co-locate key-value pairs in the same row



Bit-Serial Search

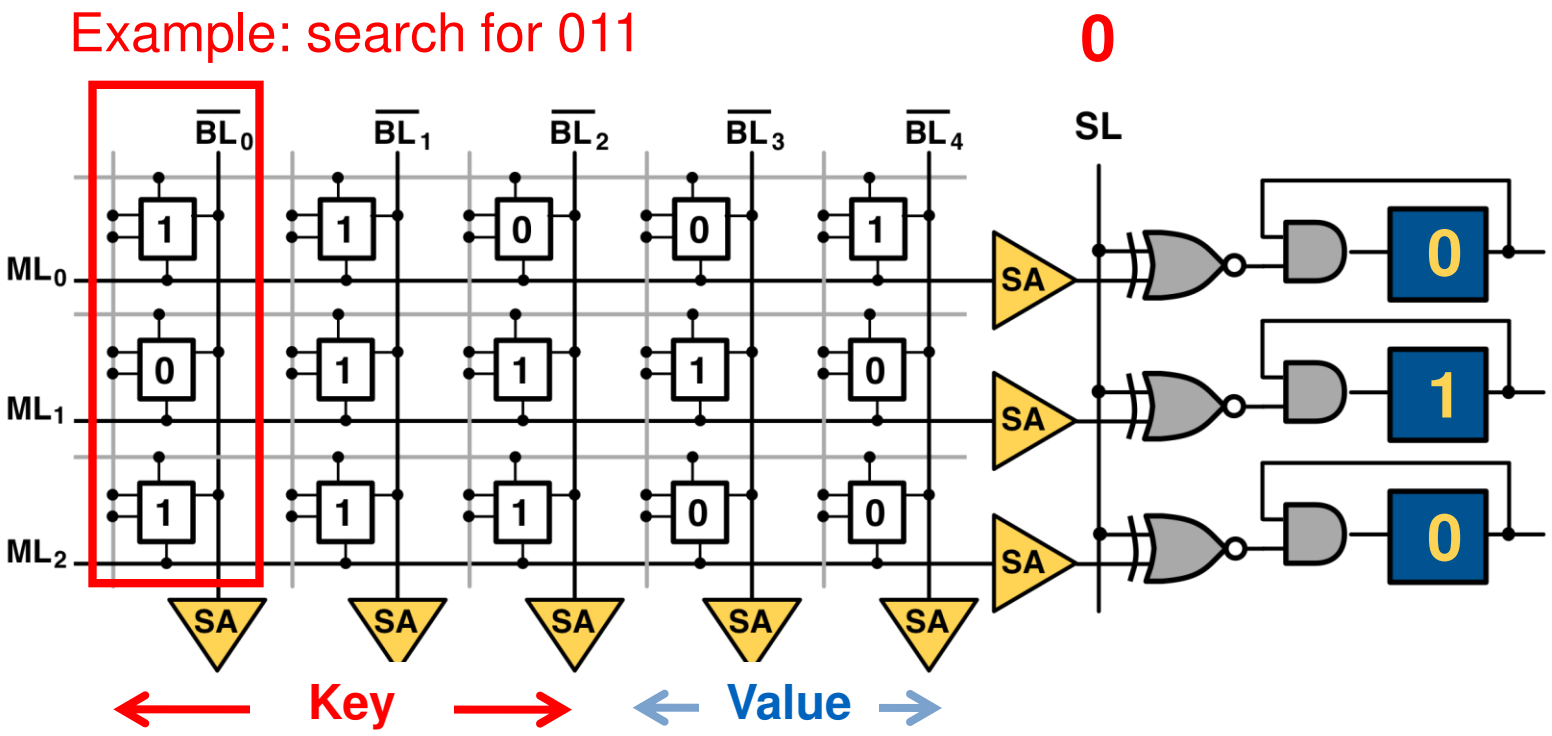
- Progressively searches column-by-column across the array
- Improves power efficiency and simplifies cell structure

Example: search for 011



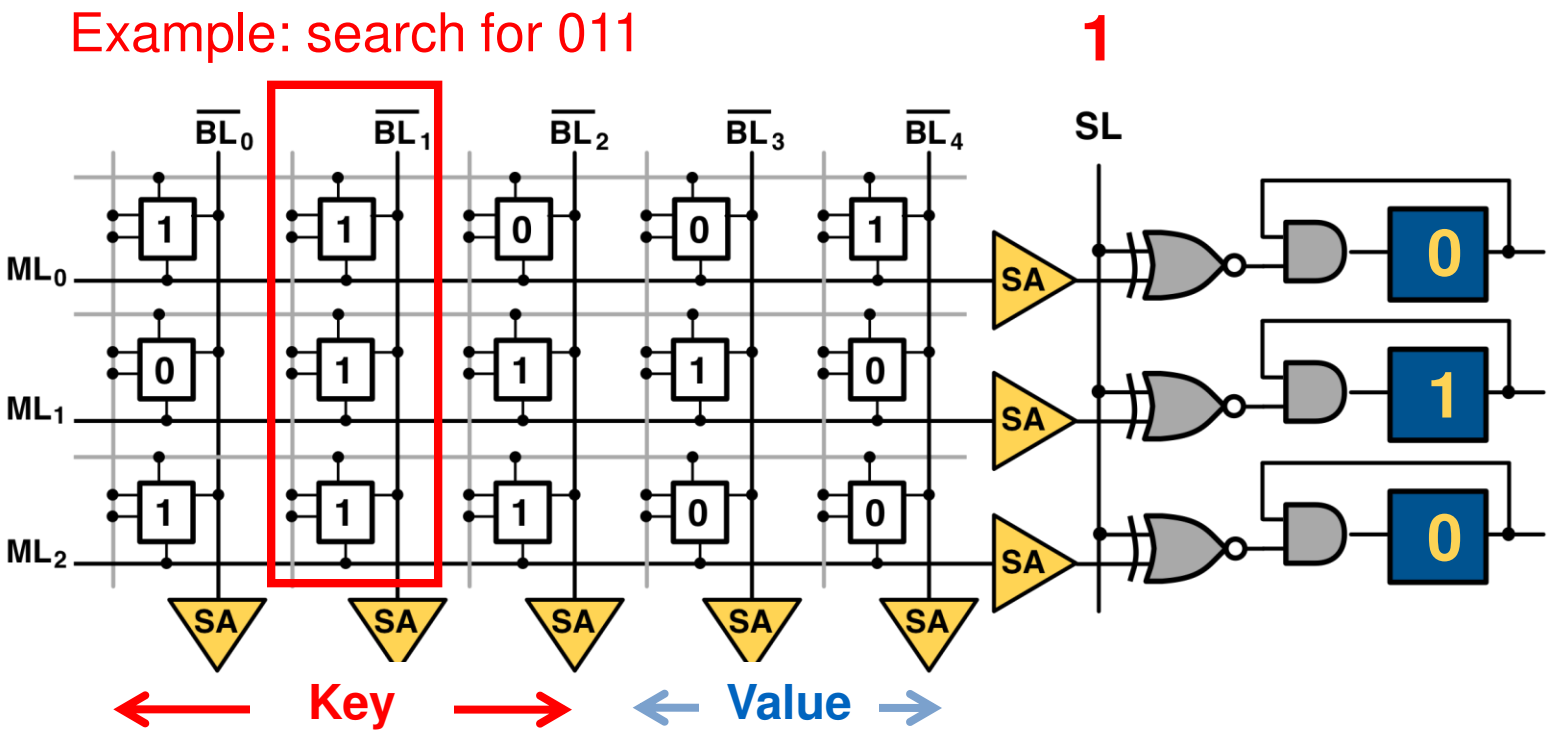
Bit-Serial Search

- Progressively searches column-by-column across the array
- Improves power efficiency and simplifies cell structure



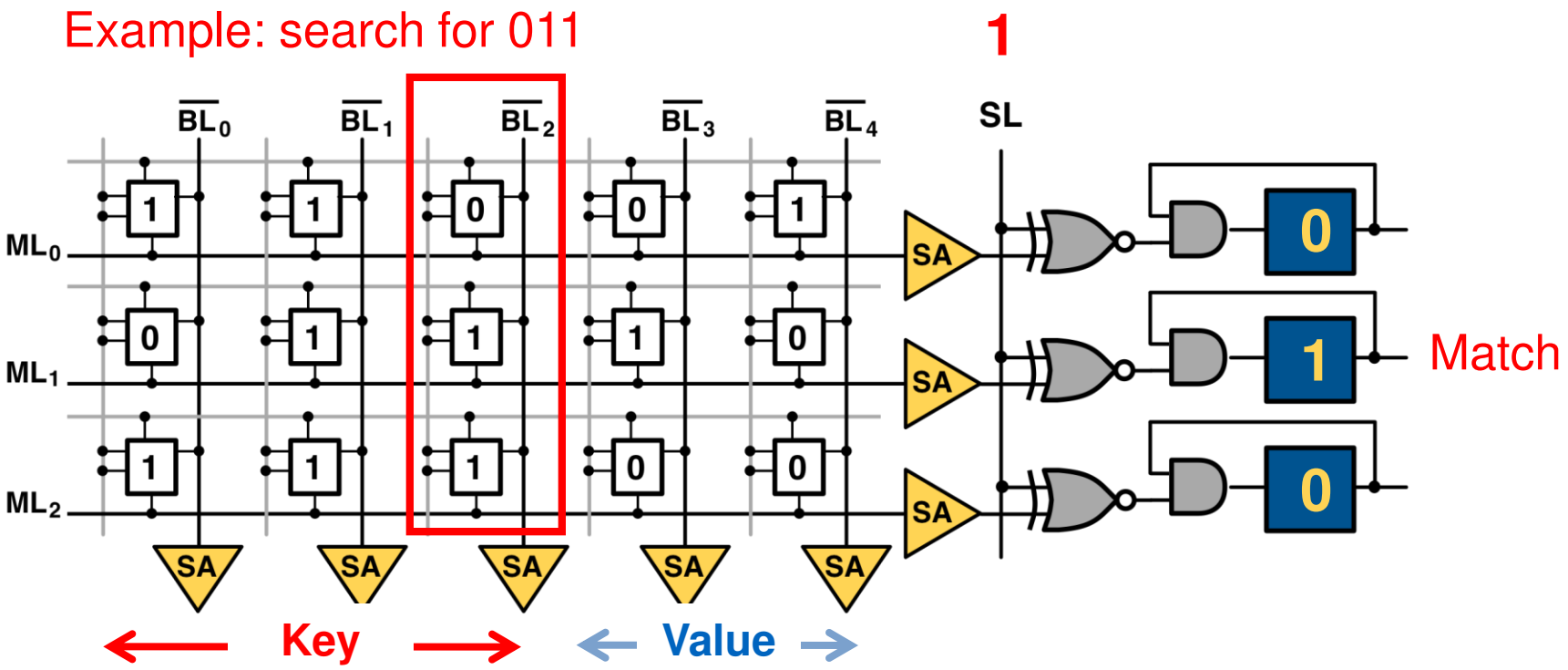
Bit-Serial Search

- Progressively searches column-by-column across the array
- Improves power efficiency and simplifies cell structure



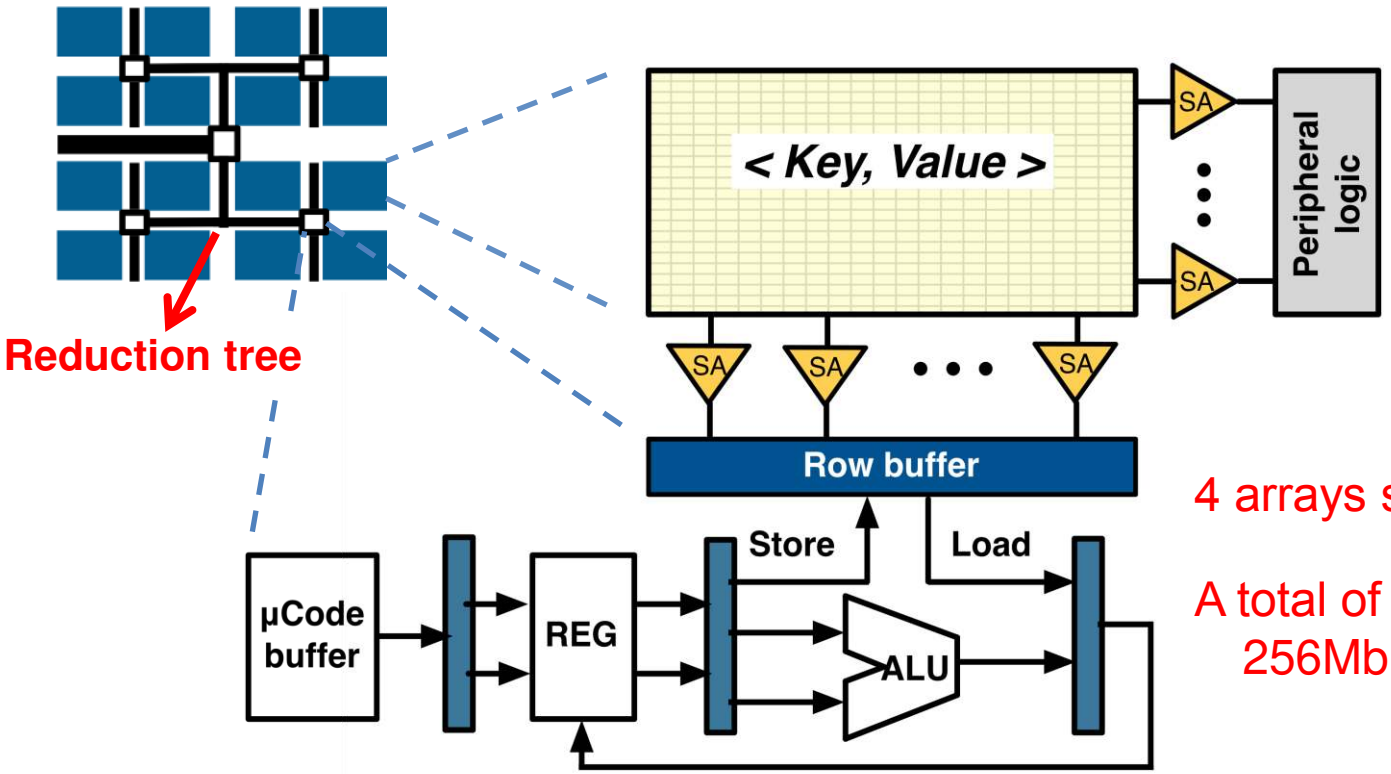
Bit-Serial Search

- Progressively searches column-by-column across the array
- Improves power efficiency and simplifies cell structure



Microcontroller

- Microcontroller runs user-defined kernel on the matching rows



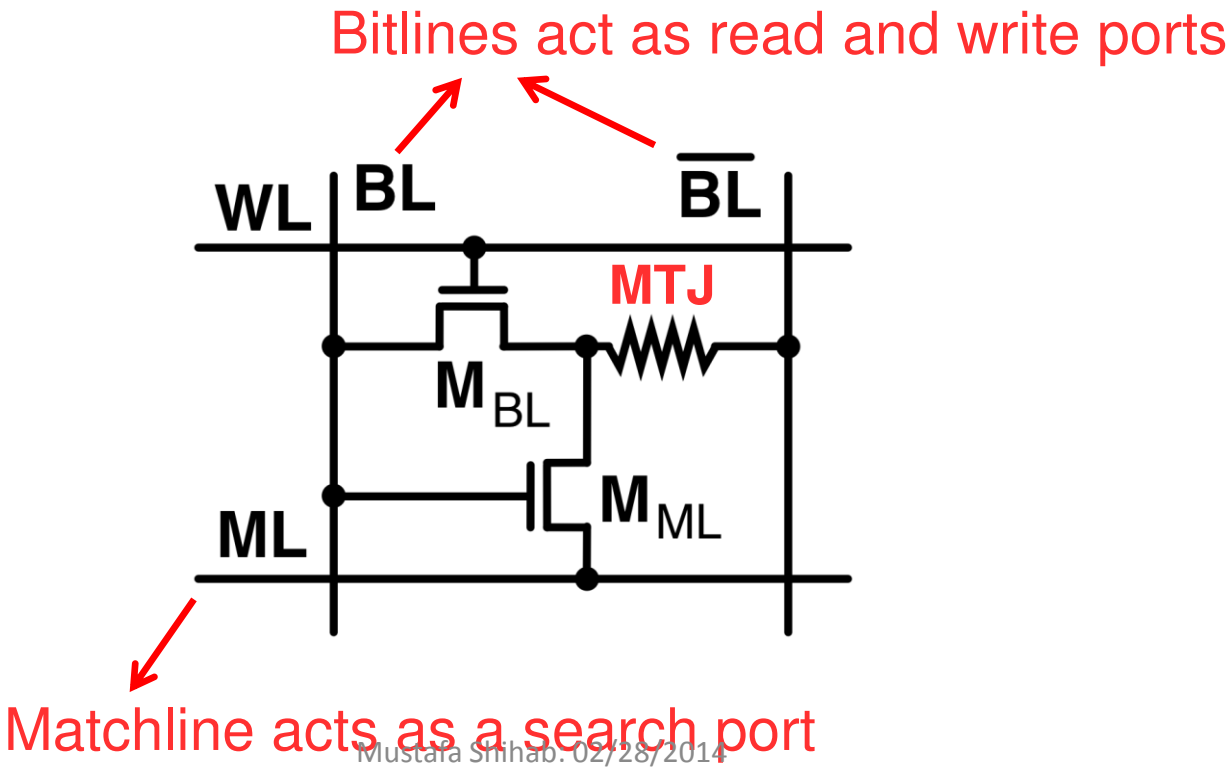
4 arrays share a µController

A total of 64 µControllers on a 256Mb chip (4% area)

AC-DIMM Cell Structure

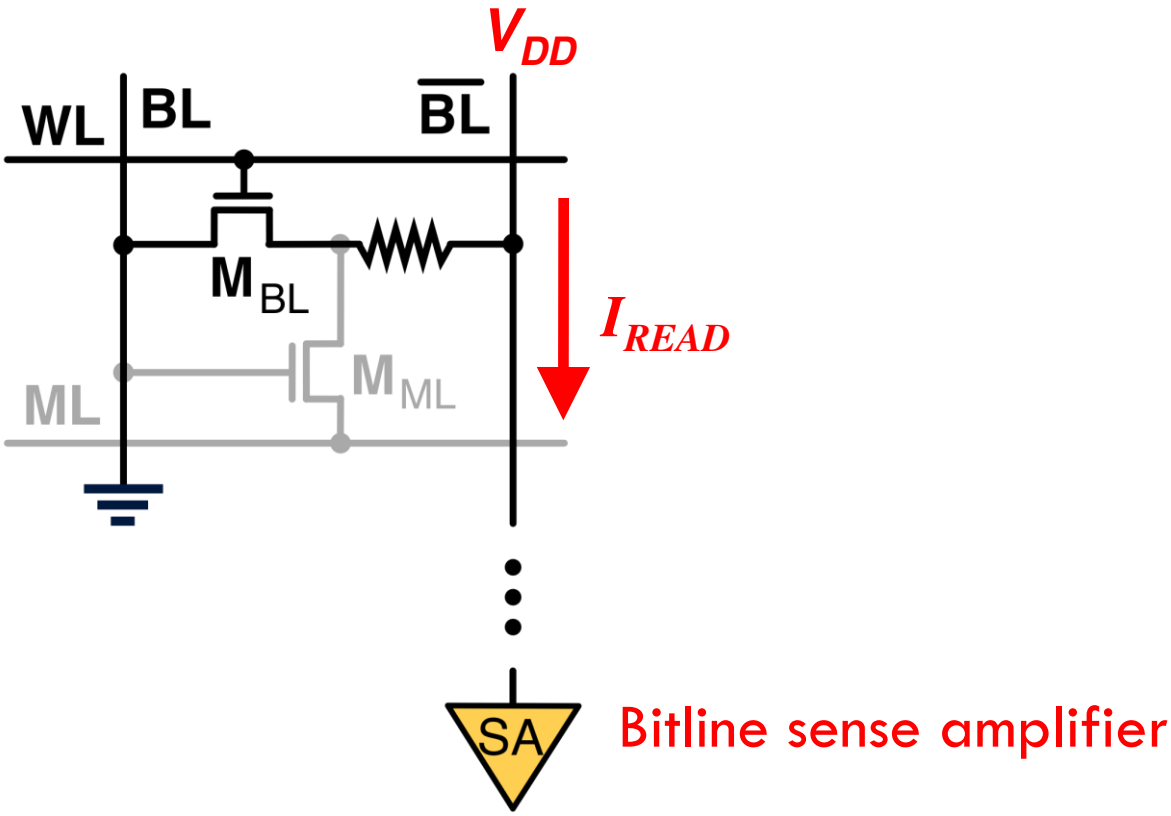
-- 2T1R CAM Cell Using STT-MRAM

- Data is stored in a magnetic tunnel junction (MTJ)



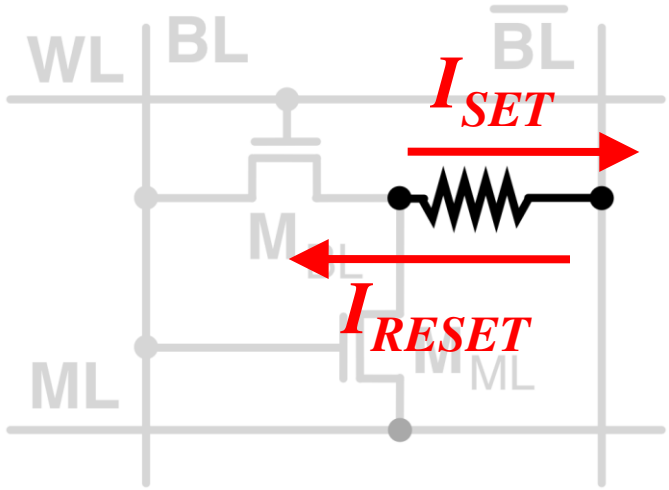
Reading

- Stored data is read by bitline sense amps



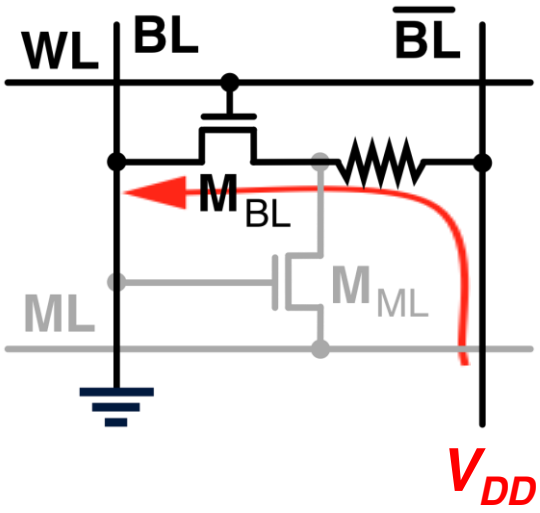
Writing

- Programming an MTJ requires a bi-directional write current

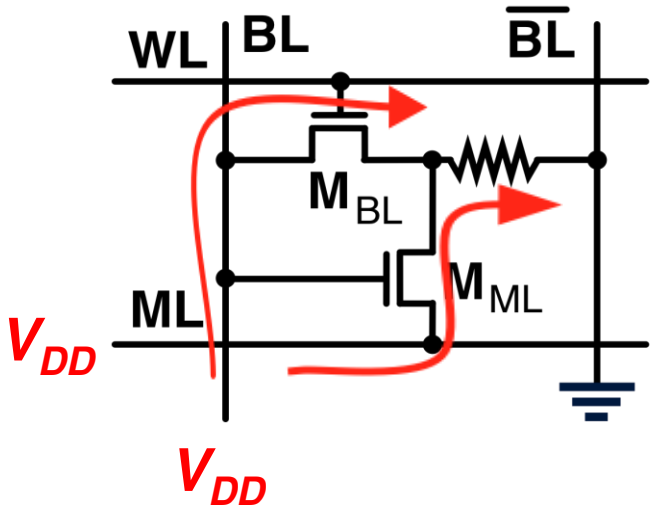


Writing

- Resetting an AC-DIMM cell

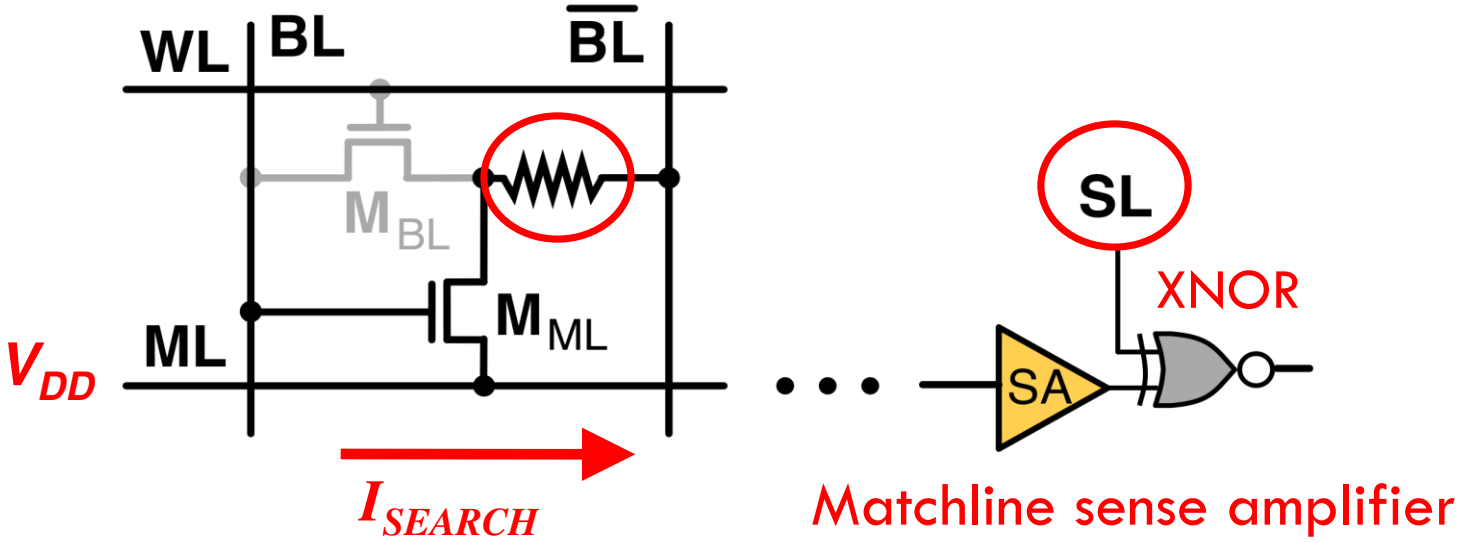


- Setting an AC-DIMM cell



Searching

- Accomplished by reading a column of bits, and comparing against the search key

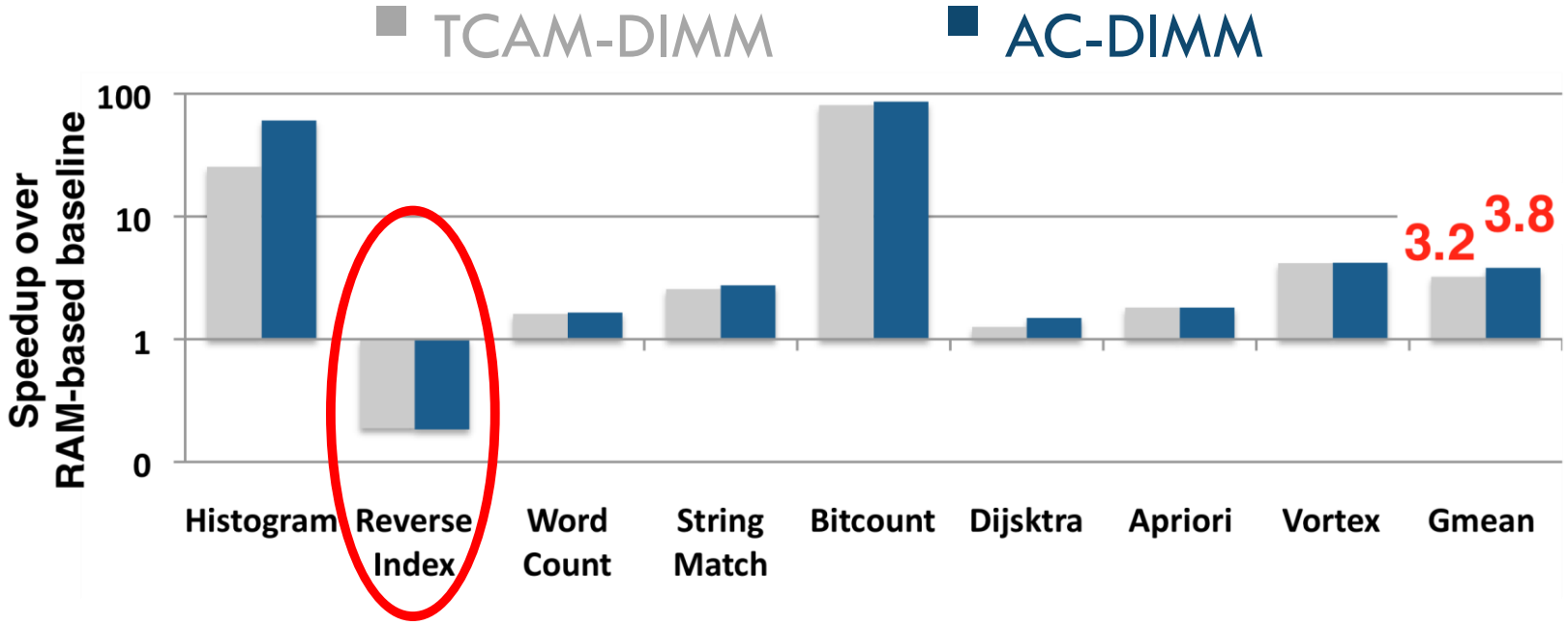


- Outputs a 1 on a match, a 0 otherwise

Experimental Setup

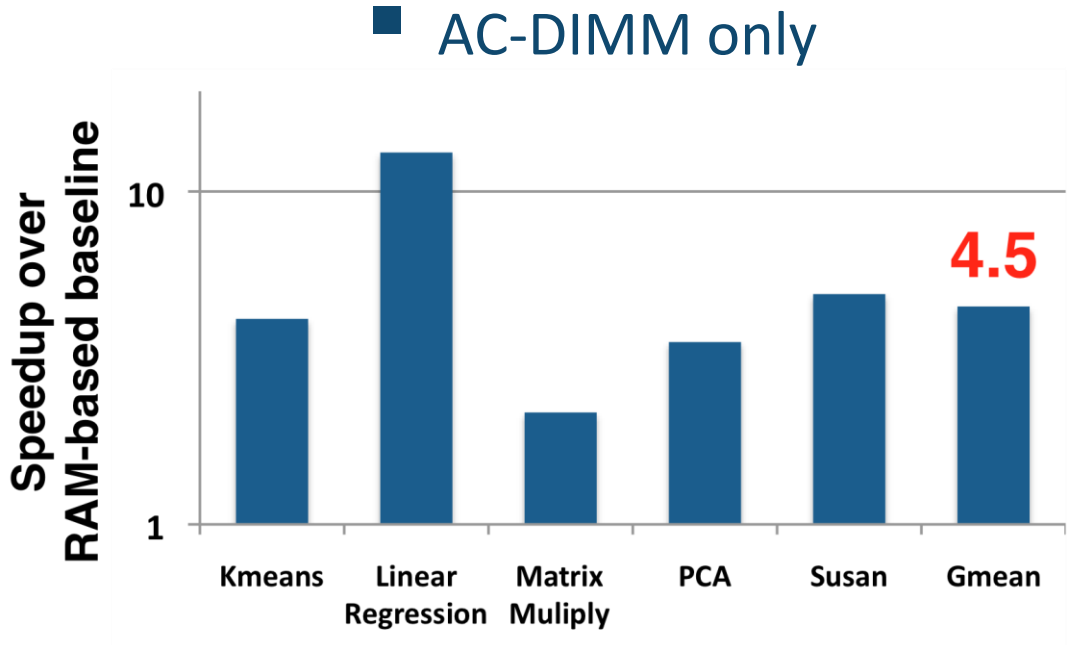
- System configuration
  - Processor: 8 cores, 4GHz
  - Memory bus: DDR3-1066
- Simulation tools
  - Cadence (Spectre), Encounter RTL Compiler with FreePDK
  - SESC simulator
- Applications
  - NuMineBench
  - MiBench
  - Phoenix
  - SPEC INT 2000

System Performance



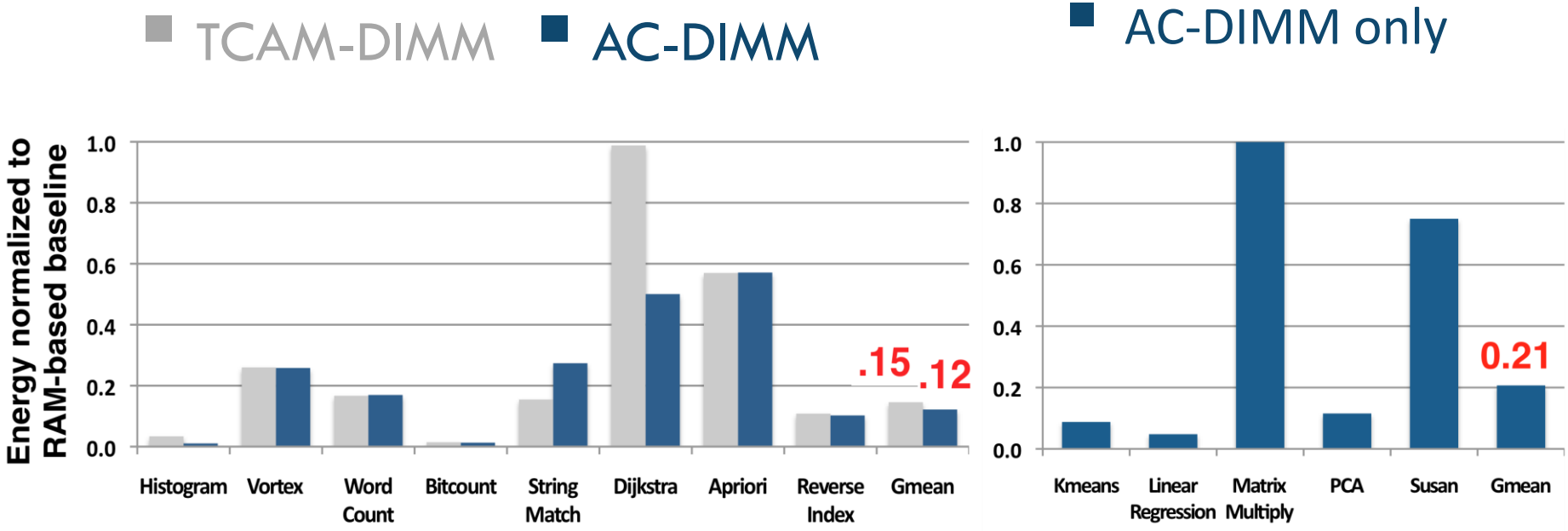
- AC-DIMM outperforms the previous TCAM-DIMM when search key is short (<32 bits)

System Performance



- AC-DIMM caters to a broader range of applications

System Energy



- Dynamic energy saved by eliminating data movement
- Leakage energy saved by reducing execution time

Summary

- AC-DIMM is an STT-MRAM based compute engine
  - DDR3 compatible module
  - Applicable to other RAM-based technologies
  - Integrates programmable microcontrollers
  - Co-locates key-value pairs
- Improves energy and bandwidth efficiency
  - Eliminates unnecessary data movement
  - Reduces instruction and address processing overheads

Thank you