

# Collaborative Research: Cross-Layer Exploration of Non-Volatile Solid-State Memories to Achieve Effective I/O Stack for High-Performance Computing Systems

Tao Li

University of Florida

Xubin He

Tennessee Tech University

Tong Zhang

Rensselaer Polytechnic Institute

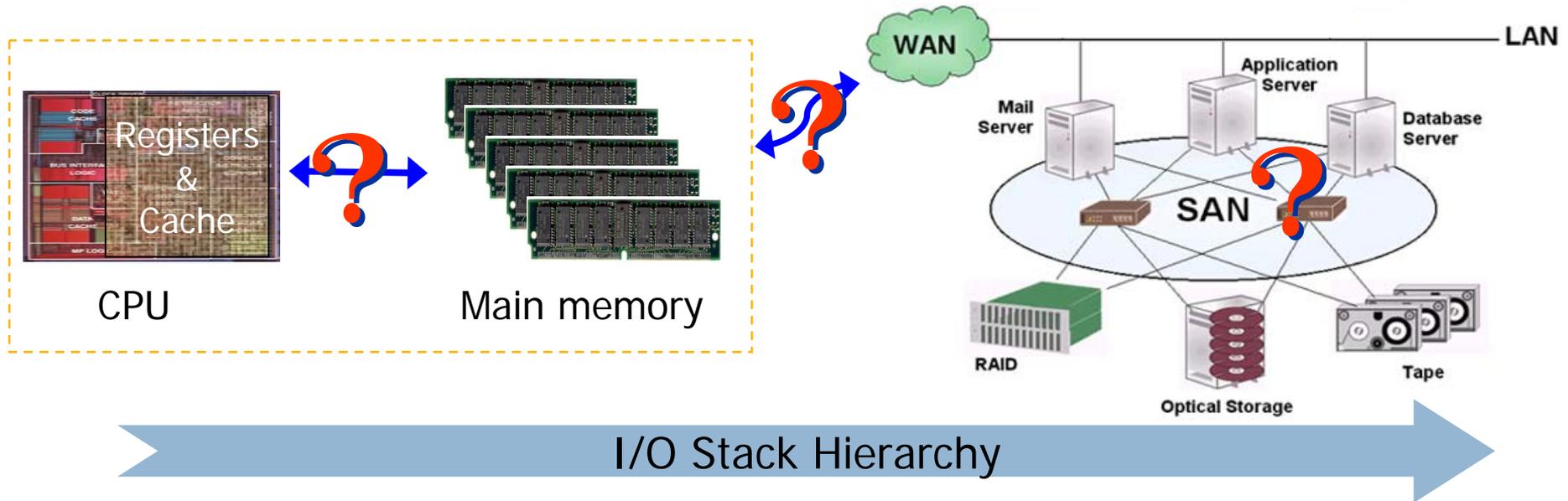


NSF HECURA (Sept. 2009 ~ Aug. 2012)

# Motivation

2

- **I/O stack:** one of the most critical performance bottlenecks in high-performance computing systems

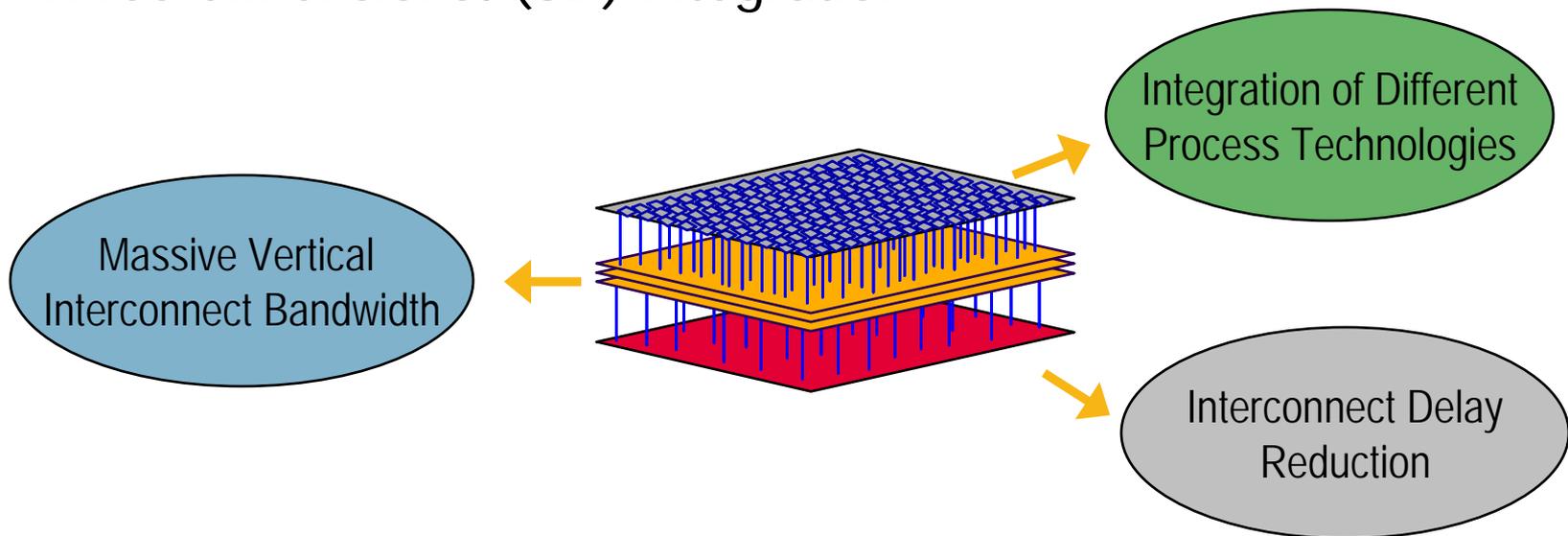


Cohesive research project exploiting emerging **storage** and **integration** technologies

# Motivation

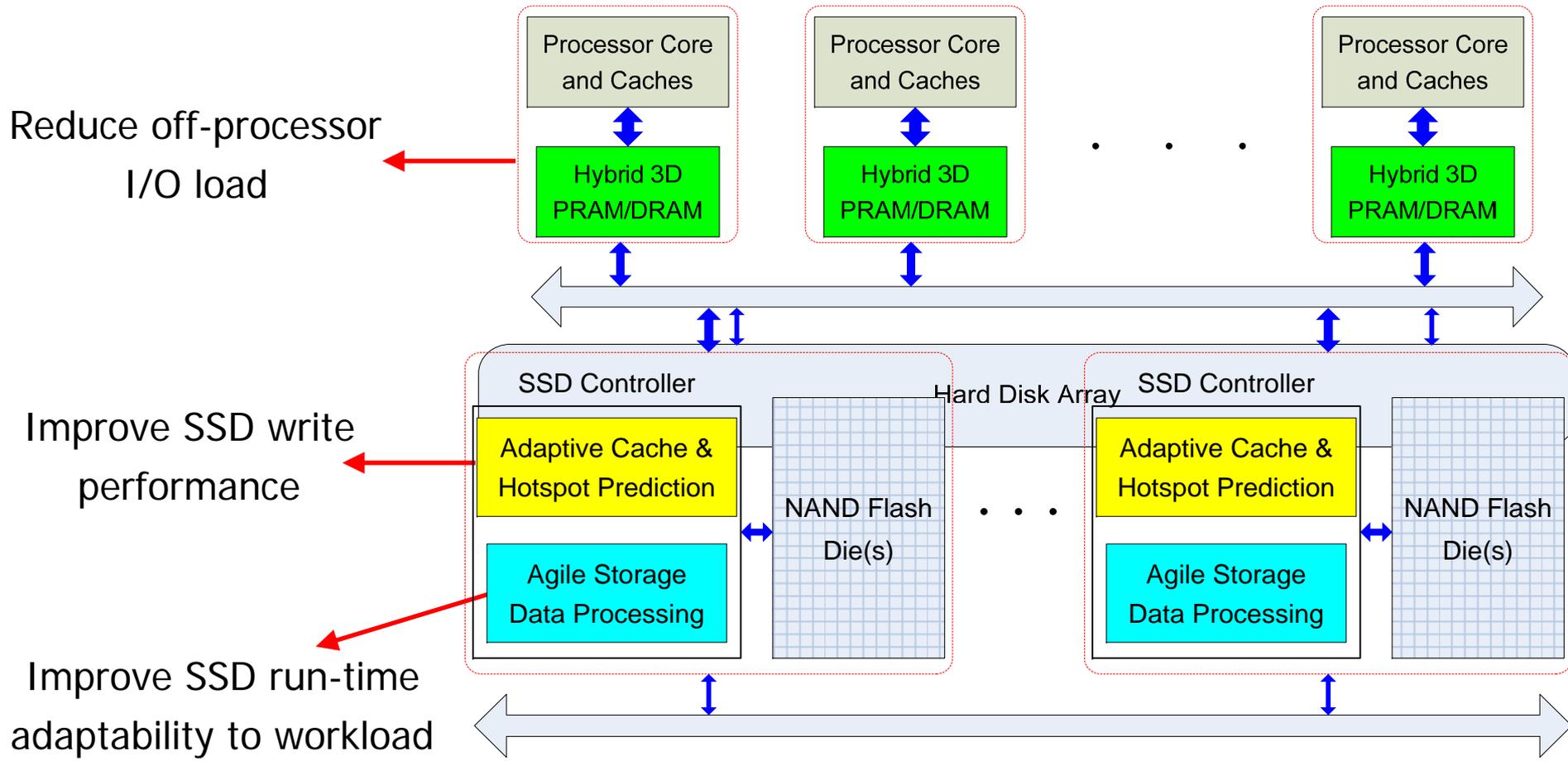
3

- High-density non-volatile solid-state memories
  - NAND flash memory: mainstream technology, scalable to  $\sim 22\text{nm}$
  - Phase-change memory: emerging technology, scalable to  $< 16\text{nm}$
  - Both support multi-bits per cell  $\rightarrow$  cost effectiveness
- Three-dimensional (3D) integration



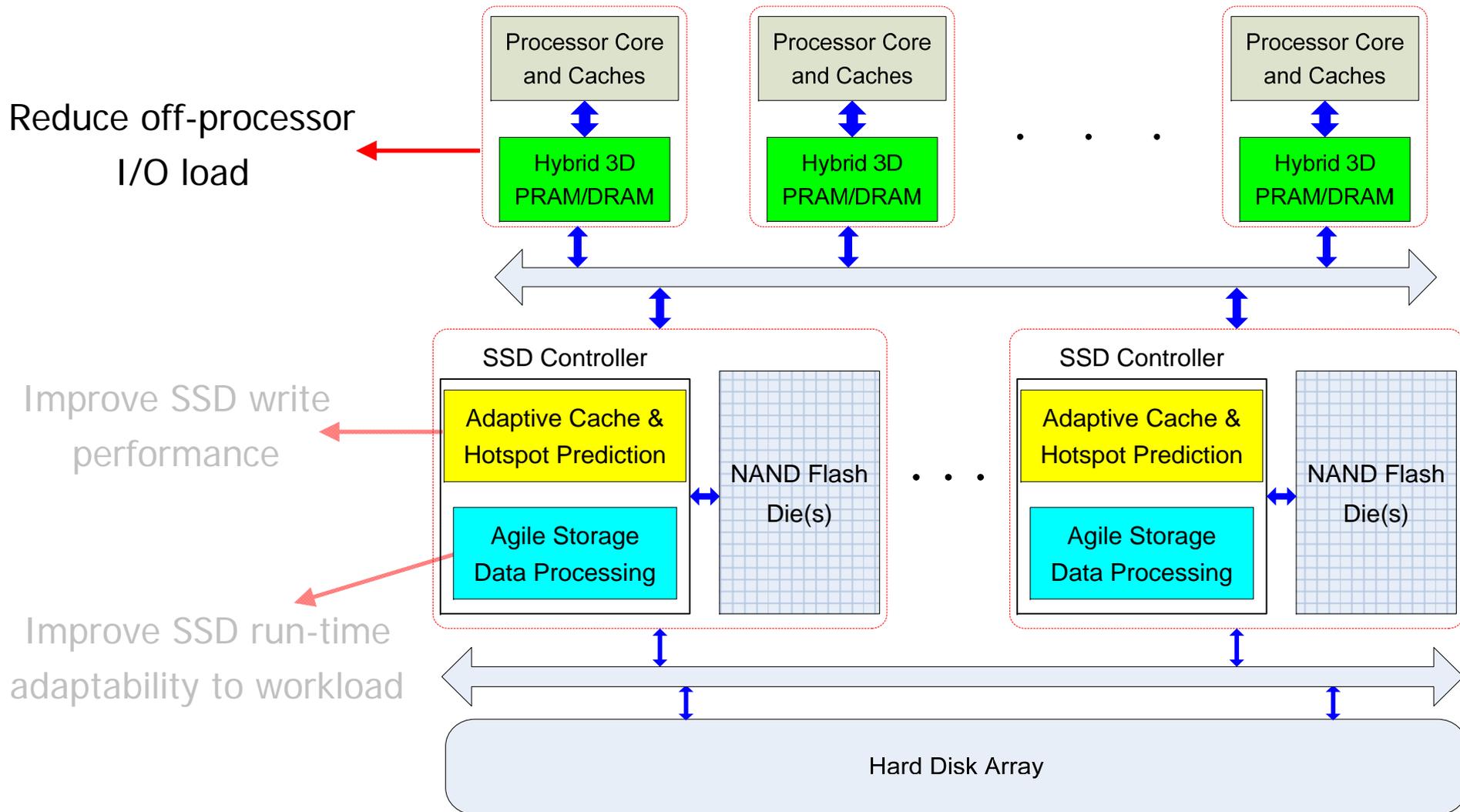
# Outline of Proposed Research

4



# Outline of Proposed Research

5



# PRAM/DRAM 3D Integration to Reduce Off-processor I/O Load

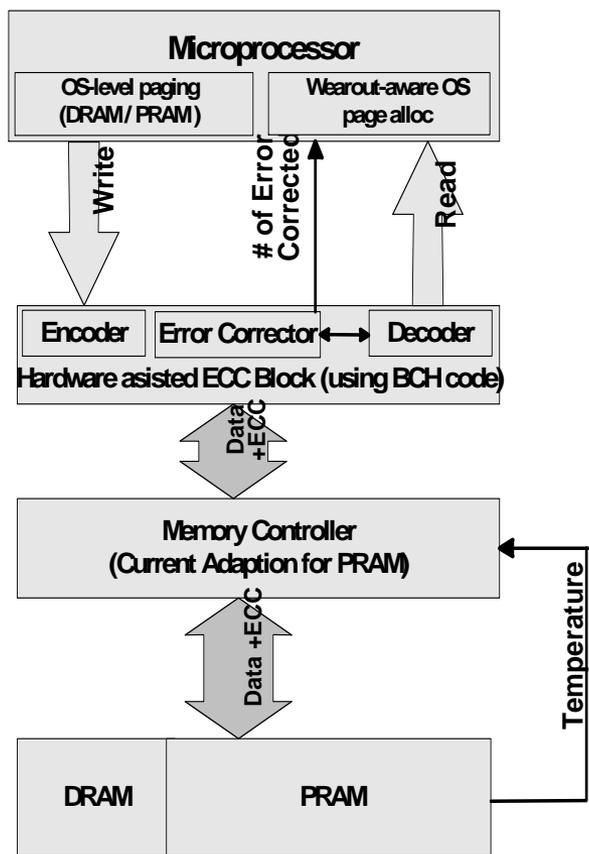
6

## □ Rationale

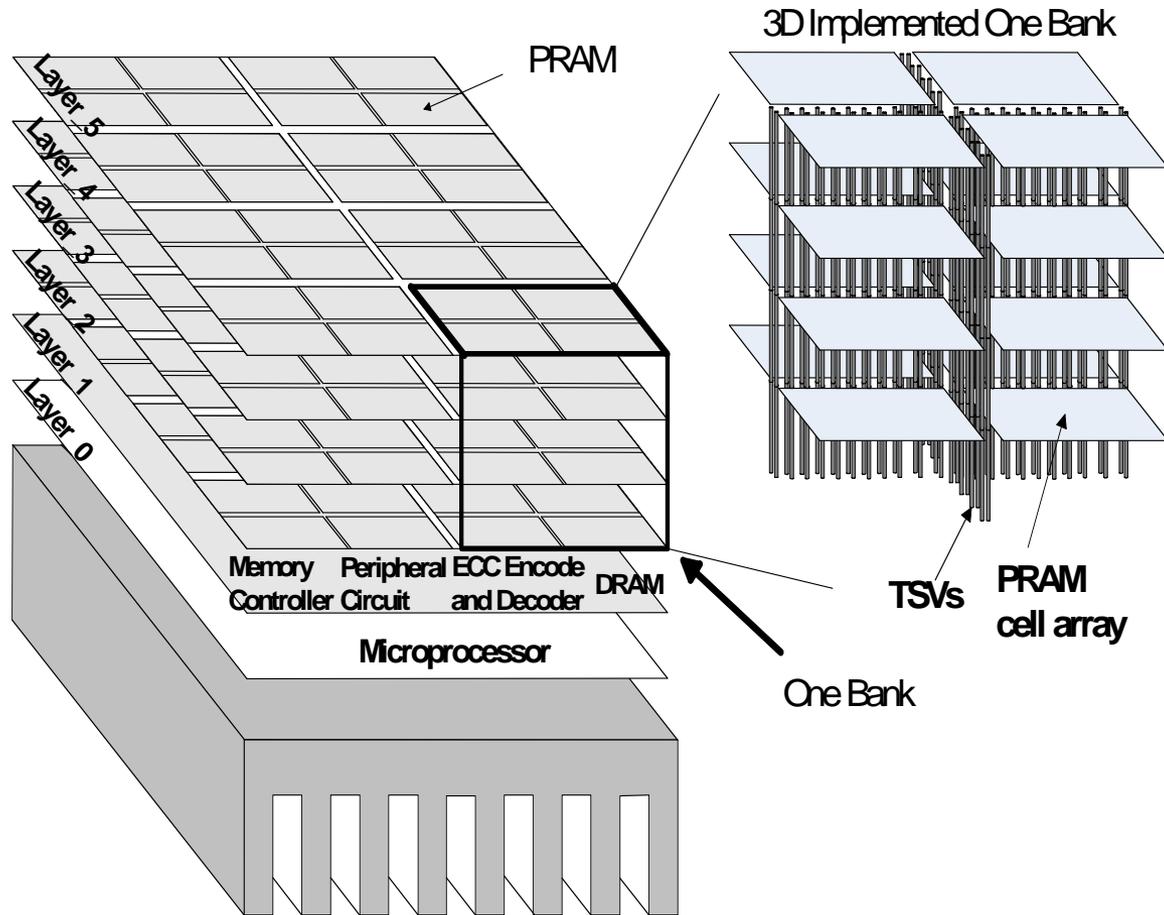
- Efficient main memory system (#of I/O operations↓ and I/O latency↓)
- Three-dimensional (3D) integration allows stacking main memory on top of the microprocessor (memory latency↓ and bandwidth constraints↓)
- DRAM technologies are facing both scalability and power issues
  - The elevated chip temperature results in an exponential rise in DRAM charge leakage
- Phase change memory (PRAM) is emerging as an attractive DRAM alternative
  - + high-density, thermal-friendly
  - - write latency, endurance

# A Hybrid 3D PRAM/DRAM Memory Architecture

7



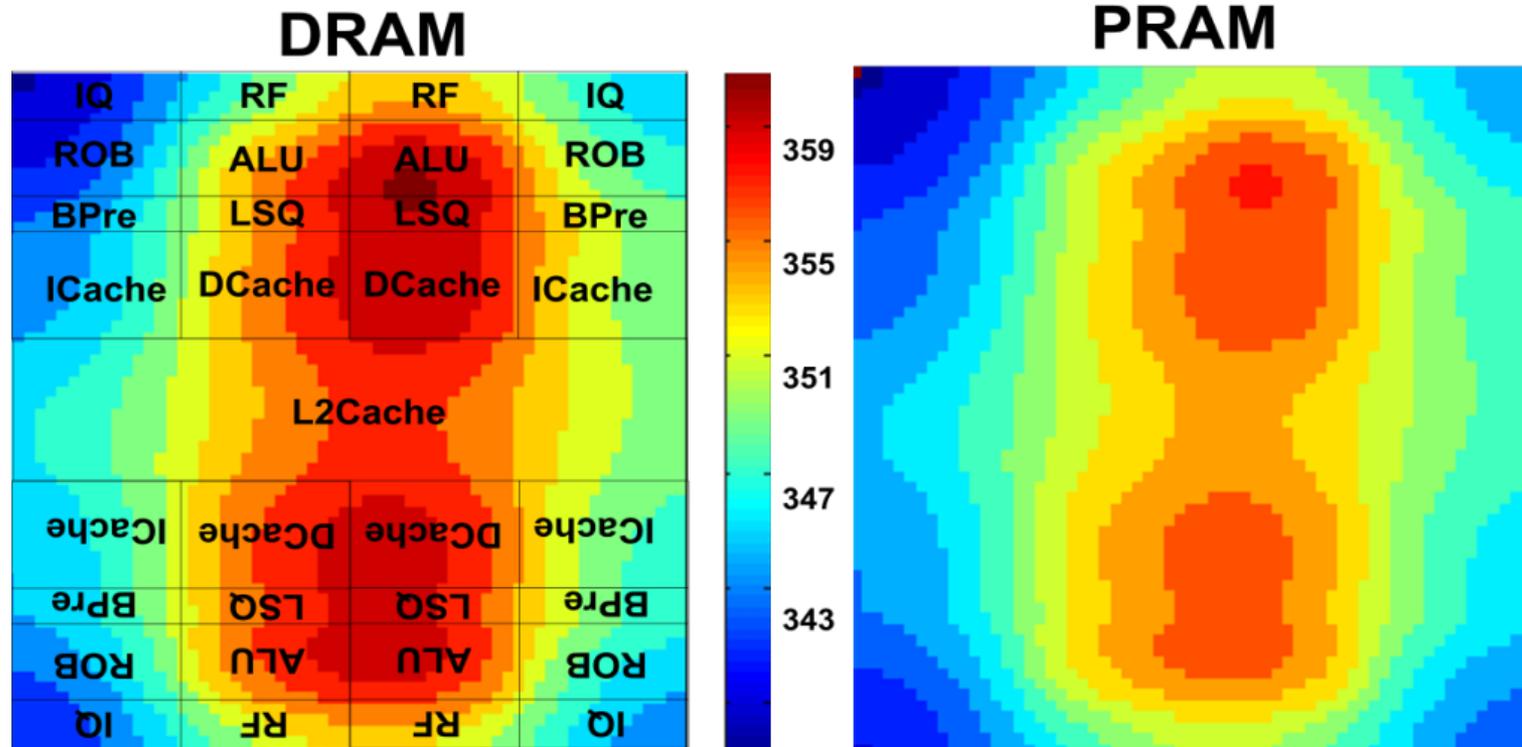
(a)



(b)

# Preliminary Results: Thermal Benefit

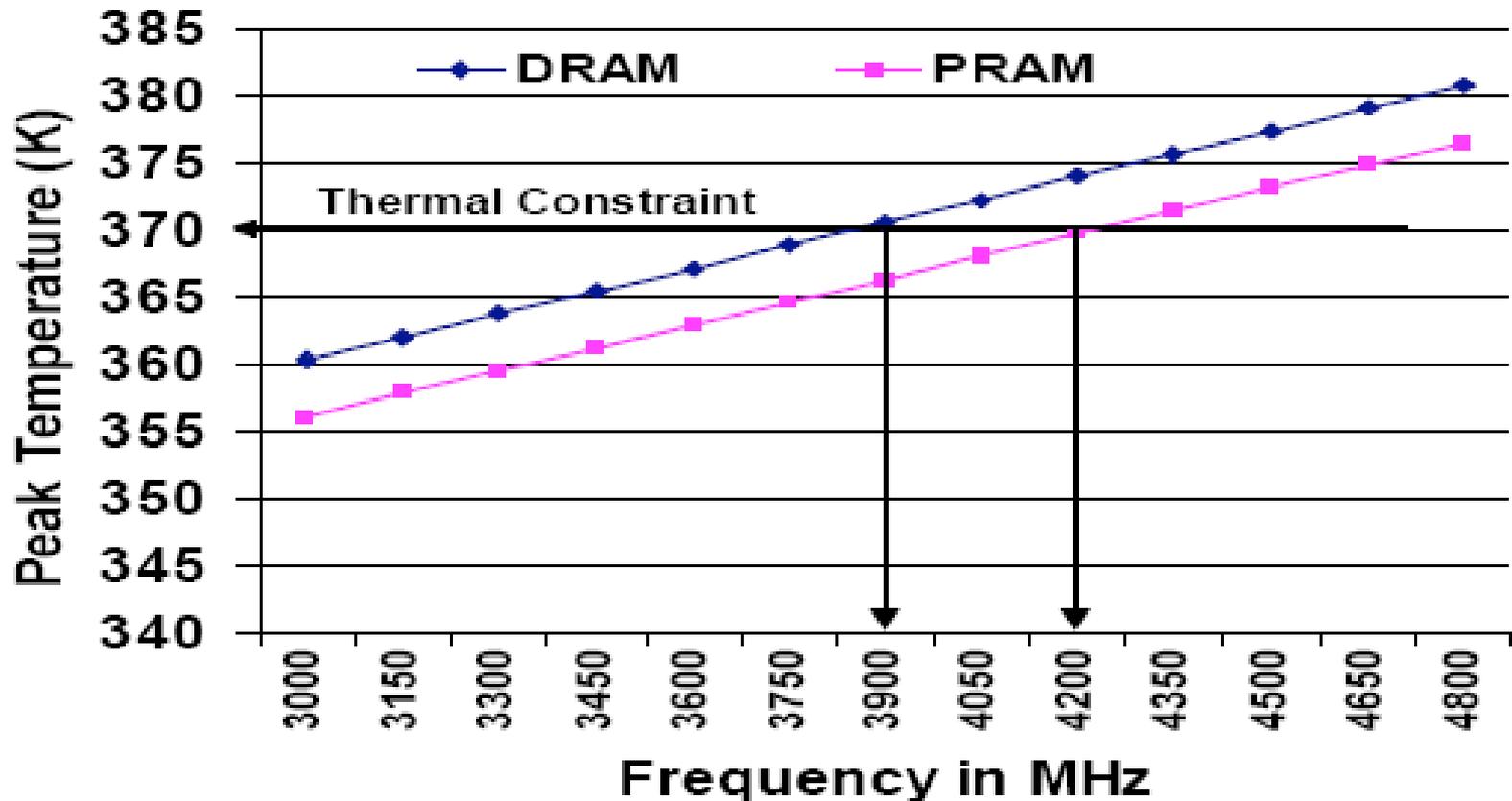
8



- ▶ PRAM alleviates thermal hotspots due to 3D die stacking

# Preliminary Results: Frequency Benefit

9



- ➡ The relief of thermal constraints increases the maximum frequency allowed

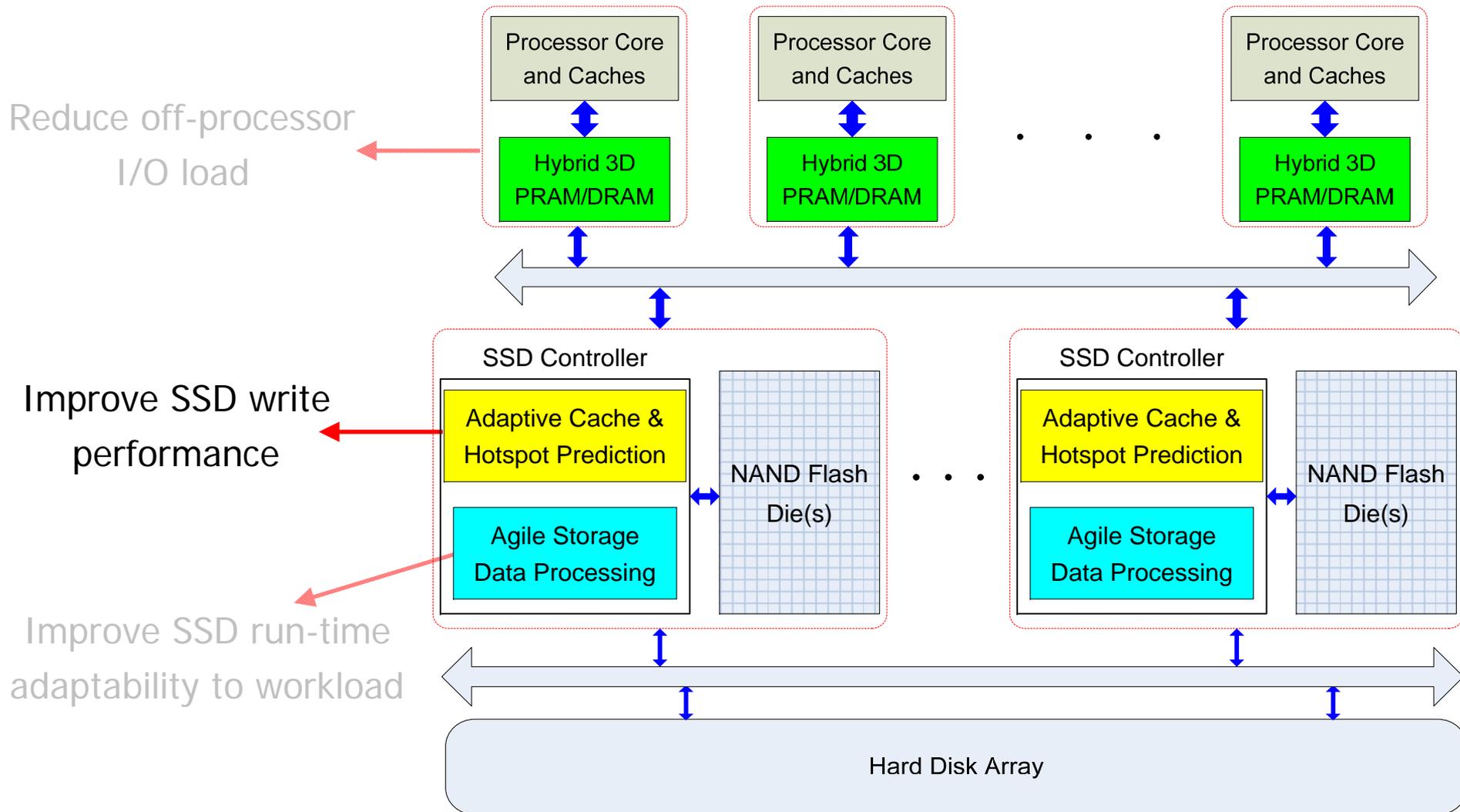
# Research Tasks

10

- ❑ Design space exploration of hybrid 3D PRAM/DRAM architecture
  - ➡ What is the appropriate memory size for each partition?
  - ➡ How to arrange the layers for processor cores/caches, DRAM, PRAM, the memory controller and peripheral circuits?
- ❑ PRAM-aware OS Paging
  - ➡ Favor PRAM over DRAM when allocating cold-modified pages which are infrequently updated; Allocating hot-modified pages to the DRAM partition to avoid the write latency and mitigate wear-out of PRAM
- ❑ Life Span Optimization using Varying ECC Strength and Wear-Leveling
  - ➡ A synergetic reliability enhancement approach that combines hardware and software wear-out optimizations

# Outline of Proposed Research

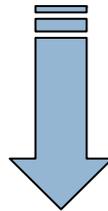
11



# Rationale

12

- Hot pages in SSDs are spread out among many different blocks
- Grouping by block only will bring along unwanted cold pages, since very few hot pages actually reside in each block

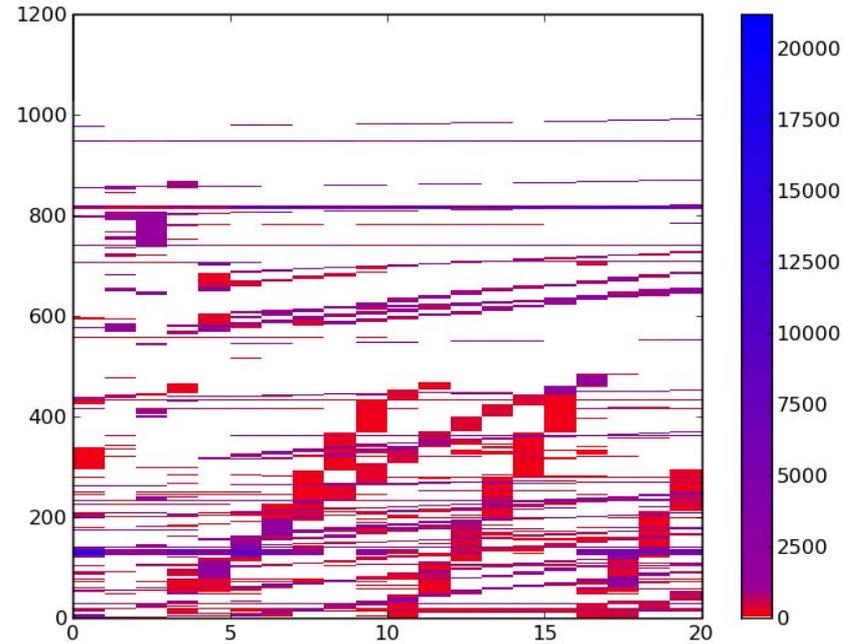
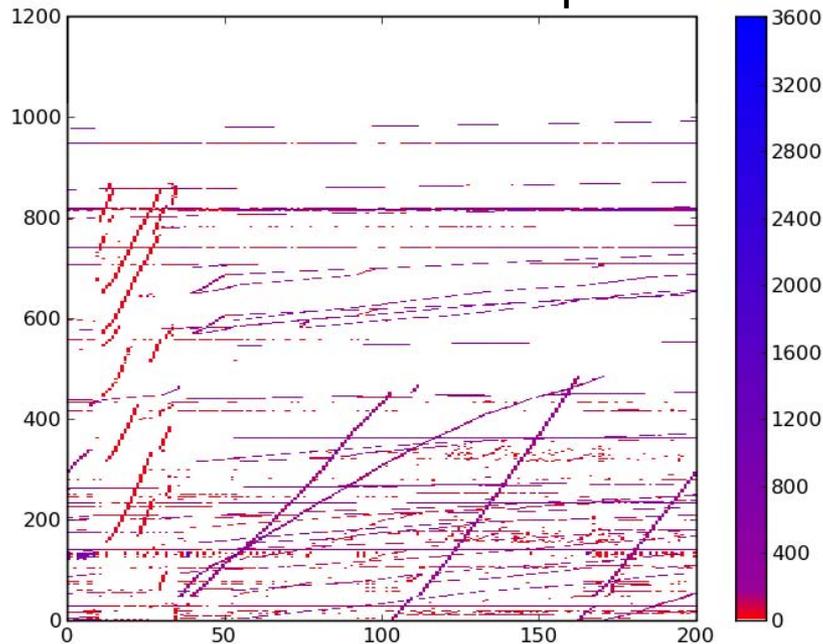


We need a new scheme that can group by pages (to keep hot pages), and evicts by blocks (to evict cold blocks).

# Some Observations

13

- 2D Histogram (heatmap) of temporal distances by block
- Financial 1 trace [courtesy of Ken Bates (HP) and Bruce McNutt (IBM), <http://traces.cs.umass.edu/index.php/Storage/Storage>]:
  - There are 1930249 write requests & 113561 unique pages in the trace
  - There are 110481 pages that get used more than once
  - There are 891 unique blocks



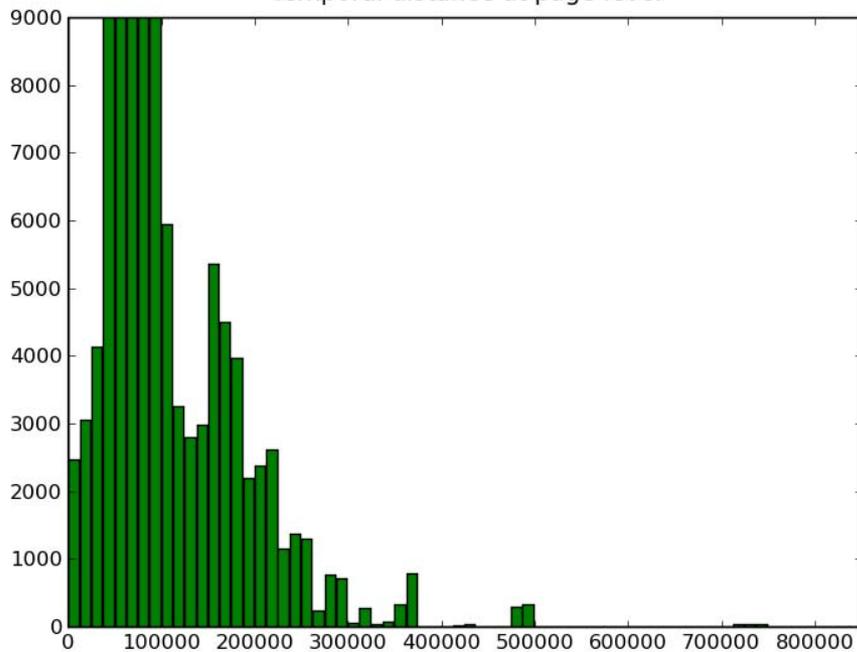
# Observation #1: Temporal Distance

14

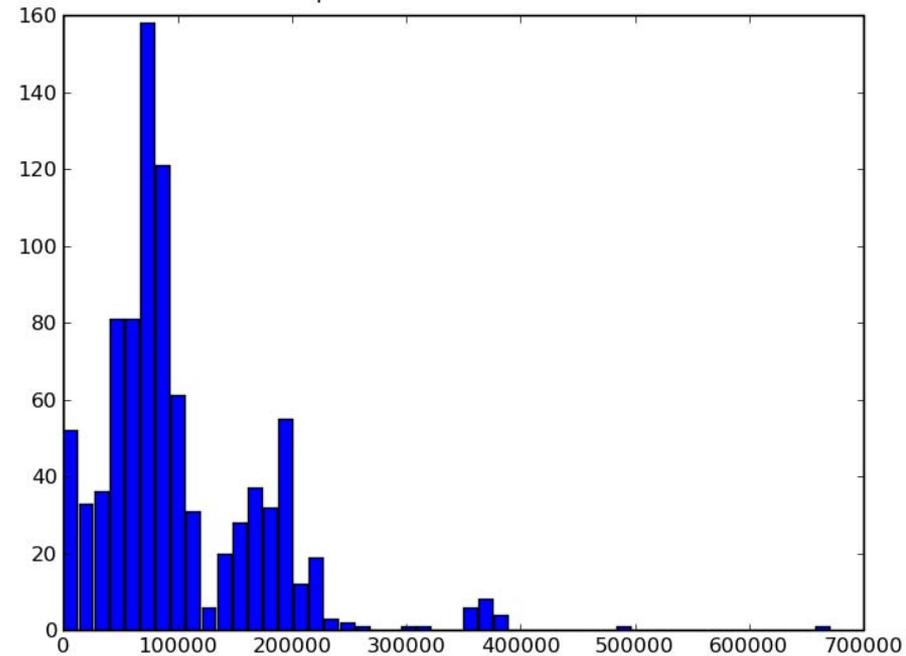
□ Avg: 110829

□ Avg: 101072

Temporal distance at page level



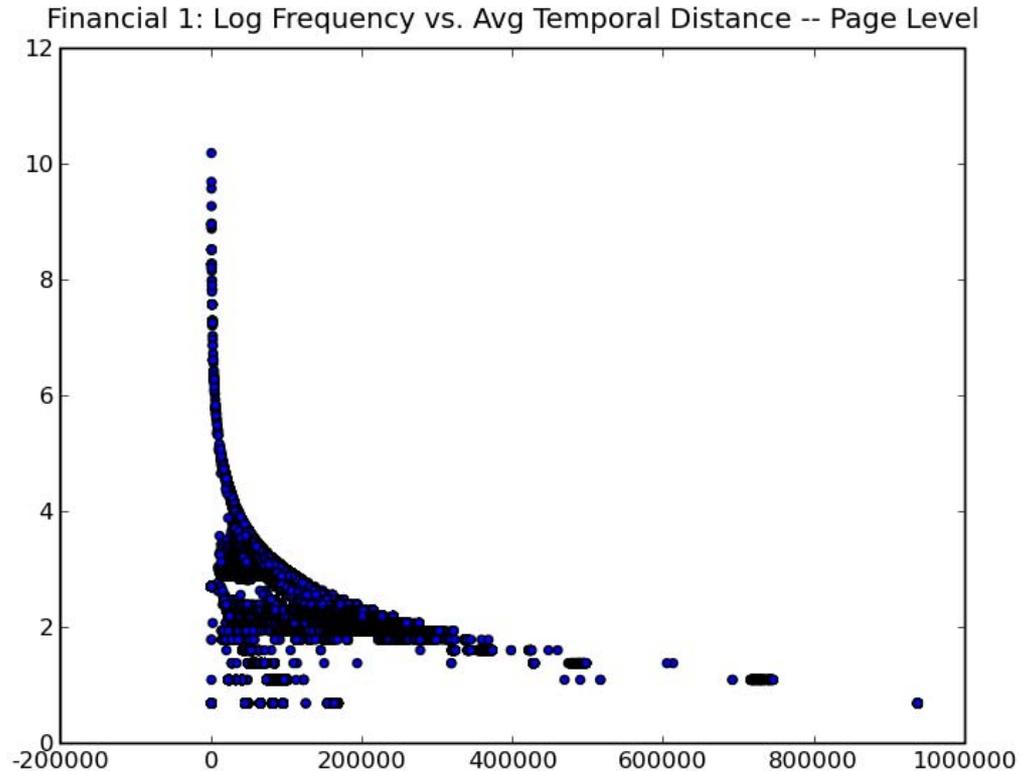
Temporal distance at block level



# Observation #2: Temporal Distance Vs. Frequency

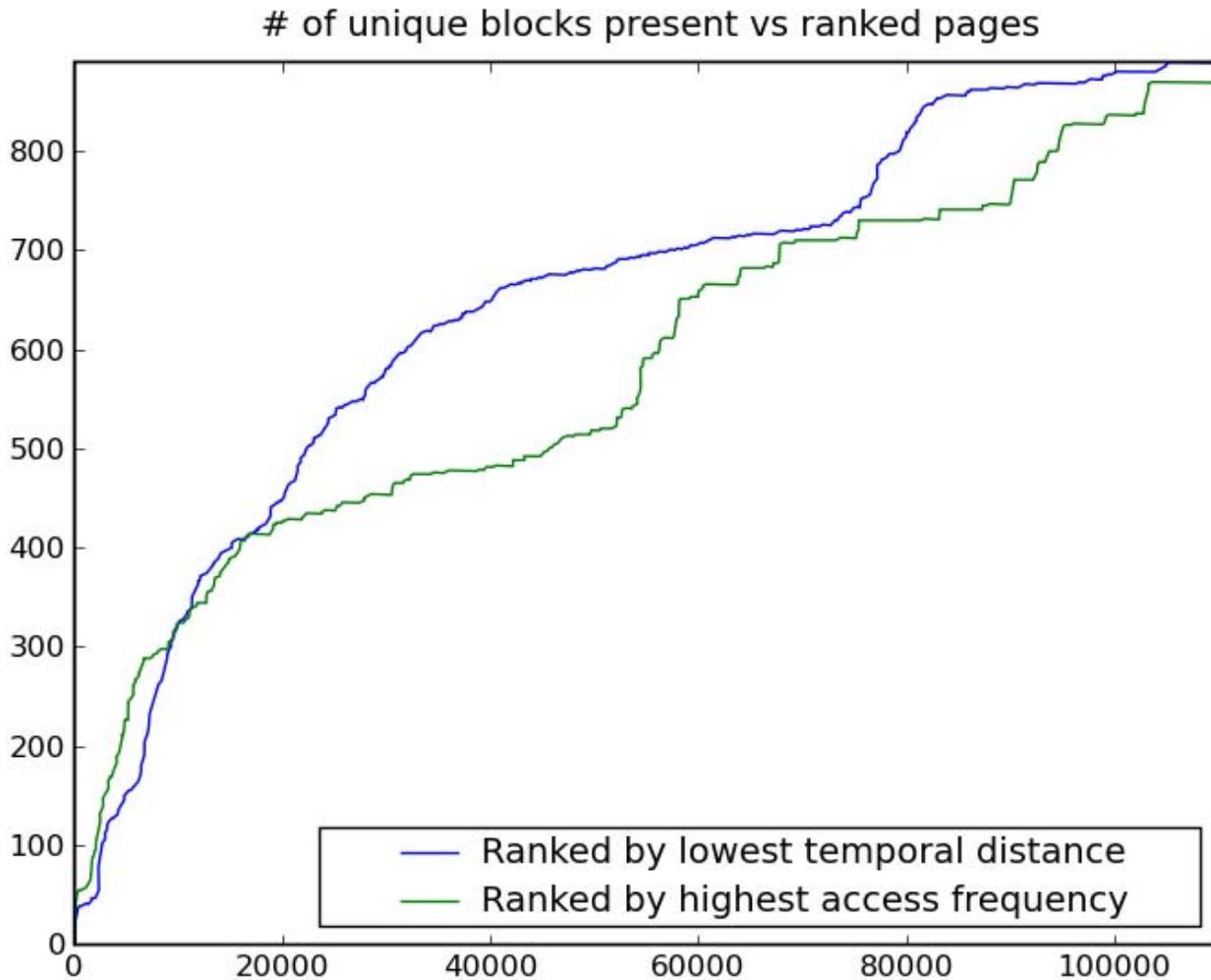
15

- Is temporal distance appropriate?
- How do frequency and temporal distance inter-relate?



# Observation #3: Spatial Correlation

16



# Research Tasks

17

## ❑ SSD adaptive cache design

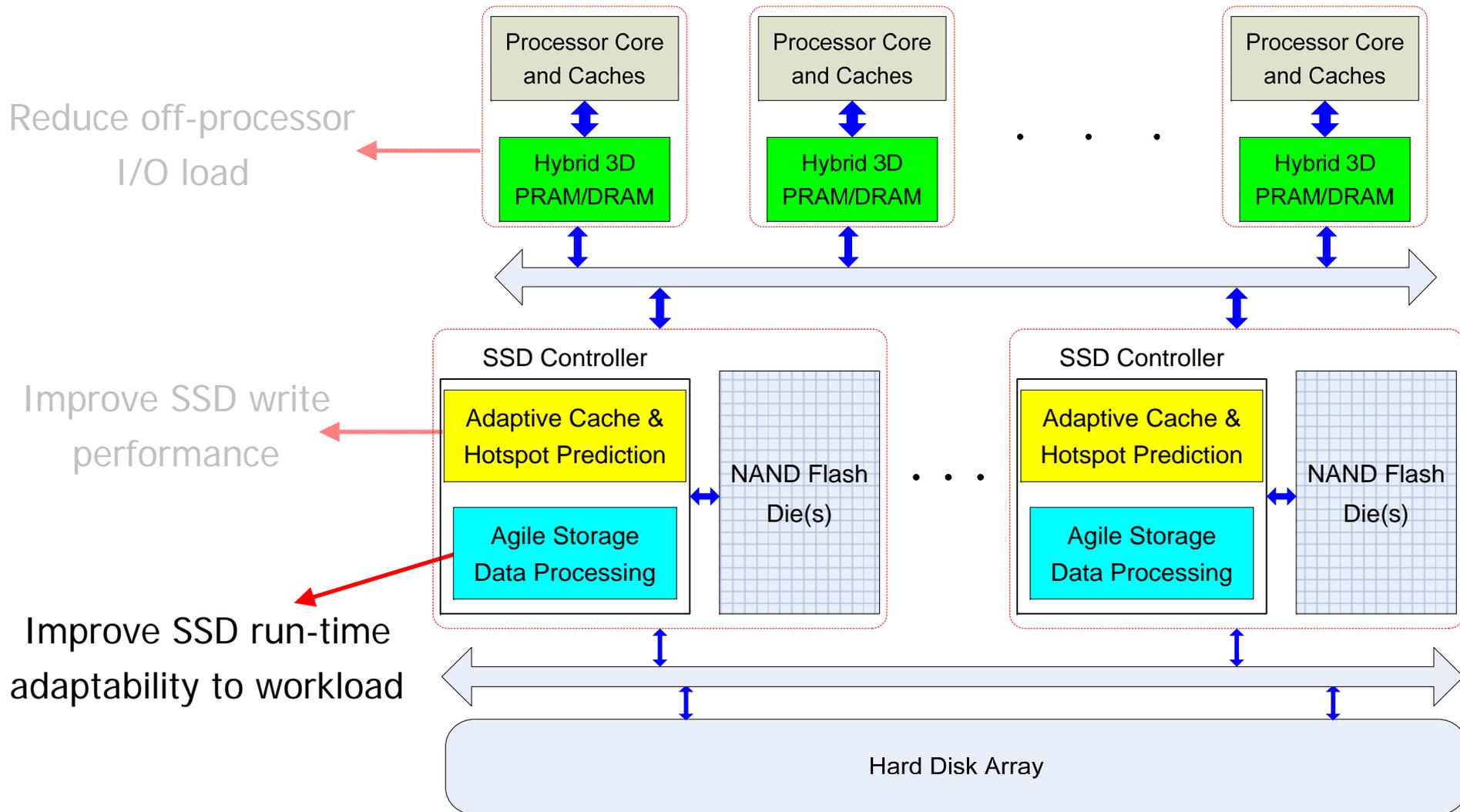
- ✓ To exploit both temporal and spatial locality at both page and block levels for SSDs.
- ✓ Exploit temporal locality → avoid grouping hot pages with cold pages in the same cluster in the page level queue
- ✓ Exploit spatial locality → hold hot clusters in the block level queue
- ✓ Fewer evictions: Less evictions improve performance because it will curtail the high latencies caused by writing and erasing
- ✓ Adaptive to workloads to take advantage of both page-level and block level locality.

## ❑ Hotspot prediction

- ✓ Investigate algorithms to detect/predict hot pages/blocks in SSDs

# Outline of Proposed Research

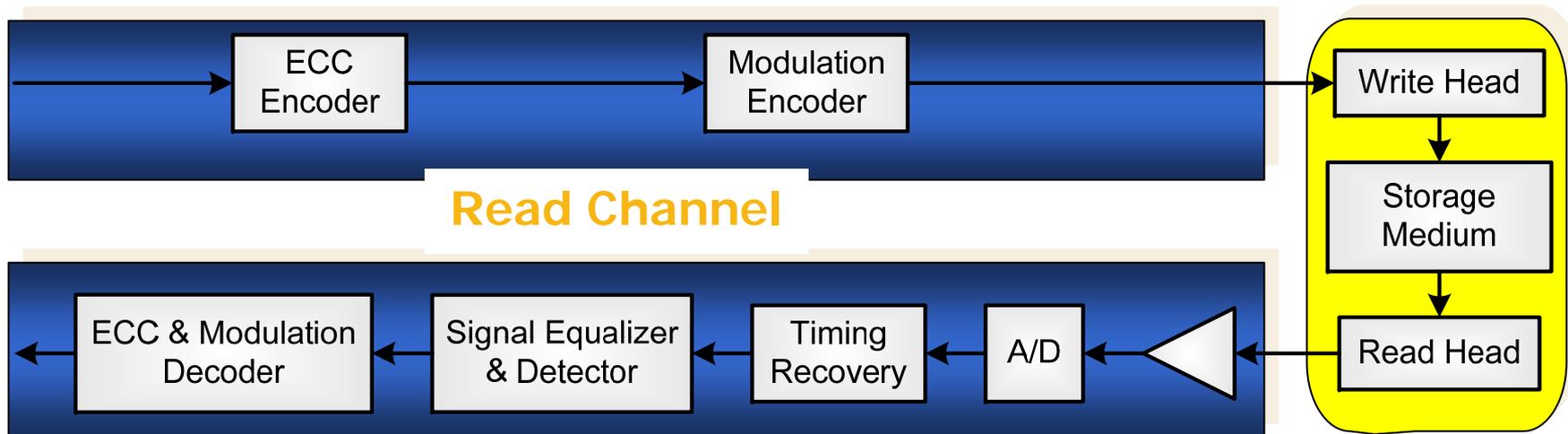
18



# Rationale

19

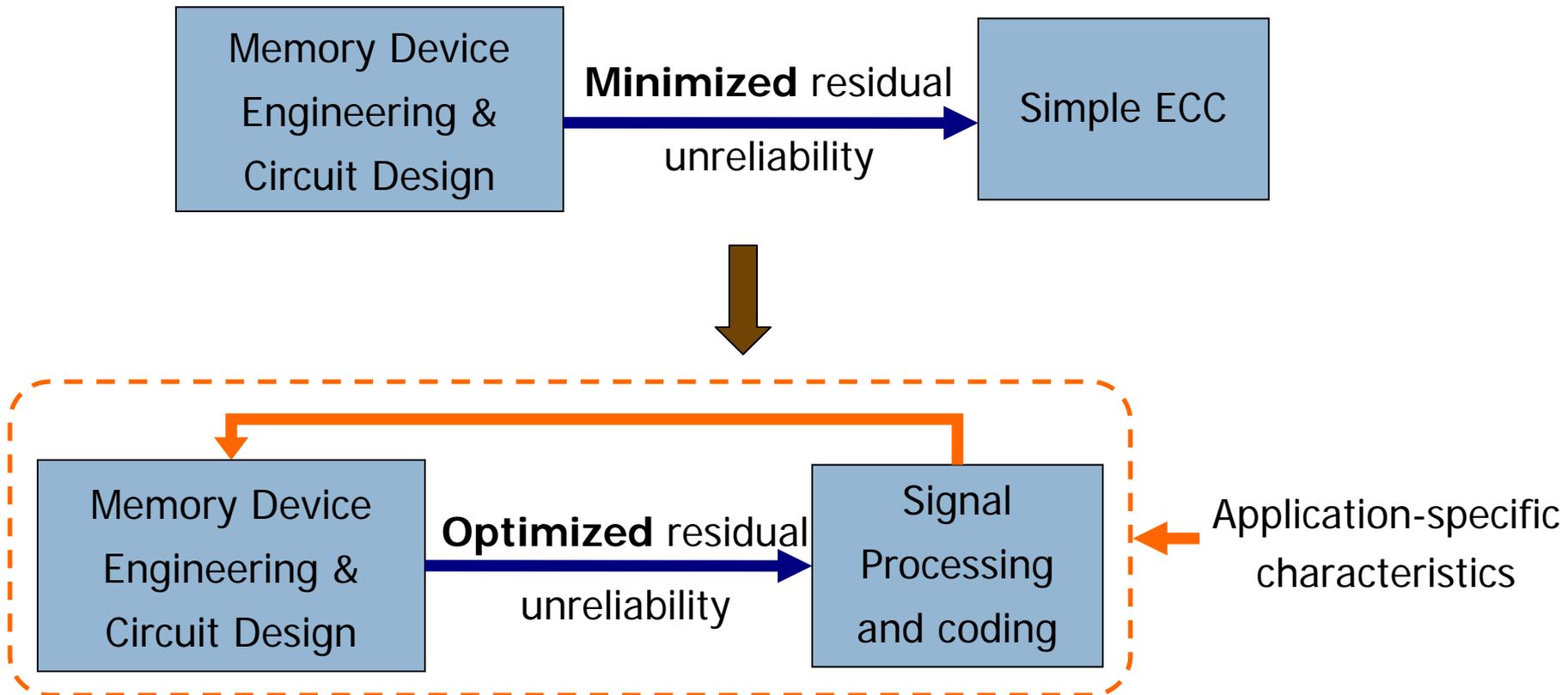
- ❑ NAND flash memory: cost reduction → scaling → reliability problem
- ❑ Hard disk drive faced the same problem 15 years ago
  - Introduce digital signal processing methods
  - Advanced signal processing and coding have been playing an ever increasingly important role in sustaining the storage density growth.



# Rationale

20

Could we repeat the success of signal processing/coding in NAND flash ?



# NAND Flash Memory

21



Communication channel

## Noisy channel

- Device noise: random telegraph noise, charge leakage, ...
- Circuit noise: cell-to-cell interference, background pattern dependency, ...

- ❑ Reliable communication through noisy channel → redundancy
- ❑ Cell storage efficiency: average number of real user bits per cell  
e.g., 28-byte redundancy for every 512-byte user data in 2bits/cell flash

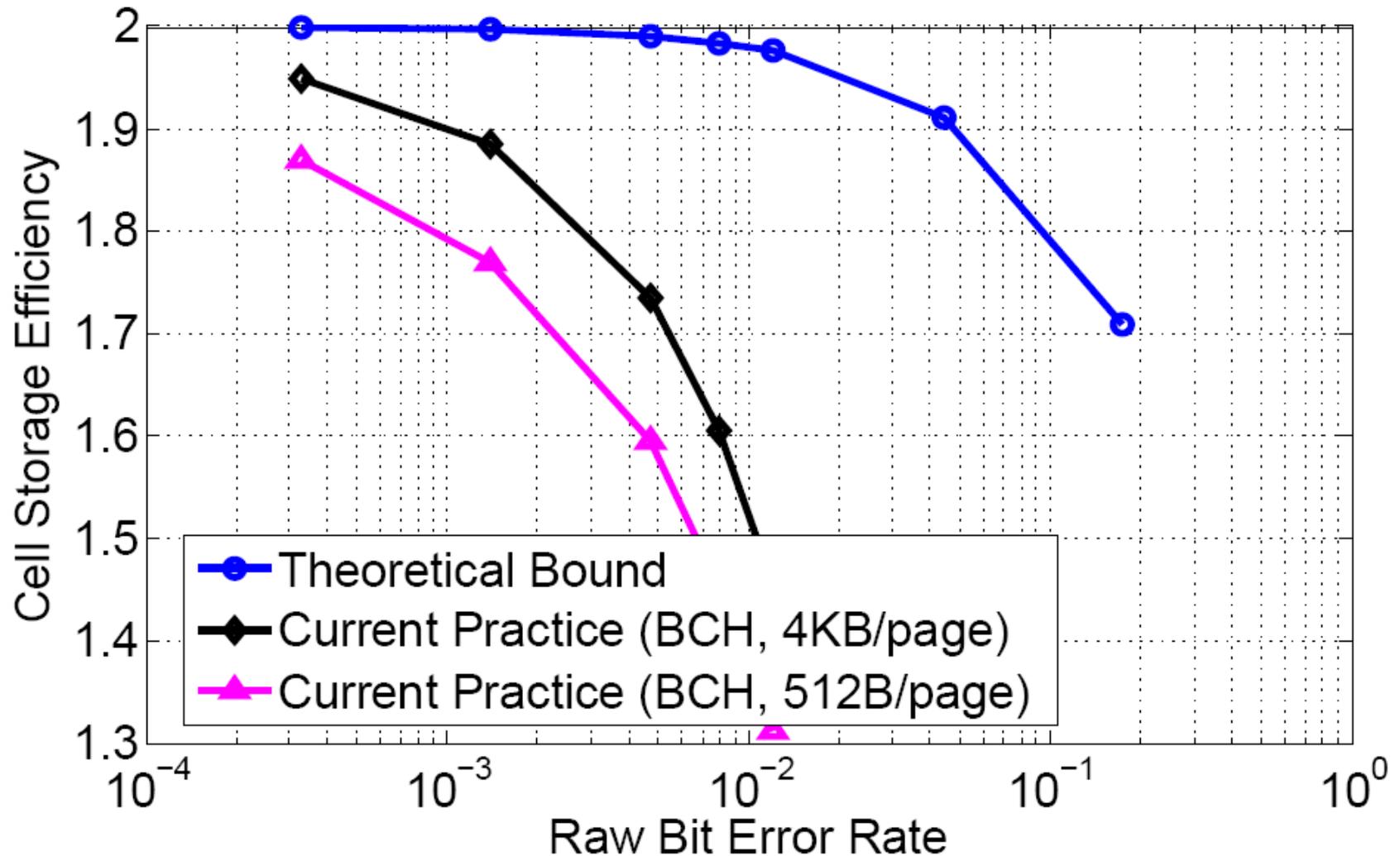
$$\frac{512}{512 + 28} \times 2 = 1.90 \text{ bits/cell}$$



Theoretical bounds of NAND flash cell storage efficiency

# Results: 2 Bits per Cell

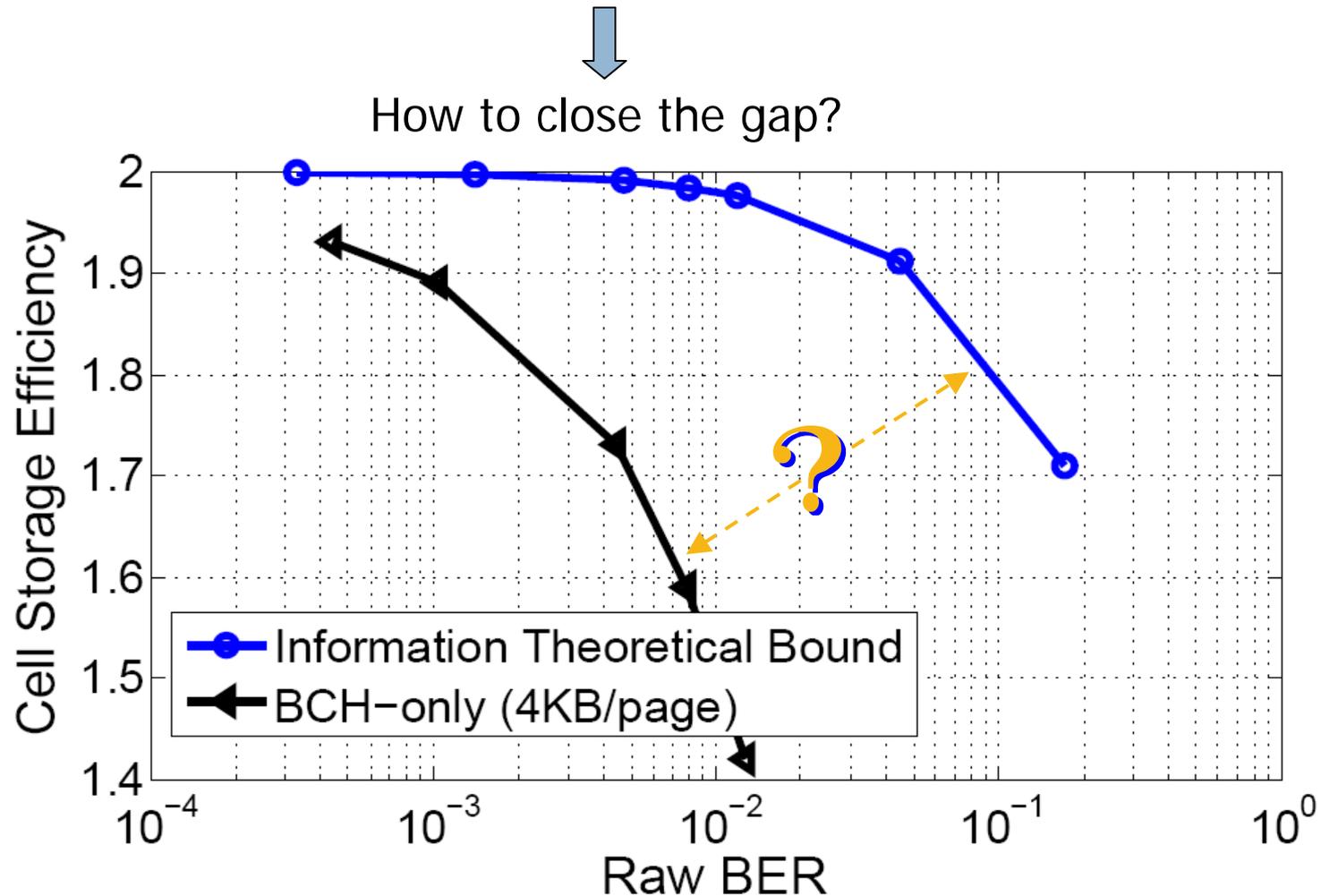
22



# Closing the Gap

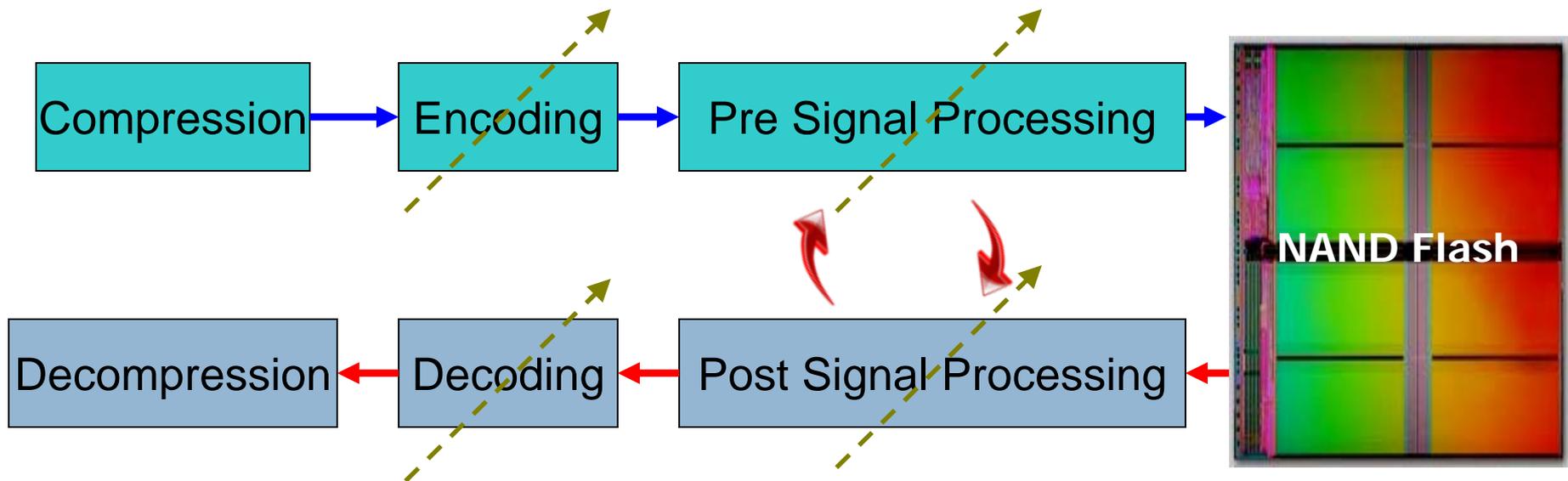
23

A **big gap** between the theoretical bound and current practice



# NAND Flash Data Processing

24



- ❑ Access latency is another critical metric
- ❑ Variable demands on write vs. read latency trade-offs



Agile SSD Data Processing

# Outline of Proposed Research

25

