

Seminar in Computer Architecture

Meeting 4: Memory Channel Partitioning

Prof. Onur Mutlu

ETH Zürich

Fall 2021

14 October 2021

Example Paper Presentations

Last Week: RowClone

Mindset: Memory as an Accelerator

The diagram illustrates a system architecture. On the left, a large grey box contains several components: four 'CPU core' blocks, a 'mini-CPU core', a 'video core', an 'imaging core', and a 'GPU (throughput) core' block. Below these is a 'LLC' (Last Level Cache) block, which is connected to a 'Memory Controller' block. The 'Memory Controller' is connected to a 'Memory Bus'. To the right of the 'Memory Bus' is a large 'Memory' block. Inside the 'Memory' block, there is a 'Specialized compute-capability in memory' block, which is highlighted by a red rounded rectangle. A red arrow points from the 'Specialized compute-capability in memory' block to the 'Memory Bus'. The video player interface at the bottom shows the title 'Memory similar to a "conventional" accelerator' and a timestamp of 43:42 / 2:03:45.

Onur Mutlu

Memory similar to a "conventional" accelerator

DEPARTMENT OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING (D-ITET)

Seminar in Computer Arch. - Meeting 3: RowClone: In-Memory Data Copy and Initialization (Fall 2021)

373 views • Streamed live on Oct 7, 2021

22 0 SHARE SAVE ...



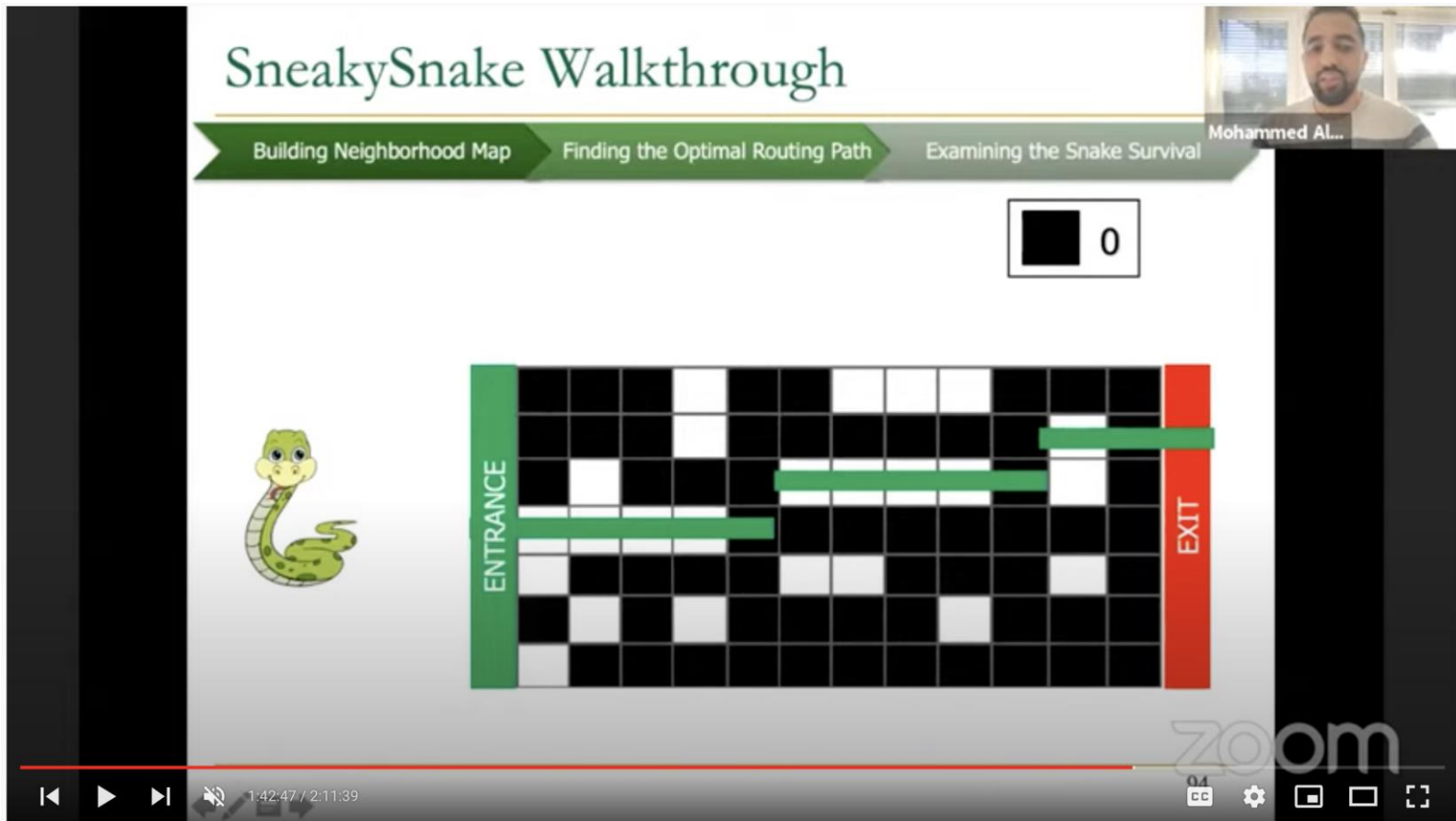
Onur Mutlu Lectures
19.2K subscribers

ANALYTICS

EDIT VIDEO

https://www.youtube.com/watch?v=n6Pwg1qax_E&list=PL5Q2soXY2Zi_7UBNmC9B8Yr5JSwTG9yH4&index=4

Prior Week: GateKeeper



DEPARTMENT OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING (D-ITET)
Seminar in Computer Architecture (Fall 2021) - Lecture 2: GateKeeper and Fast Genome Analysis

558 views • Streamed live on Sep 30, 2021

25 0 SHARE SAVE ...



Onur Mutlu Lectures
19.2K subscribers

ANALYTICS

EDIT VIDEO

2019: REAPER

Single-cell Failure Probability (Real)

operate here

Read Failure Probability

Refresh Interval (s)

hard to find

SAFARI

26

15:57 / 49:27 • Single-cell Failure Probability (Real) >

ETH ZÜRICH

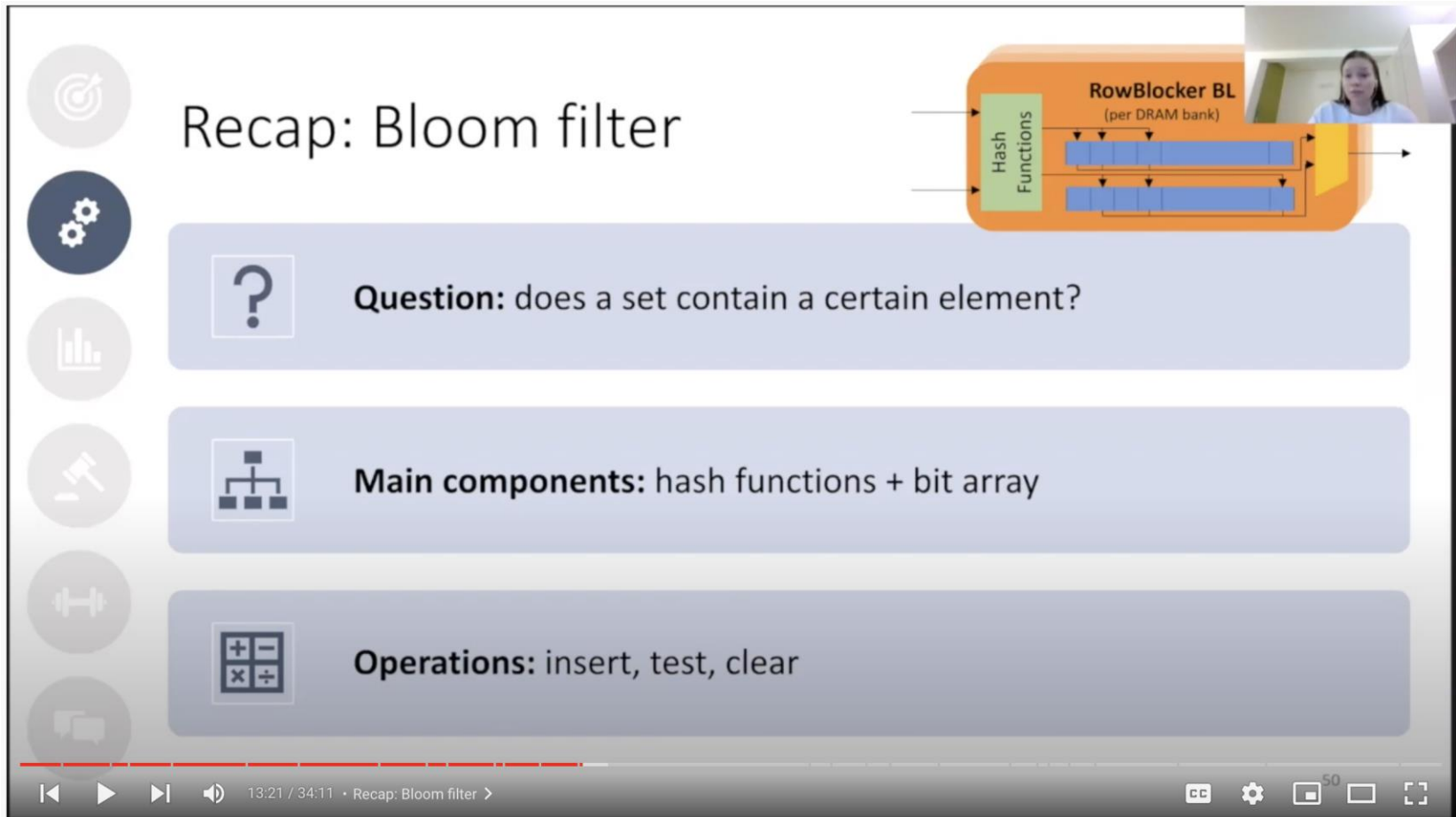
Seminar in Computer Architecture - Meeting 3b: Example Review III: REAPER (ETH Zürich, Fall 2019)

220 views • Oct 6, 2019

Onur Mutlu Lectures
19.2K subscribers

ANALYTICS EDIT VIDEO

Last Semester: BlockHammer



Recap: Bloom filter

Question: does a set contain a certain element?

Main components: hash functions + bit array

Operations: insert, test, clear

Diagram: RowBlocker BL (per DRAM bank) showing Hash Functions and bit array structure.

Video player controls: 13:21 / 34:11 • Recap: Bloom filter >

Seminar in Computer Architecture - Session 1.2: BlockHammer (ETH Zürich, Spring 2021)

365 views • Premiered Apr 26, 2021

👍 12 💬 0 ➦ SHARE ≡+ SAVE ...



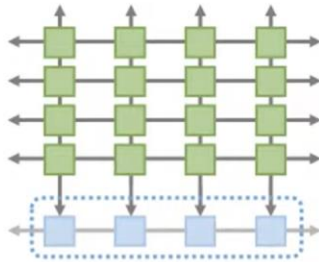
Onur Mutlu Lectures
19.2K subscribers

ANALYTICS

EDIT VIDEO

Last Semester: ComputeDRAM

DRAM Commands



Row access strobe: t_{RAS}

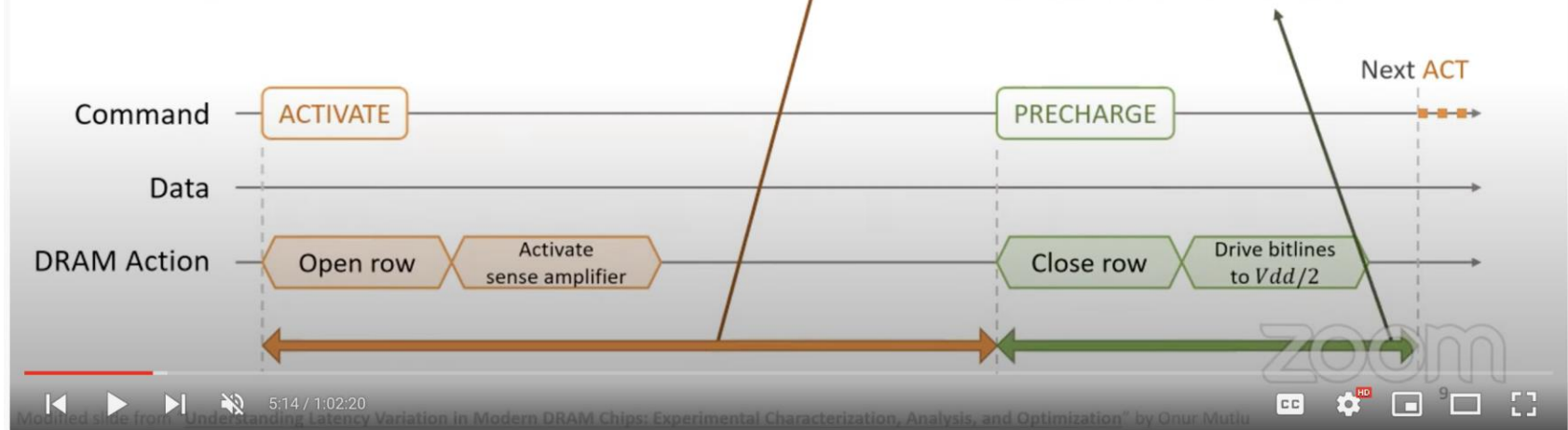
Minimum delay between ACTIVATE and the next PRECHARGE command

Row precharge: t_{RP}

Minimum delay between PRECHARGE and the next ACTIVATE command



Lorenzo Rai



Seminar in Computer Architecture - Session 2.2: ComputeDRAM (Spring 2021)

666 views • Streamed live on Apr 1, 2021

22 0 SHARE SAVE ...



Onur Mutlu Lectures
19.2K subscribers

ANALYTICS

EDIT VIDEO

Last Semester: Deep Compression & SneakySnake

Discussion (I): End-to-End Performance



- Latency/throughput is not mentioned by the paper
 - Critical for real-time processing as was targeted by the paper
- Speedup is actually... **not true**... (in my opinion)
 - Only **densely connected layers** are measured to have a significant speedup
 - Overheads are mostly in **CNN layers**
 - The overall throughput **does not** increase if the bottleneck layer is not boosted much (and so is latency)
 - How do you think that it would be fairer methodology to measure the speedup? What would you expect really from throughput by using this approach? What kind of benchmarks would make sense?



DEPARTMENT OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING (D-ITET)

Seminar in Computer Architecture - Session 6: Deep Compression & SneakySnake (Spring 2021)

572 views • Streamed live on May 6, 2021

21 0 SHARE SAVE ...



Onur Mutlu Lectures
19.2K subscribers





ANALYTICS

EDIT VIDEO

Last Semester: Alpha 21264 & Mirage Cores

ARM big.LITTLE Architecture

Released in 2011

-  Many Android Smartphones
- 
-  Apple A series (A14 used in iPhone 12s)
-  Nintendo Switch using Nvidia Tegra XI

753 views • Streamed live on May 20, 2021

Onur Mutlu Lectures
19.2K subscribers

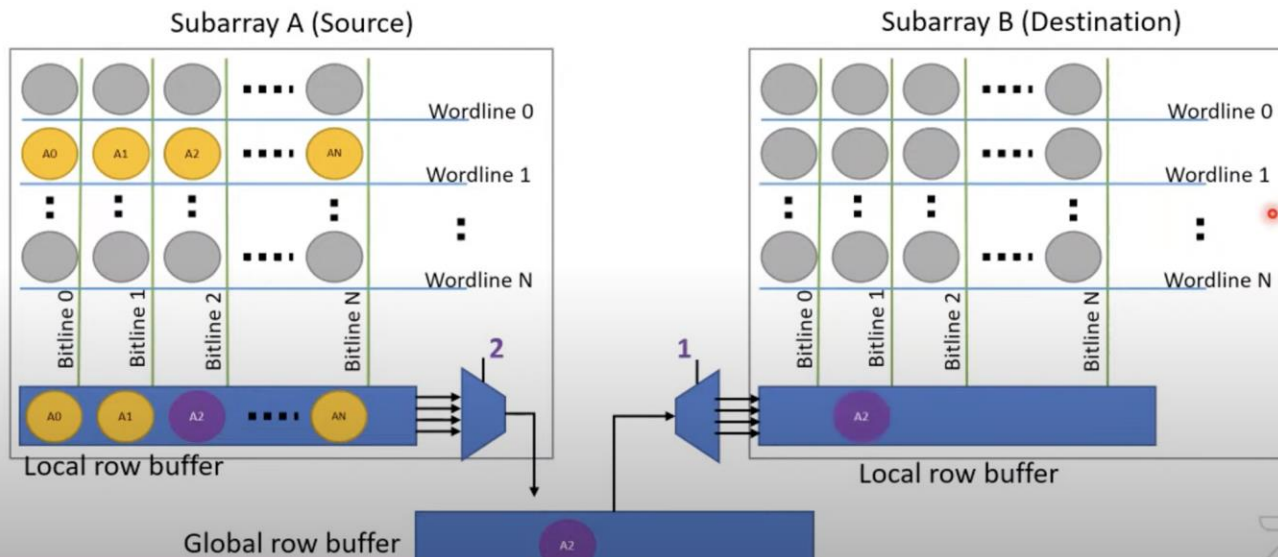
ANALYTICS EDIT VIDEO

Last Semester: FIGARO



FIGARO: working principle

Transferring data between two local row buffers



1. **ACTIVATE (A)**
2. **RELOC**

DEPARTMENT OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING (D-ITET)
Seminar in Computer Architecture - Session 9: FIGARO (Spring 2021)

386 views • Streamed live on Jun 3, 2021

12 0 SHARE SAVE ...



Onur Mutlu Lectures
19.2K subscribers

ANALYTICS

EDIT VIDEO

Today: Another Example Paper Presentation

We Will Briefly Review This Paper

- Sai Prashanth Muralidhara, Lavanya Subramanian, Onur Mutlu, Mahmut Kandemir, and Thomas Moscibroda,
"Reducing Memory Interference in Multicore Systems via Application-Aware Memory Channel Partitioning"
*Proceedings of the 44th International Symposium on Microarchitecture (**MICRO**), Porto Alegre, Brazil, December 2011. Slides (pptx)*

Reducing Memory Interference in Multicore Systems via Application-Aware Memory Channel Partitioning

Sai Prashanth Muralidhara
Pennsylvania State University
smuralid@cse.psu.edu

Lavanya Subramanian
Carnegie Mellon University
lsubrama@ece.cmu.edu

Onur Mutlu
Carnegie Mellon University
onur@cmu.edu

Mahmut Kandemir
Pennsylvania State University
kandemir@cse.psu.edu

Thomas Moscibroda
Microsoft Research Asia
moscitho@microsoft.com

Application-Aware Memory Channel Partitioning

Sai Prashanth Muralidhara § Lavanya Subramanian †

Onur Mutlu † Mahmut Kandemir §

Thomas Moscibroda ‡

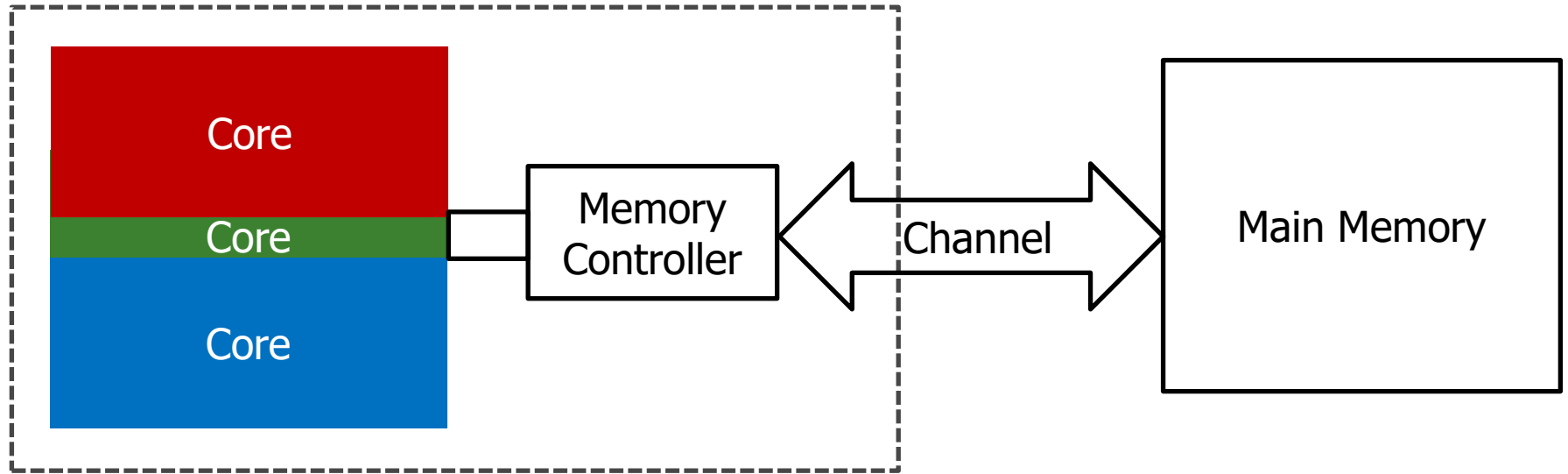
§ Pennsylvania State University † Carnegie Mellon University

‡ Microsoft Research

SAFARI Carnegie Mellon

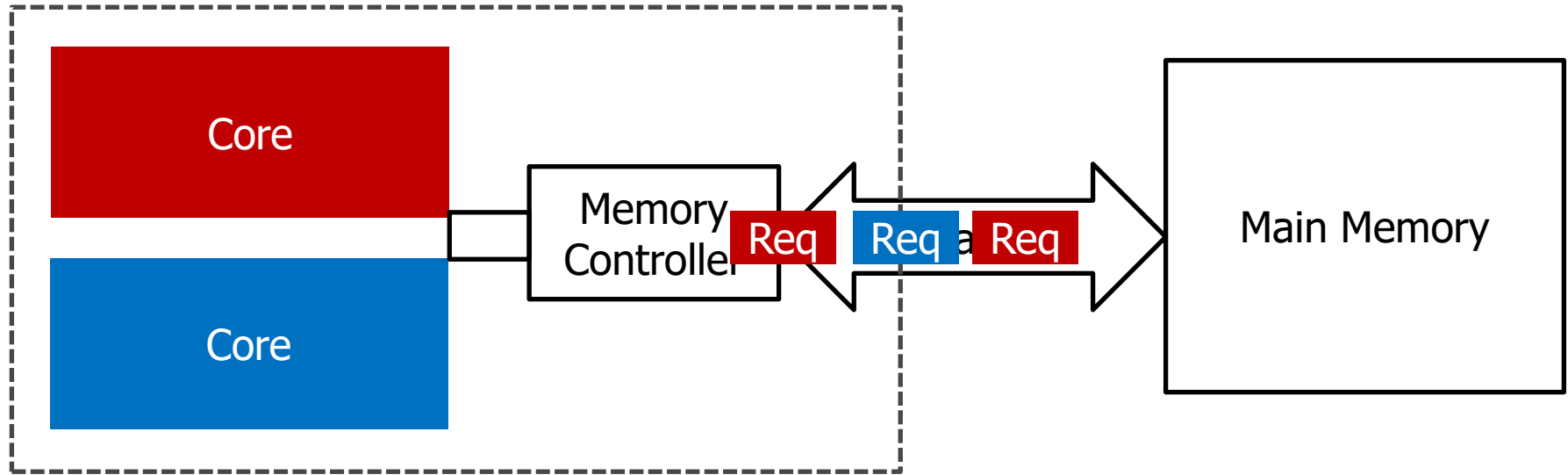
Background, Problem & Goal

Main Memory is a Bottleneck



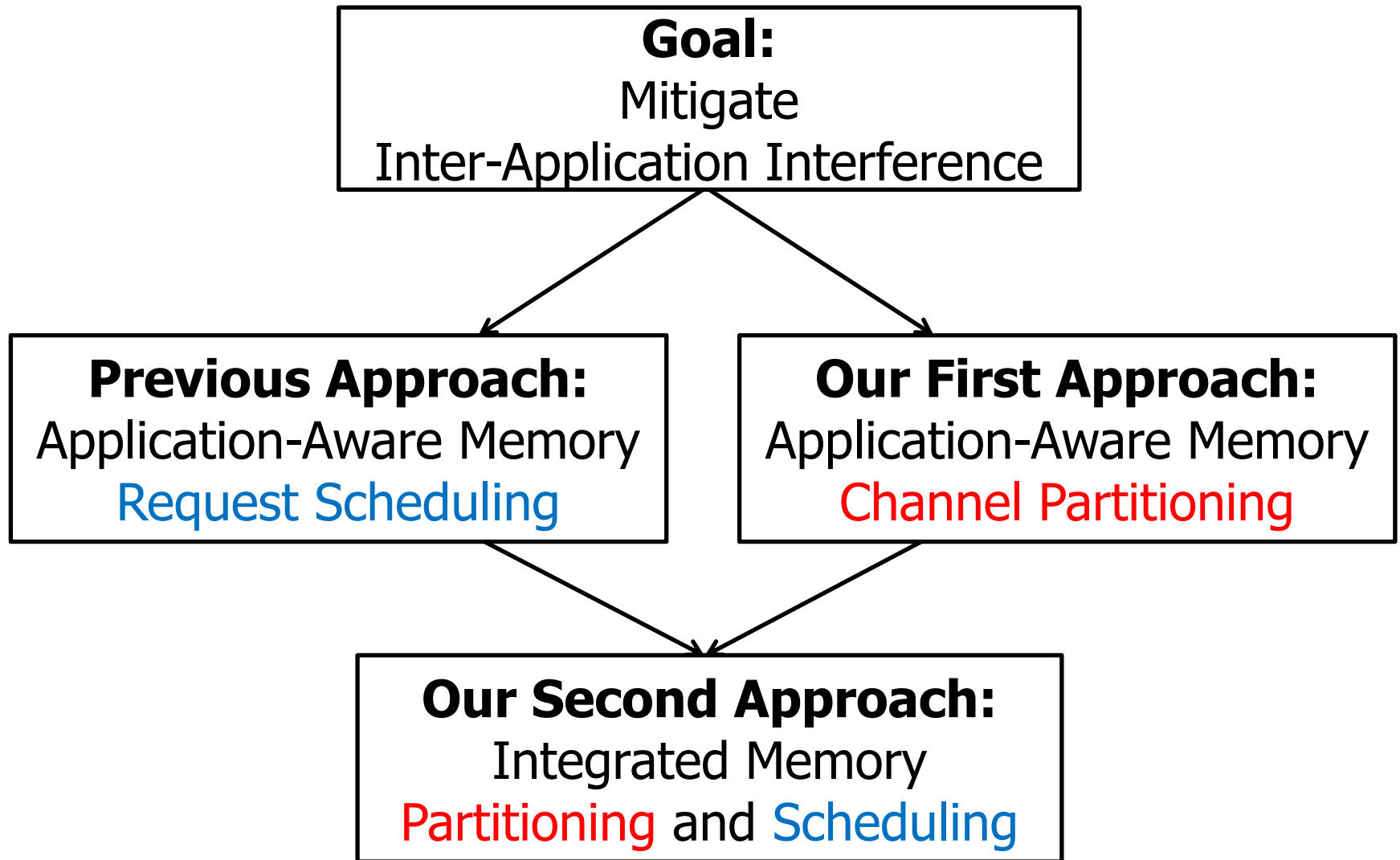
- Main memory latency is long
- Core stalls, performance degrades
- Multiple applications share the main memory

Problem of Inter-Application Interference

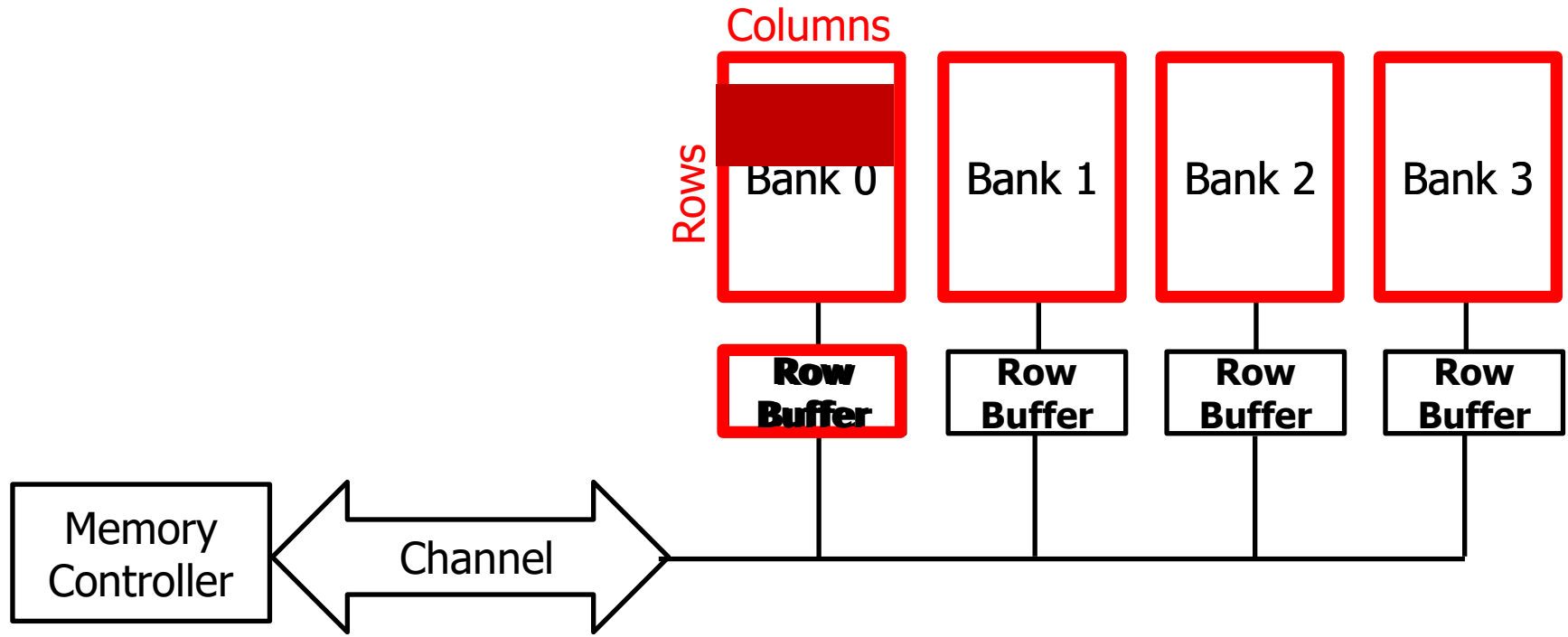


- Applications' requests interfere at the main memory
- This **inter-application interference** degrades system performance
- Problem further exacerbated due to
 - ❑ Increasing number of cores
 - ❑ Limited off-chip pin bandwidth

Outline



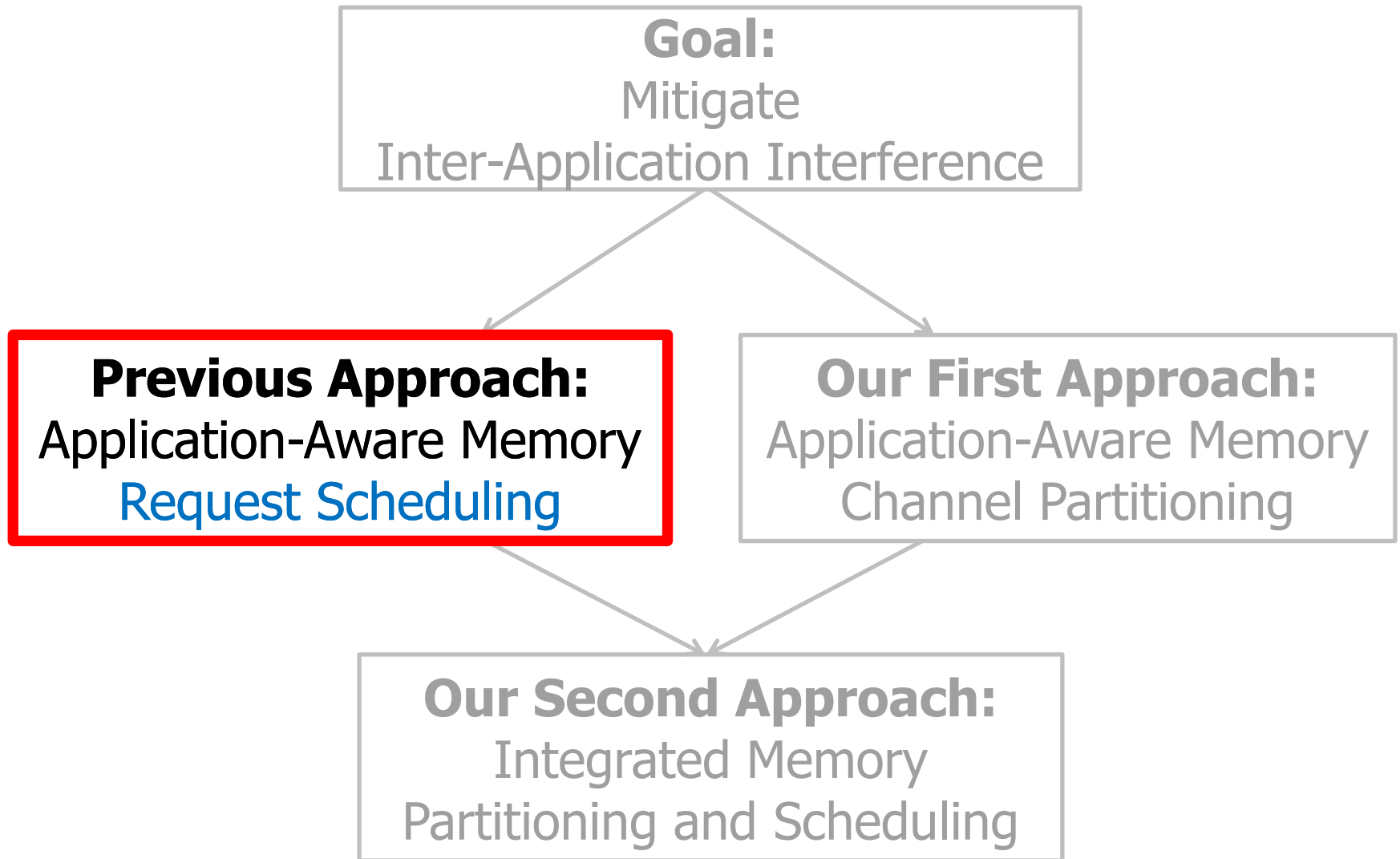
Background: Main Memory



- FR-FCFS memory scheduling policy [Zuravleff et al., US Patent '97; Rixner et al., ISCA '00]
 - ❑ Row-buffer hit first
 - ❑ Oldest request first
- Unaware of inter-application interference

Novelty

Previous Approach



Application-Aware Memory Request Scheduling

- **Monitor** application memory access characteristics
- **Rank** applications based on memory access characteristics
- **Prioritize** requests at the memory controller, based on ranking

An Example: Thread Cluster Memory Scheduling

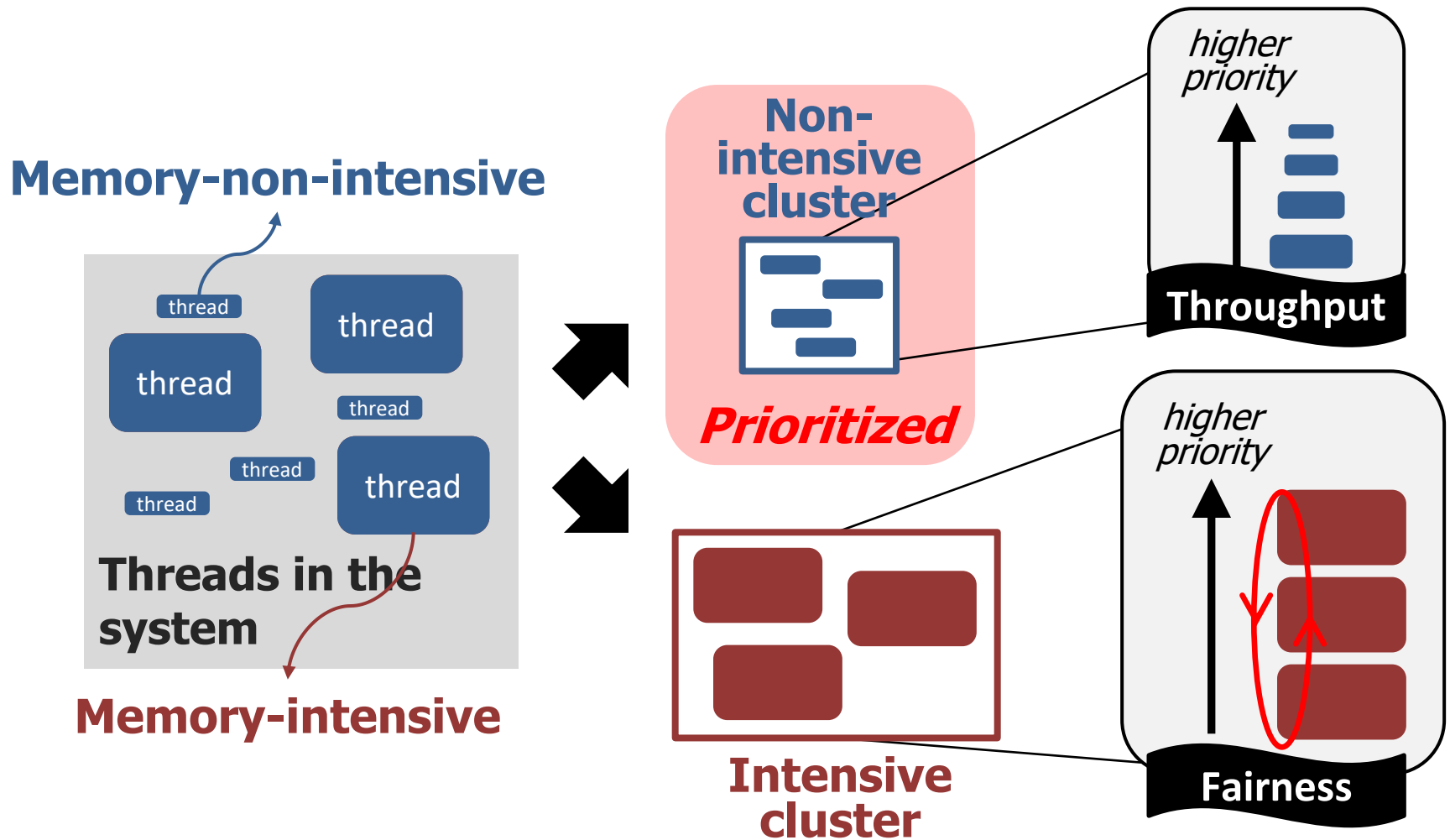


Figure: Kim et al., MICRO 2010

Application-Aware Memory Request Scheduling

Advantages

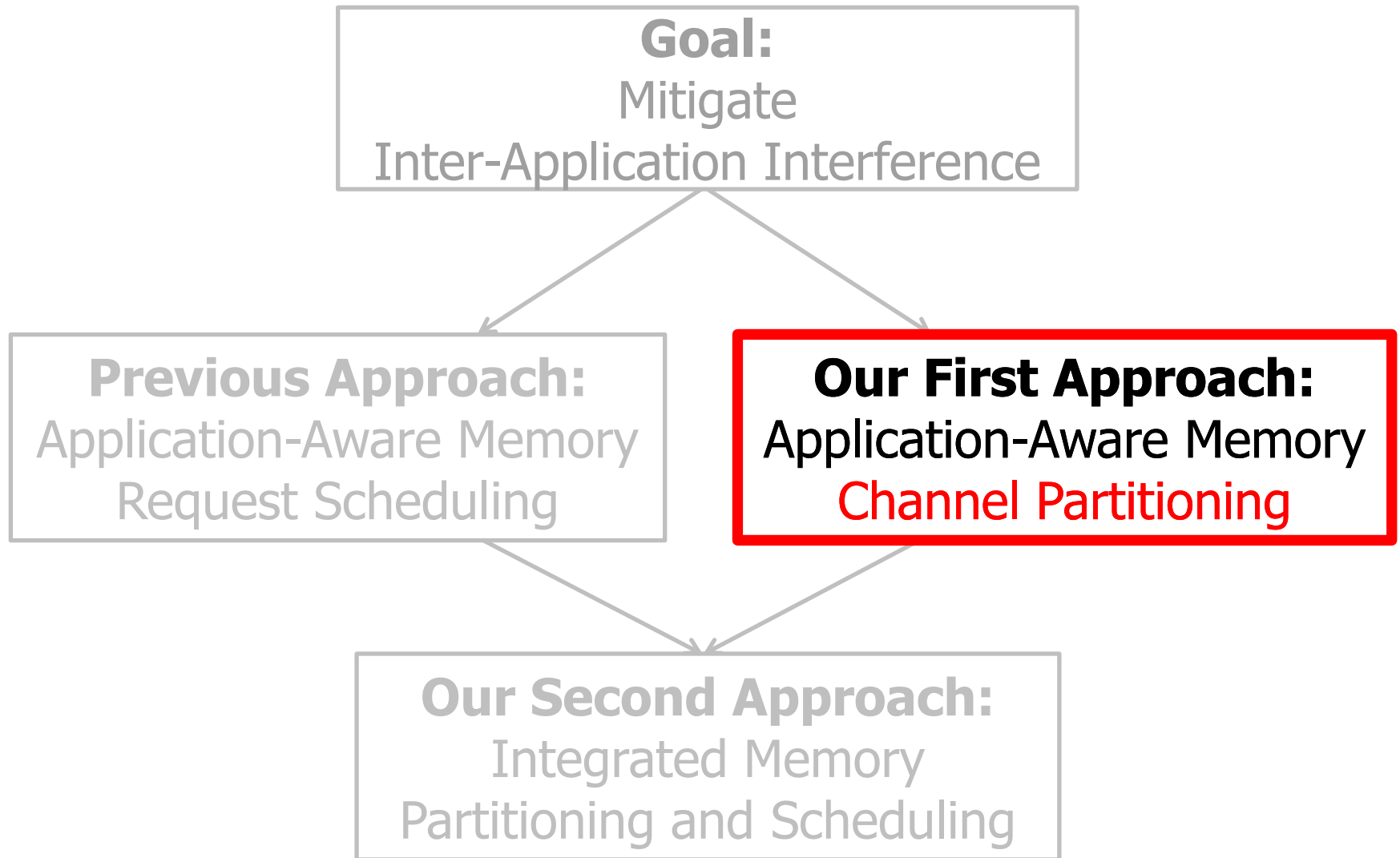
- Reduces interference between applications by request reordering
- Improves system performance

Disadvantages

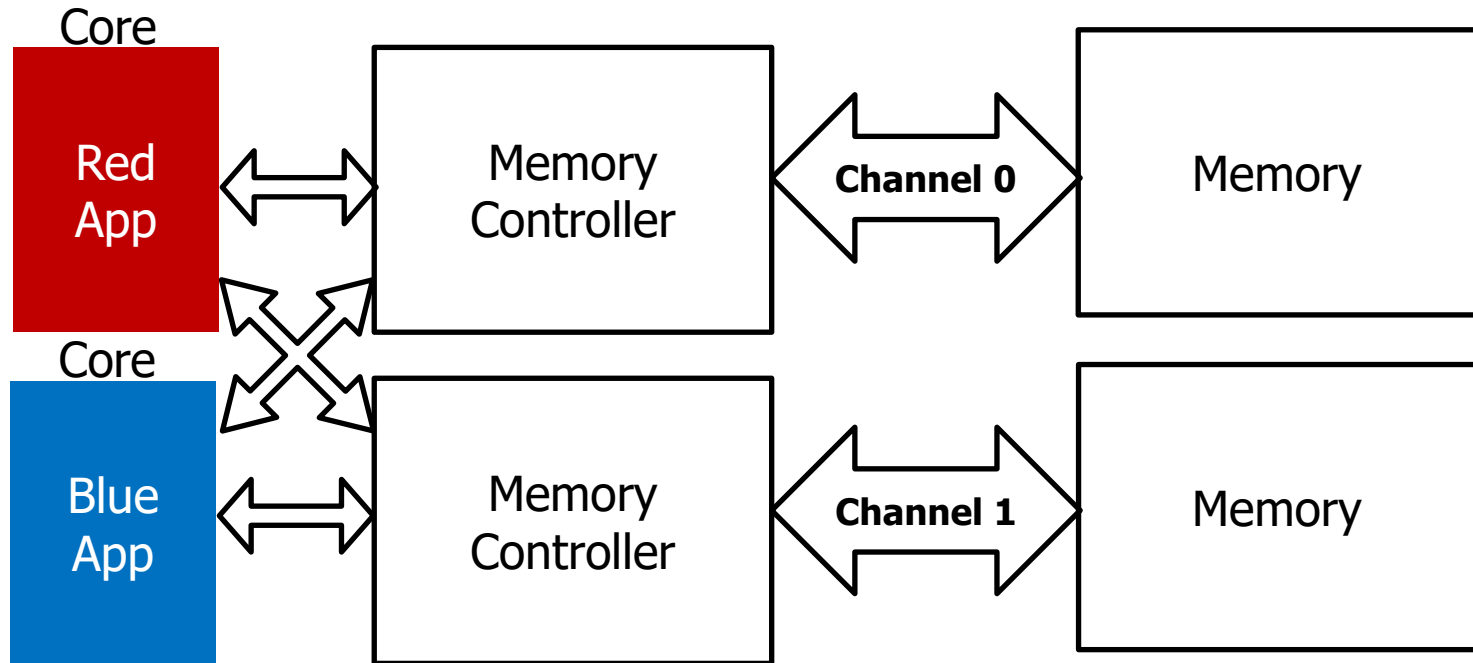
- Requires modifications to memory scheduling logic for
 - Ranking
 - Prioritization
- Cannot completely eliminate interference by request reordering

Key Approach and Ideas

The Paper's Approach

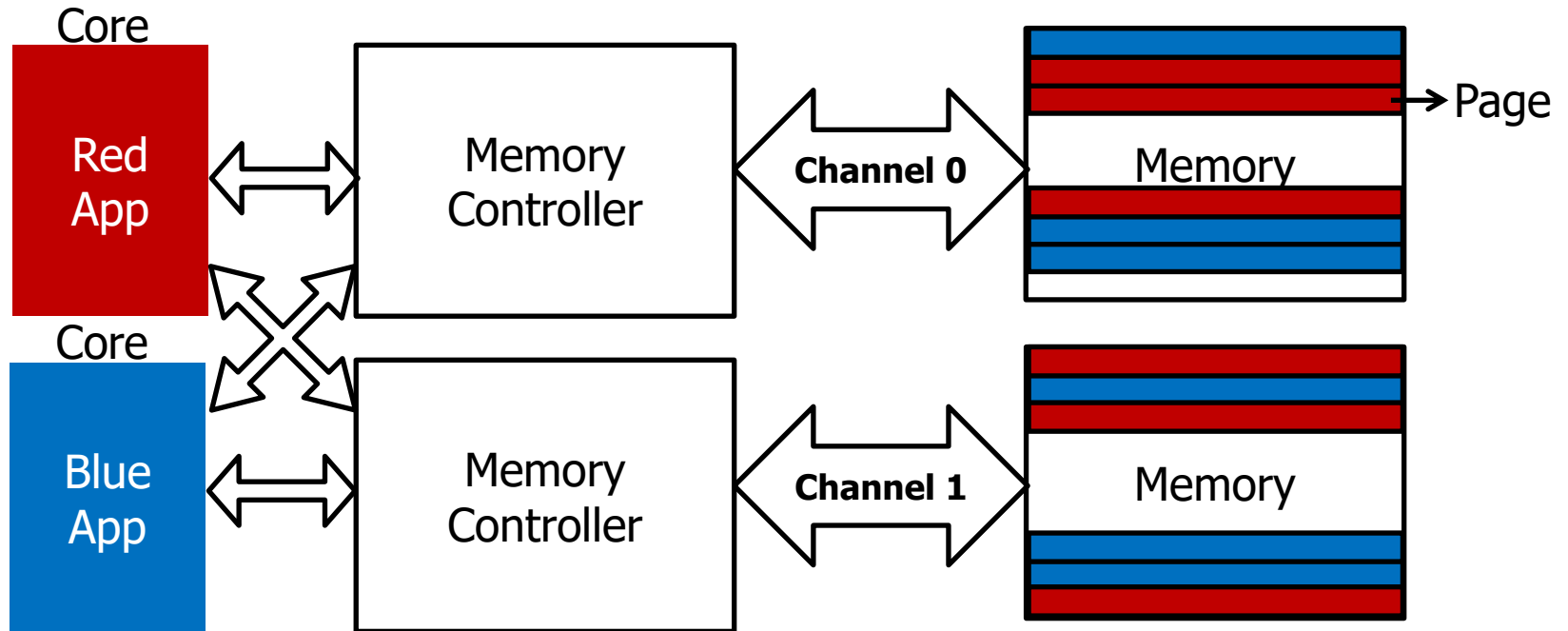


Observation: Modern Systems Have Multiple Channels



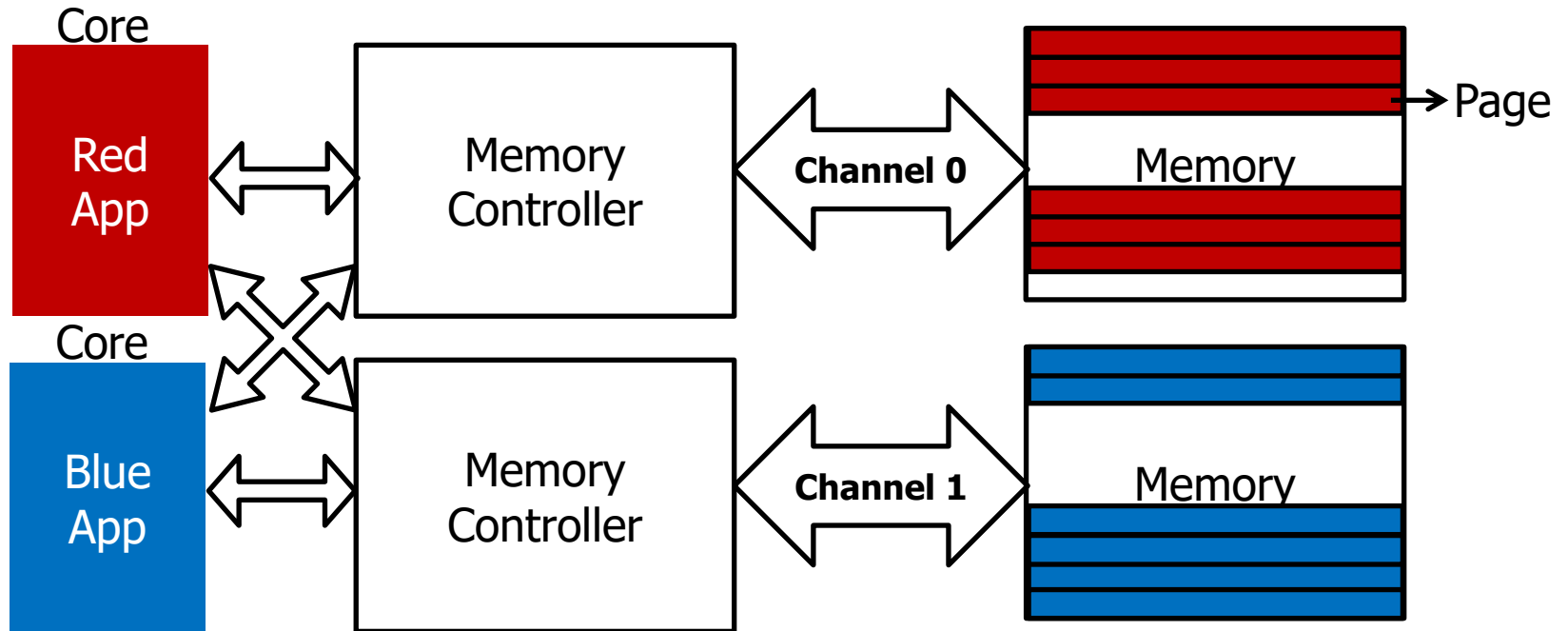
A new degree of freedom
Mapping data across multiple channels

Data Mapping in Current Systems



Causes interference between applications' requests

Partitioning Channels Between Applications



Eliminates interference between applications' requests

Overview: Memory Channel Partitioning (MCP)

■ Goal

- Eliminate harmful interference between applications

■ Basic Idea

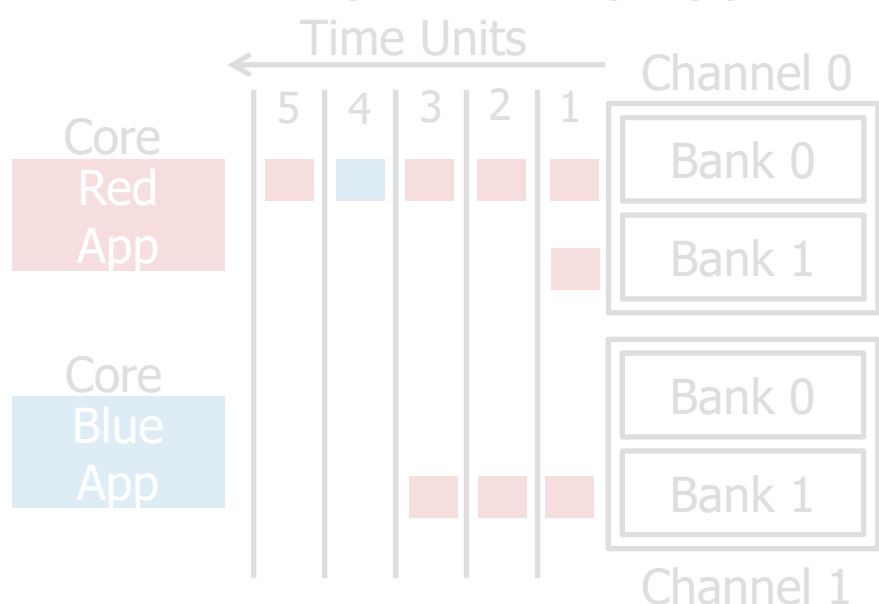
- Map the data of **badly-interfering applications** to different channels

■ Key Principles

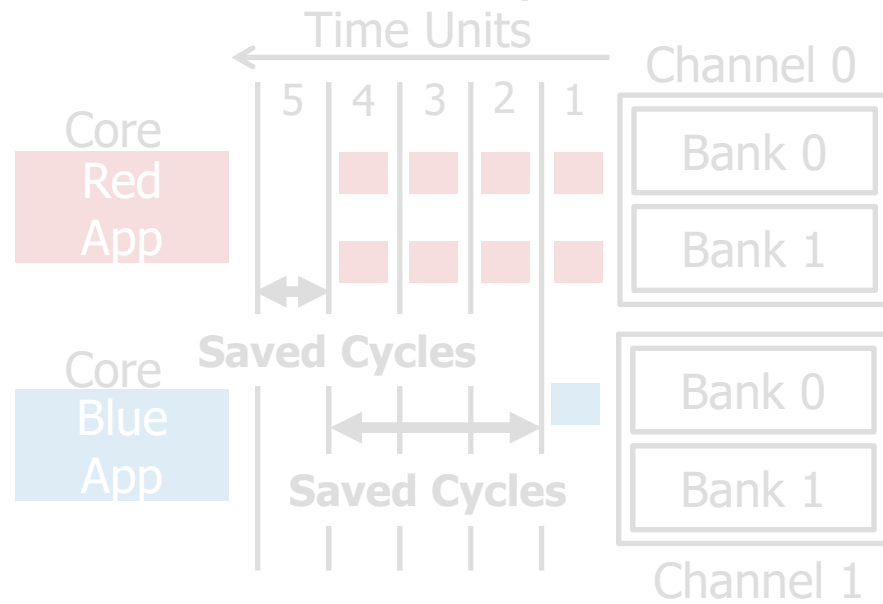
- Separate **low and high memory-intensity applications**
- Separate **low and high row-buffer locality applications**

Key Insight 1: Separate by Memory Intensity

High memory-intensity applications interfere with low memory-intensity applications in shared memory channels



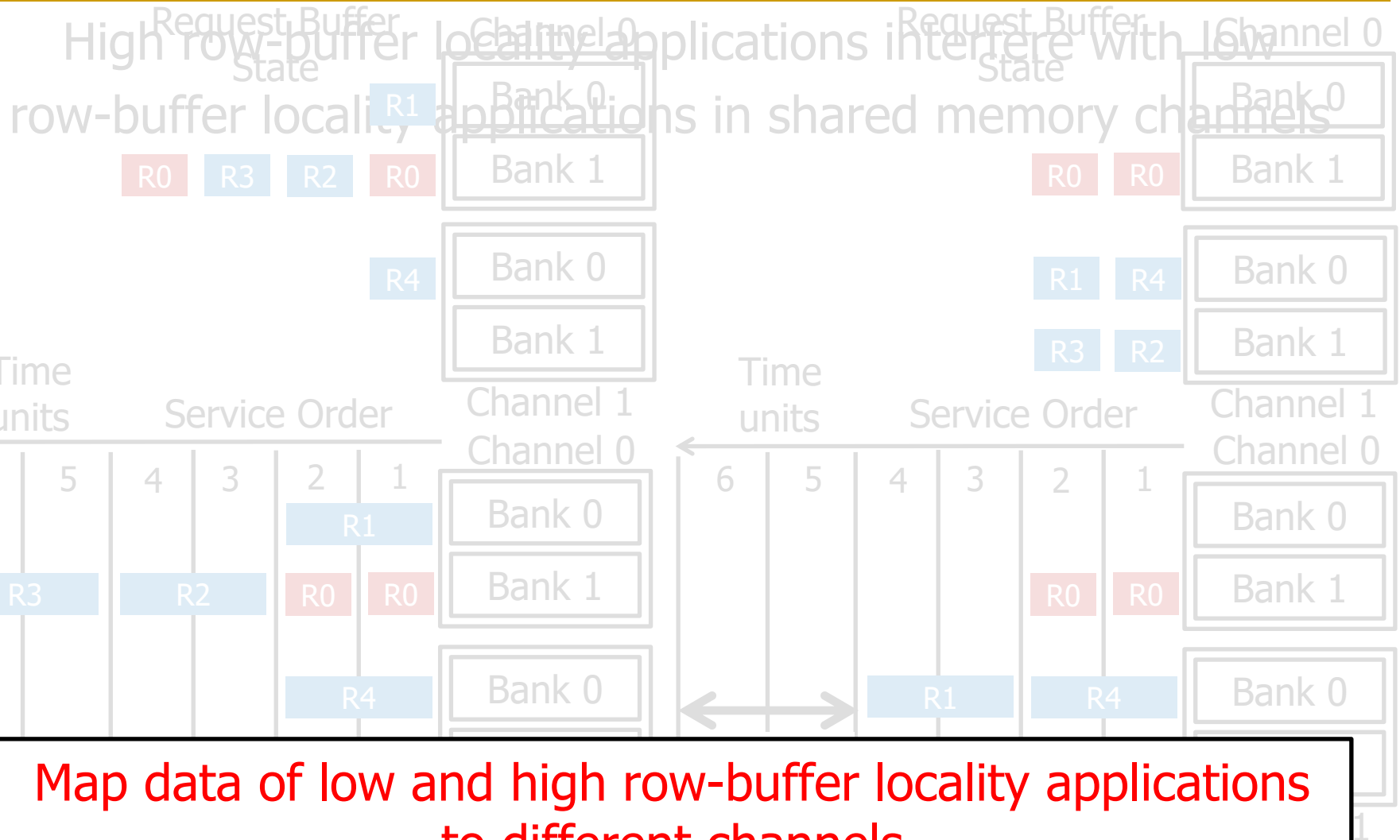
Conventional Page Mapping



Channel Partitioning

Map data of low and high memory-intensity applications to different channels

Key Insight 2: Separate by Row-Buffer Locality



Map data of low and high row-buffer locality applications to different channels

Mechanisms (in some detail)

Memory Channel Partitioning (MCP) Mechanism

Hardware

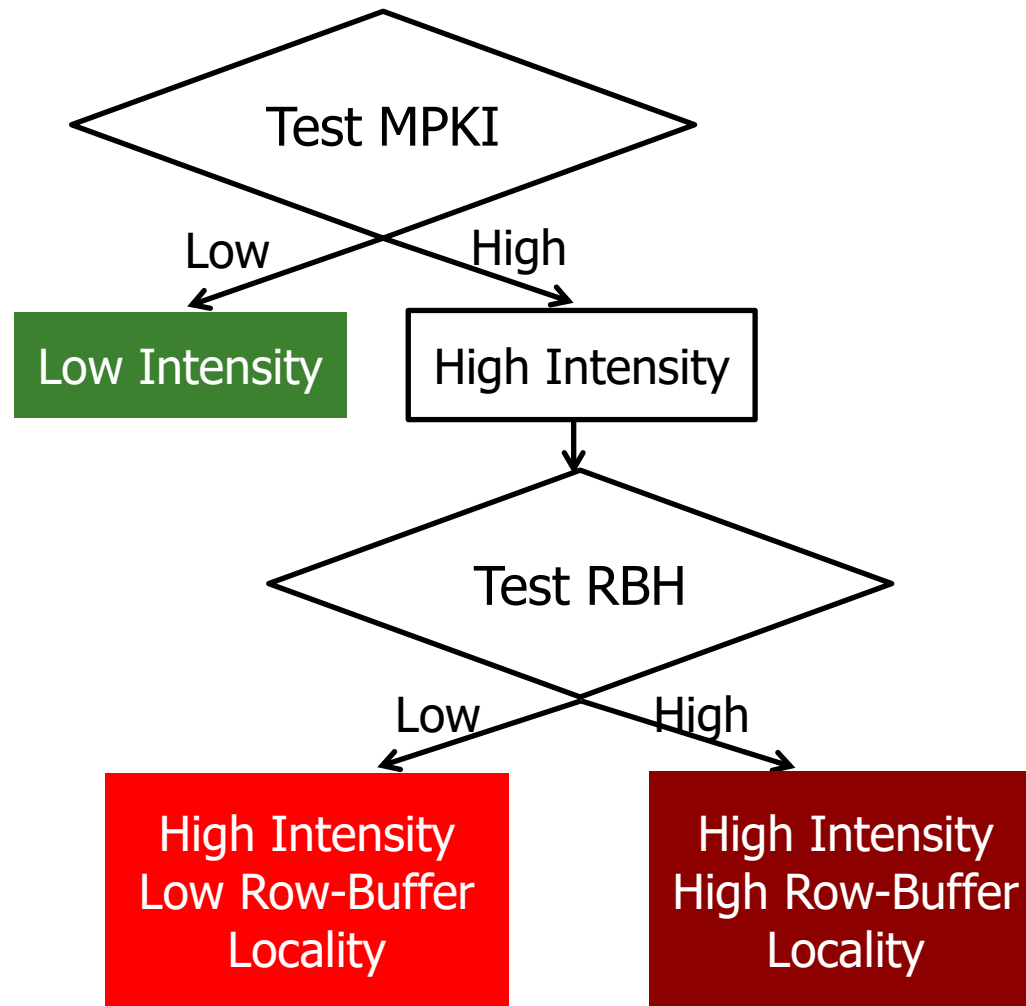
1. Profile applications
2. Classify applications into groups
3. Partition channels between application groups
4. Assign a preferred channel to each application
5. Allocate application pages to preferred channel

**System
Software**

1. Profile Applications

- Hardware counters collect application memory access characteristics
- Memory access characteristics
 - **Memory intensity:**
Last level cache **Misses Per Kilo Instruction (MPKI)**
 - **Row-buffer locality:**
Row-buffer Hit Rate (RBH) - percentage of accesses that hit in the row buffer

2. Classify Applications



3. Partition Channels Among Groups: Step 1

Low Intensity

High Intensity
Low Row-Buffer
Locality

High Intensity
High Row-Buffer
Locality

Assign number of channels
proportional to number of
applications in group

Channel 1

Channel 2

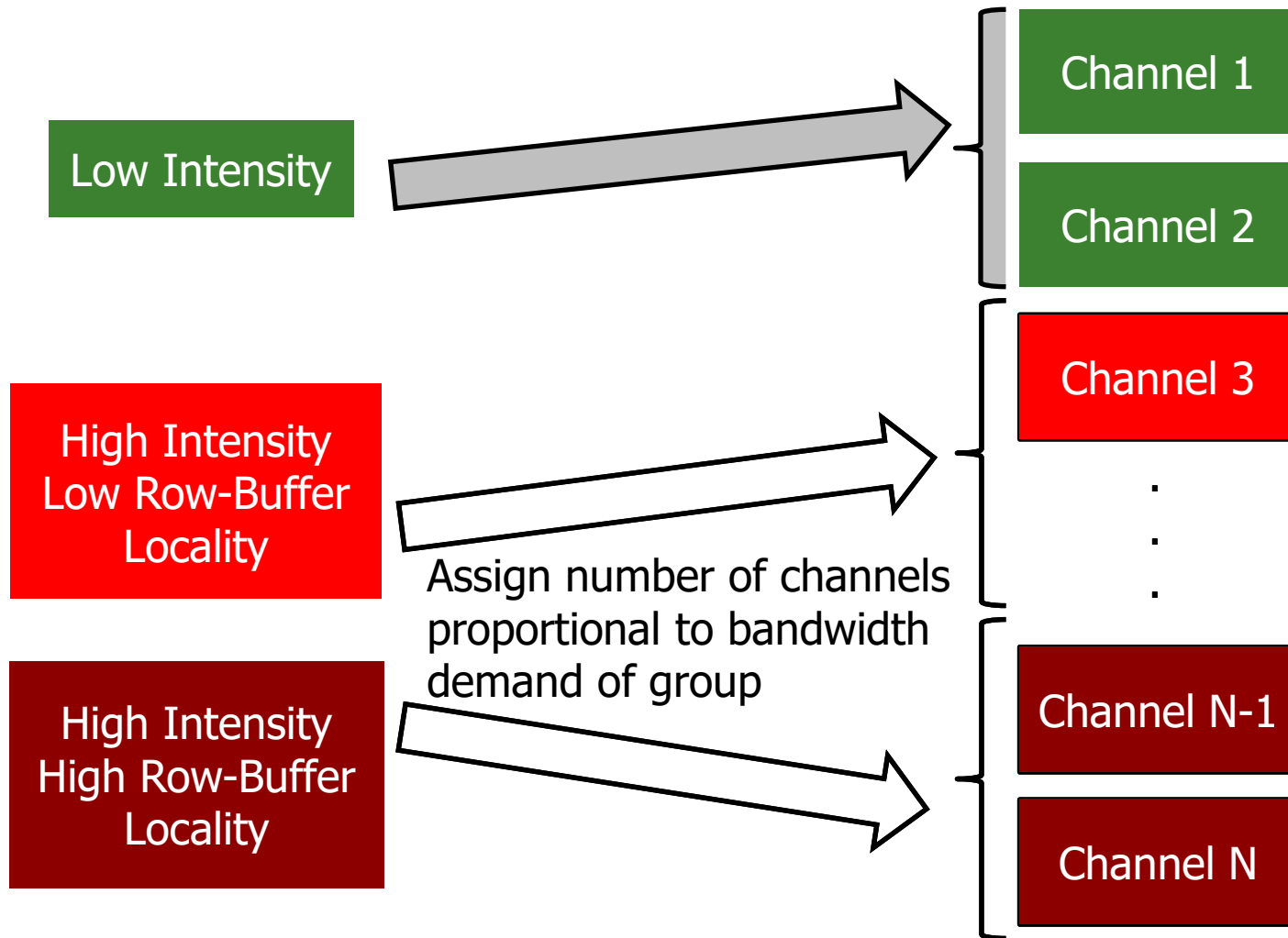
Channel 3

⋮

Channel N-1

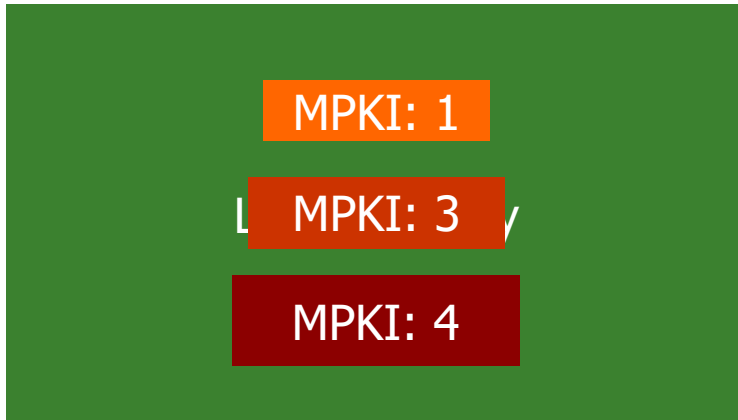
Channel N

3. Partition Channels Among Groups: Step 2



4. Assign Preferred Channel to Application

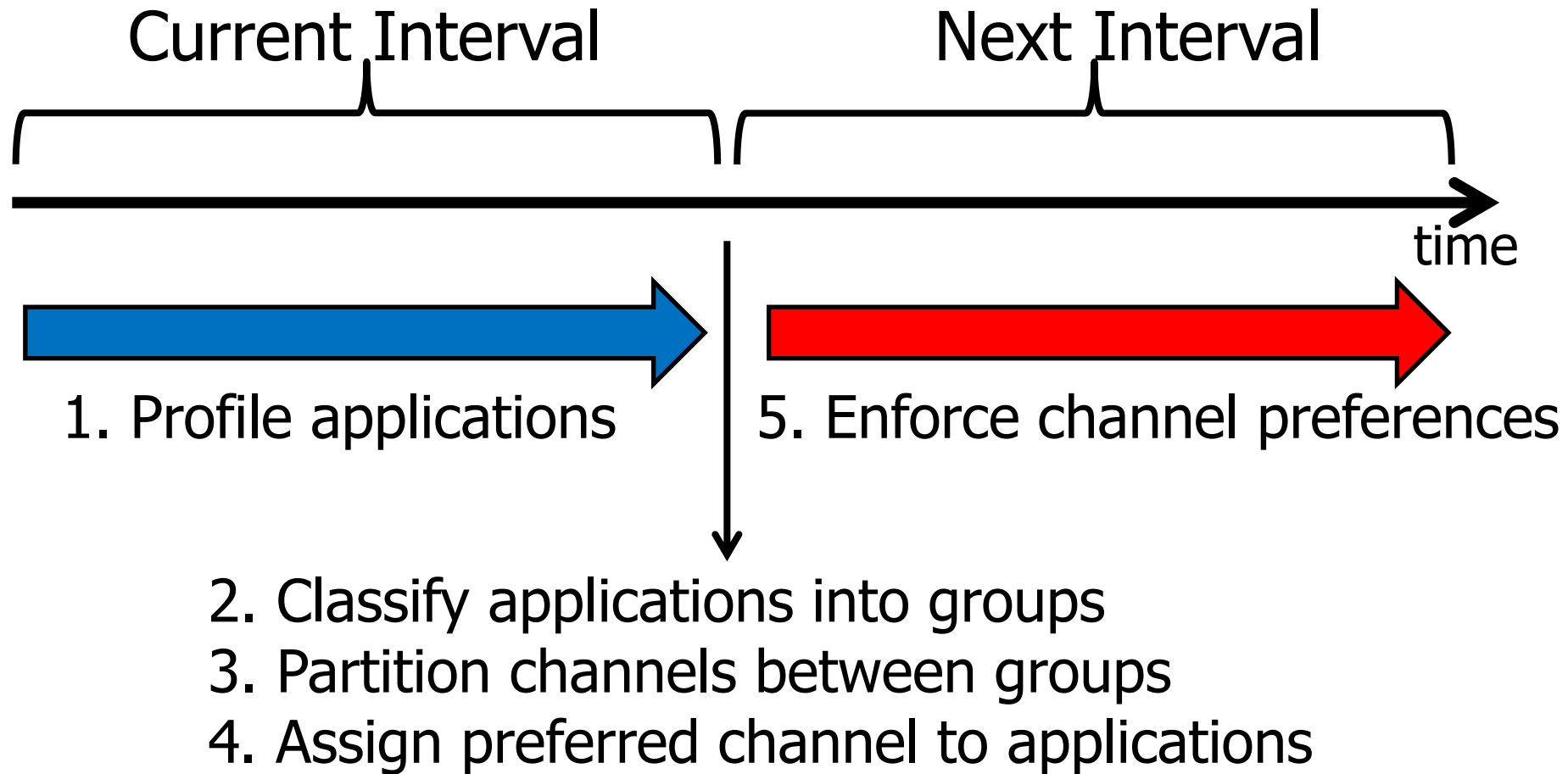
- Assign **each application a preferred channel** from its group's allocated channels
- Distribute applications to channels such that **group's bandwidth demand is balanced** across its channels



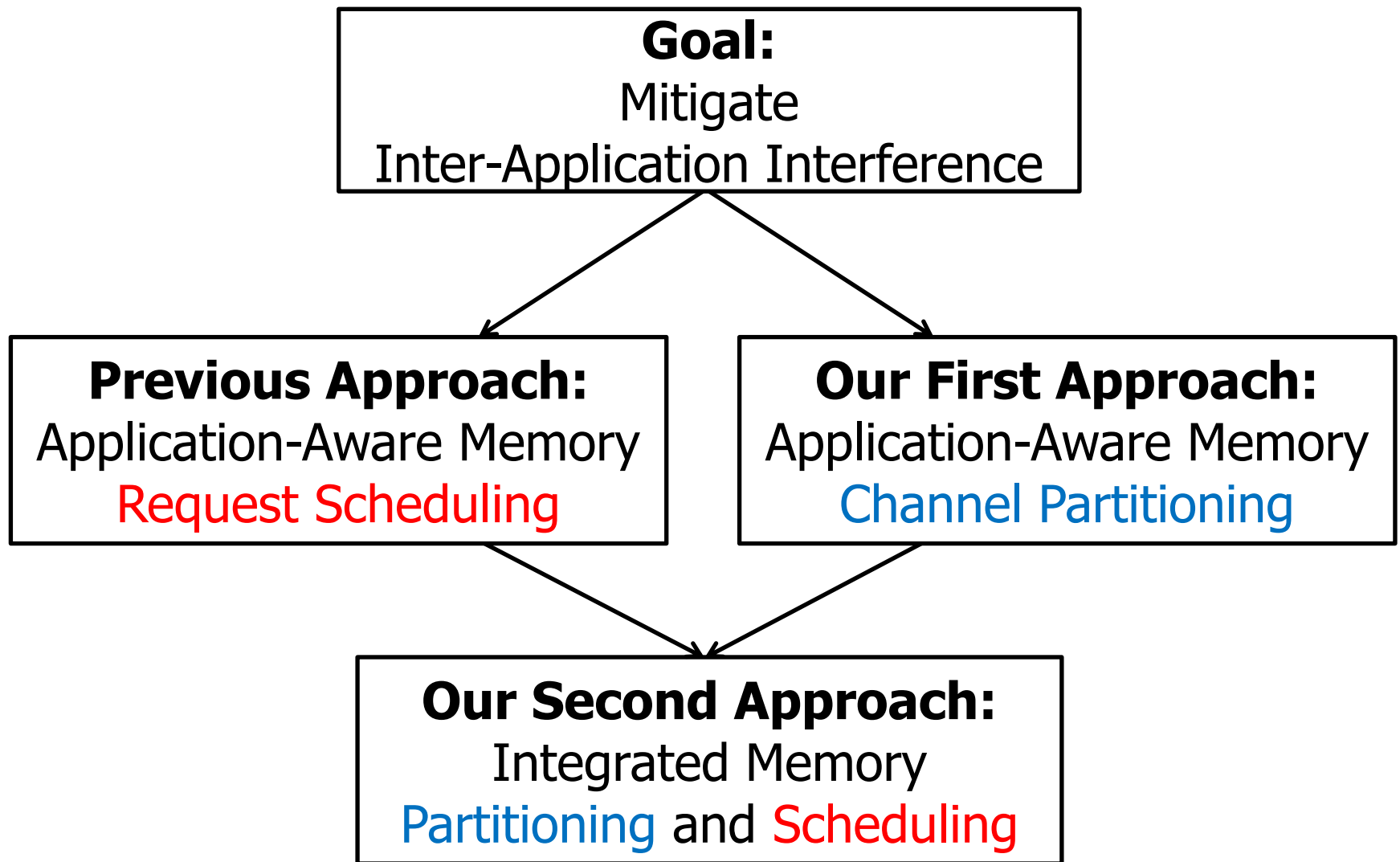
5. Allocate Page to Preferred Channel

- **Enforce channel preferences**
computed in the previous step
- On a page fault, the operating system
 - ❑ allocates page to preferred channel **if free page available** in preferred channel
 - ❑ **if free page not available**, replacement policy tries to allocate page to preferred channel
 - ❑ **if it fails**, allocate page to another channel

Interval Based Operation



Integrating Partitioning and Scheduling



Observations

- Applications with very low memory-intensity rarely access memory
 - Dedicating channels to them results in precious memory bandwidth waste
- They have the most potential to keep their cores busy
 - We would really like to prioritize them
- They interfere minimally with other applications
 - Prioritizing them does not hurt others

Integrated Memory Partitioning and Scheduling (IMPS)

- Always prioritize very low memory-intensity applications in the memory scheduler
- Use memory channel partitioning to mitigate interference between other applications

Key Results:

Methodology and Evaluation

Hardware Cost

- Memory Channel Partitioning (MCP)
 - ❑ Only profiling counters in hardware
 - ❑ No modifications to memory scheduling logic
 - ❑ 1.5 KB storage cost for a 24-core, 4-channel system
- Integrated Memory Partitioning and Scheduling (IMPS)
 - ❑ A single bit per request
 - ❑ Scheduler prioritizes based on this single bit

Methodology

■ Simulation Model

- ❑ 24 cores, 4 channels, 4 banks/channel
- ❑ Core Model
 - Out-of-order, 128-entry instruction window
 - 512 KB L2 cache/core
- ❑ Memory Model – DDR2

■ Workloads

- ❑ 240 SPEC CPU 2006 multiprogrammed workloads (categorized based on memory intensity)

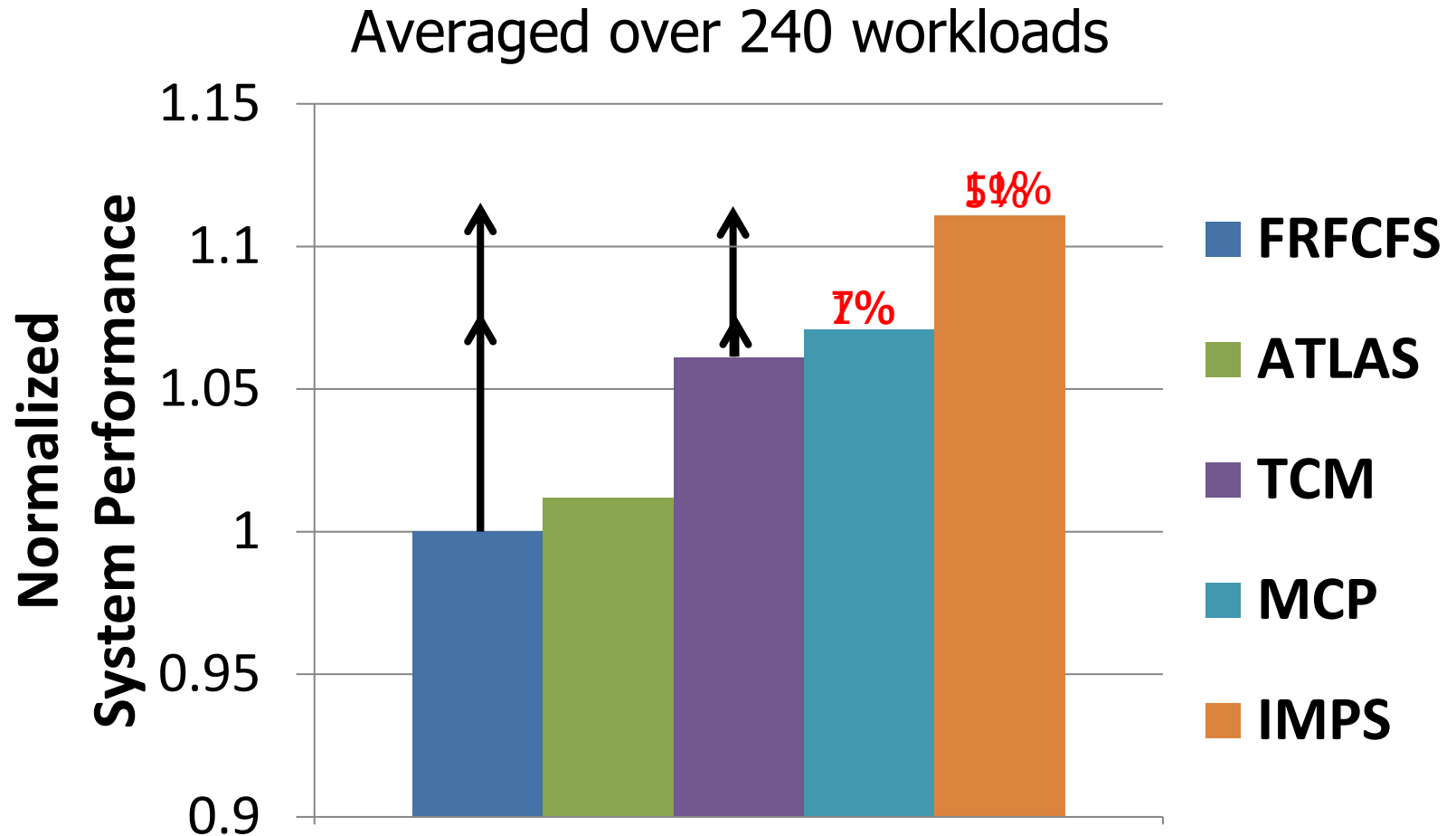
■ Metrics

- ❑ System Performance $WeightedSpeedup = \sum_i \frac{IPC_i^{shared}}{IPC_i^{alone}}$

Previous Work on Memory Scheduling

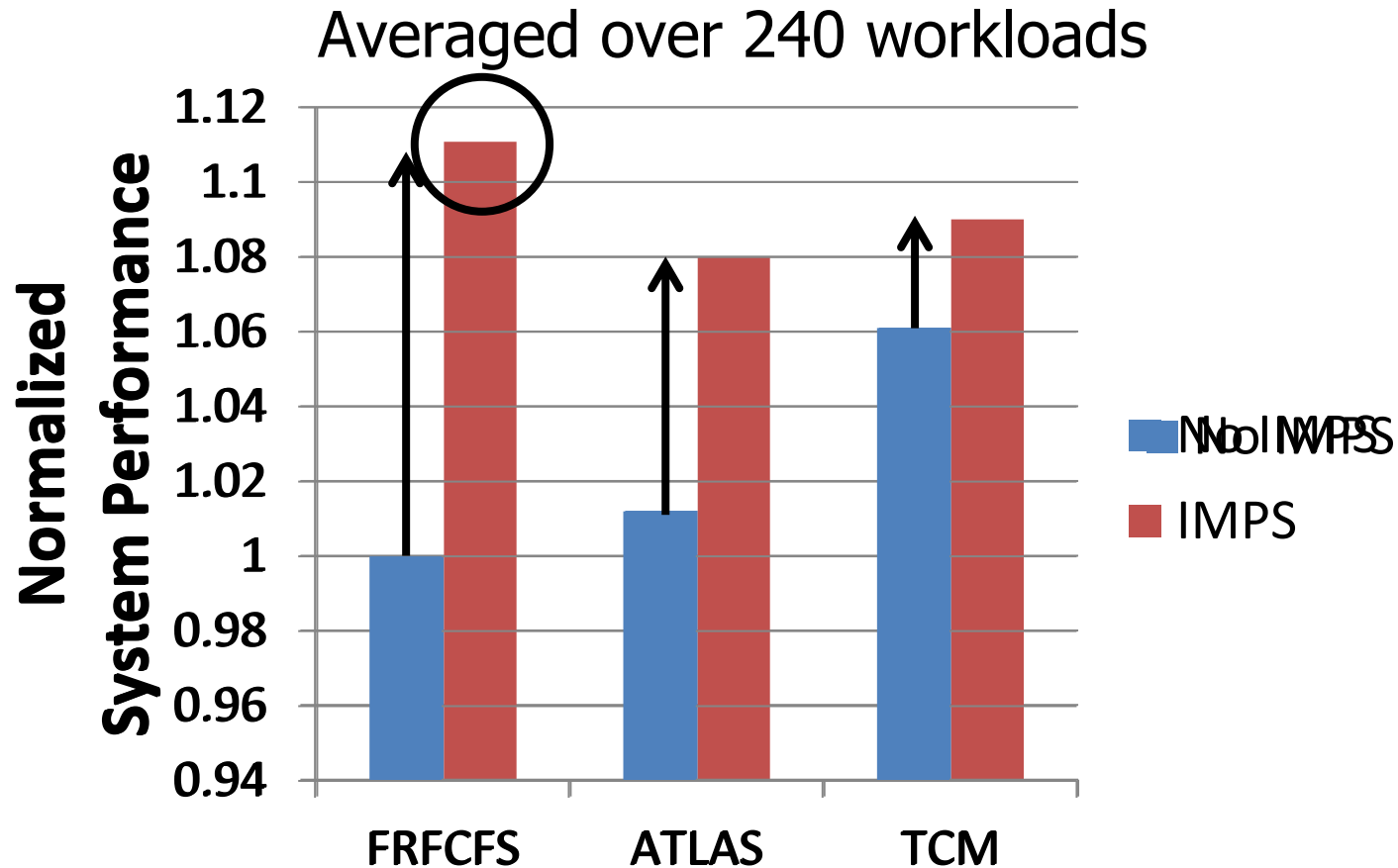
- **FR-FCFS** [Zuravleff et al., US Patent 1997, Rixner et al., ISCA 2000]
 - ❑ Prioritizes row-buffer hits and older requests
 - ❑ Application-unaware
- **ATLAS** [Kim et al., HPCA 2010]
 - ❑ Prioritizes applications with low memory-intensity
- **TCM** [Kim et al., MICRO 2010]
 - ❑ Always prioritizes low memory-intensity applications
 - ❑ Shuffles request priorities of high memory-intensity applications

Comparison to Previous Scheduling Policies



Better system performance than the best previous scheduler
Significant performance improvement over baseline FRFCFS
at lower hardware cost

Interaction with Memory Scheduling



IMPS improves performance regardless of scheduling policy
Highest improvement over FRFCFS as IMPS designed for FRFCFS

Summary

Summary

- Uncontrolled inter-application interference in main memory degrades system performance
- Application-aware memory channel partitioning (MCP)
 - Separates the data of badly-interfering applications to different channels, eliminating interference
- Integrated memory partitioning and scheduling (IMPS)
 - Prioritizes very low memory-intensity applications in scheduler
 - Handles other applications' interference by partitioning
- MCP/IMPS provide better performance than application-aware memory request scheduling at lower hardware cost

Strengths

Strengths of the Paper

- Novel solution to a key problem in multi-core systems, memory interference; the importance of problem will increase over time
- Keeps the memory scheduling hardware simple
- Combines multiple interference reduction techniques
- Can provide performance isolation across applications mapped to different channels
- General idea of partitioning can be extended to smaller granularities in the memory hierarchy: banks, subarrays, etc.
- Well-written paper
- Thorough simulation-based evaluation

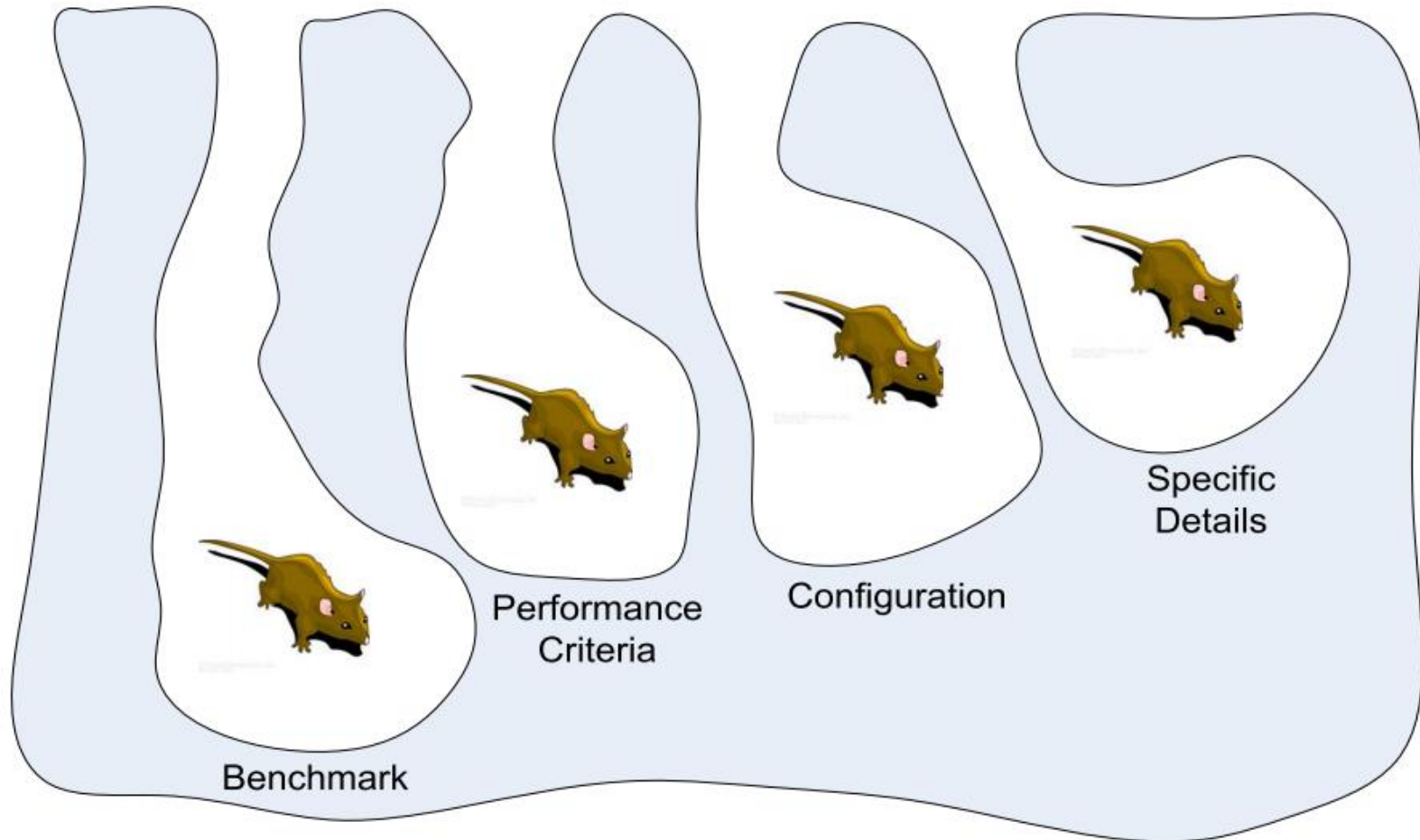
Weaknesses

Weaknesses/Limitations of the Paper

- Mechanism may not work effectively if workload changes behavior after profiling
- Overhead of moving pages between channels restricts mechanism's benefits
- Small number of memory channels reduces the scope of partitioning
- Load imbalance across channels can reduce performance
 - The paper addresses this and compares to another mechanism
- Software-hardware cooperative solution might not always be easy to adopt
- Evaluation is done solely in simulation
- Evaluation does not consider multi-chip systems
- Are these the best workloads to evaluate?

Recall: Try to Avoid Rat Holes

Performance Analysis Rat Holes



Thoughts and Ideas

Extensions (I)

- Can this idea be extended to different granularities in memory?
 - Partition banks, subarrays, mats across workloads
- Can this idea be extended to provide performance predictability and performance isolation? How?
- How can MCP be combined effectively with other interference reduction techniques?
 - E.g., source throttling methods [Ebrahimi+, ASPLOS 2010]
 - E.g., thread scheduling methods
- Can this idea be evaluated on a real system? How?

Aside: Source Throttling

- Eiman Ebrahimi, Chang Joo Lee, Onur Mutlu, and Yale N. Patt, **"Fairness via Source Throttling: A Configurable and High-Performance Fairness Substrate for Multi-Core Memory Systems"**

Proceedings of the 15th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), pages 335-346, Pittsburgh, PA, March 2010. [Slides](#) [\(pdf\)](#)

Best paper award.

Fairness via Source Throttling: A Configurable and High-Performance Fairness Substrate for Multi-Core Memory Systems

Eiman Ebrahimi[†] Chang Joo Lee[†] Onur Mutlu[§] Yale N. Patt[†]

[†]Department of Electrical and Computer Engineering
The University of Texas at Austin
{ebrahimi, cjlee, patt}@ece.utexas.edu

[§]Computer Architecture Laboratory (CALCM)
Carnegie Mellon University
onur@cmu.edu

Takeaways

Key Takeaways

- A novel method to reduce memory interference
- Simple and effective
- Hardware/software cooperative
- Good potential for work building on it to extend it
 - To different structures
 - To different metrics
 - Multiple works have already built on the paper (see bank partitioning works in PACT 2012, HPCA 2012 + HPCA 2013)
- Easy to read and understand paper

Example: Application to Core Mapping

- Reetuparna Das, Rachata Ausavarungnirun, Onur Mutlu, Akhilesh Kumar, and Mani Azimi,

"Application-to-Core Mapping Policies to Reduce Memory System Interference in Multi-Core Systems"

Proceedings of the 19th International Symposium on High-Performance Computer Architecture (HPCA), Shenzhen, China, February 2013. Slides (pptx)

Application-to-Core Mapping Policies to Reduce Memory System Interference in Multi-Core Systems

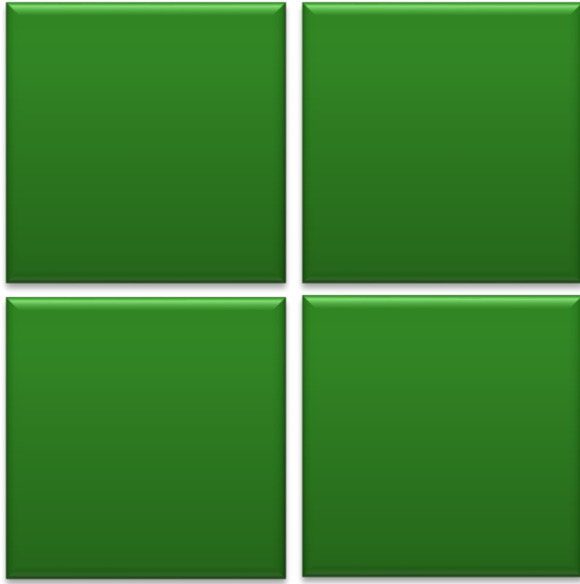
Reetuparna Das* Rachata Ausavarungnirun† Onur Mutlu† Akhilesh Kumar‡ Mani Azimi‡
University of Michigan* Carnegie Mellon University† Intel Labs‡

Application-to-Core Mapping to Reduce Interference

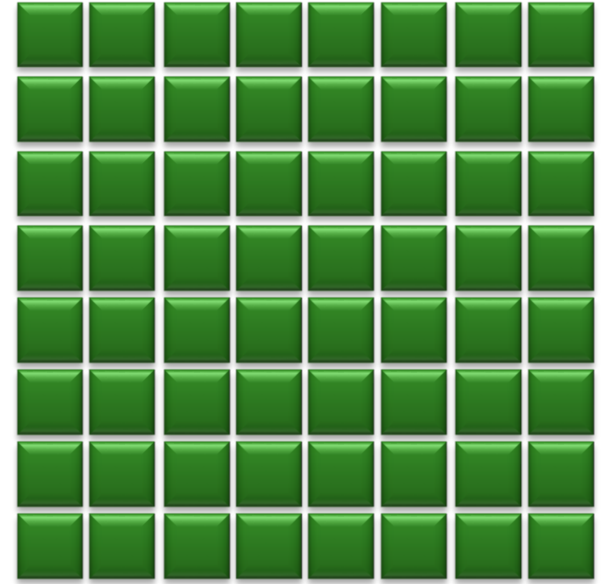
- Reetuparna Das, Rachata Ausavarungnirun, Onur Mutlu, Akhilesh Kumar, and Mani Azimi,
"Application-to-Core Mapping Policies to Reduce Memory System Interference in Multi-Core Systems"
Proceedings of the 19th International Symposium on High-Performance Computer Architecture (HPCA), Shenzhen, China, February 2013.
Slides (pptx)

- Key ideas:
 - ❑ Cluster threads to memory controllers (to reduce across chip interference)
 - ❑ Isolate interference-sensitive (low-intensity) applications in a separate cluster (to reduce interference from high-intensity applications)
 - ❑ Place applications that benefit from memory bandwidth closer to the controller (to improve performance)

Multi-Core to Many-Core



Multi-Core



Many-Core

Many-Core On-Chip Communication

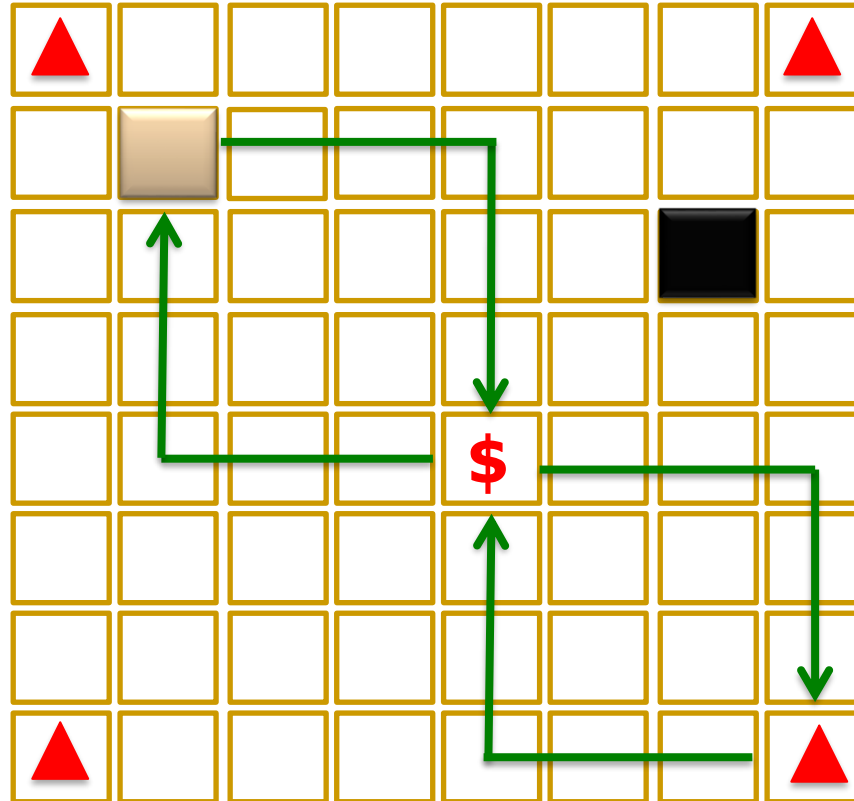
Applications



Light



Heavy



**Memory
Controller**

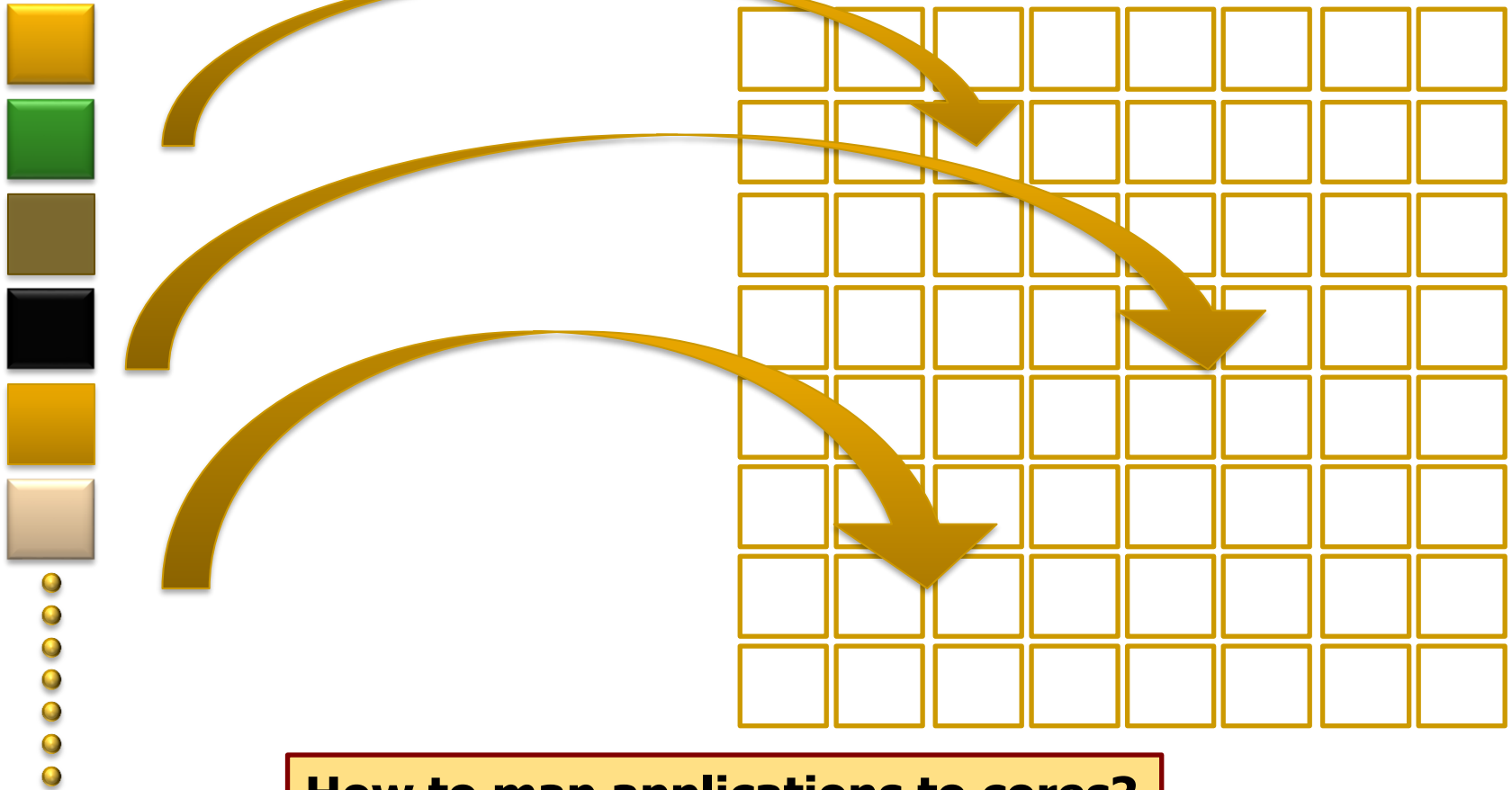


**Shared
Cache Bank**

Problem: Spatial Task Scheduling

Applications

Cores



How to map applications to cores?

Challenges in Spatial Task Scheduling

Applications

Cores

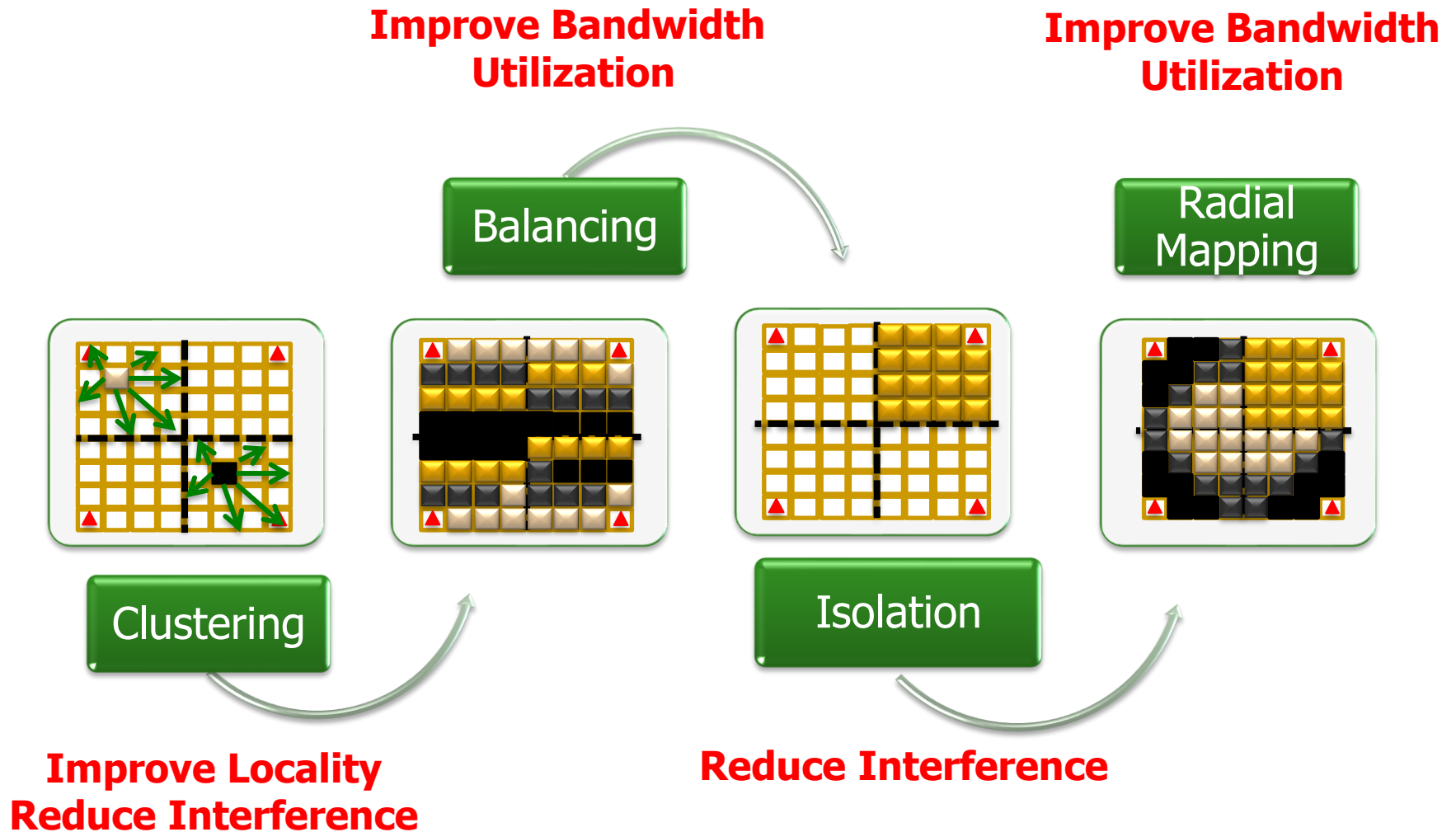


How to reduce communication distance?

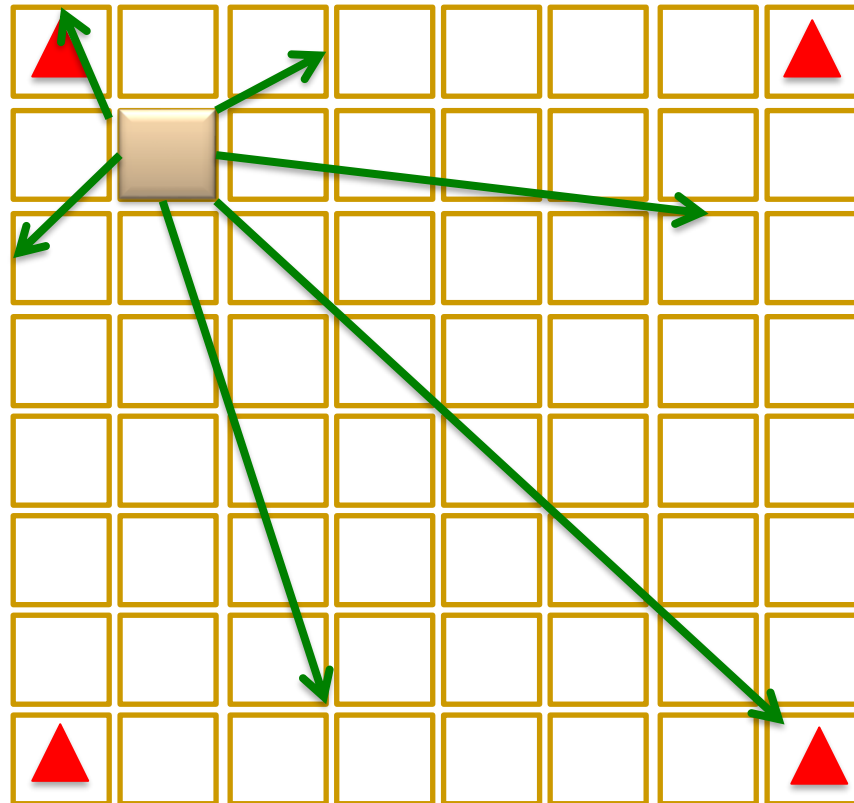
How to reduce destructive interference between applications?

How to prioritize applications to improve throughput?

Application-to-Core Mapping



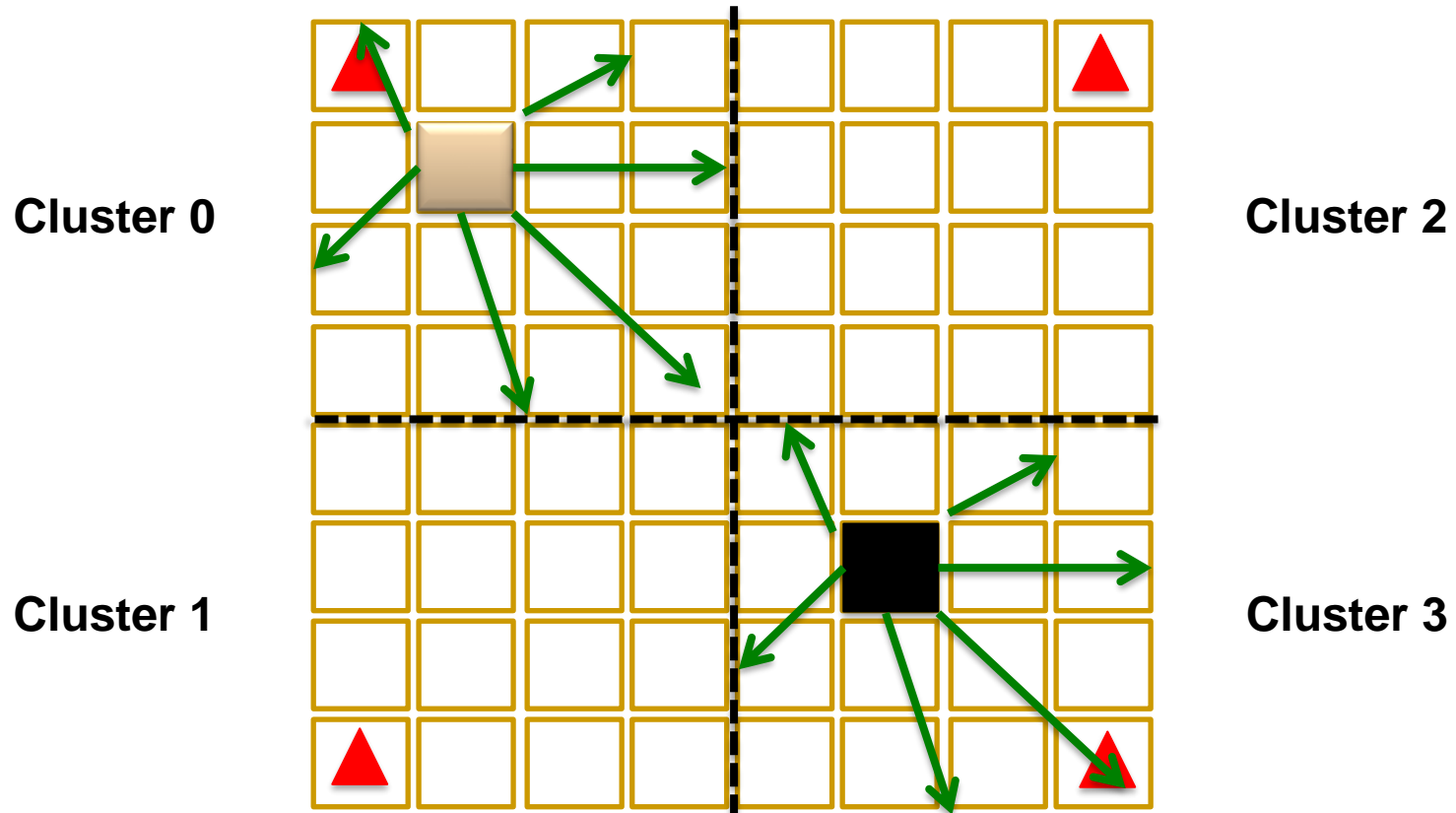
Step 1 — Clustering



 **Memory
Controller**

Inefficient data mapping to memory and caches

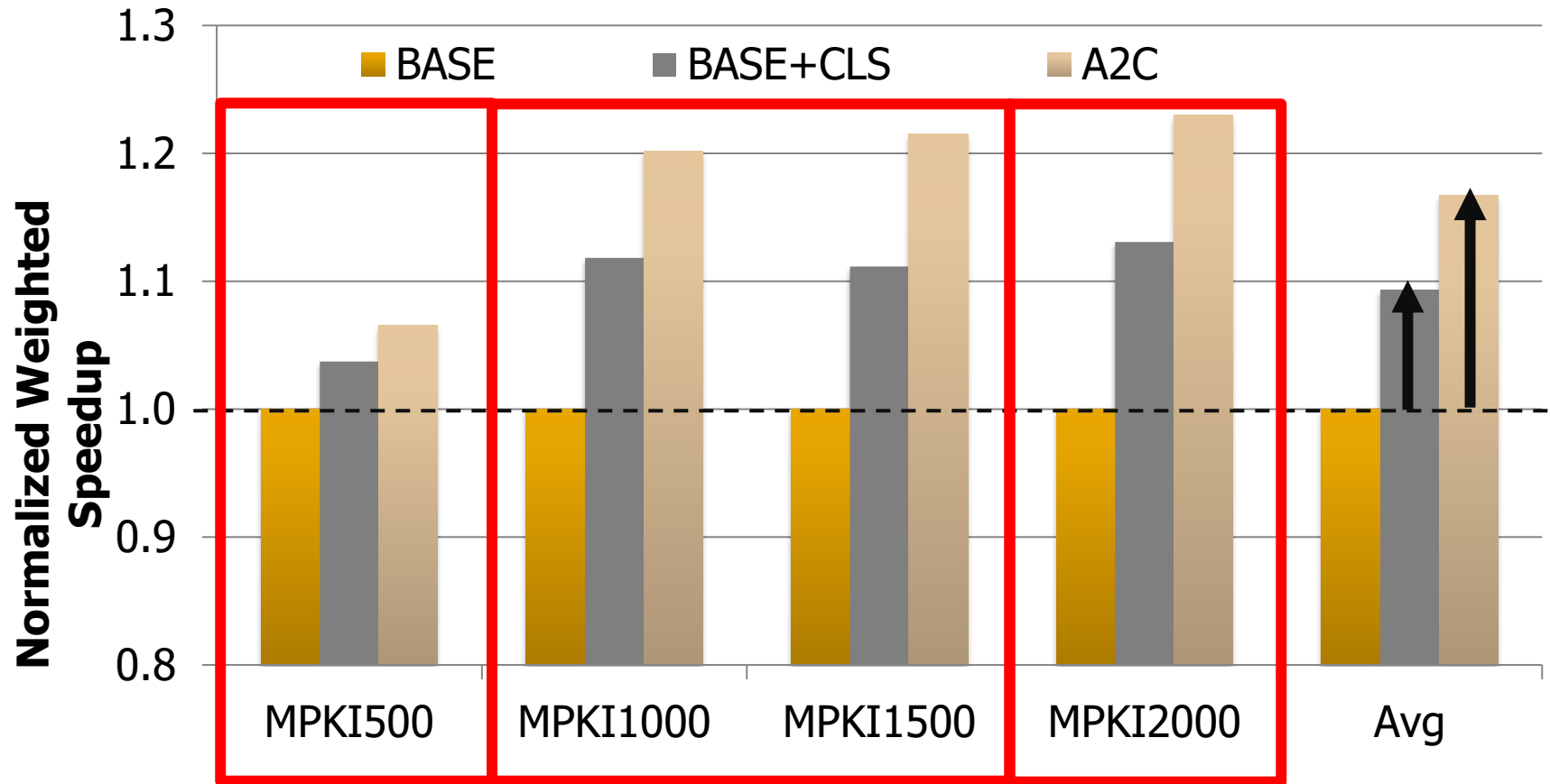
Step 1 — Clustering



Improved Locality

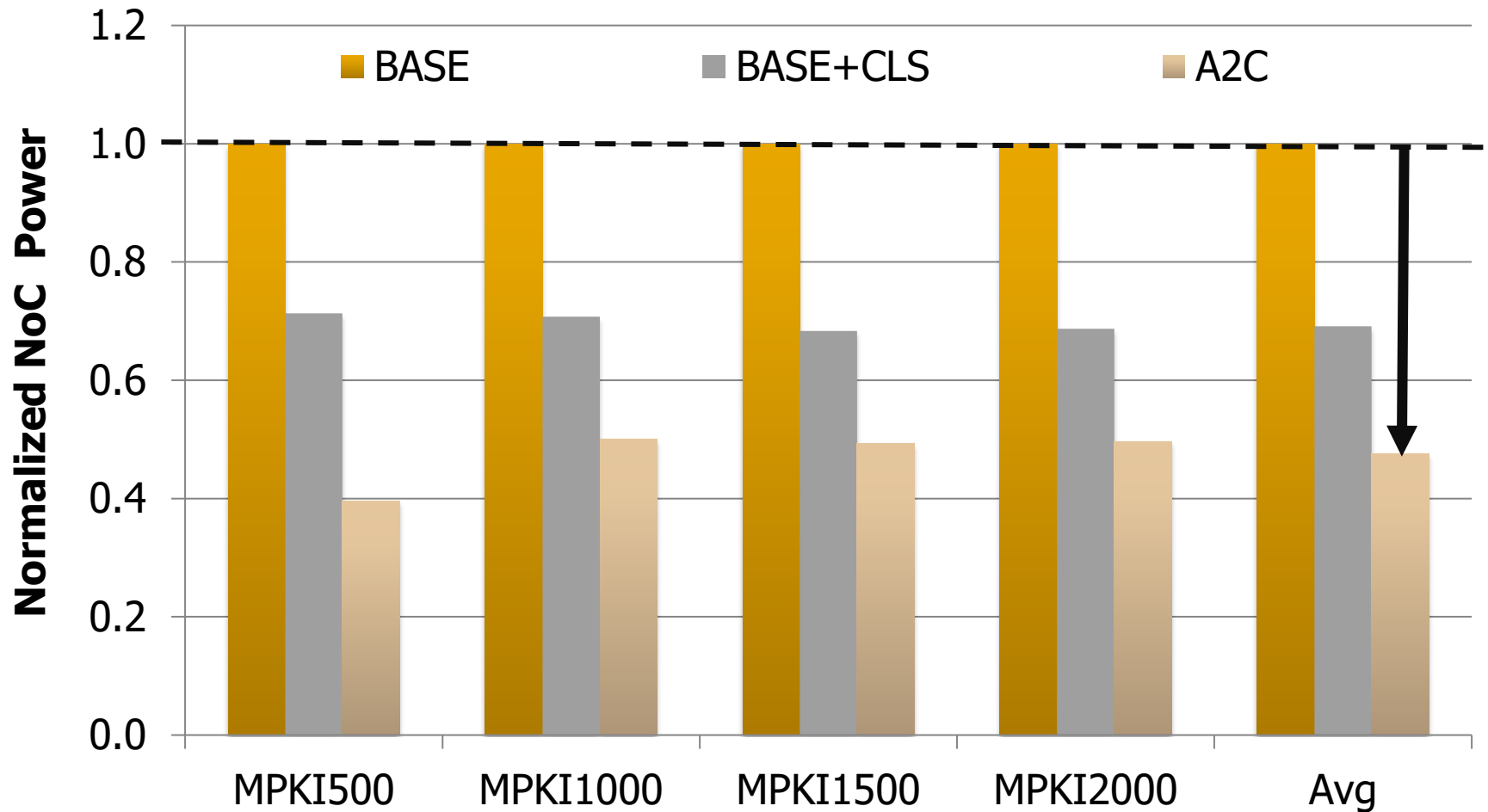
Reduced Interference

System Performance



System performance improves by 17%

Network Power



Average network power consumption reduces by 52%

Example: Application to Core Mapping

- Reetuparna Das, Rachata Ausavarungnirun, Onur Mutlu, Akhilesh Kumar, and Mani Azimi,

"Application-to-Core Mapping Policies to Reduce Memory System Interference in Multi-Core Systems"

Proceedings of the 19th International Symposium on High-Performance Computer Architecture (HPCA), Shenzhen, China, February 2013. Slides (pptx)

Application-to-Core Mapping Policies to Reduce Memory System Interference in Multi-Core Systems

Reetuparna Das* Rachata Ausavarungnirun† Onur Mutlu† Akhilesh Kumar‡ Mani Azimi‡
University of Michigan* Carnegie Mellon University† Intel Labs‡

Example Follow-On Works (II)

- https://lph.ece.utexas.edu/merez/uploads/MattanErez/bpart_hpca12.pdf

Balancing DRAM Locality and Parallelism in Shared Memory CMP Systems

Min Kyu Jeong^{*}, Doe Hyun Yoon[†], Dam Sunwoo[‡], Michael Sullivan^{*}, Ikhwan Lee^{*}, and Mattan Erez^{*}

^{*} *Dept. of Electrical and Computer Engineering, The University of Texas at Austin*

[†] *Intelligent Infrastructure Lab, Hewlett-Packard Labs*

[‡] *ARM Inc.*

`{mkjeong, mbsullivan, ikhwan, mattan.erez}@mail.utexas.edu`
`doe-hyun.yoon@hp.com dam.sunwoo@arm.com`

Example Follow-On Work (III)

- <https://liulei-sys-inventor.github.io/files/pact140-liu-final.pdf>

A Software Memory Partition Approach for Eliminating Bank-level Interference in Multicore Systems

Lei Liu, Zehan Cui, Mingjie Xing and Chengyong Wu

State Key Laboratory of Computer Architecture, Institute of Computing Technology,
Chinese Academy of Science

(Revised 2016-01-01)

Open Discussion

Discussion Starters

- Thoughts on the previous ideas?
- How practical is this?
- Will the problem become bigger and more important over time?
- Will the solution become more important over time?
- Are other solutions better?
- Is this solution clearly advantageous in some cases?

Seminar in Computer Architecture

Meeting 4: Memory Channel Partitioning

Prof. Onur Mutlu

ETH Zürich

Fall 2021

14 October 2021