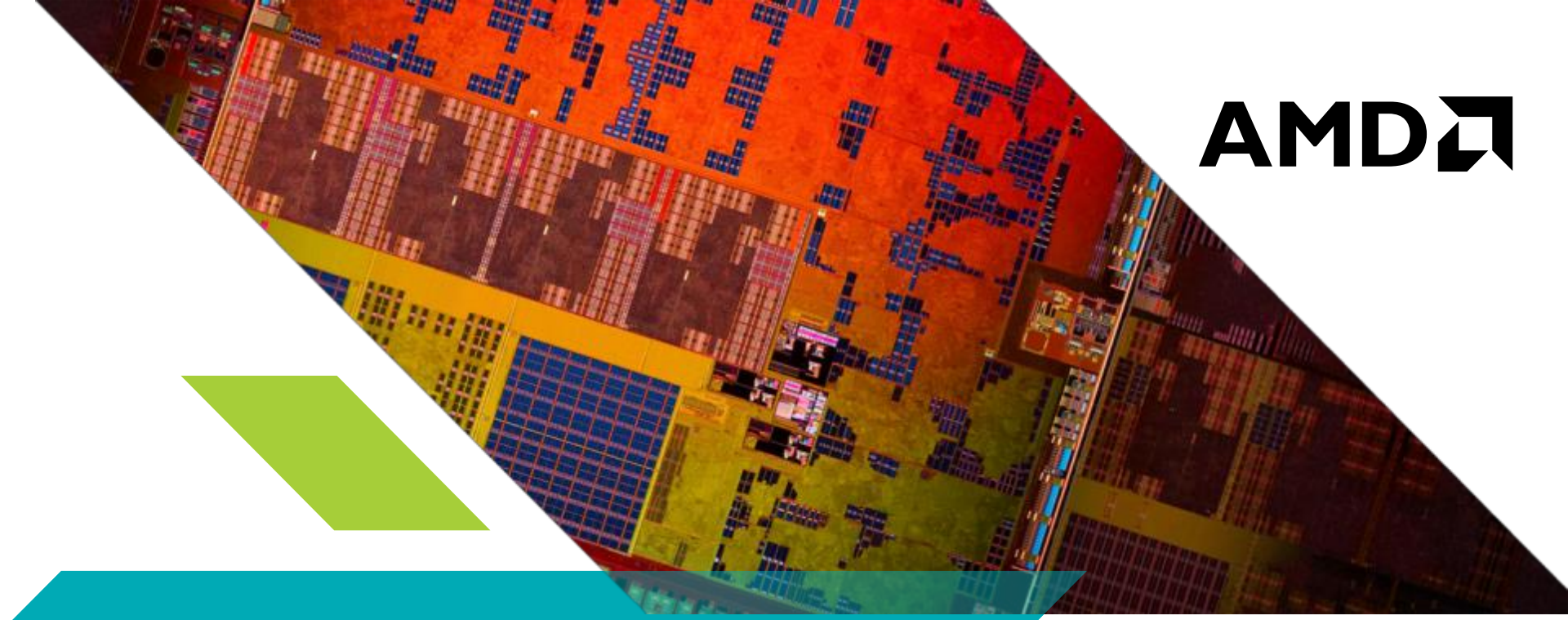


HETEROGENEOUS SYSTEM COHERENCE
FOR INTEGRATED CPU-GPU SYSTEMS

JASON POWER*, ARKAPRAVA BASU*, JUNLI GU†, SOORAJ PUTHOOR†,
BRADFORD M BECKMANN†, MARK D HILL*†, STEVEN K REINHARDT†, DAVID A WOOD*†

*University of Wisconsin-Madison
†Advanced Micro Devices, Inc.

AMD

Powerpoint version available on:

http://pages.cs.wisc.edu/~powerjg/

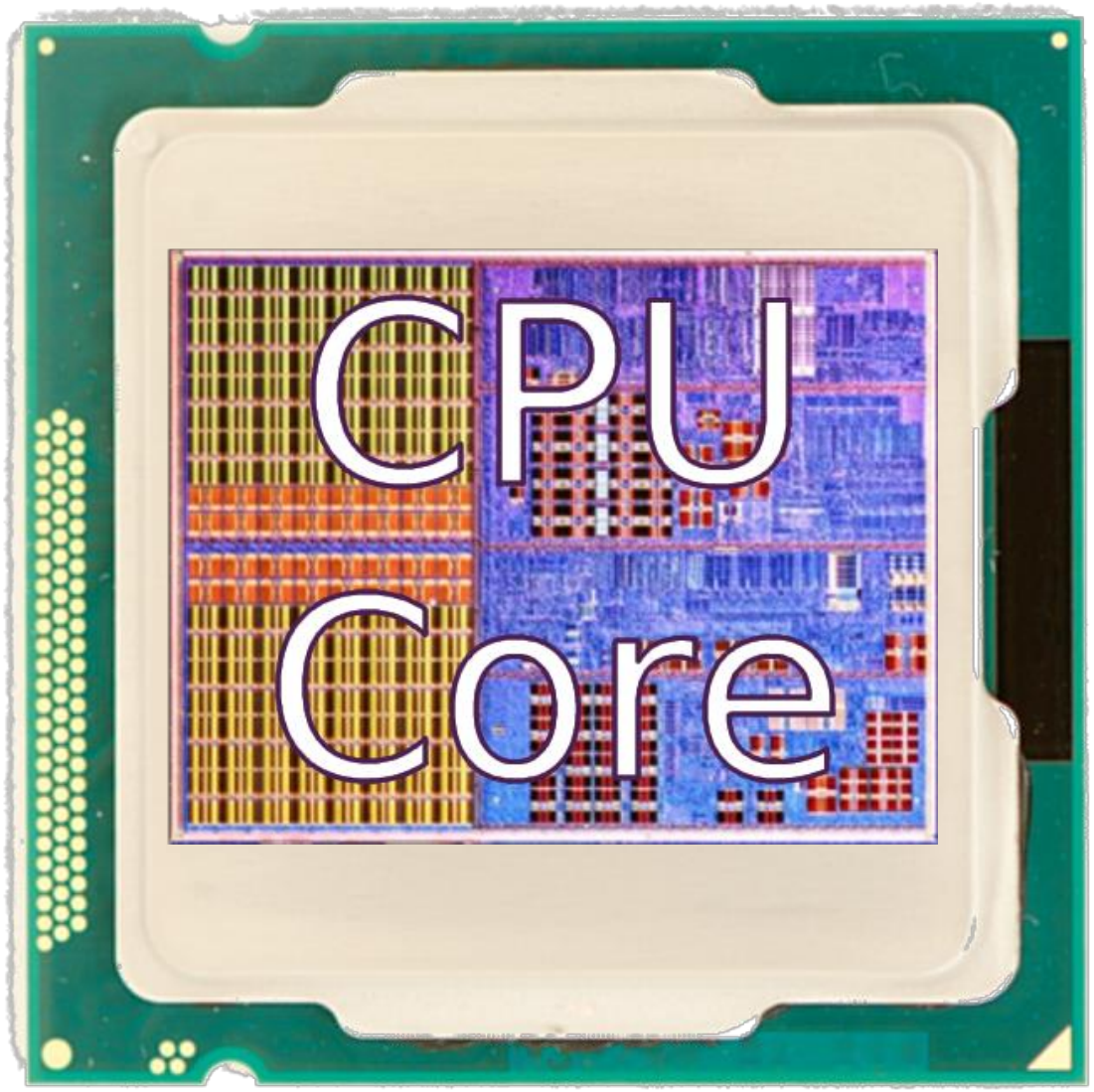
WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

ABSTRACT

- Hardware coherence can increase the utility of heterogeneous systems
- Major bottlenecks in current coherence implementations
  - High bandwidth difficult to support at directory
  - Extreme resource requirements
- We propose Heterogeneous System Coherence
  - Leverages spatial locality and region coherence
  - Reduces bandwidth by 94%
  - Reduces resource requirements by 95%

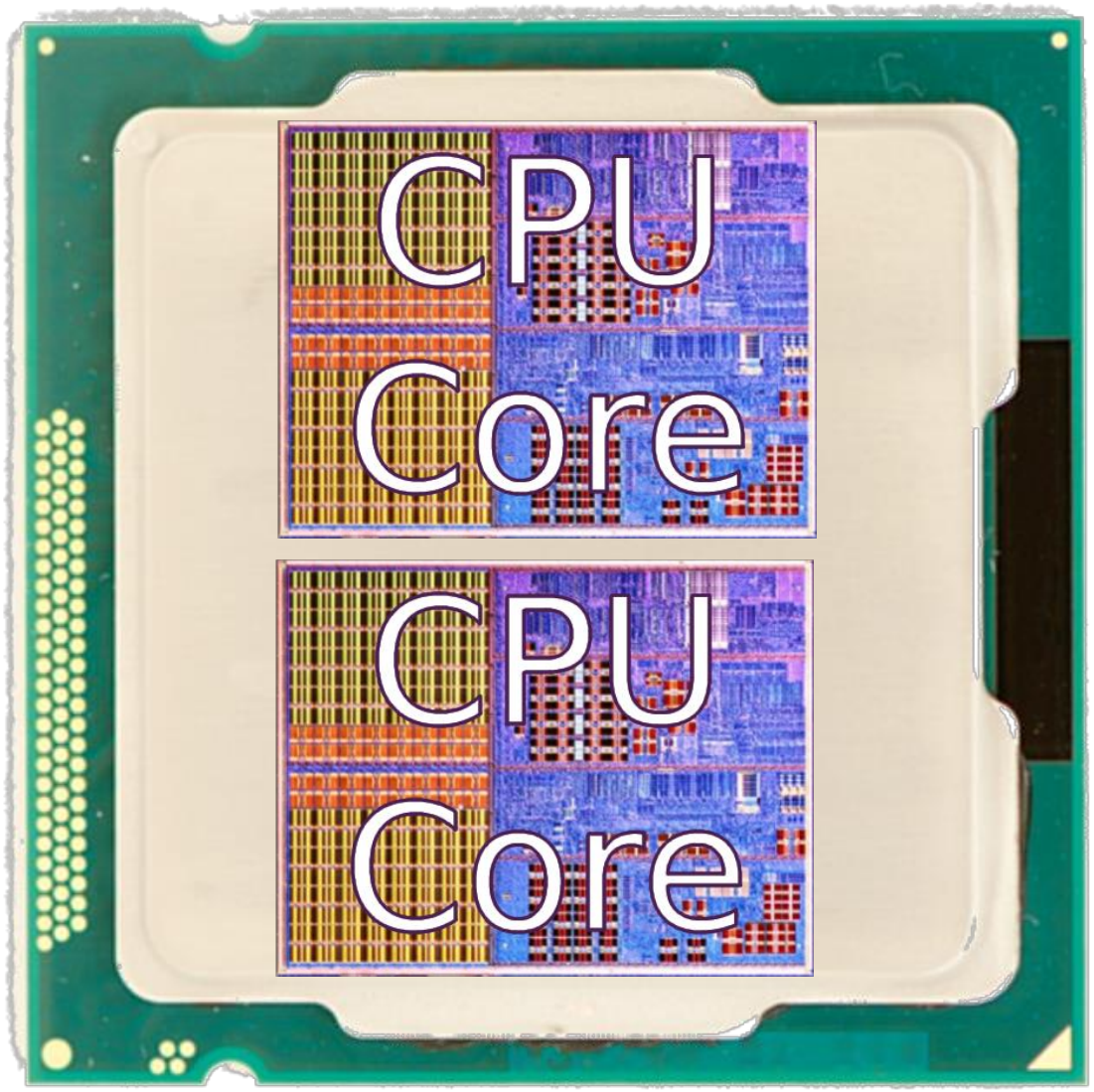
PHYSICAL INTEGRATION

AMD



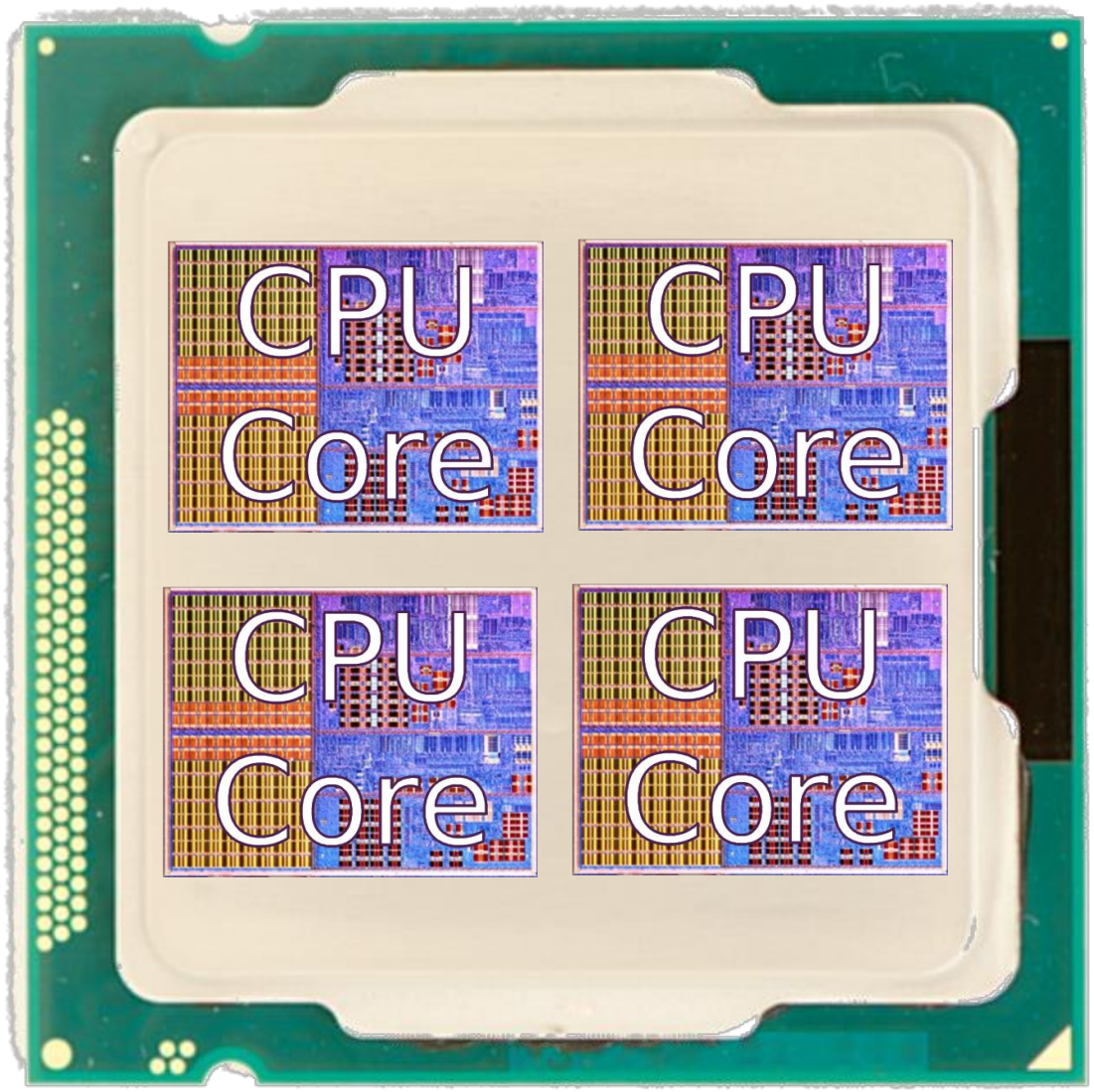
WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

PHYSICAL INTEGRATION



PHYSICAL INTEGRATION

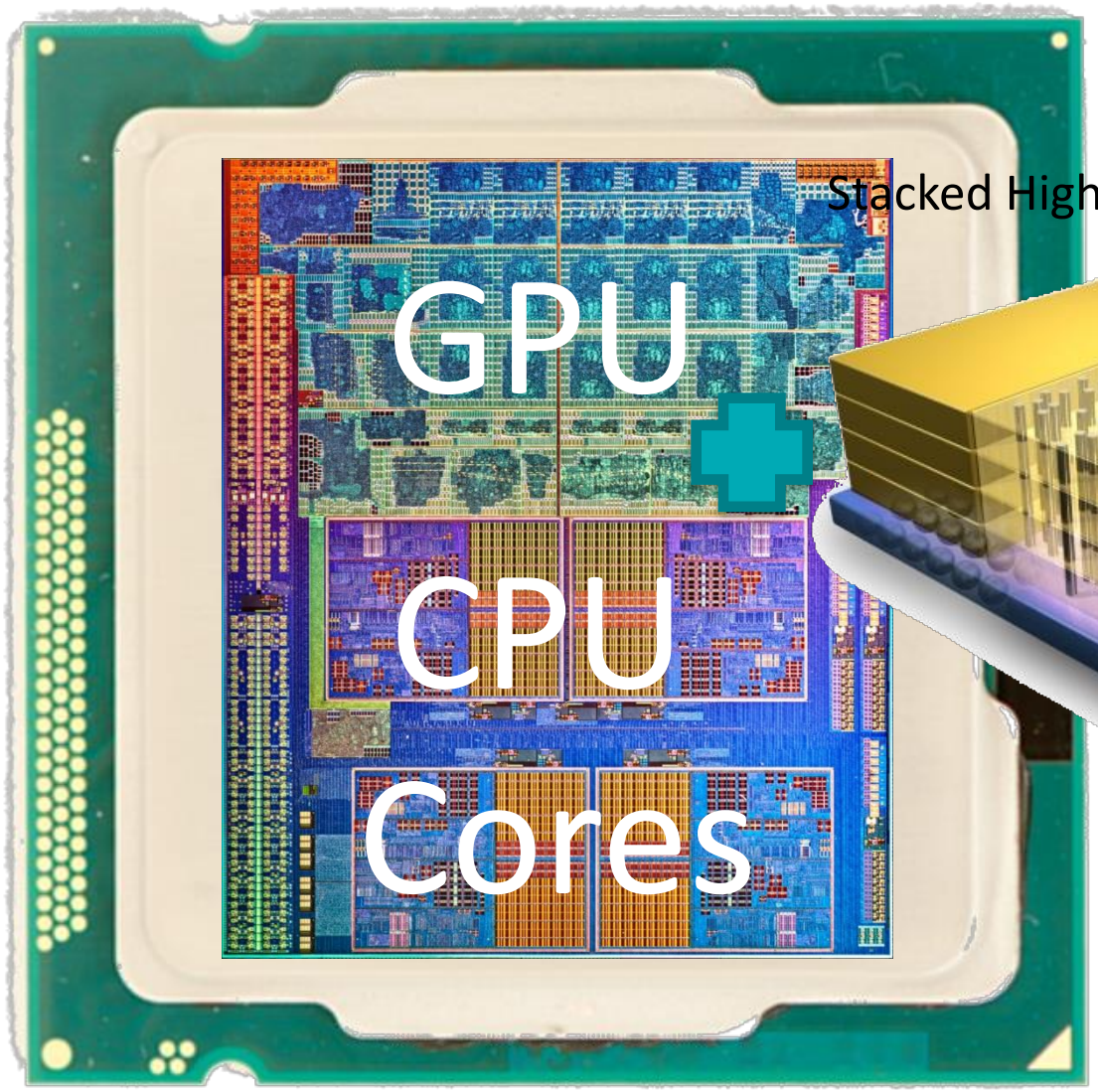
AMD



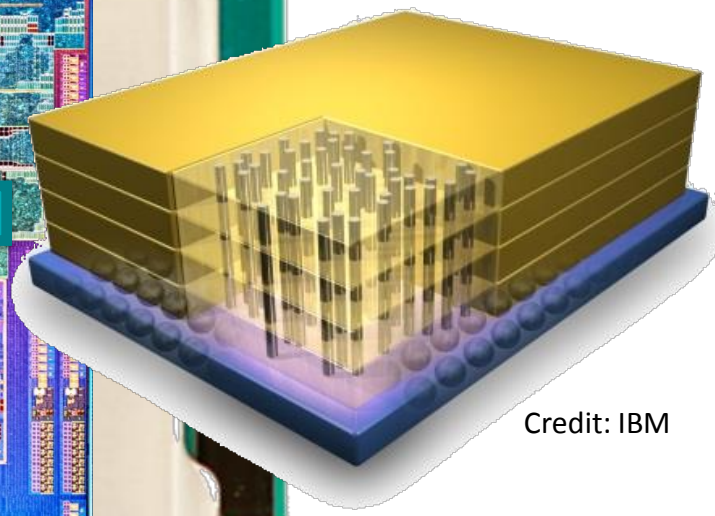
WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

PHYSICAL INTEGRATION

AMD



Stacked High-bandwidth DRAM



Credit: IBM

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

LOGICAL INTEGRATION

AMD

- General-purpose GPU computing
  - OpenCL
  - CUDA
- Heterogeneous Uniform Memory Access (hUMA)
  - Shared virtual address space
  - **Cache coherence**
- Allows new heterogeneous apps

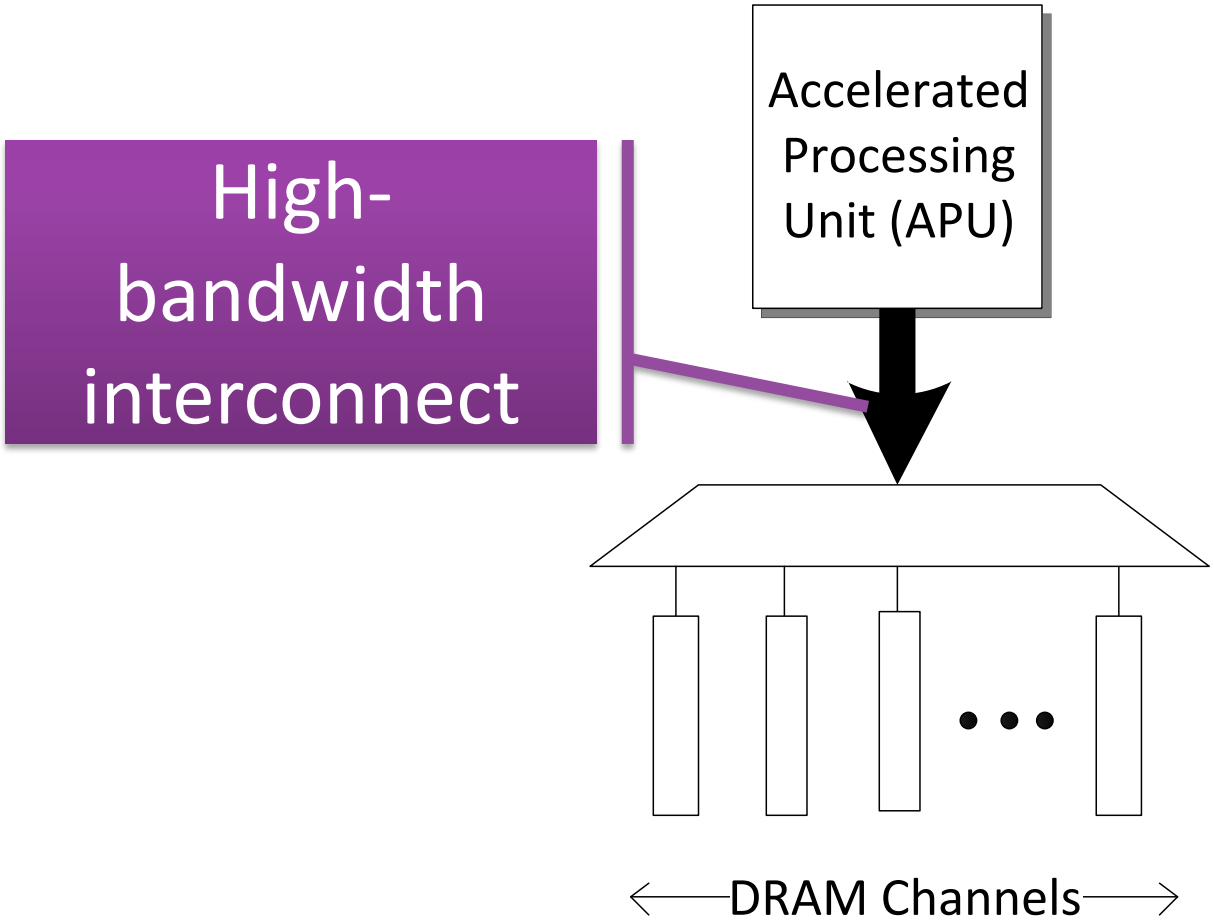
OUTLINE

AMD

- Motivation
- **Background**
  - System overview
  - Cache architecture reminder
- Heterogeneous System Bottlenecks
- Heterogeneous System Coherence Details
- Results
- Conclusions

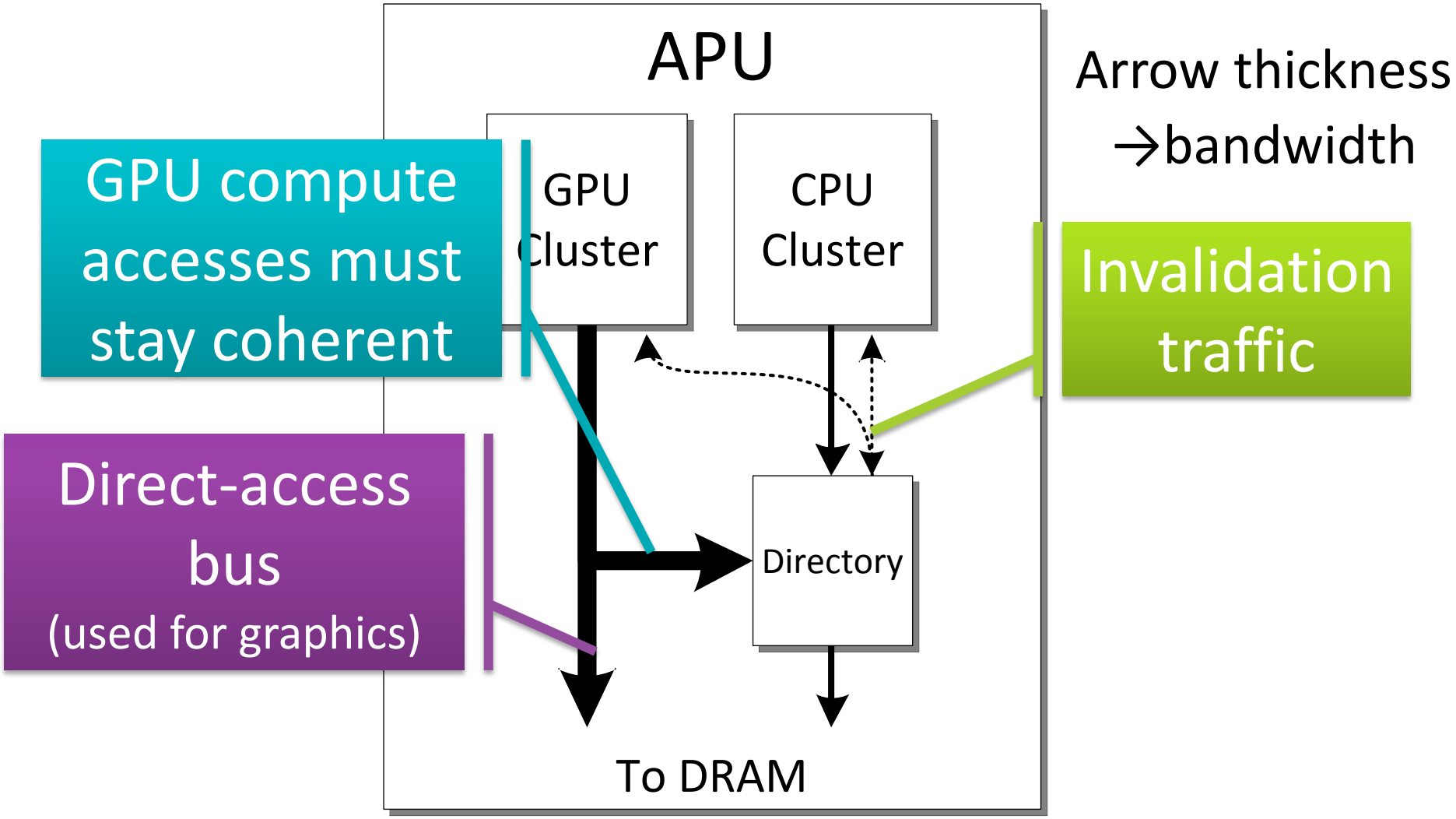
SYSTEM OVERVIEW

SYSTEM LEVEL



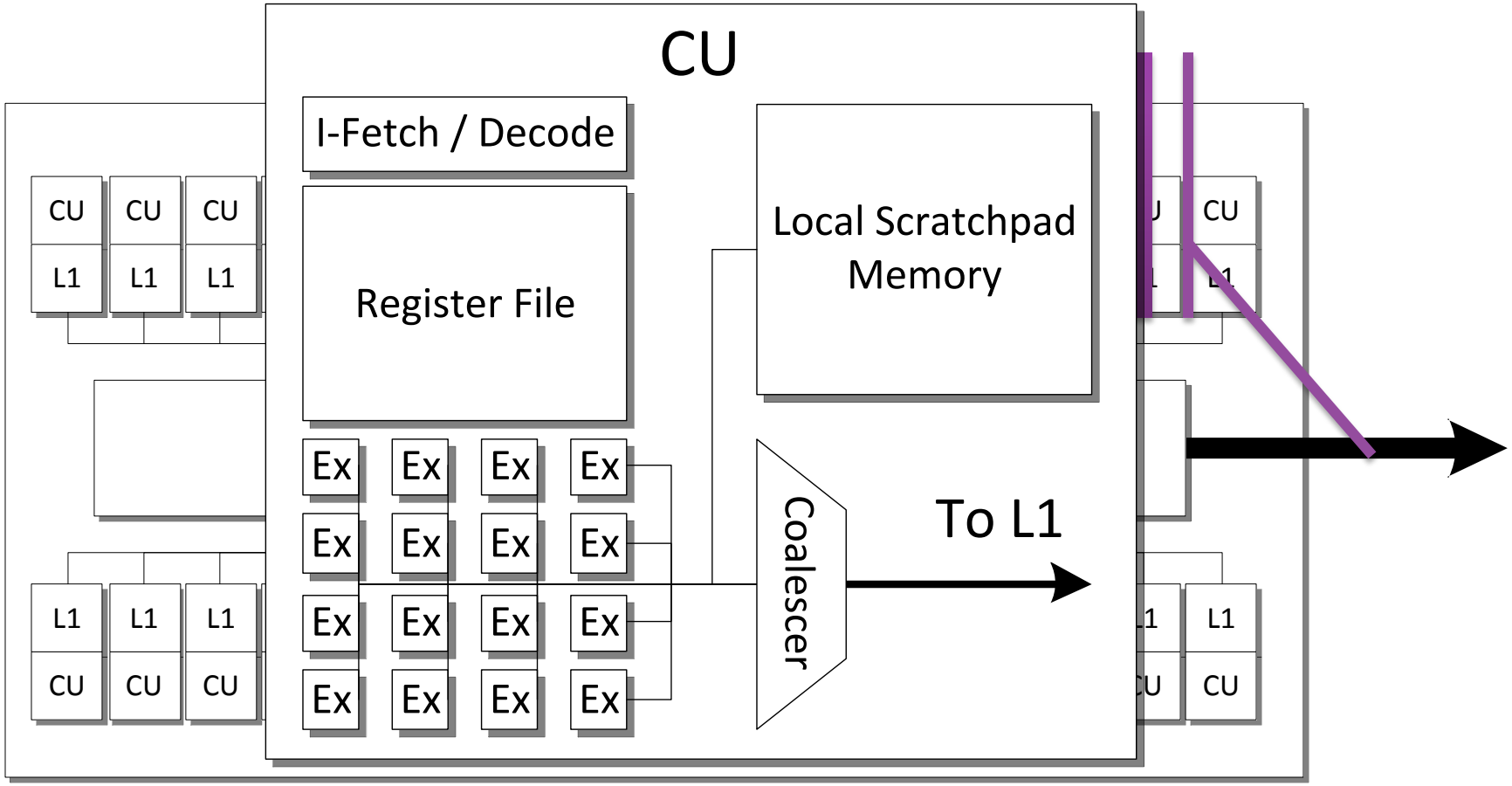
SYSTEM OVERVIEW

APU

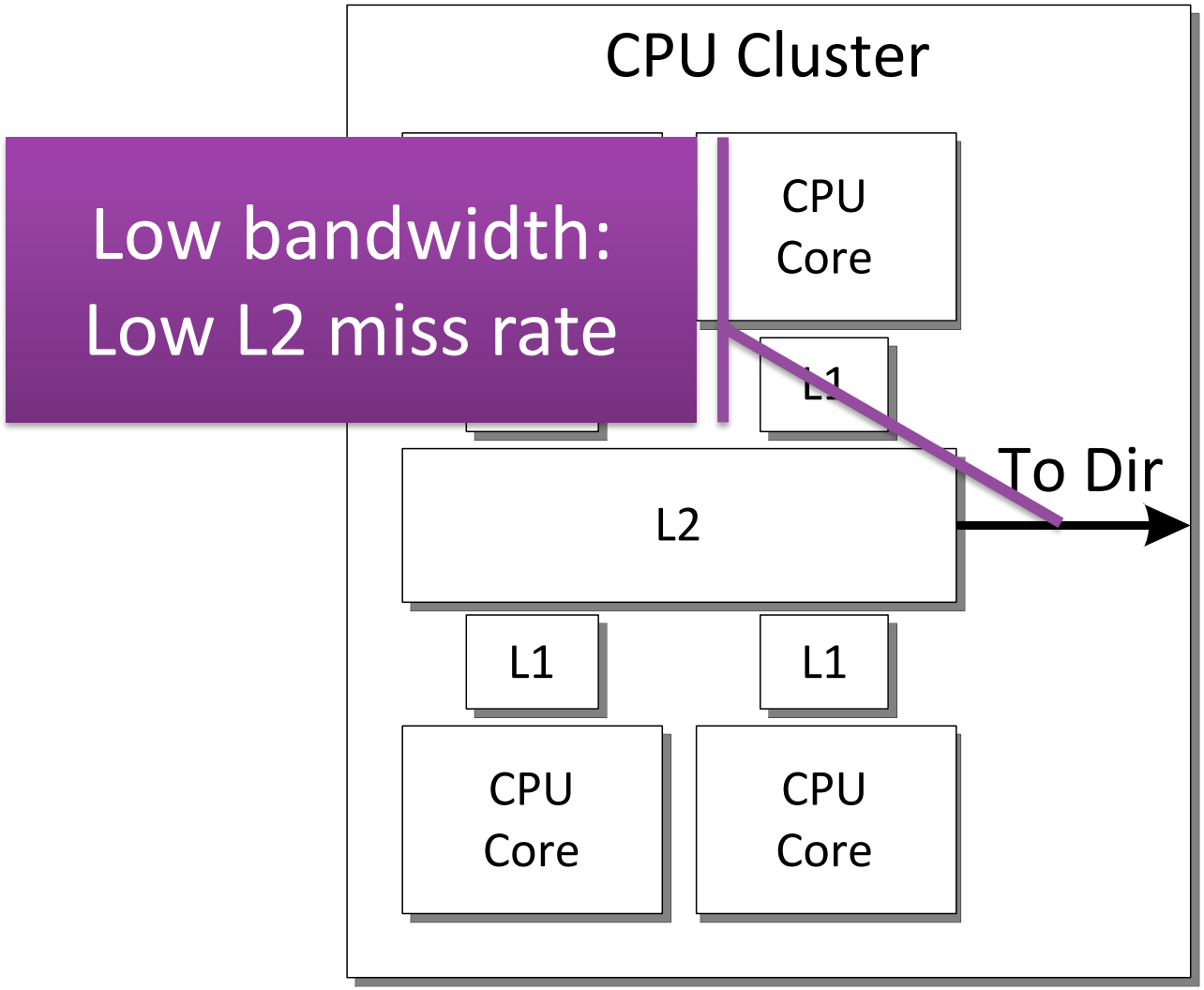


SYSTEM OVERVIEW

GPU

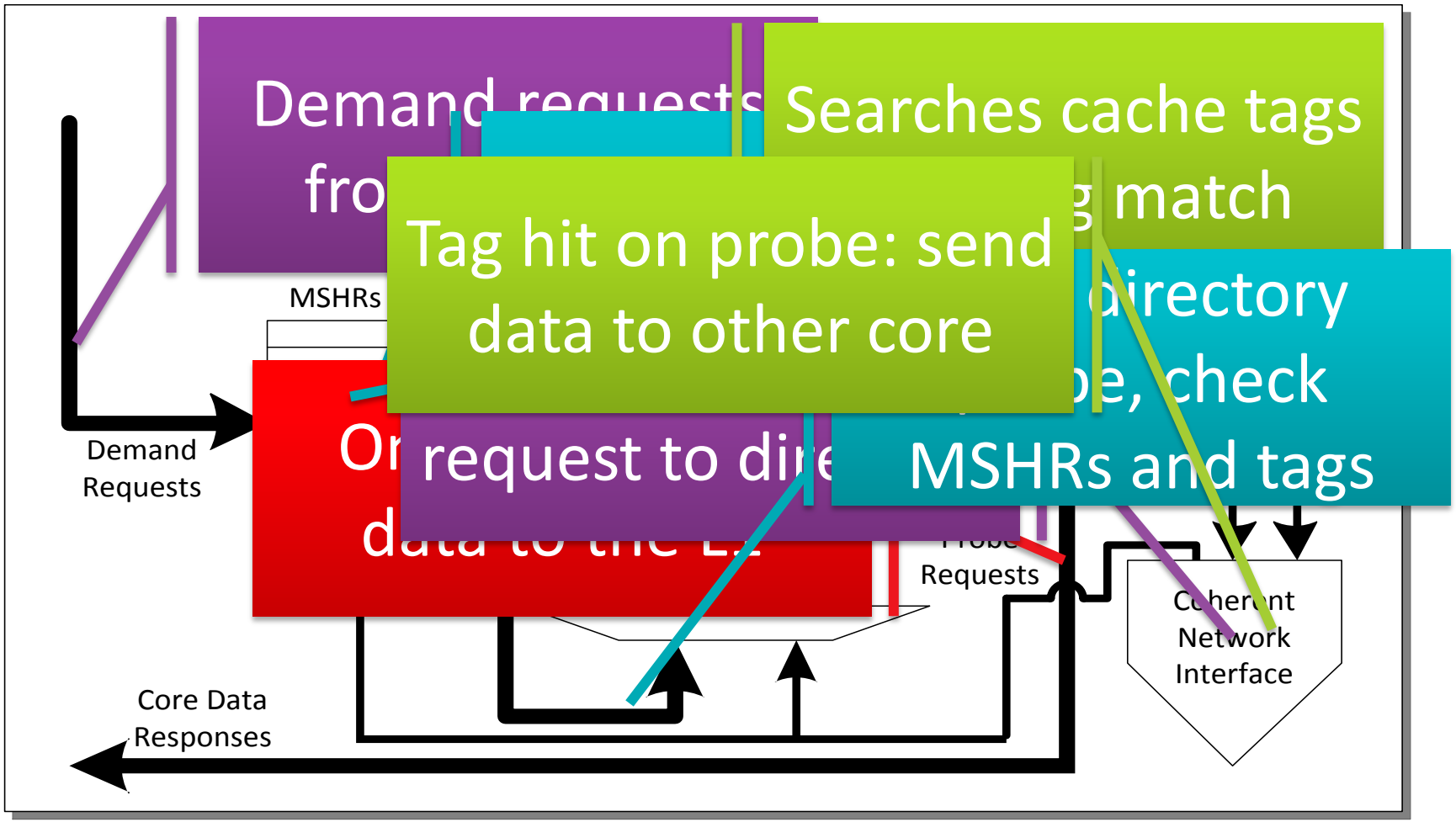


SYSTEM OVERVIEW



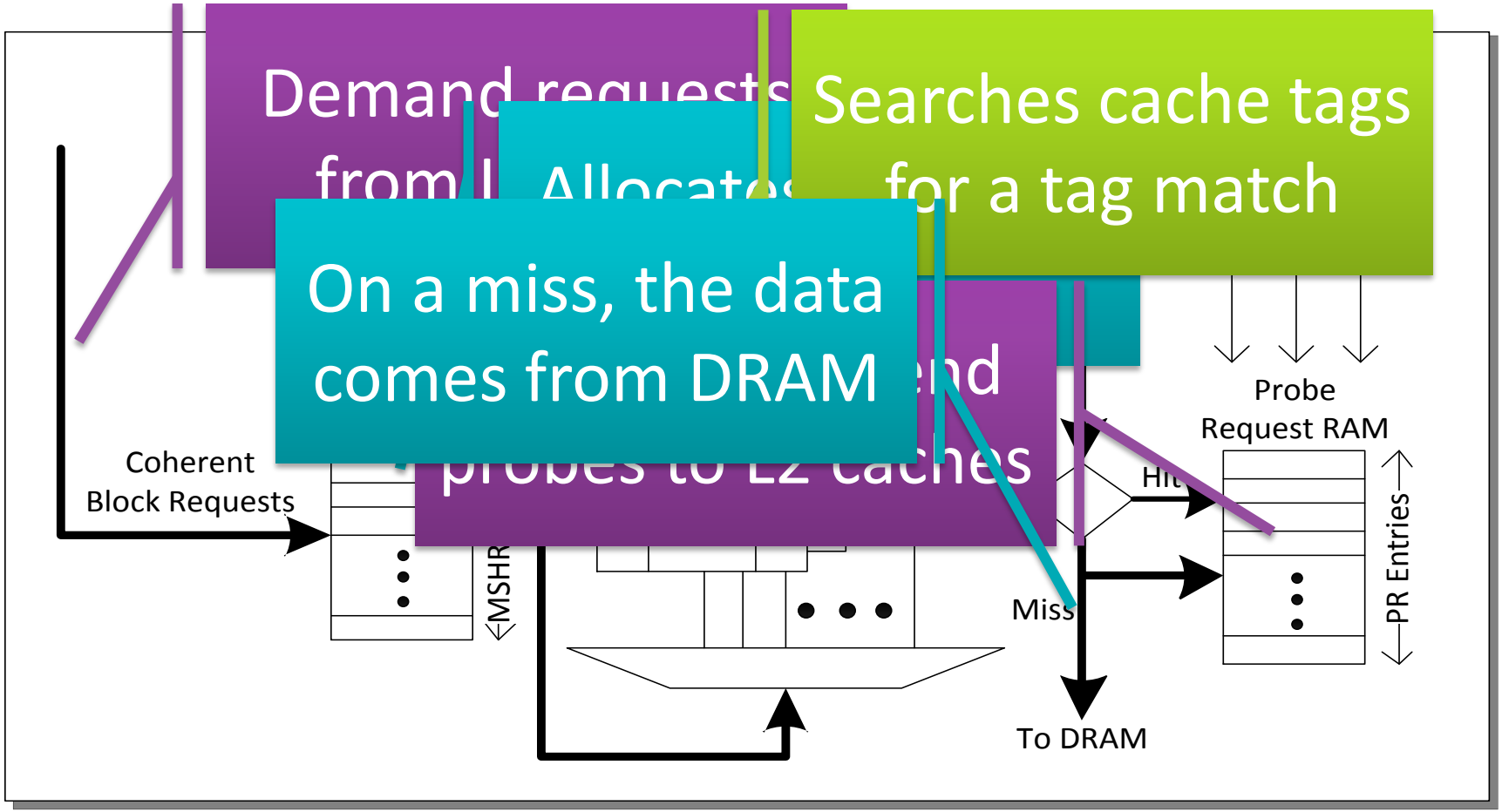
CACHE ARCHITECTURE REMINDER

CPU/GPU L2 CACHE



DIRECTORY ARCHITECTURE REMINDER

DIRECTORY



BACKGROUND SUMMARY

AMD

- System under investigation
  - Heterogeneous CPU-GPU on chip
  - High-bandwidth DRAM
- Directory pipeline complex
  - MSHR array is associative
  - Difficult to pipeline with more than 1 request per cycle
  - Important resources: MSHR entries

OUTLINE

AMD

- Motivation
- Background
- **Heterogeneous System Bottlenecks**
  - Simulation overview
  - Directory bandwidth
  - MSHRs
  - Performance is significantly affected
- Heterogeneous System Coherence Details
- Results
- Conclusions

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

SIMULATION DETAILS

AMD

- gem5 simulator
  - Simple CPU
  - GPU simulator based on AMD GCN
  - All memory requests through gem5
- Workloads
  - Modified to use hUMA
  - Rodinia & AMD APP SDK

CPU Clock	2 GHz
CPU Cores	2
CPU Shared L2	2 MB (16-way banked)
GPU Clock	1 GHz
Compute Units	32
GPU Shared L2	4 MB (64-way banked)
L3 (Memory-side)	16 MB (16-way banked)
DRAM	DDR3, 16 channels
Peak Bandwidth	700 GB/s
Baseline Directory	256k entries (8-way banked)

GPGPU BENCHMARKS

AMD

Rodinia benchmarks

- **bp** trains the connection weights on a neural network
- **bfs** breadth-first search
- **hs** performs a transient 2D thermal simulation (5-point stencil)
- **lud** matrix decomposition
- **nw** performs a global optimization for DNA sequence alignment
- **km** does k-means clustering
- **sd** speckle-reducing anisotropic diffusion

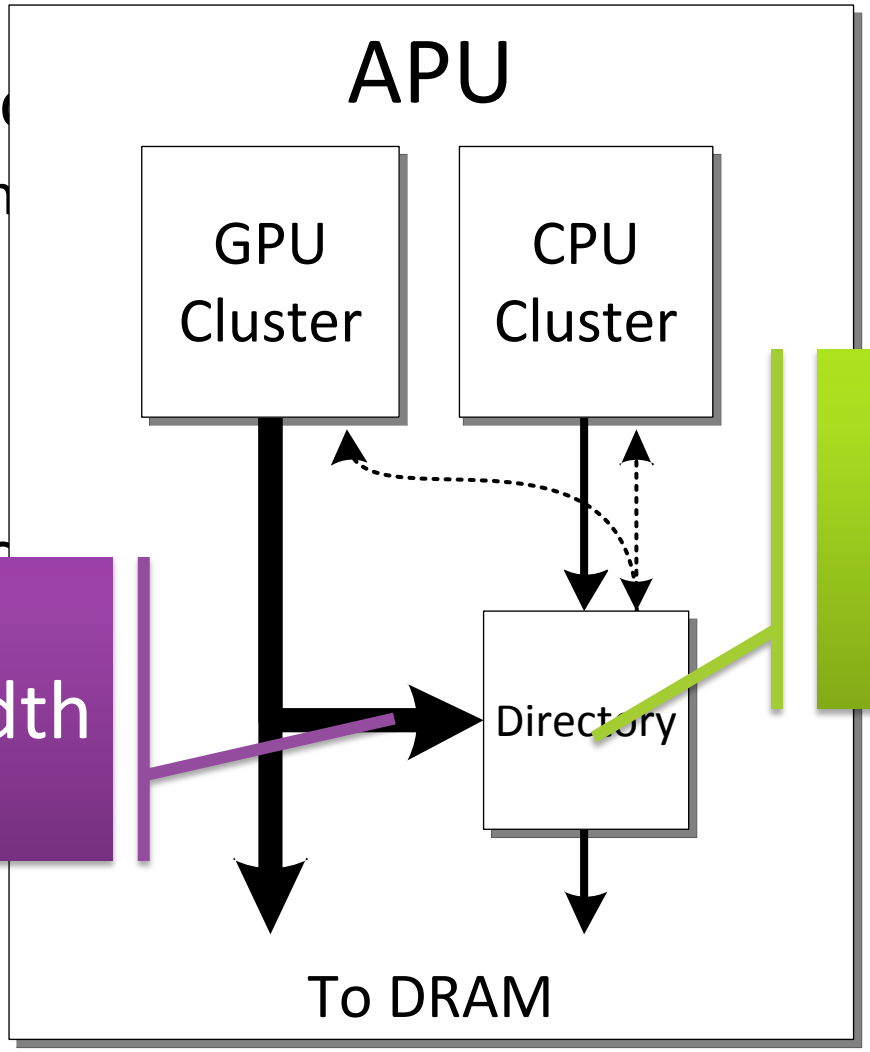
AMD SDK

- **bn** bitonic sort
- **dct** discrete cosine transform
- **hg** histogram
- **mm** matrix multiplication

SYSTEM BOTTLENECKS

AMD

- Difficult to s
  - Difficult to m
  - Complicated
- High resourc

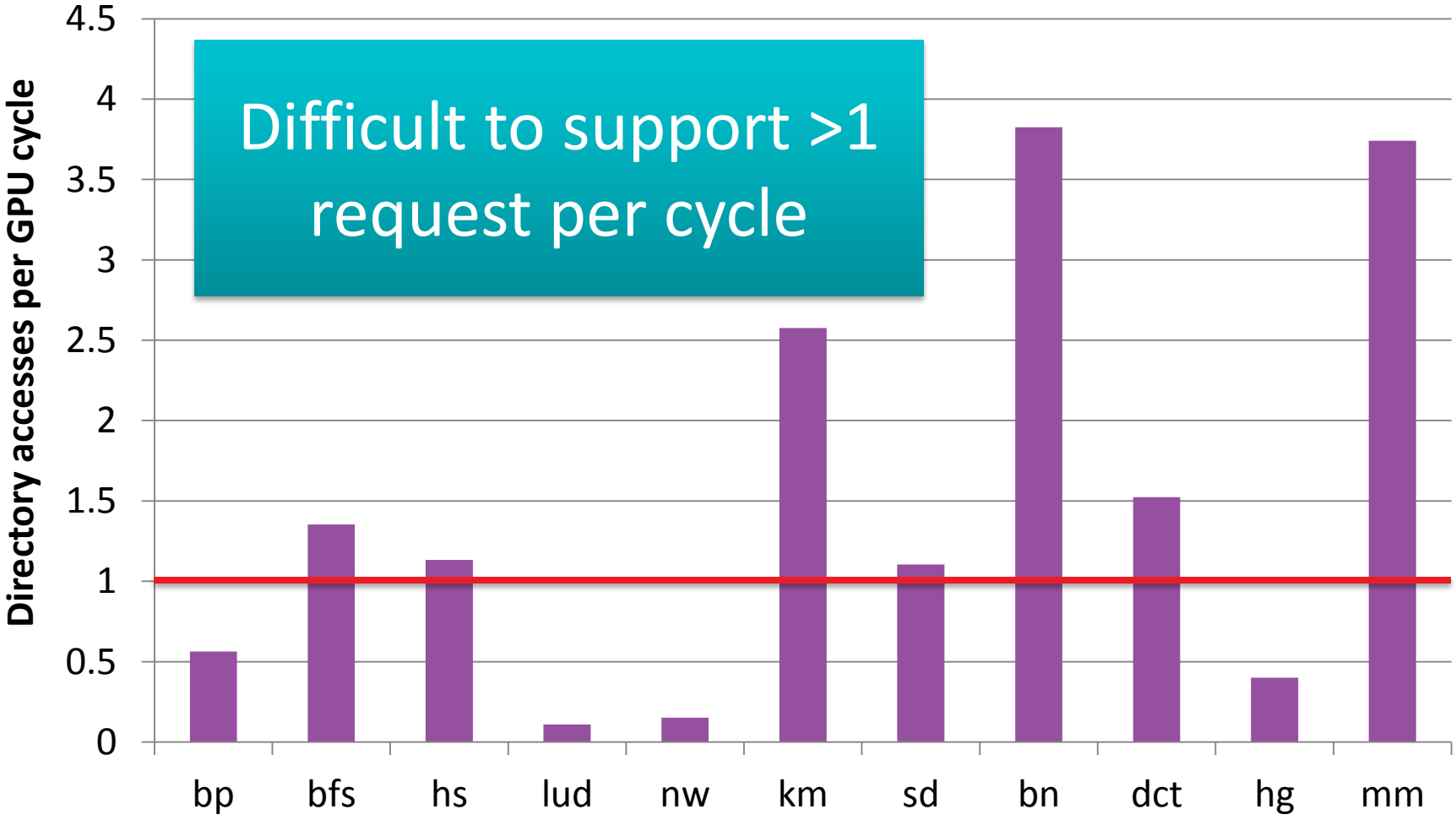


High bandwidth

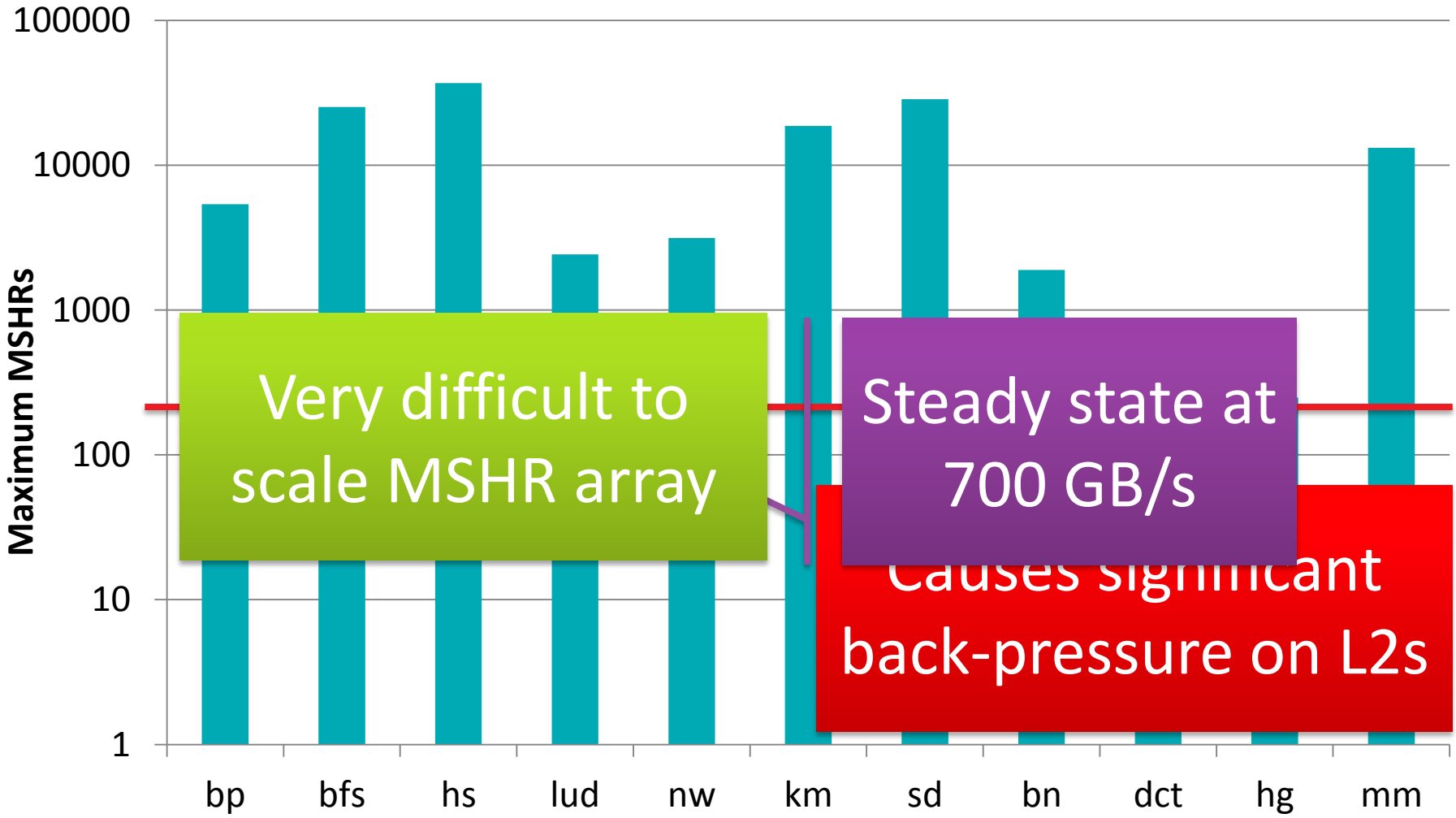
Designed to support CPU bandwidth

WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

DIRECTORY TRAFFIC



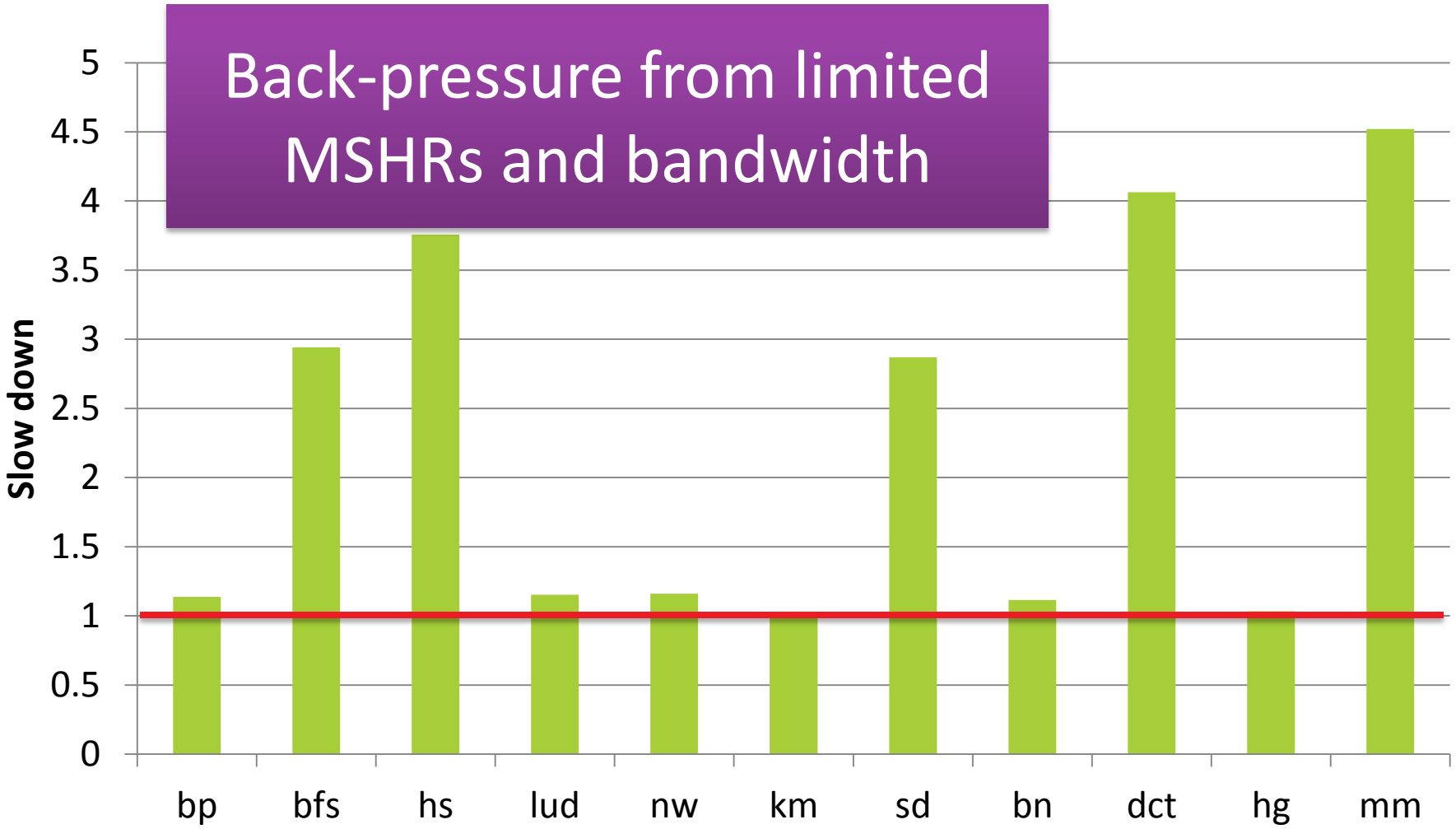
RESOURCE USAGE



PERFORMANCE OF BASELINE

COMPARED TO UNCONSTRAINED RESOURCES

AMD



WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

BOTTLENECKS SUMMARY

AMD

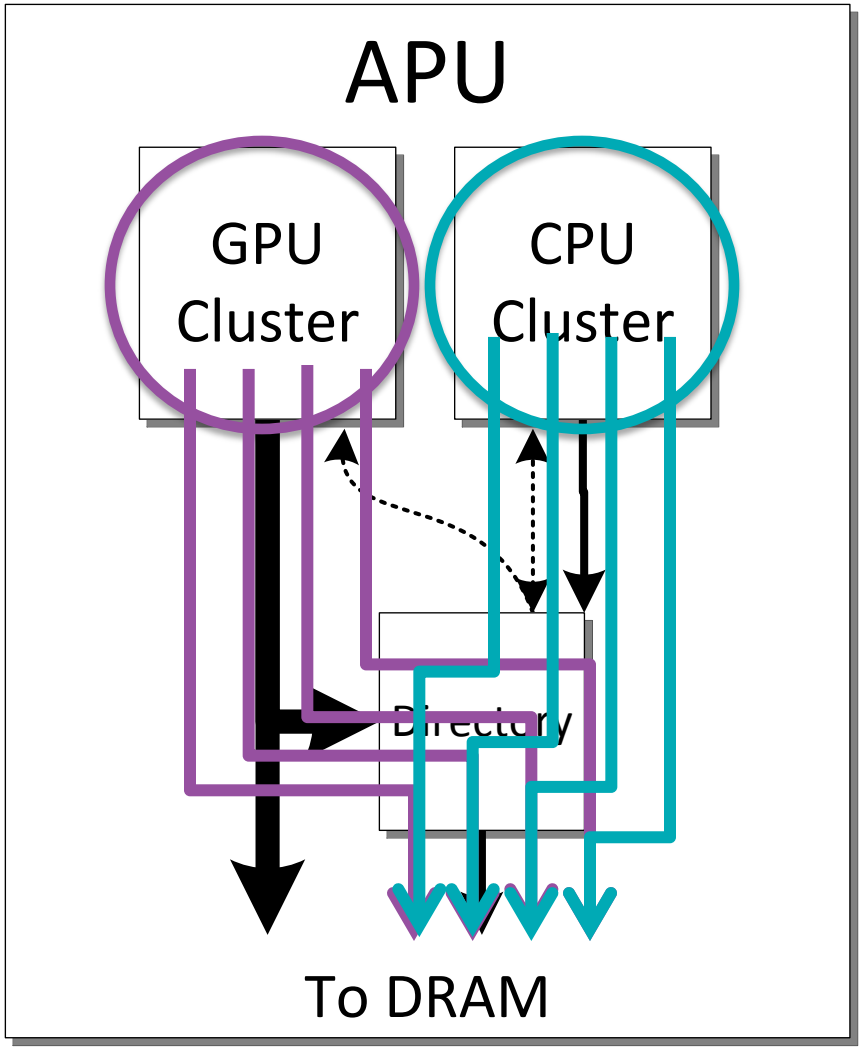
- Directory bandwidth
  - Must support up to 4 requests per cycle
  - Difficult to construct pipeline
- Resource usage
  - MSHRs are a constraining resource
  - Need more than 10,000
  - Without resource constraints, up to 4x better performance

OUTLINE

AMD

- Motivation
- Background
- Heterogeneous System Bottlenecks
- **Heterogeneous System Coherence Details**
  - Overall system design
  - Region buffer design
  - Region directory design
  - Example
  - Hardware complexity
- Results
- Conclusions

BASELINE DIRECTORY COHERENCE

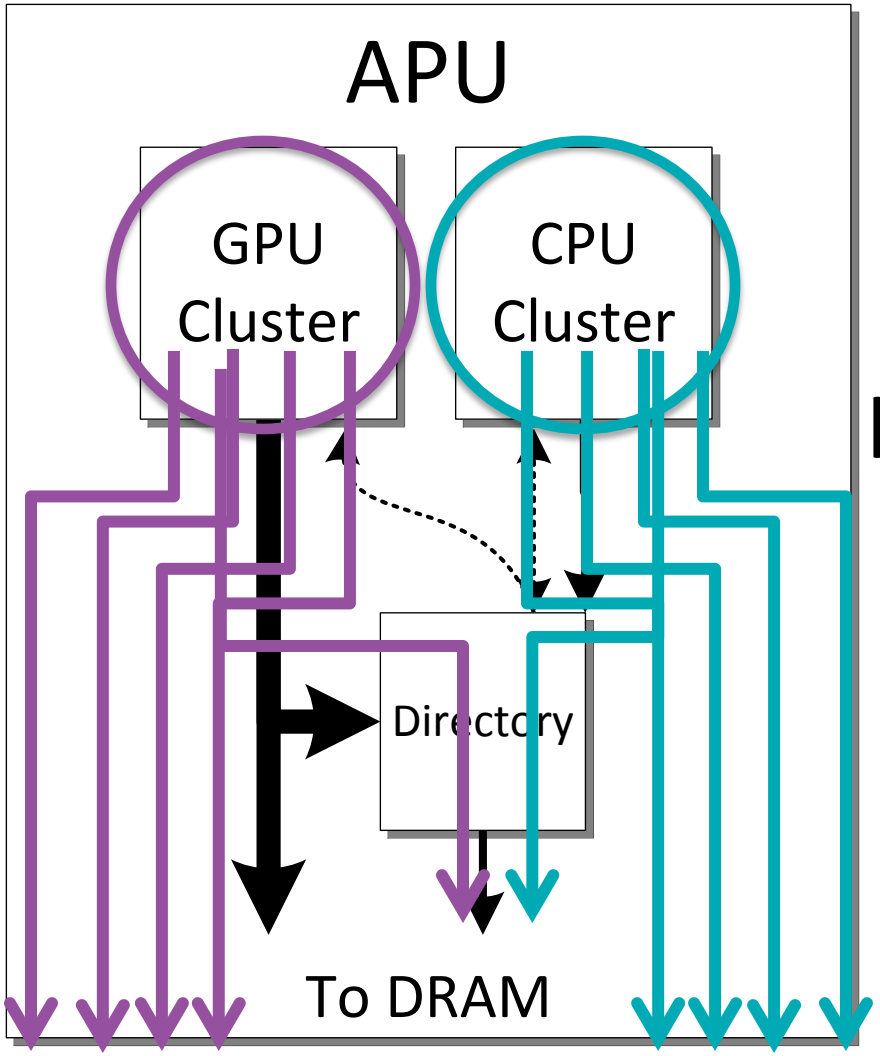


Initialization

Kernel Launch

Read result

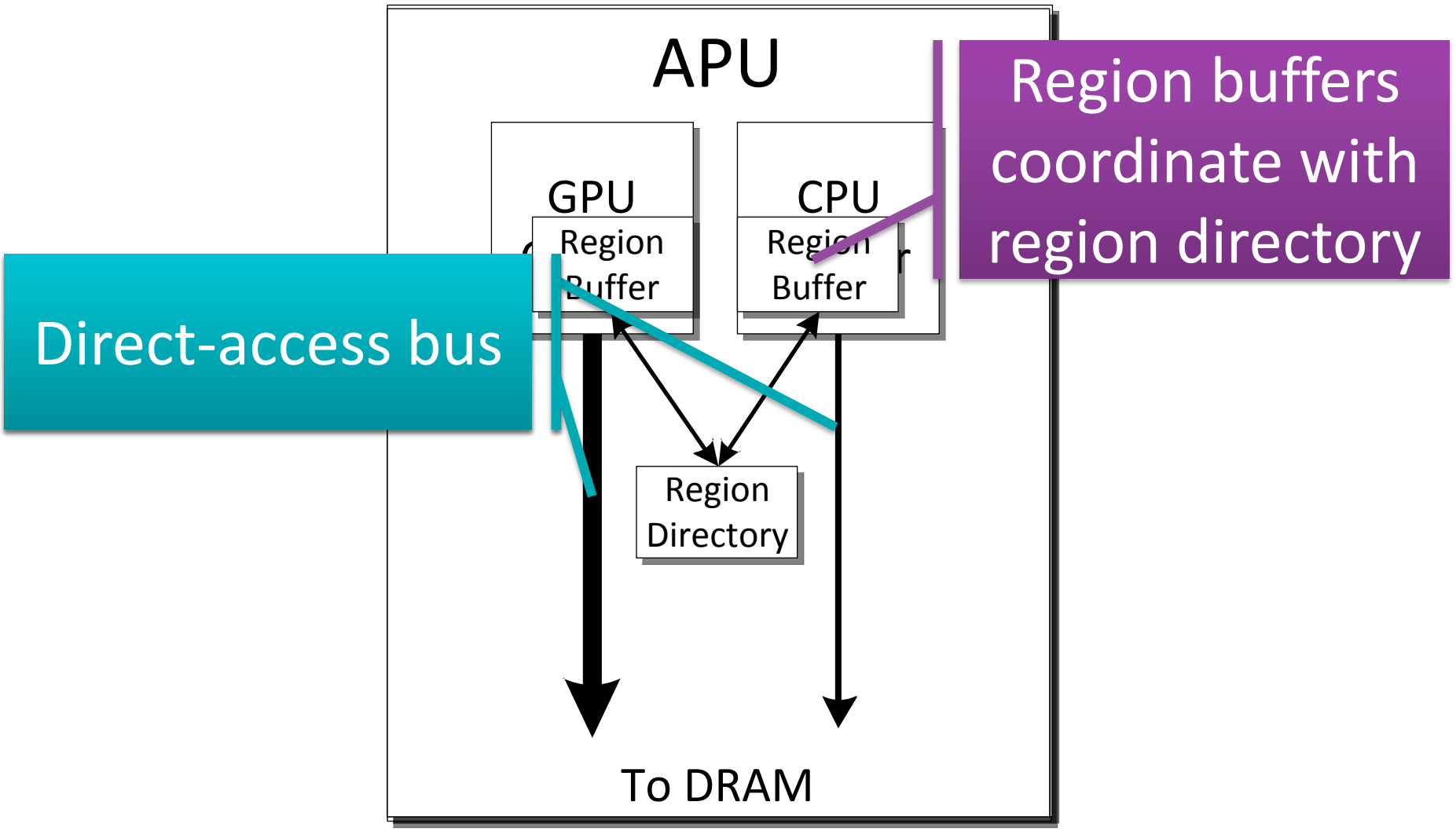
HETEROGENEOUS SYSTEM COHERENCE (HSC) AMD



Initialization

Kernel Launch

HETEROGENEOUS SYSTEM COHERENCE (HSC)

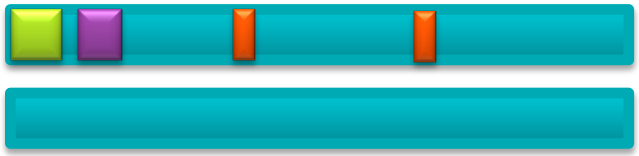


HSC: EXAMPLE MEMORY REQUEST

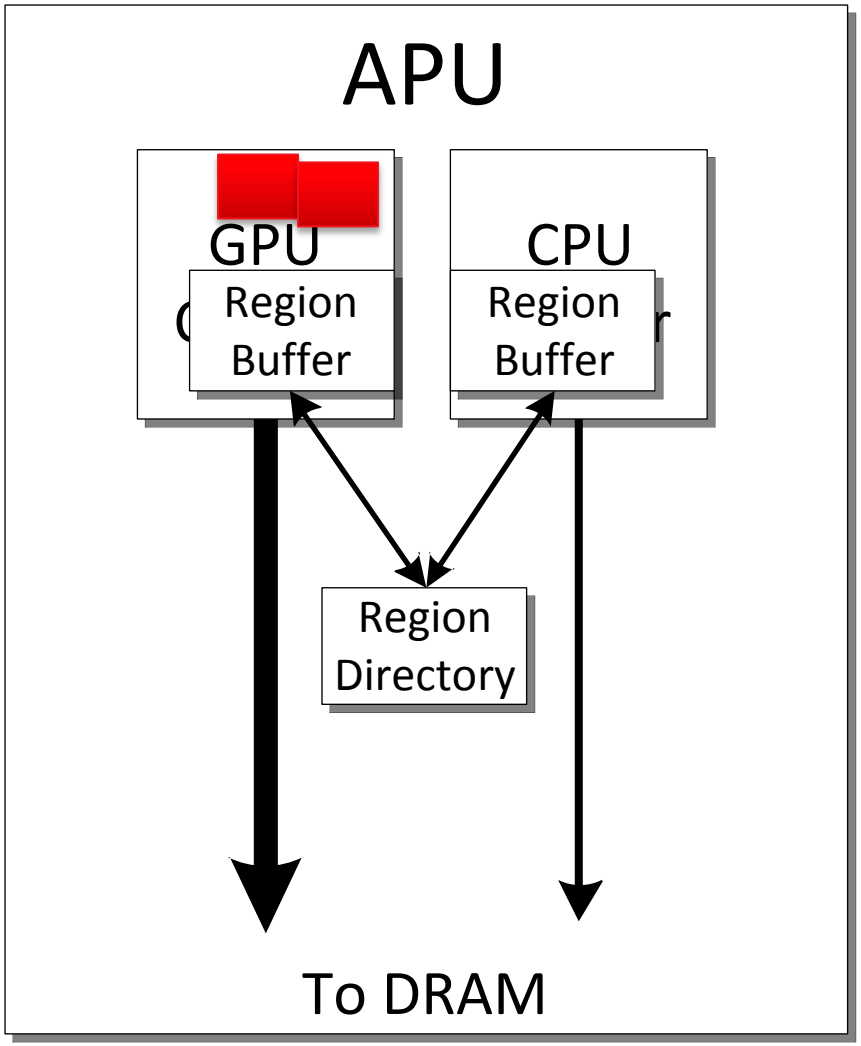
AMD

GPU L2 Cache

GPU Region Buffer

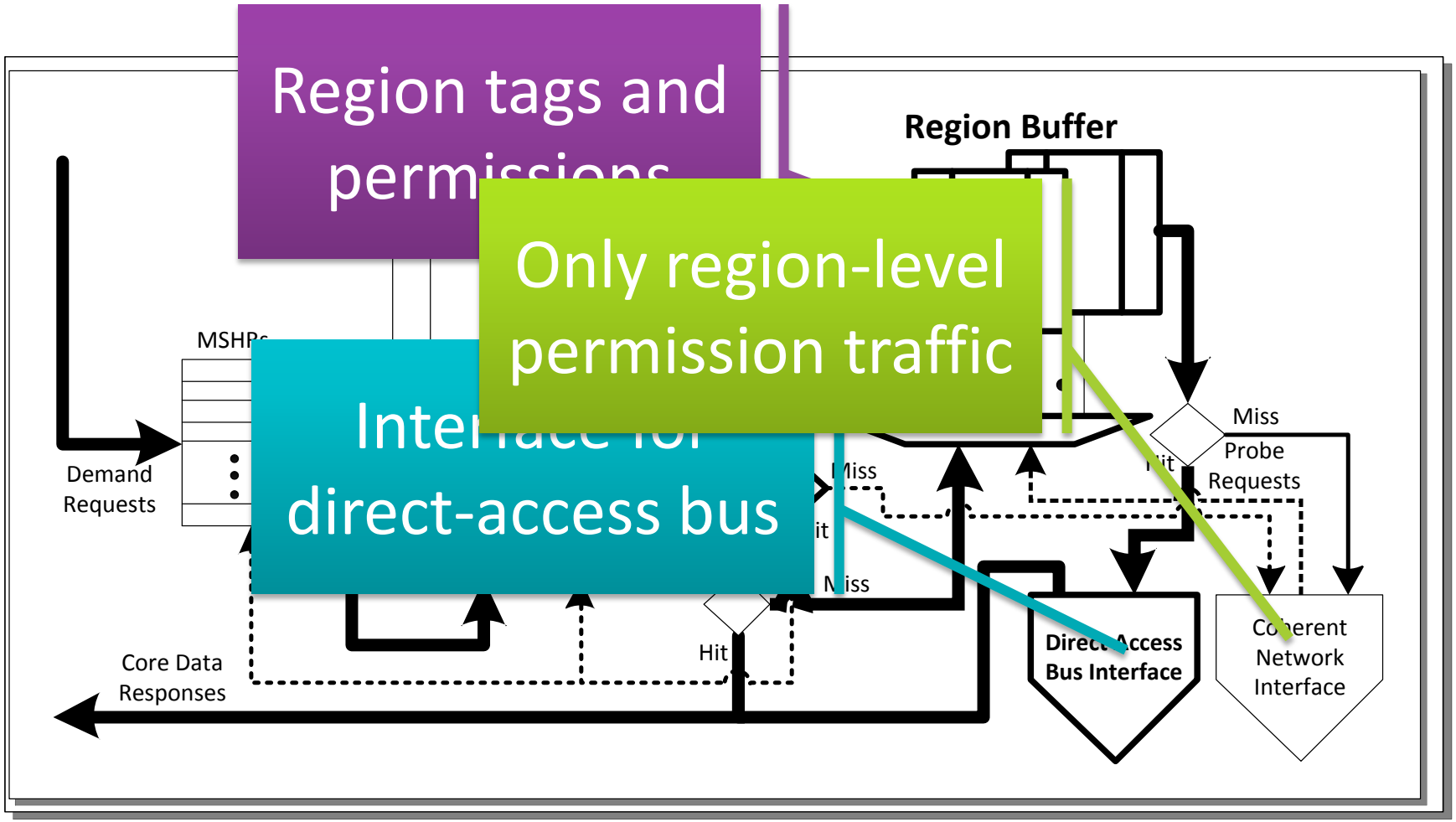


Region Directory

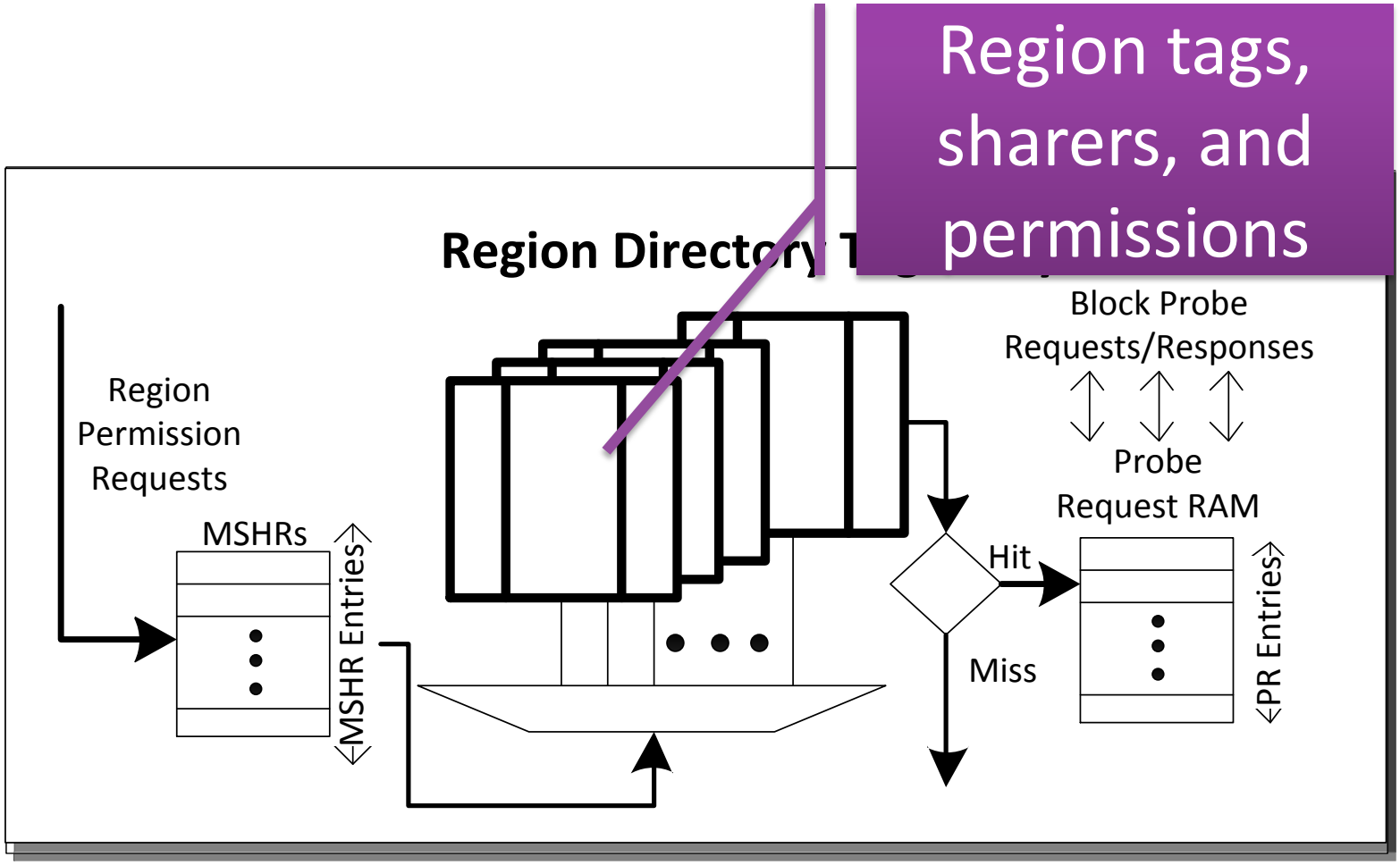


WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

HSC: L2 CACHE & REGION BUFFER



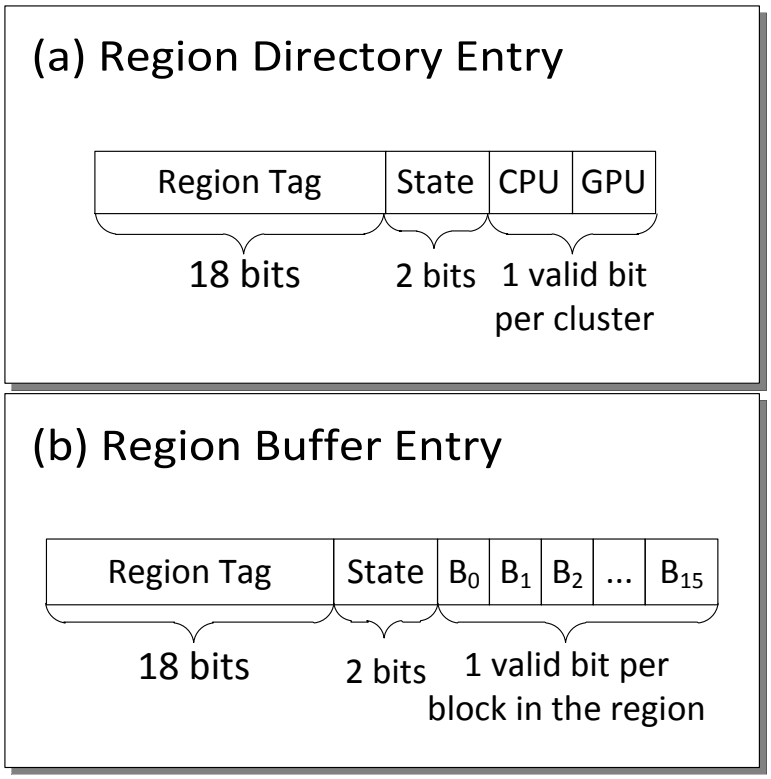
HSC: REGION DIRECTORY



HSC: HARDWARE COMPLEXITY

AMD

- Region protocols reduce directory size
  - Region directory: 8x fewer entries
- Region buffers
  - At each L2 cache
  - 1-KB region (16 64-B blocks)
  - 16-K region entries
  - Overprovisioned for low-locality workloads



WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

HSC SUMMARY

AMD

- Key insight
  - GPU-CPU applications exhibit high spatial locality
  - Use direct-access bus present in systems
  - Offload bandwidth onto direct-access bus
- Use coherence network only for permission
- Add region buffer to track region information
  - At each L2 cache
  - Bypass coherence network and directory
- Replace directory with region directory
  - Significantly reduces total size needed

OUTLINE

AMD

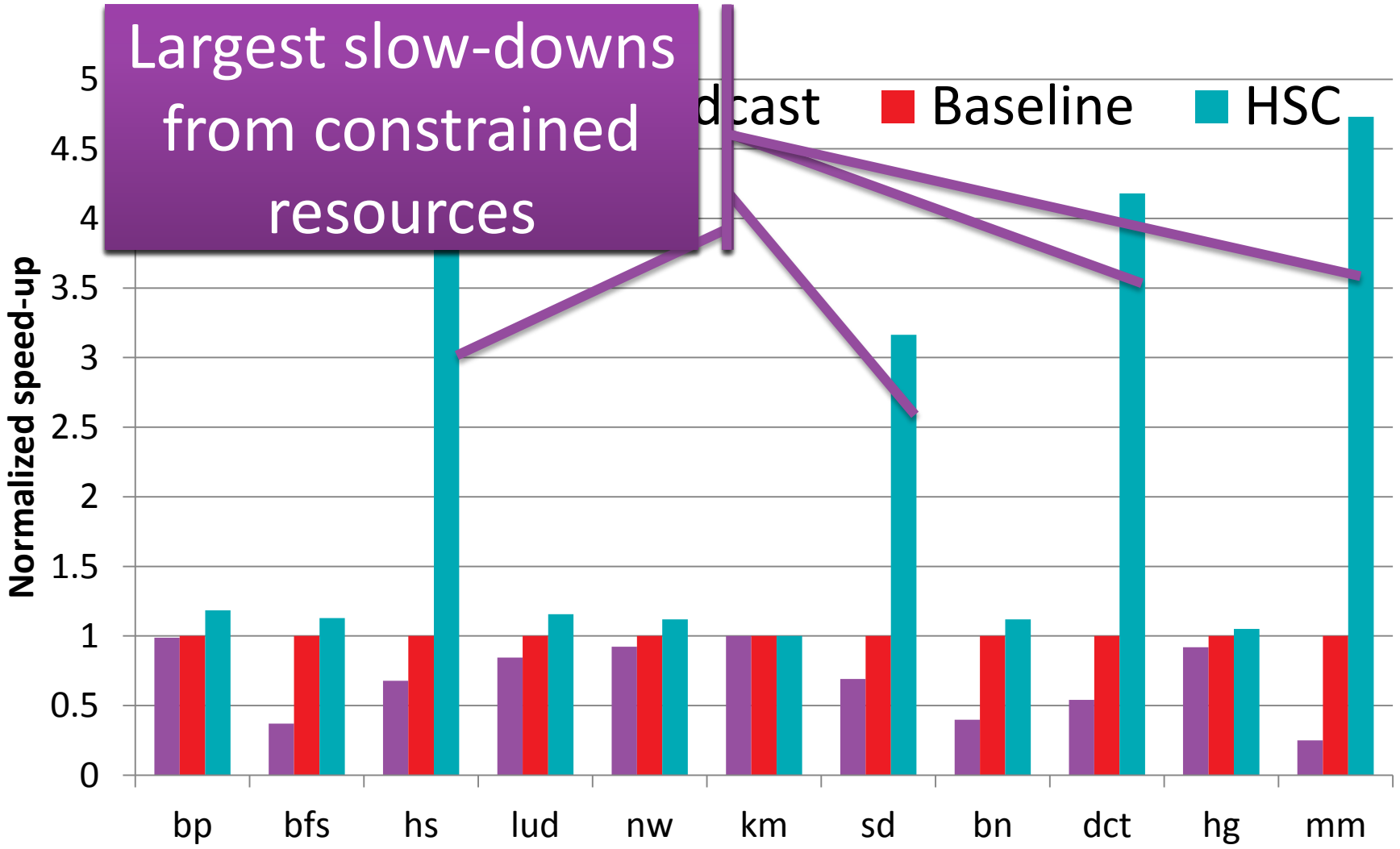
- Motivation
- Background
- Heterogeneous System Bottlenecks
- Heterogeneous System Coherence Details
- **Results**
  - Speed-up
  - Latency of loads
  - Bandwidth
  - MSHR usage
- Conclusions

THREE CACHE-COHERENCE PROTOCOLS

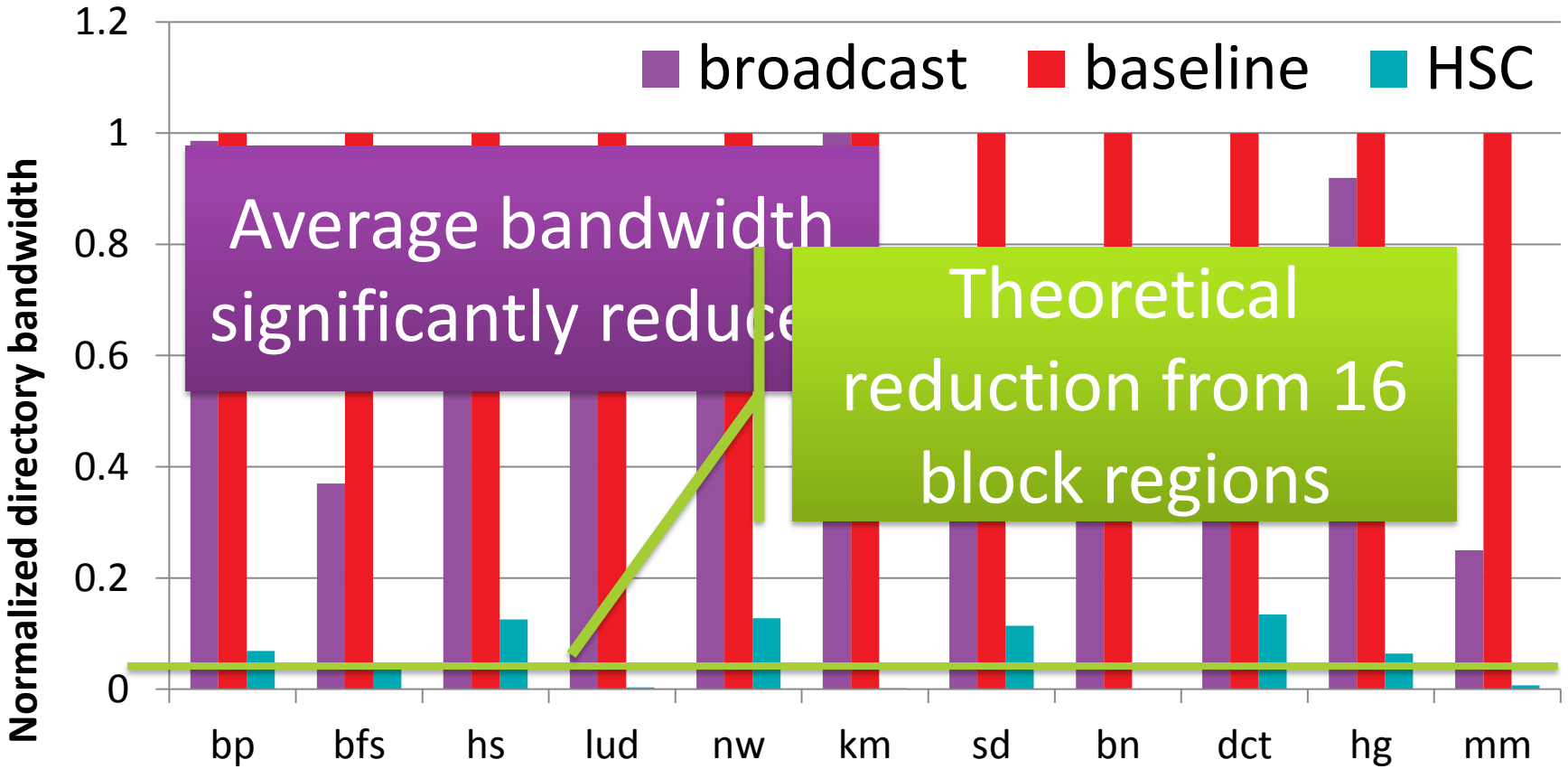
AMD

- **Broadcast**: Null-directory that broadcasts on all requests
- **Baseline**: Block-based, mostly inclusive, directory
- **HSC**: Region-based directory with 1-KB region size

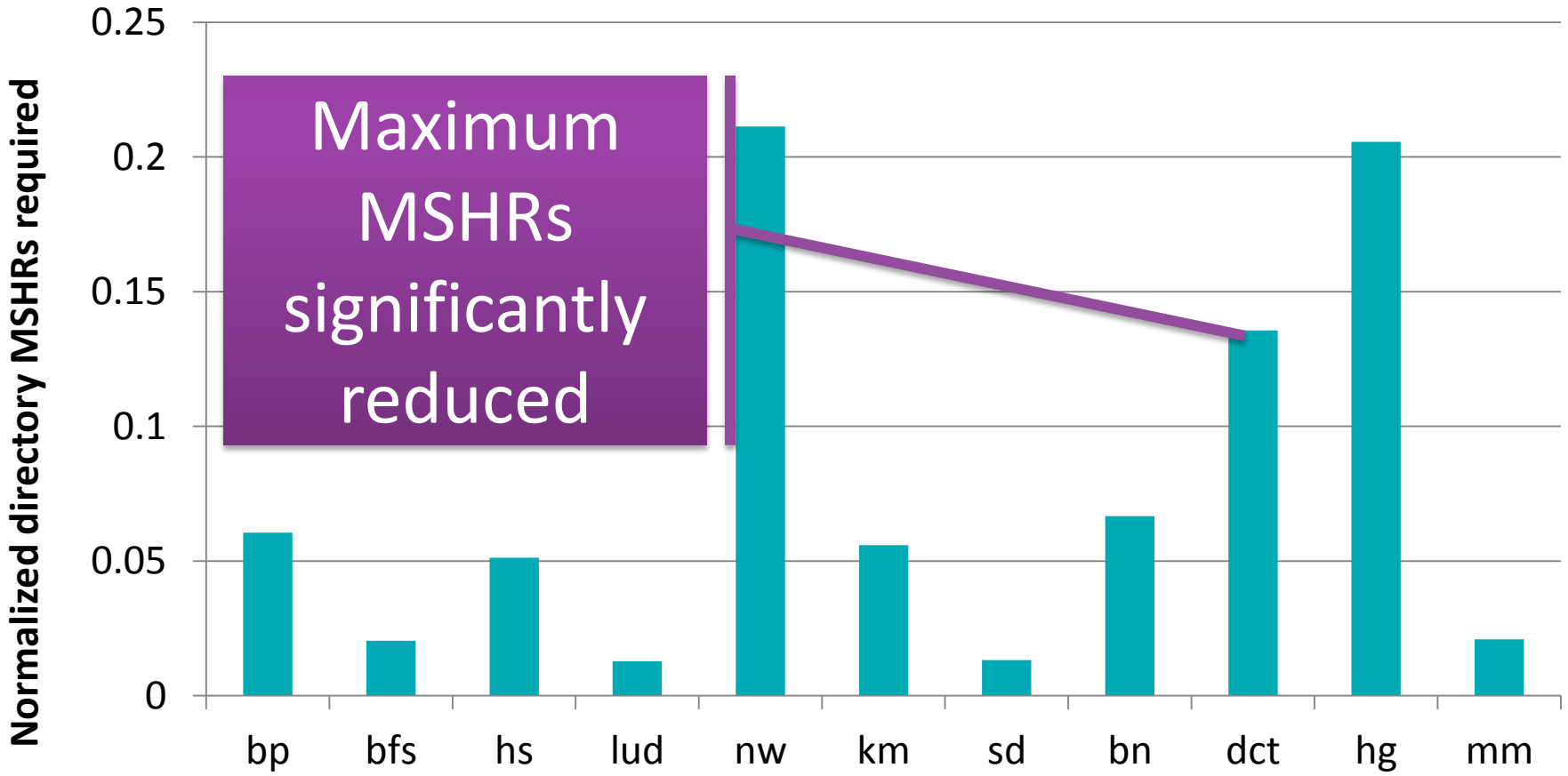
HSC PERFORMANCE



DIRECTORY TRAFFIC REDUCTION



HSC RESOURCE USAGE



RESULTS SUMMARY

AMD

- Used a detailed timing simulator for CPU and GPU
- HSC significantly improves performance
  - Reduces the average load latency
  - Decreases bandwidth requirement of directory
- HSC reduces the required MSHRs at the directory

RELATED WORK

AMD

- Coarse-grained coherence
  - Region coherence
    - Applied to snooping systems [Cantin, ISCA 2005] [Moshovos, ISCA 2005] [Zebchuk, MICRO 2007]
    - Extended to directories [Fang, PACT 2013] [Zebchuk, MICRO 2013]
  - Spatiotemporal coherence [Alisafaee, MICRO 2012]
  - Dual-grain directory coherence [Basu, UW-TR 2013]
    - Primarily focused on directory size
- GPU coherence [Singh et al. HPCA 2013]
  - Intra-GPU coherence

CONCLUSIONS

AMD

- Hardware coherence can increase the utility of heterogeneous systems
- Major bottlenecks in current coherence implementations
  - High bandwidth difficult to support at directory
  - Extreme resource requirements
- We propose Heterogeneous System Coherence
  - Leverages spatial locality and region coherence
  - Reduces bandwidth by 94%
  - Reduces resource requirements by 95%

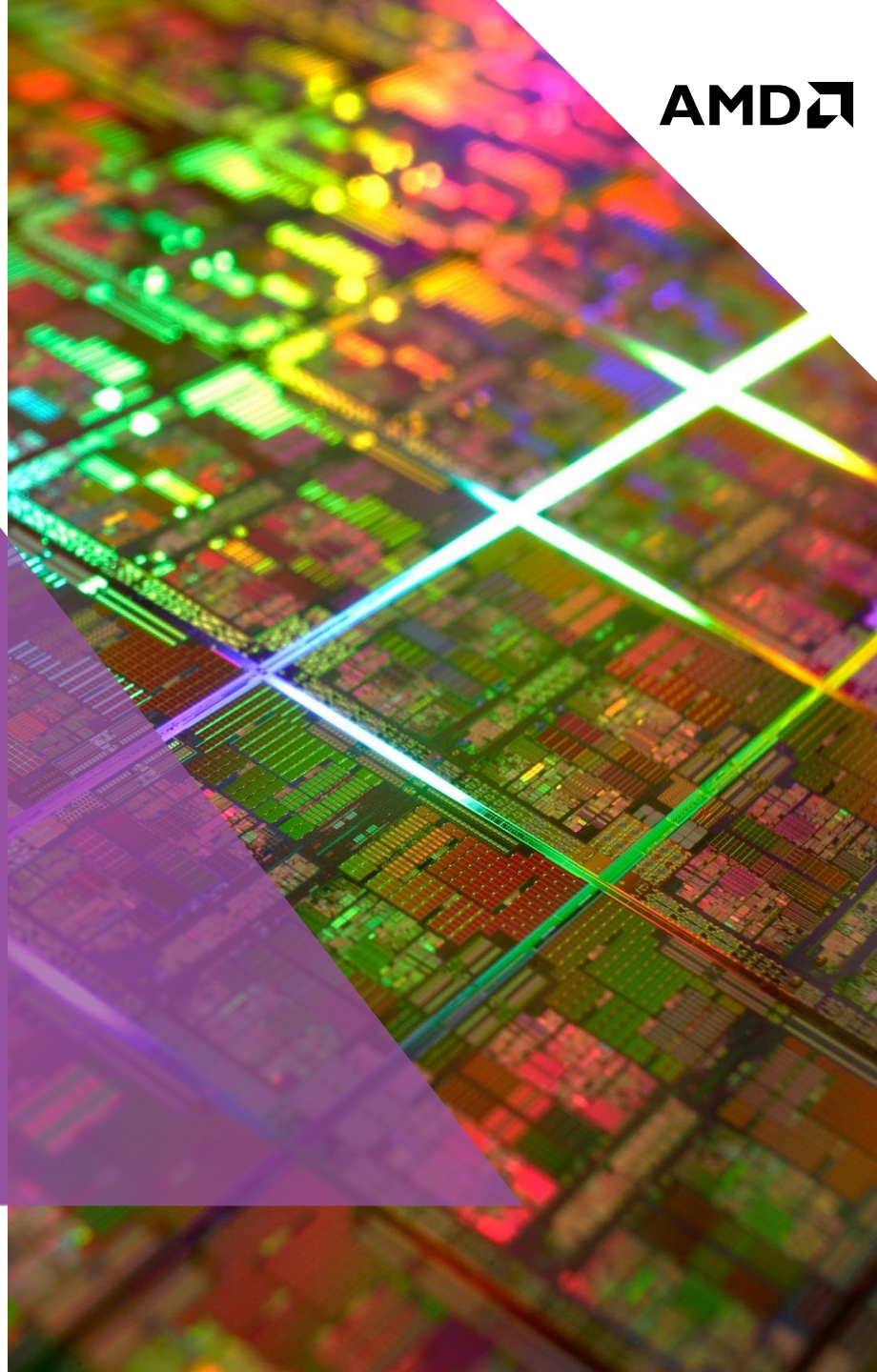
WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

AMD

Questions?

Contact:

powerjg@cs.wisc.edu



DISCLAIMER & ATTRIBUTION

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

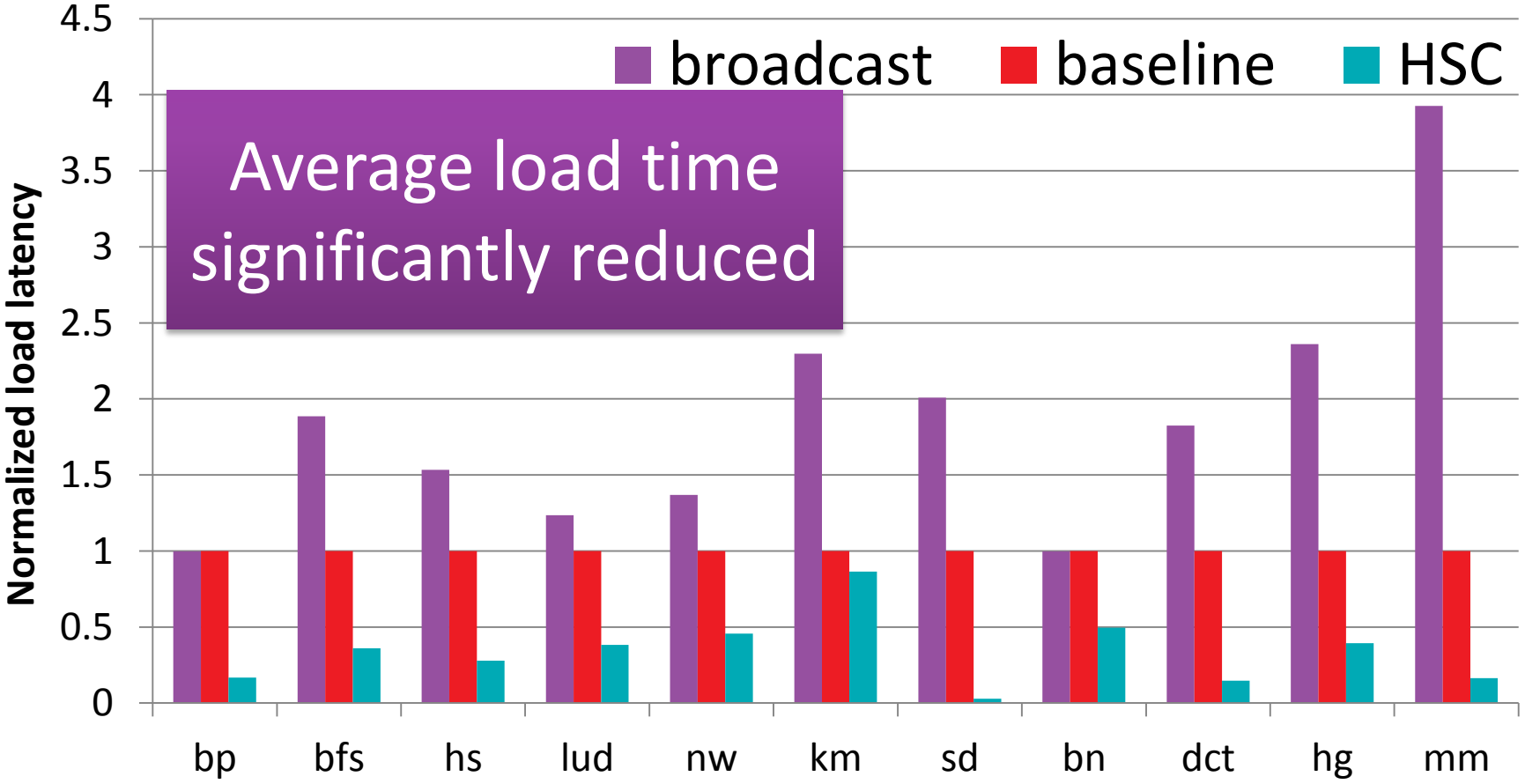
AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

Backup Slides

LOAD LATENCY



EXECUTION TIME BREAKDOWN