Heterogeneous System Coherence for Integrated CPU-GPU Systems

WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

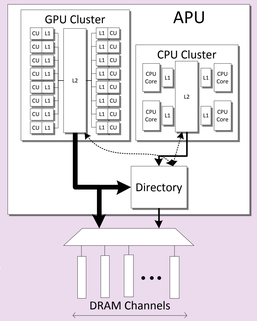
Contact: powerjg@cs.wisc.edu
bit.ly/JasonHSC

AMD

Jason Power*, Arkaprava Basu*, Junli Gu†, Sooraj Puthoor†, Bradford M Beckmann†, Mark D Hill*†, Steven K Reinhardt†, David A Wood*†

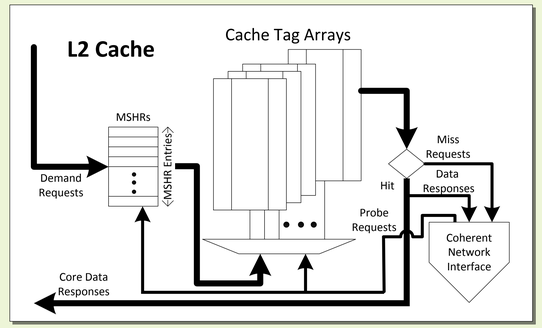
Summary

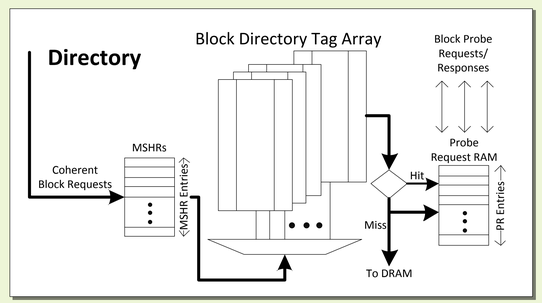
- Technology drivers
  - Increasing memory BW
  - Increasing integration



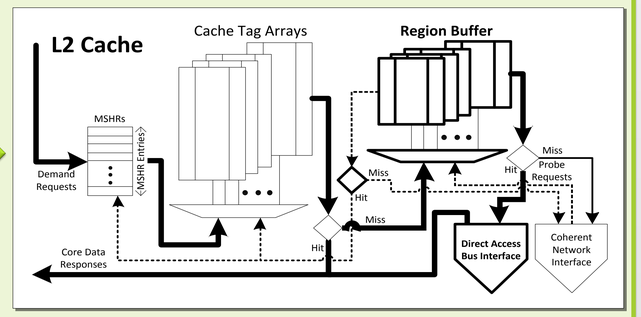
- System drivers: hUMA
  - Shared virtual address space
  - **Cache coherence**
- Key bottlenecks in today's systems
  - Directory BW
  - Resource usage
- ***Heterogeneous System Coherence*** (HSC)
  - Leverage coarse-grained sharing between CPU & GPU
  - Move coherent traffic onto direct-access bus
  - Bandwidth ↓ 94%, resources ↓ 95%
  - Average 2x speedup over conventional directory

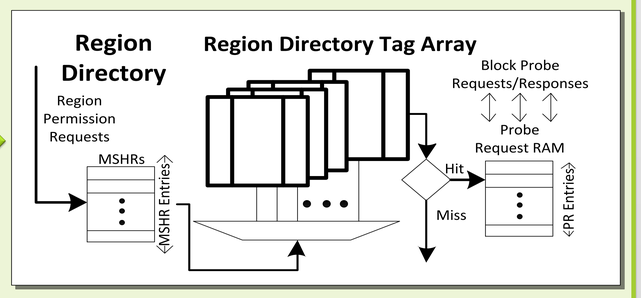
Baseline Design



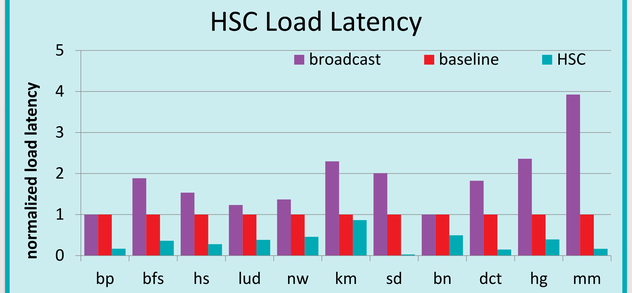


Heterogeneous System Coherence Design

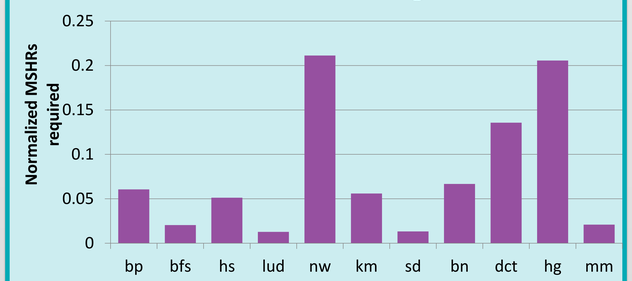




Discussion

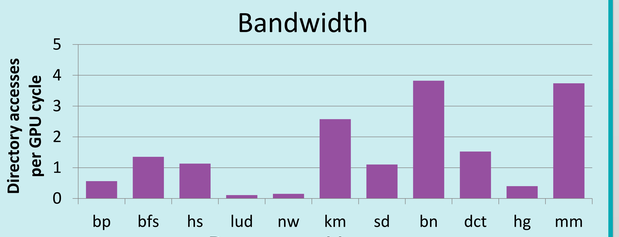


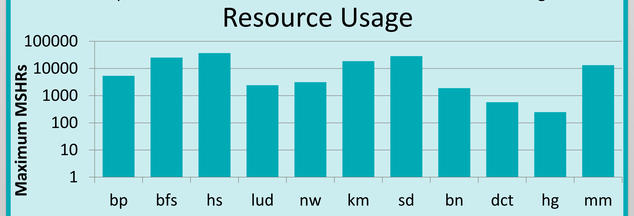
- HSC reduces the average load latency

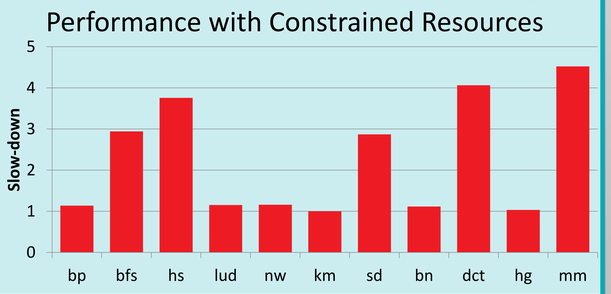


- HSC reduces the required MSHRs at the directory
  - By more than 4x in all cases
  - By more than 20x in many cases

Heterogeneous System Bottlenecks







Methodology

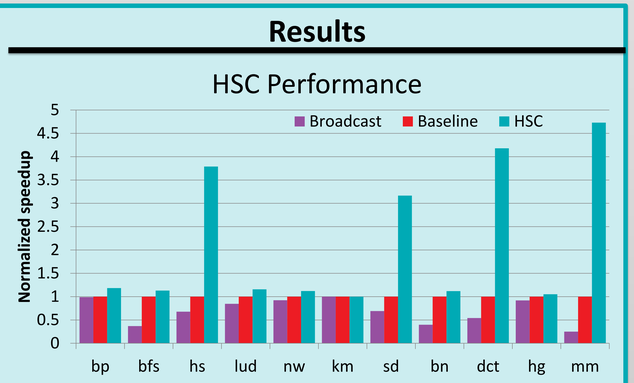
- Simulation
  - gem5 for CPU and memory system
  - Ruby for caches
  - GPU modeled off of GCN
- Workloads
  - Subset of Rodinia
  - AMD APP SDK

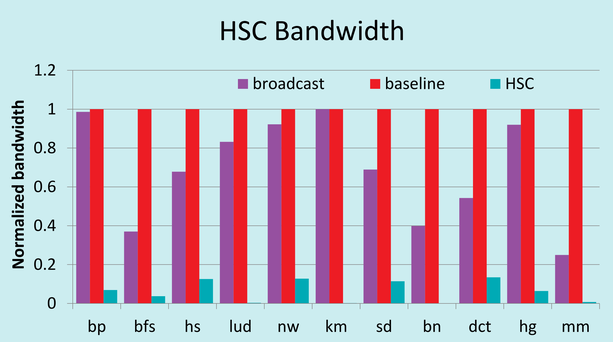
CPU Clock	2 GHz
CPU Cores	2
CPU Shared L2 Cache	2 MB
GPU Clock	1 GHz
Compute Units	32
GPU L1 Data Cache	32 KB
GPU Shared L2 Cache	4 MB
L3 Memory-side Cache	16 MB
Peak Memory Bandwidth	700 GB/s
Baseline Directory	262,144 entries
Region Directory	32,768 entries
MSHRs	32 entries
Region Buffer	16,384 entries

Results Summary

- Largest speedup for workloads which constrained resources hurt the most
- Massive bandwidth reduction
  - Due to offloading data onto direct-access bus
  - More than theoretical max of 94% in some cases
    - Region buffers can "prefetch" cache permissions
    - HSC significantly improves performance over the baseline design
- Decreases bandwidth requirement of directory

Results





Conclusions

- Hardware coherence can increase the utility of heterogeneous systems
- Major bottlenecks in current coherence implementations
  - High BW 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%
- **Come to our talk, Session 7, 11am on Wednesday!**

Acknowledgements

This work was performed at AMD Research, including student internships. Wisconsin authors improved the paper's presentation while being partially supported with NSF grants CCF-0916725, SHF-1017650, CNS-1117280, and CCF-1218323. The views expressed herein are not necessarily those of the NSF. Professors Hill and Wood have significant financial interests in AMD.