Enabling Multi-GPU High Performance Computing with Memory System Design

Yifan Sun
Northeastern University
Problems

- Low interconnect bandwidth utilization
Problems

• Low interconnect bandwidth utilization

• Not utilizing spatial locality

• Not utilizing temporal locality
Progressive Page-Splitting Migration (PASI)
Memory Coherence Protocol

E  S  I

Exclusive
Write

Shared
Read

Invalid
False sharing
Page Splitting

DRAM 0

Write

S

Acquire E

S

DRAM 1

Write

DRAM 0

Write

ES

Split

IS

DRAM 1
Progressive Page-Splitting Migration

Page Migration

ESI Protocol

Page Splitting
Akita: A Next-Gen Simulator

AMD GCN3 ISA
Multi-GPU Modeling
Parallel Simulation
Go

https://gitlab.com/akita/gcn3
Takeaways

• Memory management in multi-GPU system
• Progressive Page-Splitting Migration
• Akita: A next-gen, flexible, high-performance computer architecture simulator [https://gitlab.com/akita/gcn3](https://gitlab.com/akita/gcn3)
Thanks!

Yifan Sun
Northeastern University

https://syifan.github.io

@_syifan_