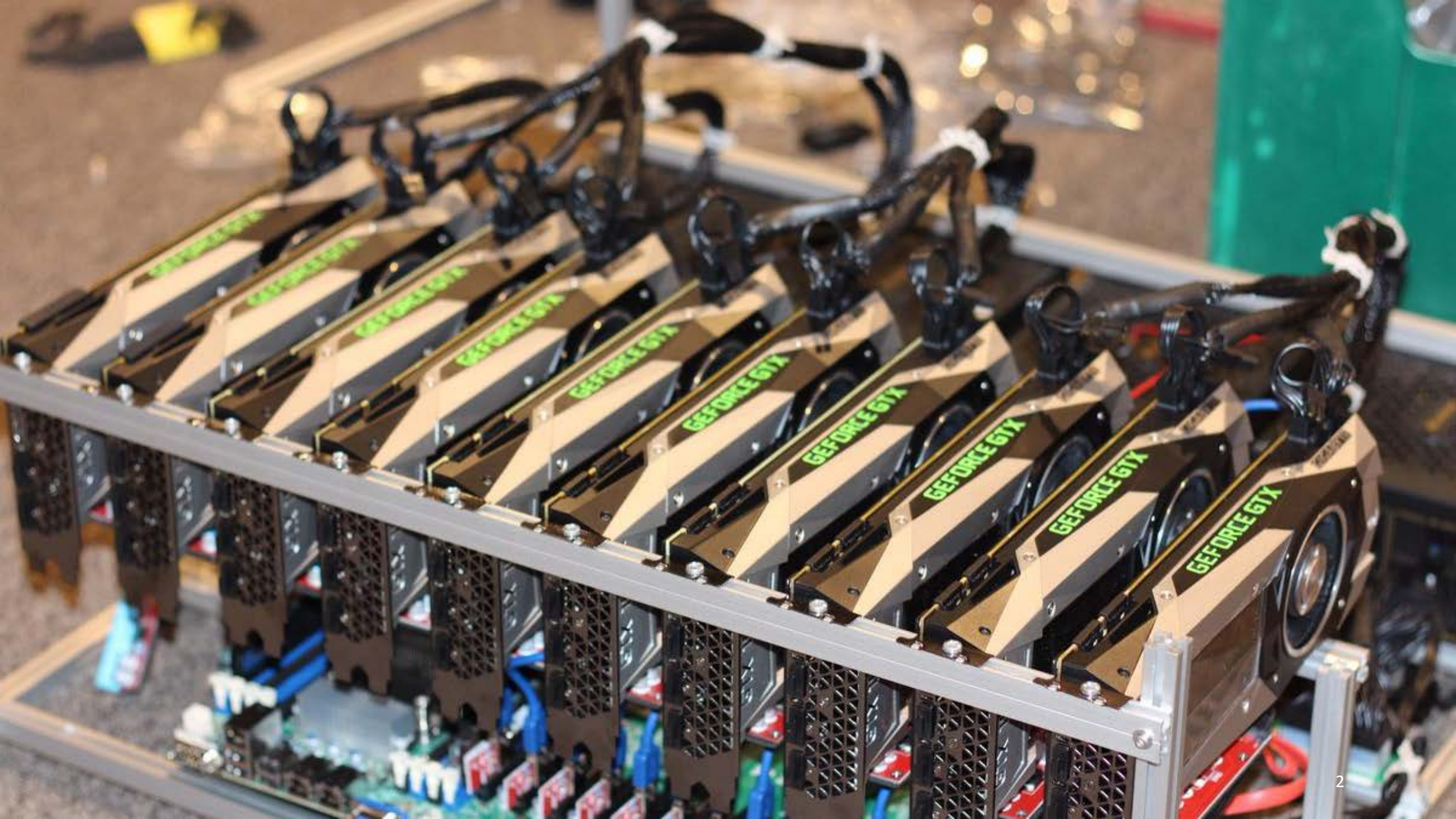
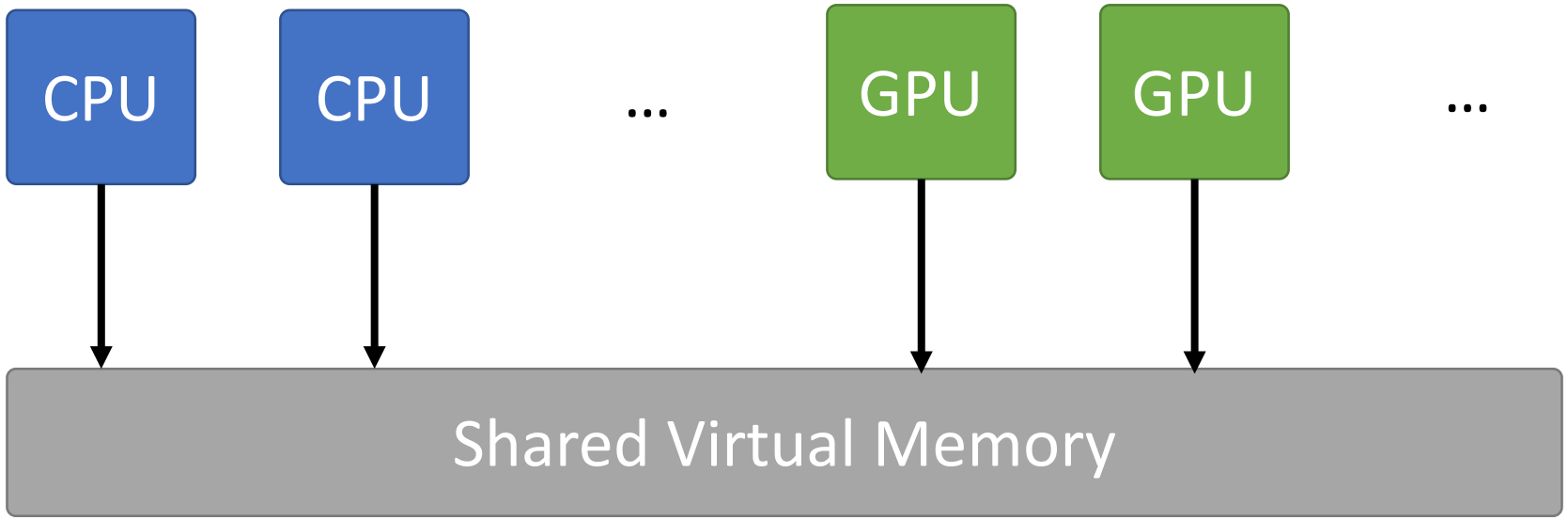
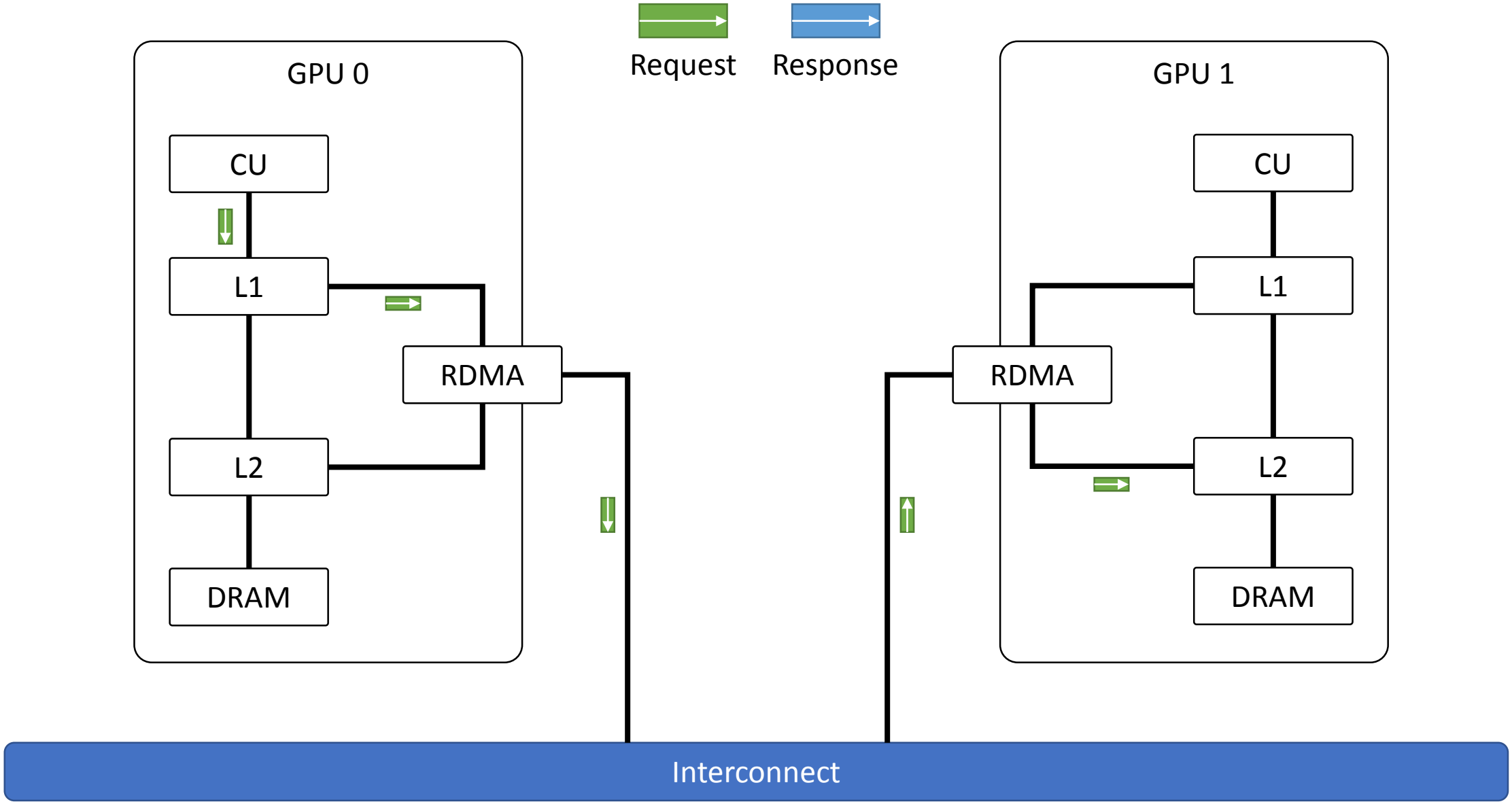
Enabling Multi-GPU High Performance Computing with Memory System Design

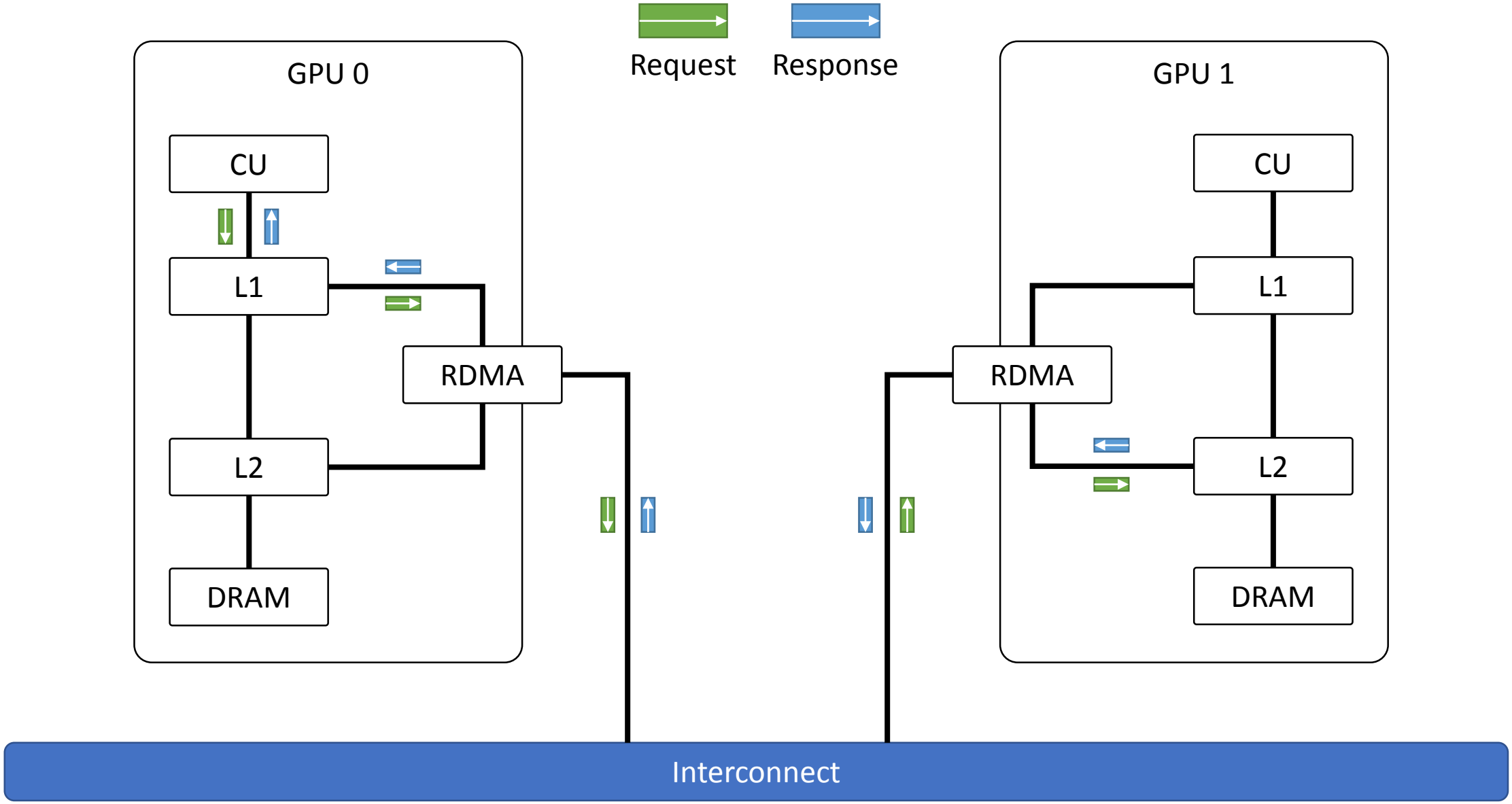
Yifan Sun

Northeastern University


GEFORCE GTX
GEFORCE GTX
GEFORCE GTX
GEFORCE GTX
GEFORCE GTX
GEFORCE GTX
GEFORCE GTX
GEFORCE GTX


CPU
CPU
...
GPU
GPU
...
Shared Virtual Memory


Request
Response
GPU 0
CU
L1
RDMA
L2
DRAM
GPU 1
CU
L1
RDMA
L2
DRAM
Interconnect

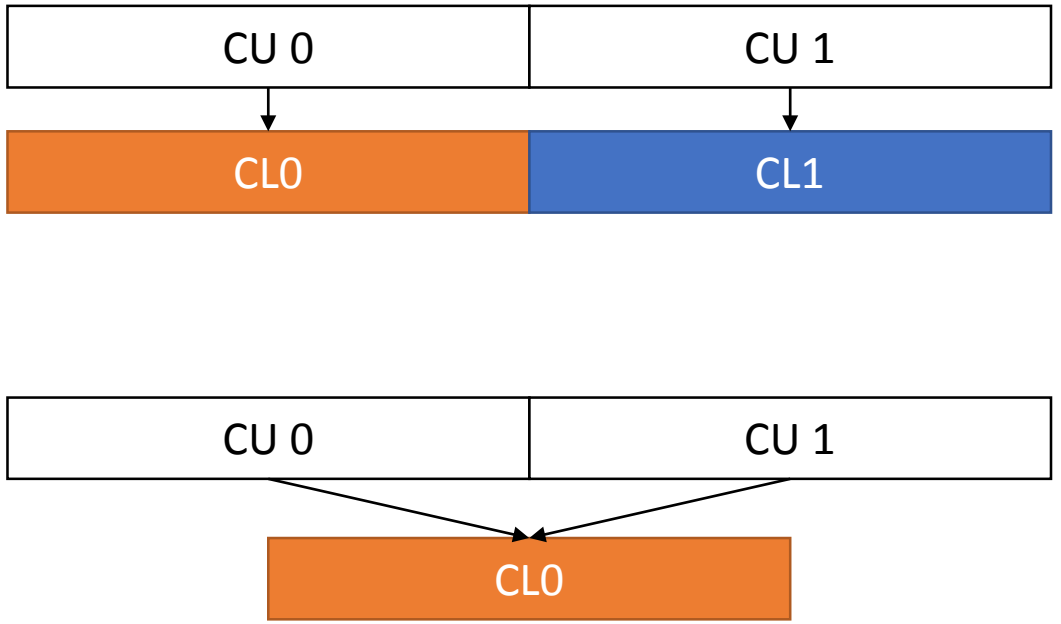

Request
Response
GPU 0
CU
L1
RDMA
L2
DRAM
GPU 1
CU
L1
RDMA
L2
DRAM
Interconnect

Problems

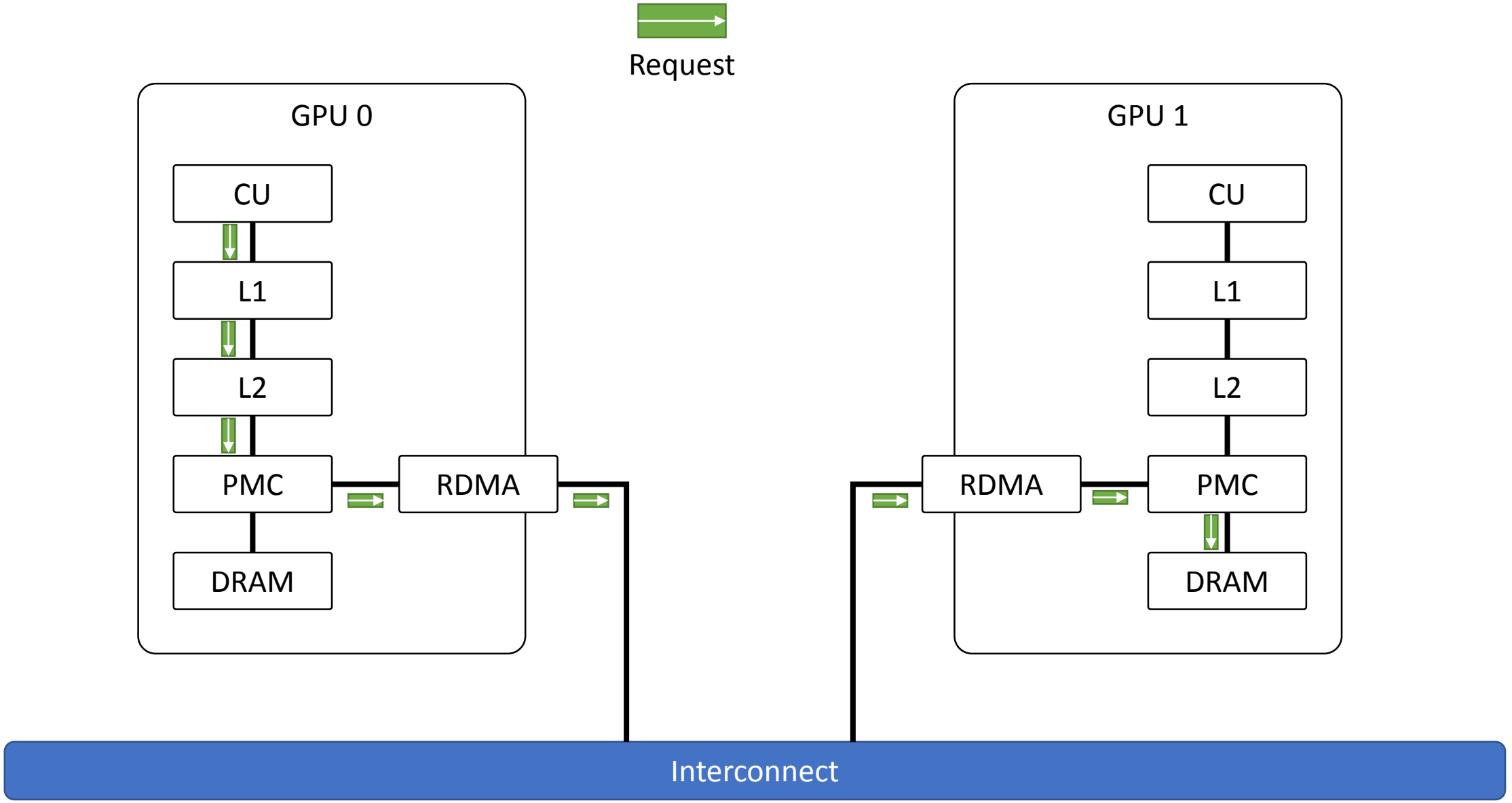
- Low interconnect bandwidth utilization

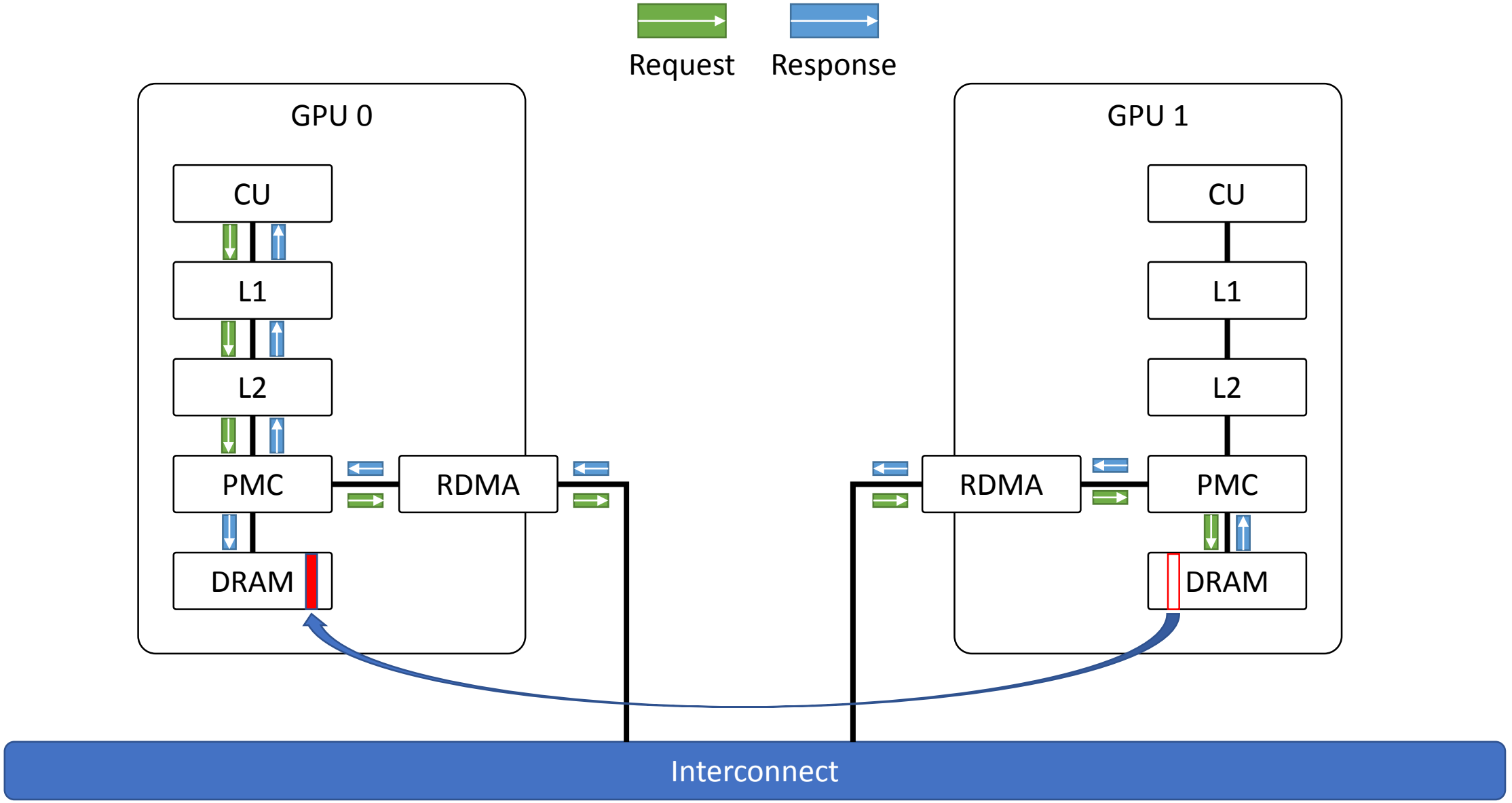
Problems

- Low interconnect bandwidth utilization
- Not utilizing spatial locality
- Not utilizing temporal locality

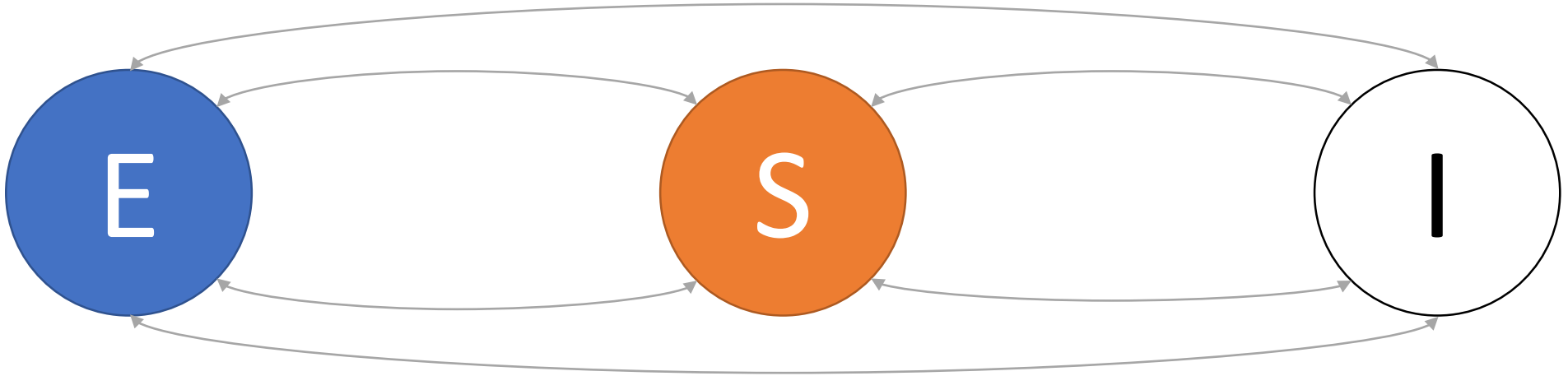


Progressive Page-Splitting Migration (PASI)


Request
GPU 0
CU
L1
L2
PMC
RDMA
DRAM
GPU 1
CU
L1
L2
RDMA
PMC
DRAM
Interconnect


Request
Response
GPU 0
CU
L1
L2
PMC
RDMA
DRAM
GPU 1
CU
L1
L2
RDMA
PMC
DRAM
Interconnect

Memory Coherence Protocol

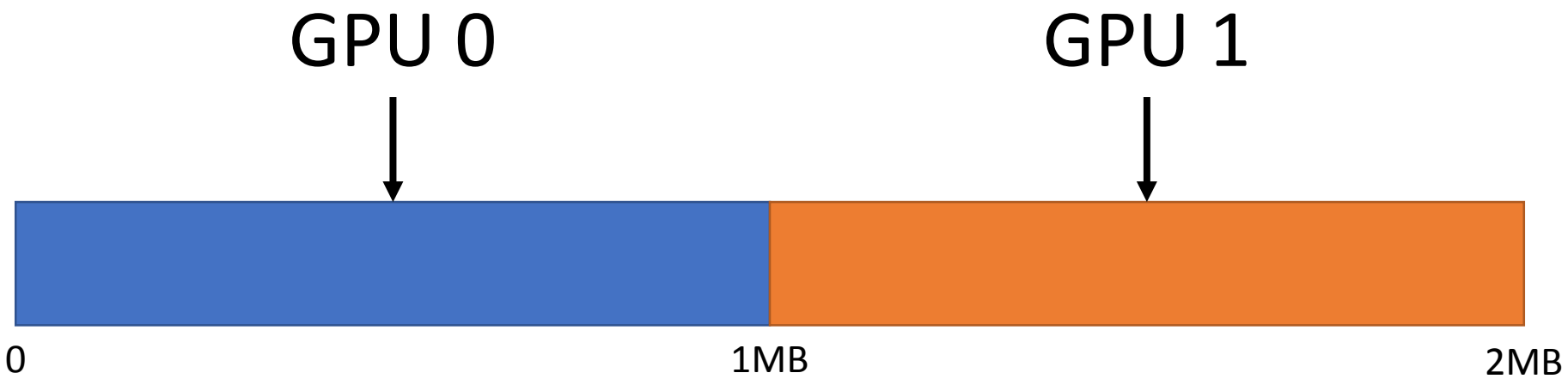


Exclusive
Write

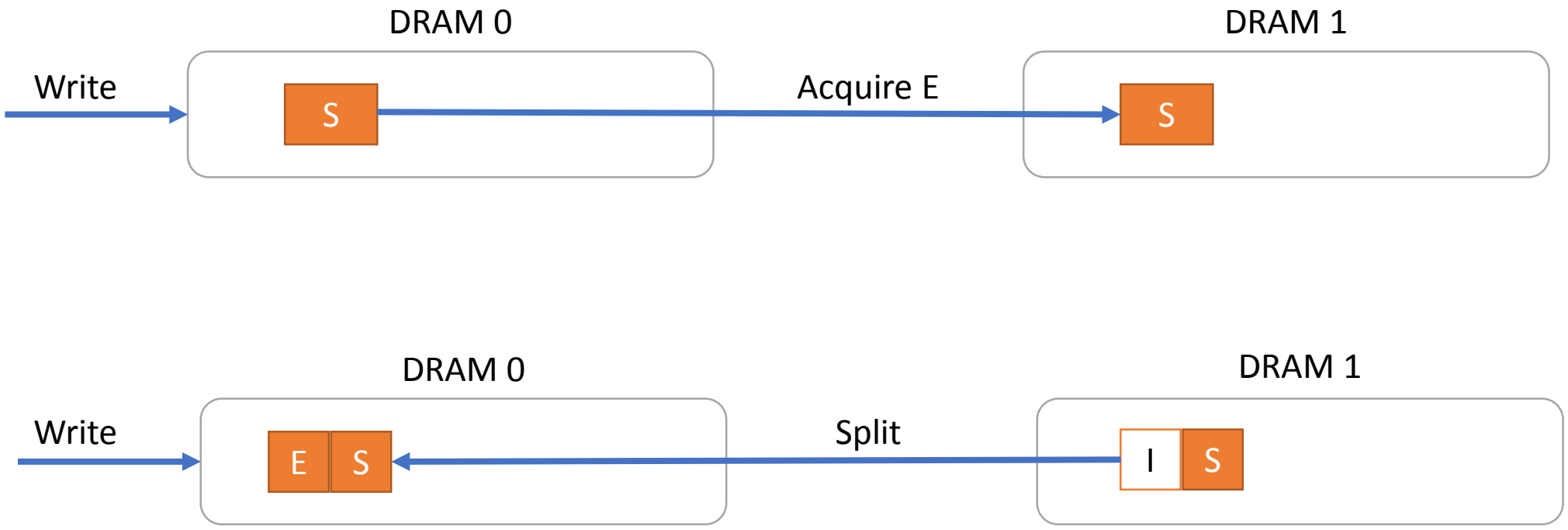
Shared
Read

Invalid

False sharing

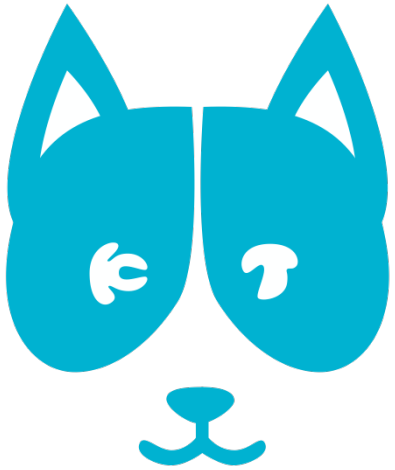


Page Splitting



Progressive Page-Splitting Migration

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

Thanks!

Yifan Sun

Northeastern University

https://syifan.github.io

@_syifan_