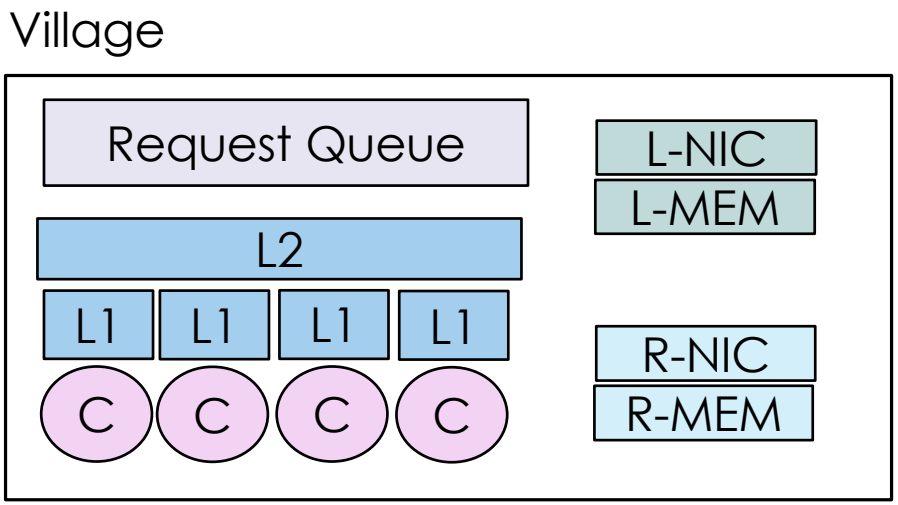
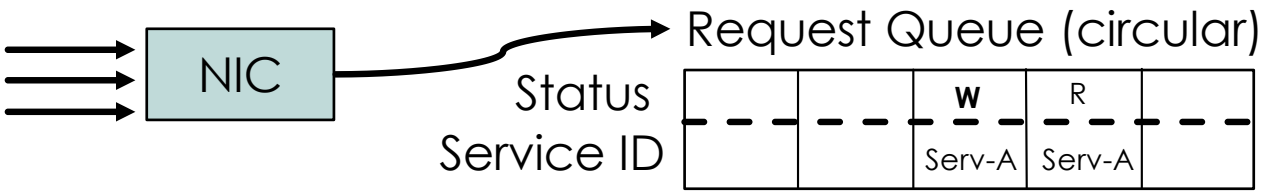
µManycore: A Cloud-Native CPU for Tail at Scale

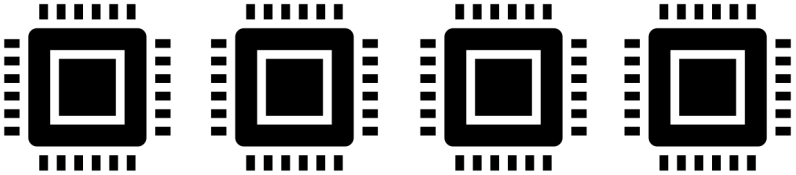
Basic unit of µManycore: a hardware cache-coherent Village



Hardware for Request Scheduling

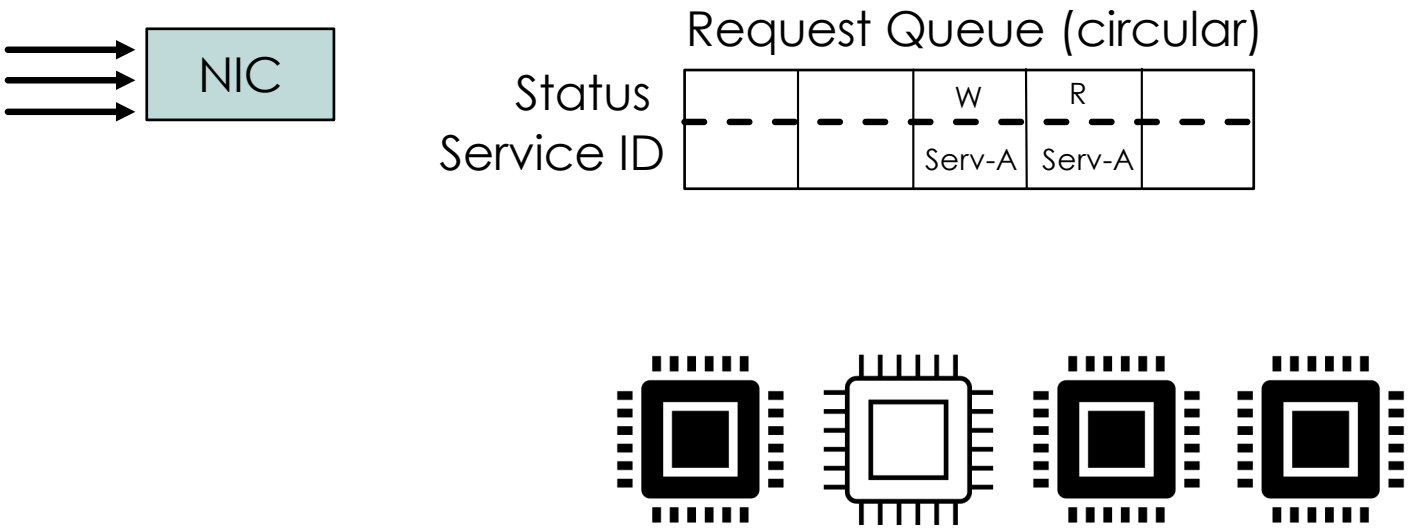
- NIC deposits ready requests to the queue
- Cores spin on *Work* flag, execute *Dequeue* instruction, finish with *Complete* instruction





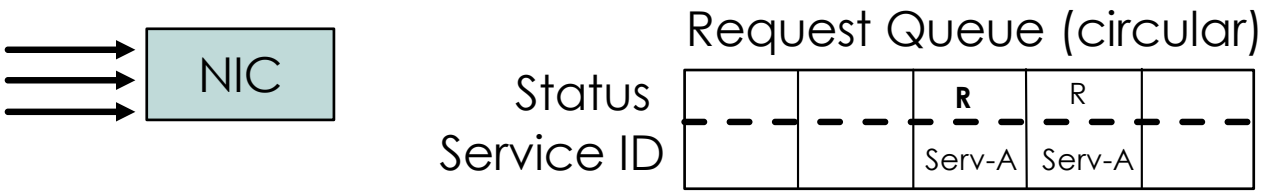
Hardware for Request Scheduling

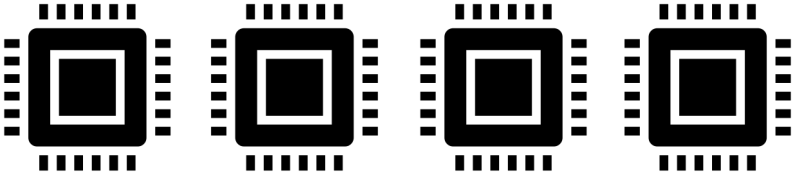
- NIC deposits ready requests to the queue
- Cores spin on *Work* flag, execute *Dequeue* instruction, finish with *Complete* instruction



Hardware for Request Scheduling

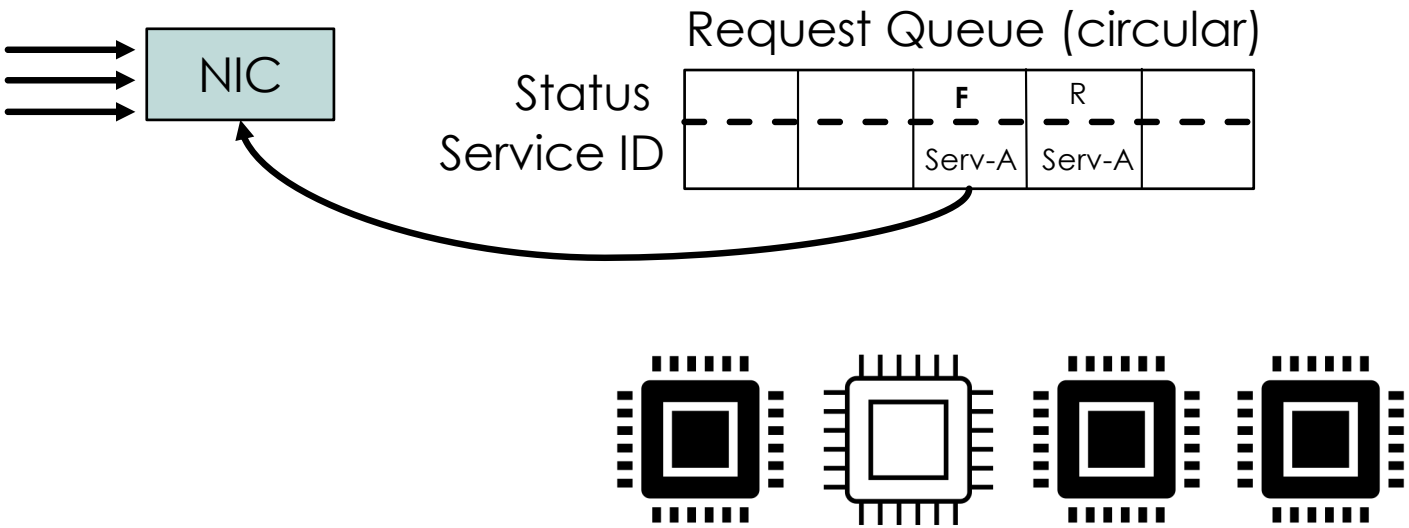
- NIC deposits ready requests to the queue
- Cores spin on *Work* flag, execute *Dequeue* instruction, finish with *Complete* instruction





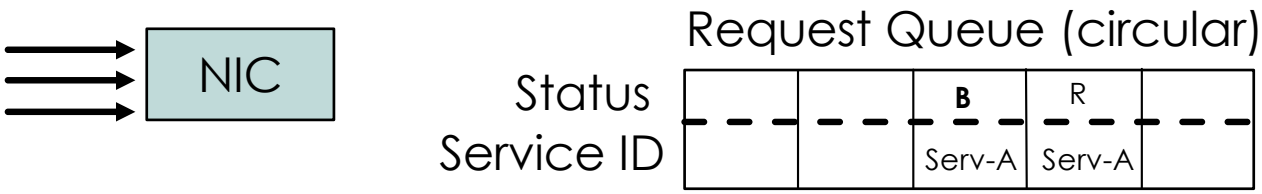
Hardware for Request Scheduling

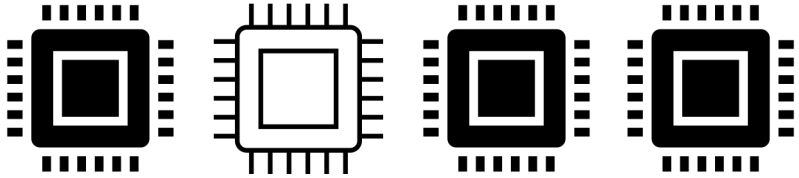
- NIC deposits ready requests to the queue
- Cores spin on *Work* flag, execute *Dequeue* instruction, finish with *Complete* instruction



Hardware for Context Switching

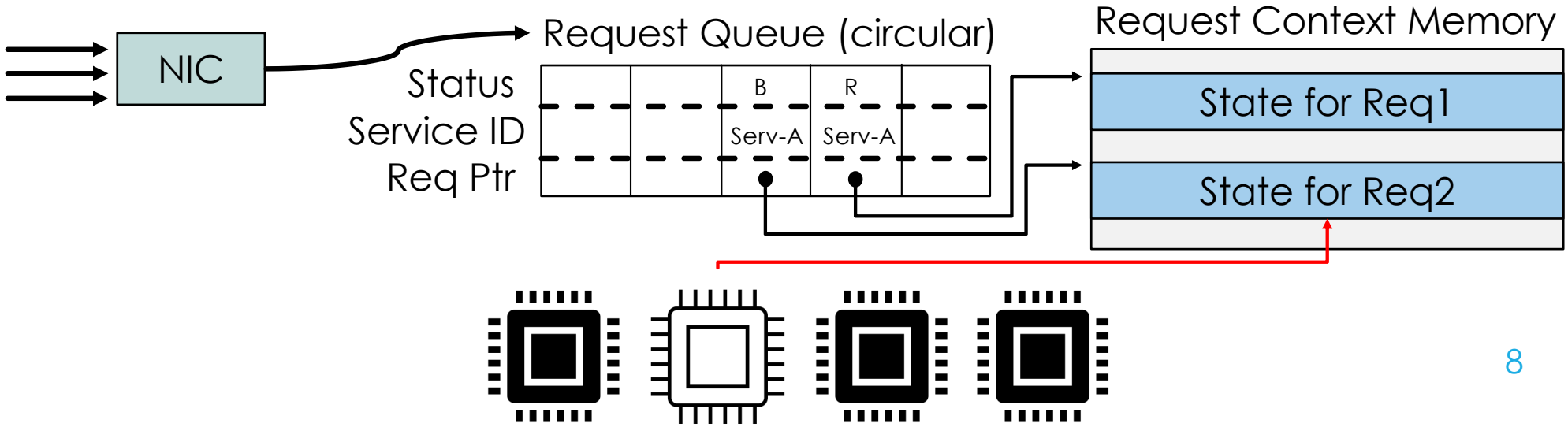
- Requests can get blocked during execution – need to context switch





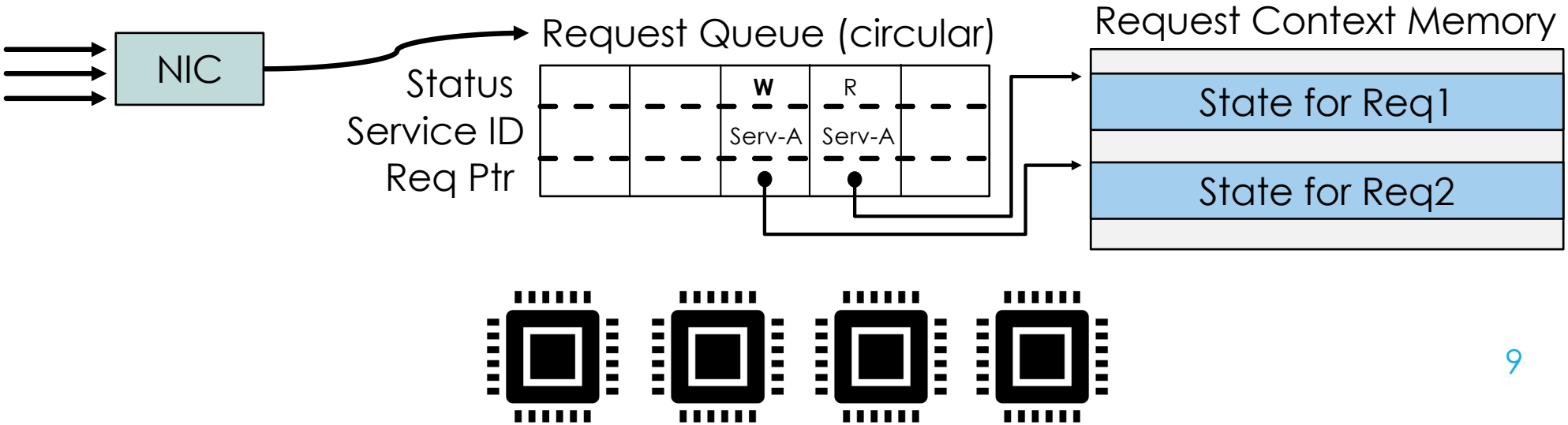
Hardware for Context Switching

- Avoid OS invocations and software overheads
- Core saves and restores context in hardware



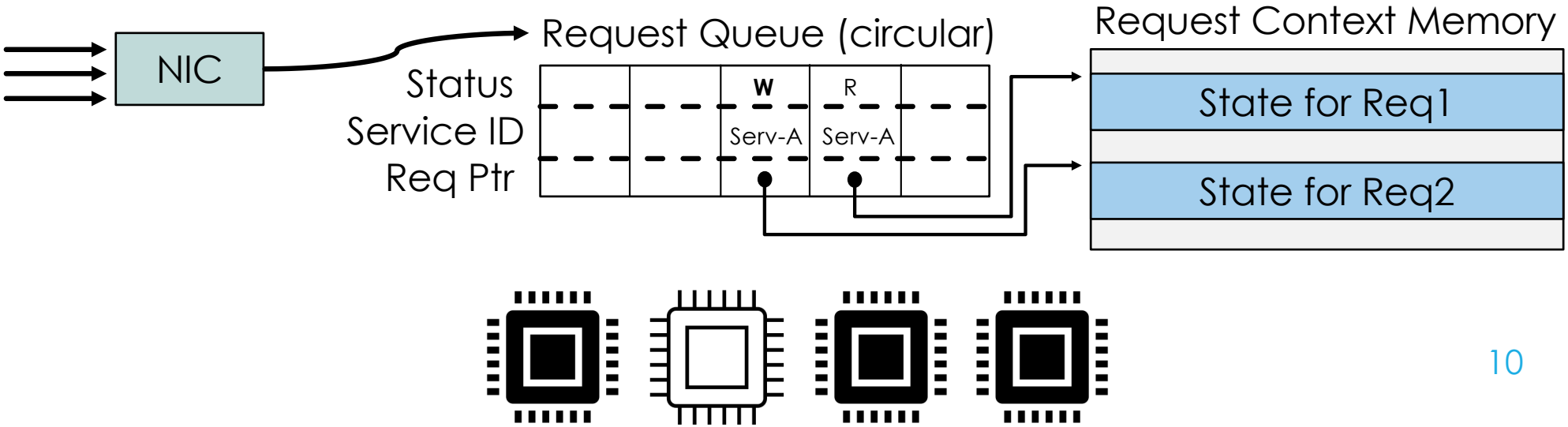
Hardware for Context Switching

- Avoid OS invocations and software overheads
- Core saves and restores context in hardware



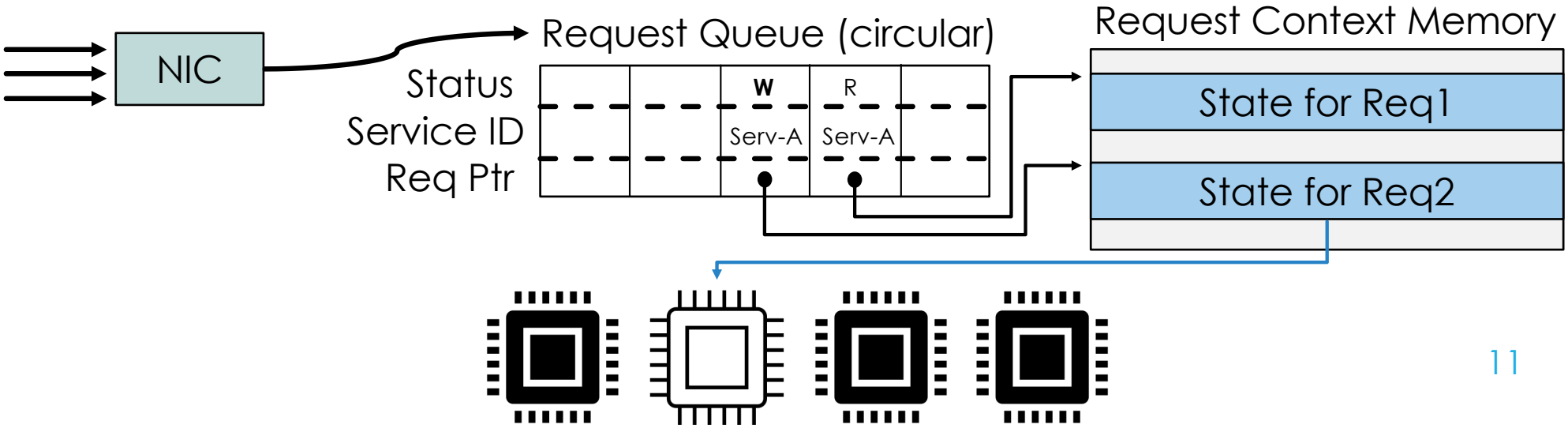
Hardware for Context Switching

- Avoid OS invocations and software overheads
- Core saves and restores context in hardware



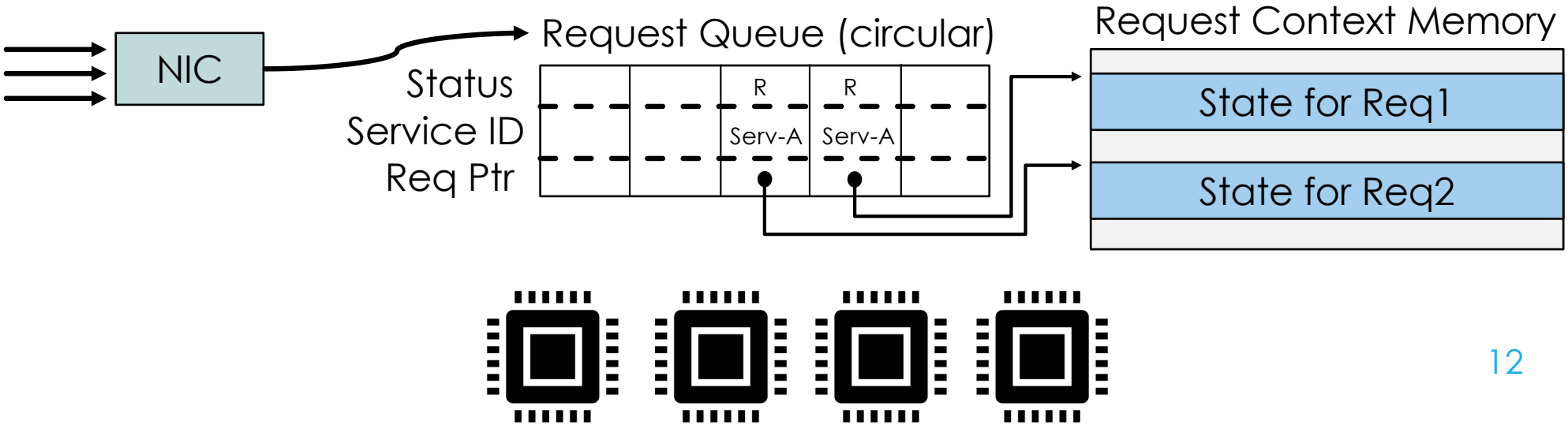
Hardware for Context Switching

- Avoid OS invocations and software overheads
- Core saves and restores context in hardware



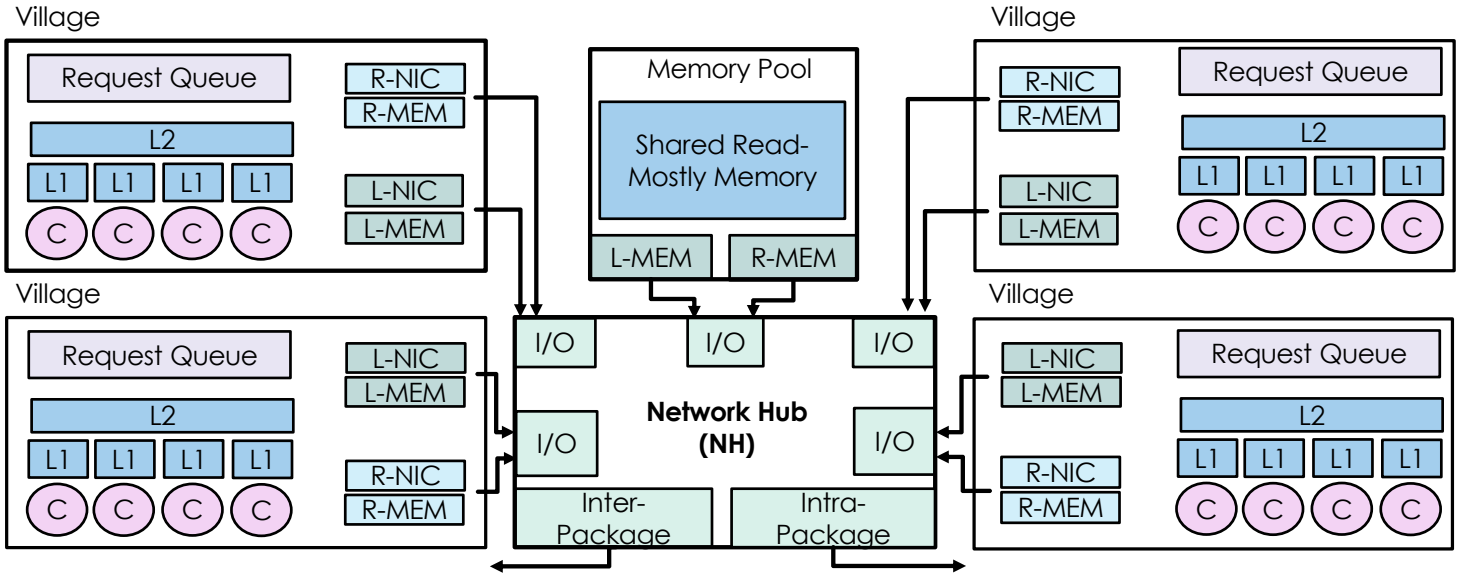
Hardware for Context Switching

- Avoid OS invocations and software overheads
- Core saves and restores context in hardware



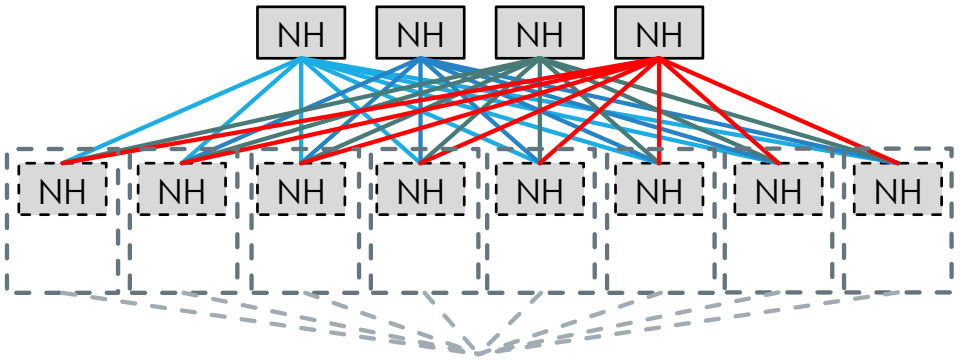
Villages grouped into clusters

- The combination of a few villages, a memory pool, and a network hub → a cluster



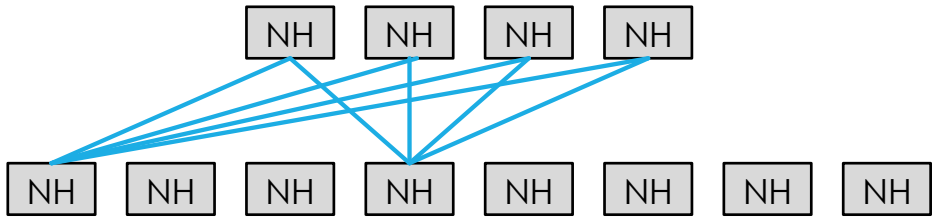
Leaf-spine on-package network

- Many redundant, low-hop count paths between any two clusters



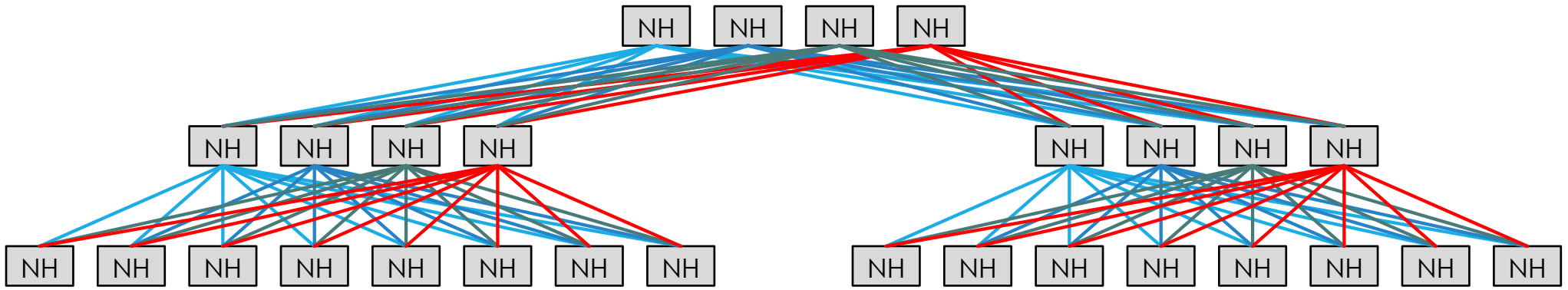
Leaf-spine on-package network

- Many redundant, low-hop count paths between any two clusters
  - Even between the same source and destination multiple parallel links



Hierarchical leaf-spine on-package network

- Many redundant, low-hop count paths between any two clusters

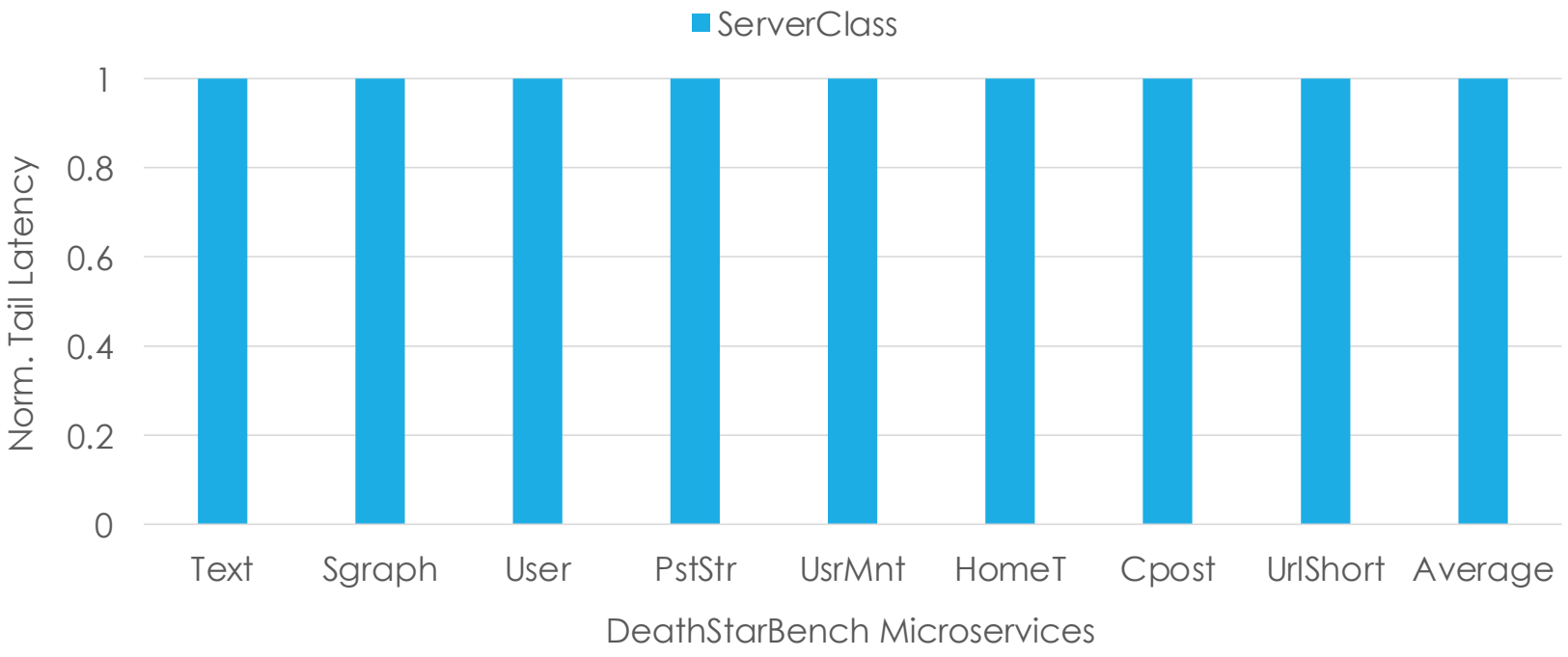


Evaluation Setup

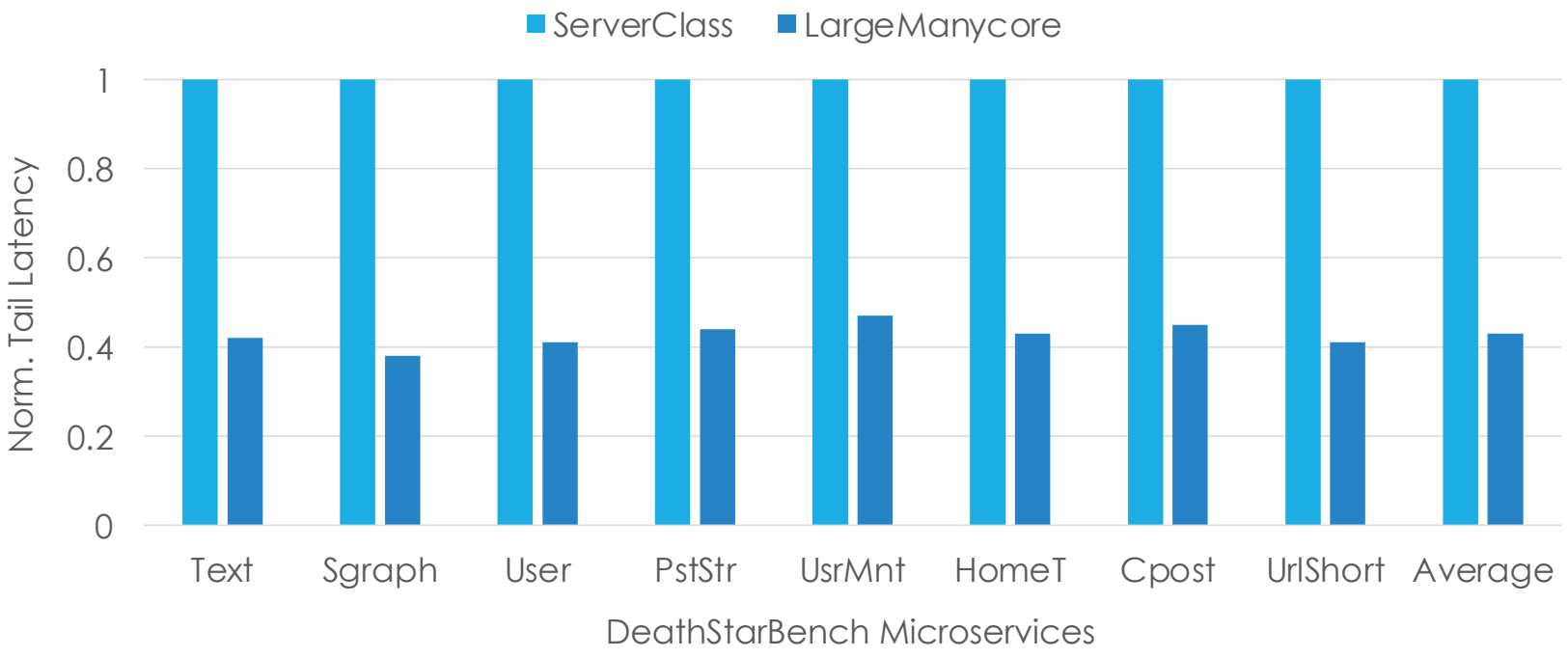
- 1024-core µManycore
- DeathStarBench microservices
- PinTool to extract traces
- SST for cycle-accurate timing measurements
- McPAT + Cacti for power/area measurements
- Two baselines

Baseline	Number of cores	Modeled After	Design Point
ServerClass	40	Intel Ice-Lake	Same Power as µManycore
LargeManycore	1024	ARM A15	Same Area as µManycore

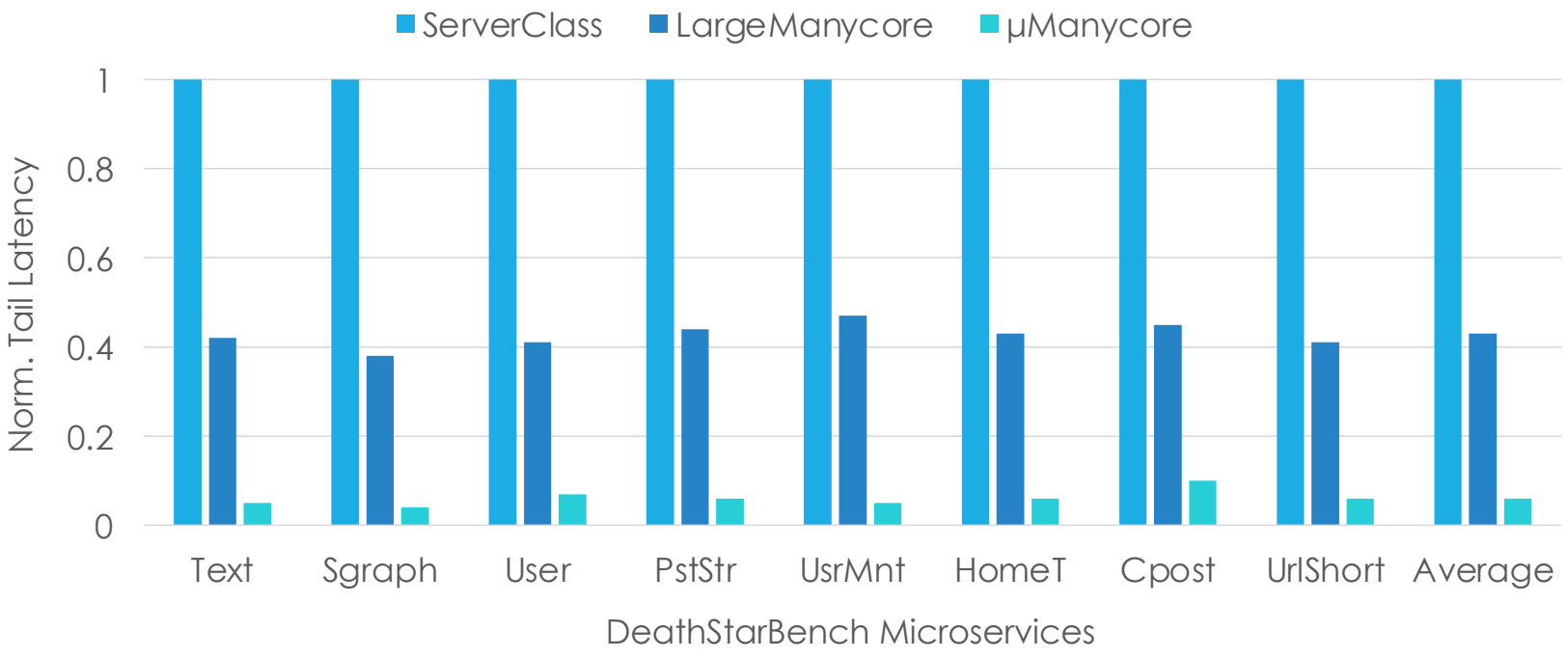
µManycore Significantly Reduces Tail



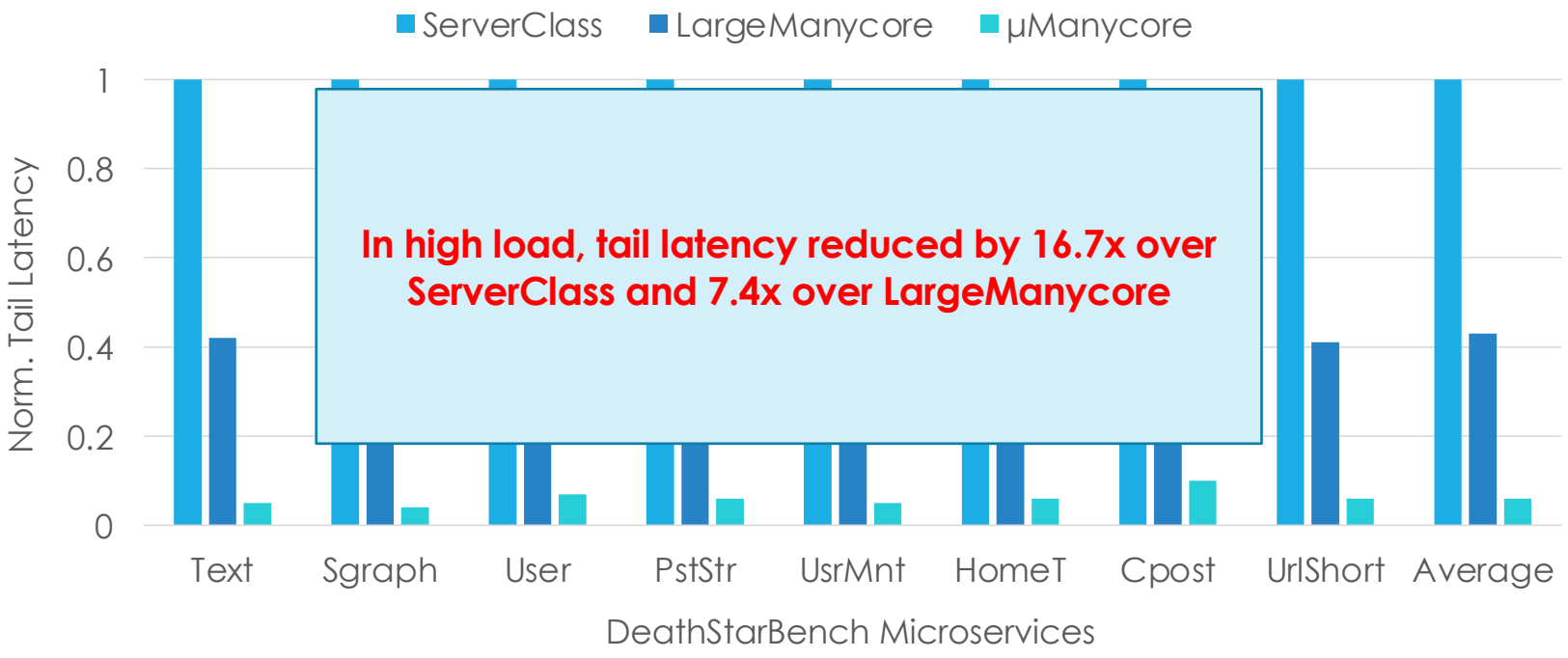
µManycore Significantly Reduces Tail



µManycore Significantly Reduces Tail



µManycore Significantly Reduces Tail



Conclusion

- Imbalance between current processors and emerging microservice environments
- μManycore → an architecture optimized for microservice environments
- μManycore delivers high performance for microservice workloads
  - 10.4X reduced tail latency
  - 15.5X improved throughput