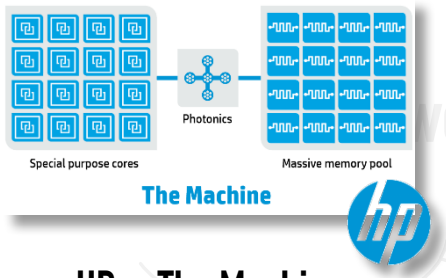
Network Requirements for Resource Disaggregation

Peter Gao (Berkeley), Akshay Narayan (MIT), Sagar Karandikar (Berkeley), Joao Carreira (Berkeley), Sangjin Han (Berkeley), Rachit Agarwal (Cornell), Sylvia Ratnasamy (Berkeley), Scott Shenker (Berkeley/ICSI)

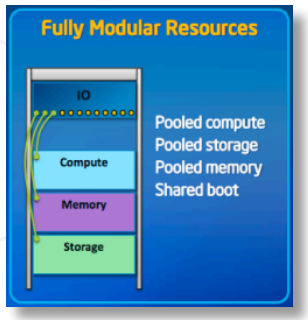
Disaggregated Datacenters

Current Datacenter: Server-Centric

Future datacenter: Disaggregated?



HP – The Machine

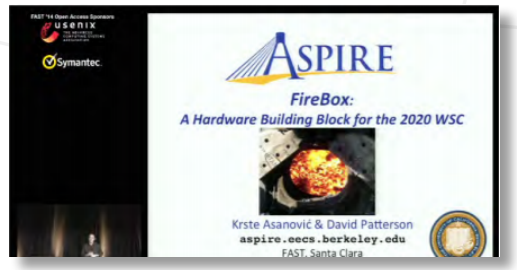


Intel - RSD

Facebook

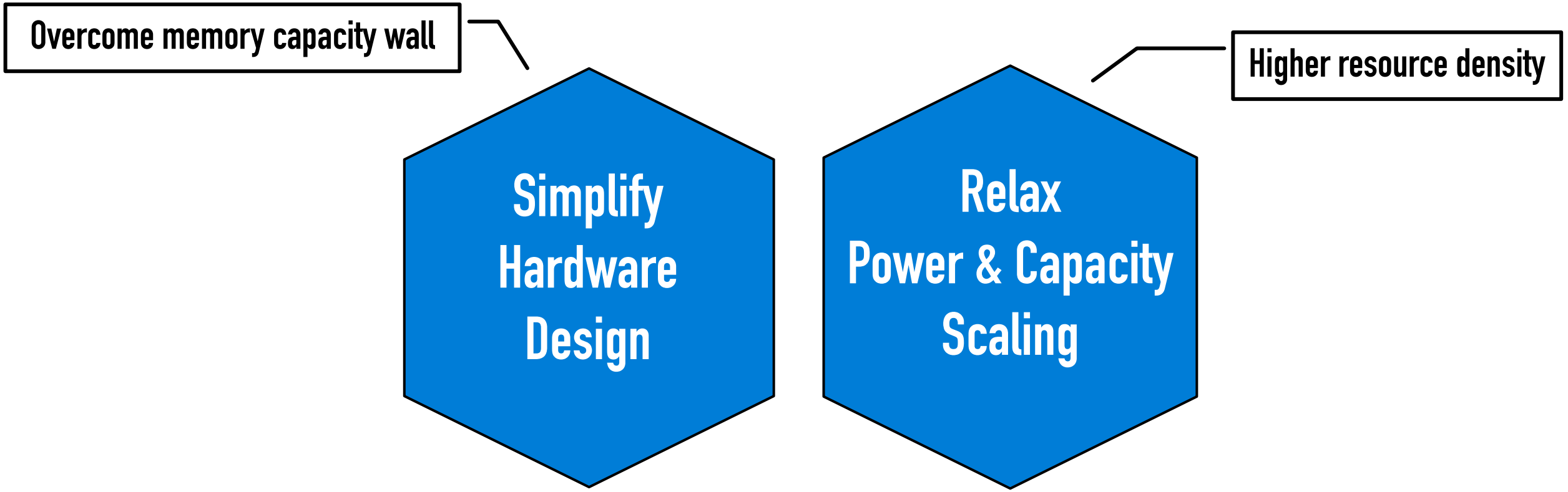
Huawei - NUWA

SeaMicro

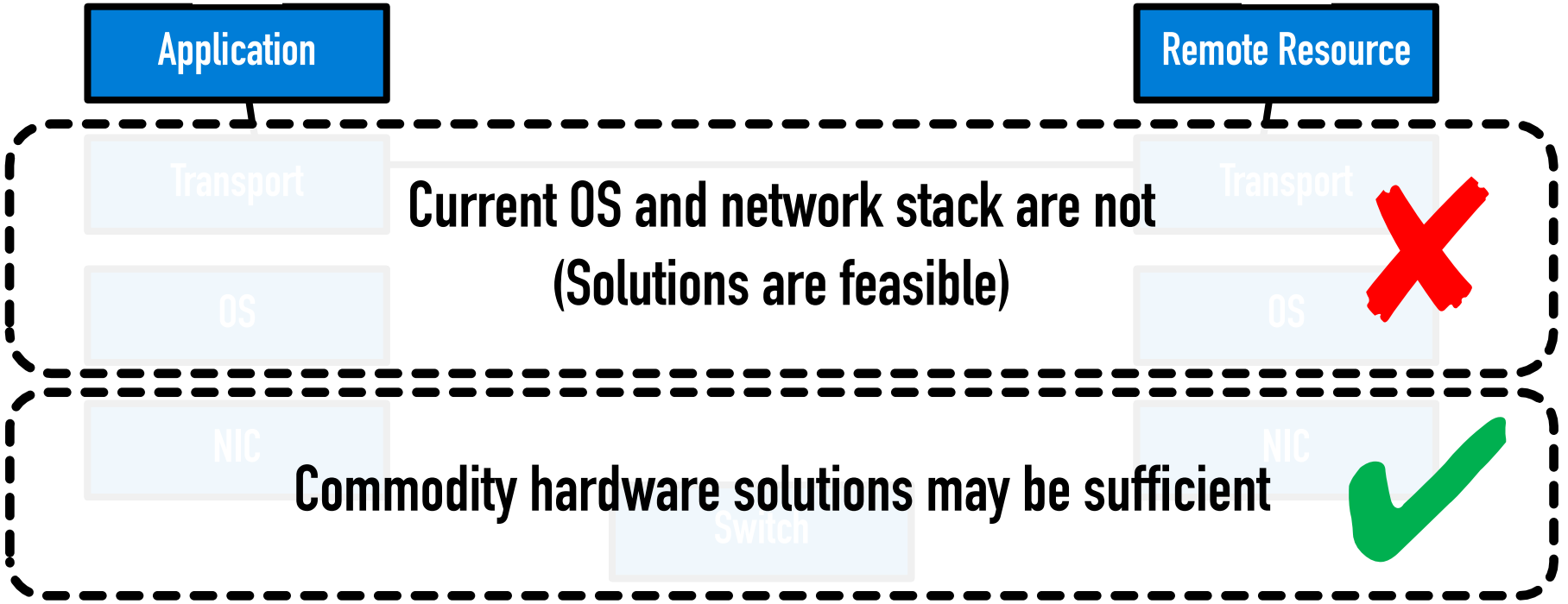


Berkeley - FireBox

Disaggregation Benefits (Architecture Community)



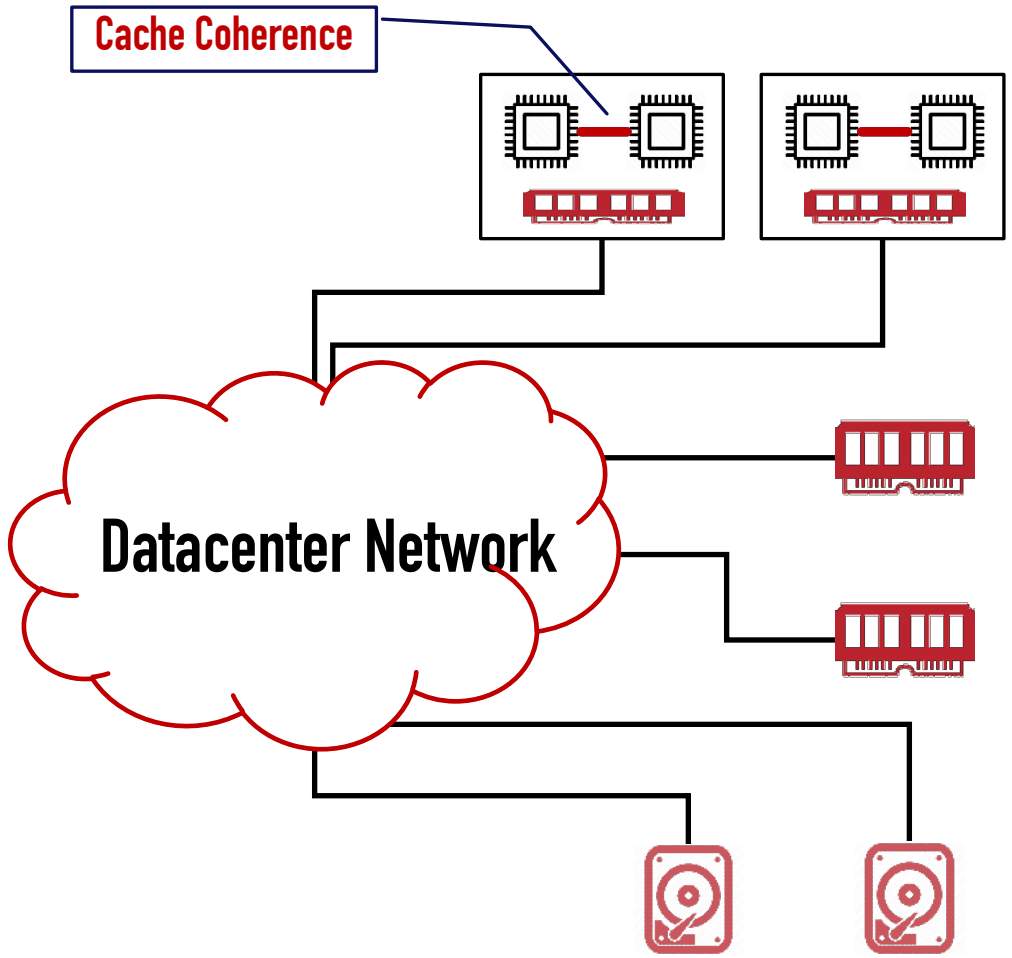
Do we need specialized hardware?
e.g.: Silicon photonics, PCI-e



- What end-to-end latency and bandwidth must the network provide for legacy apps?
  - Do existing transport protocols meet these requirements?
  - Do existing OS network stacks meet these requirements?
  - Can commodity network hardware meet these requirements?

Worst case performance degradation

Assumptions



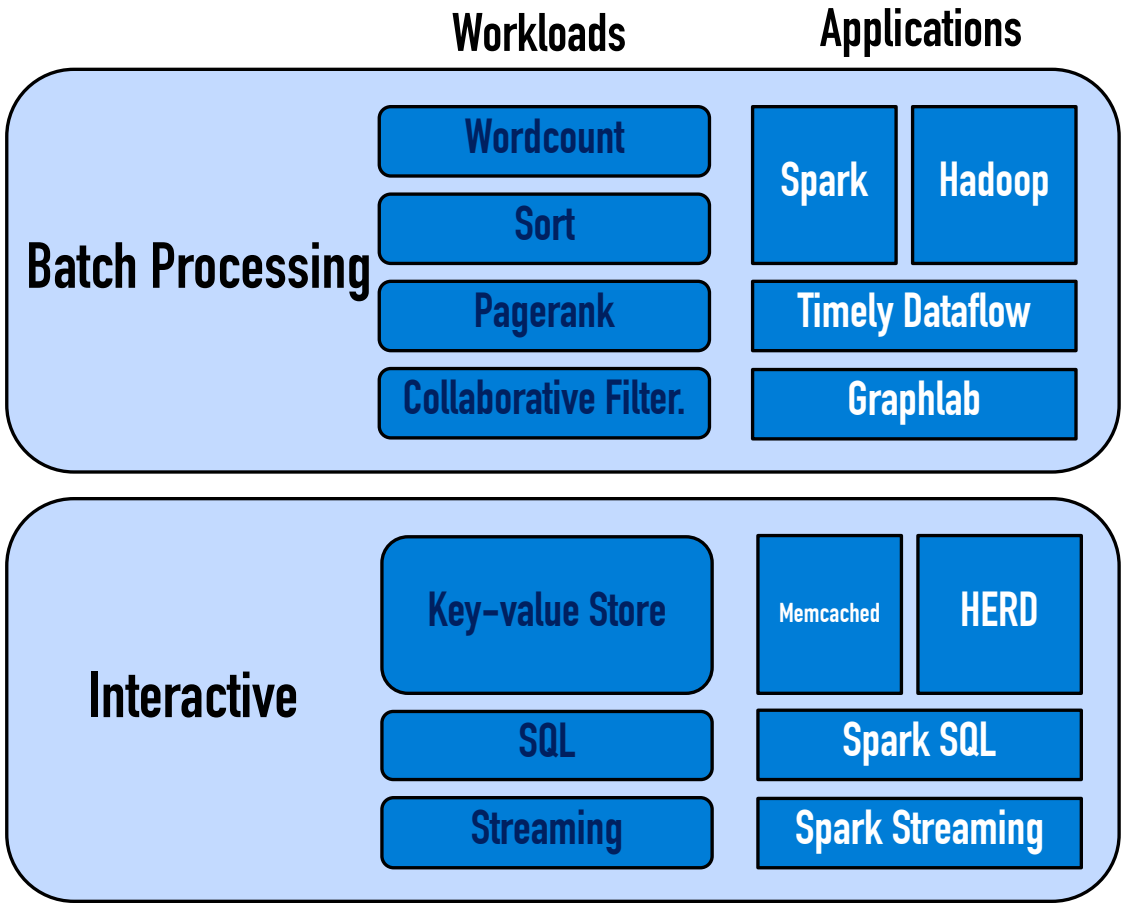
CPU	• Limited cache coherence domain • Small amount of local cache (how much?)
Memory	• Page-level remote memory access
Storage	• Block-level distributed data placement
Scale	• Rack-scale? • Datacenter-scale?

Latency and Bandwidth Requirements

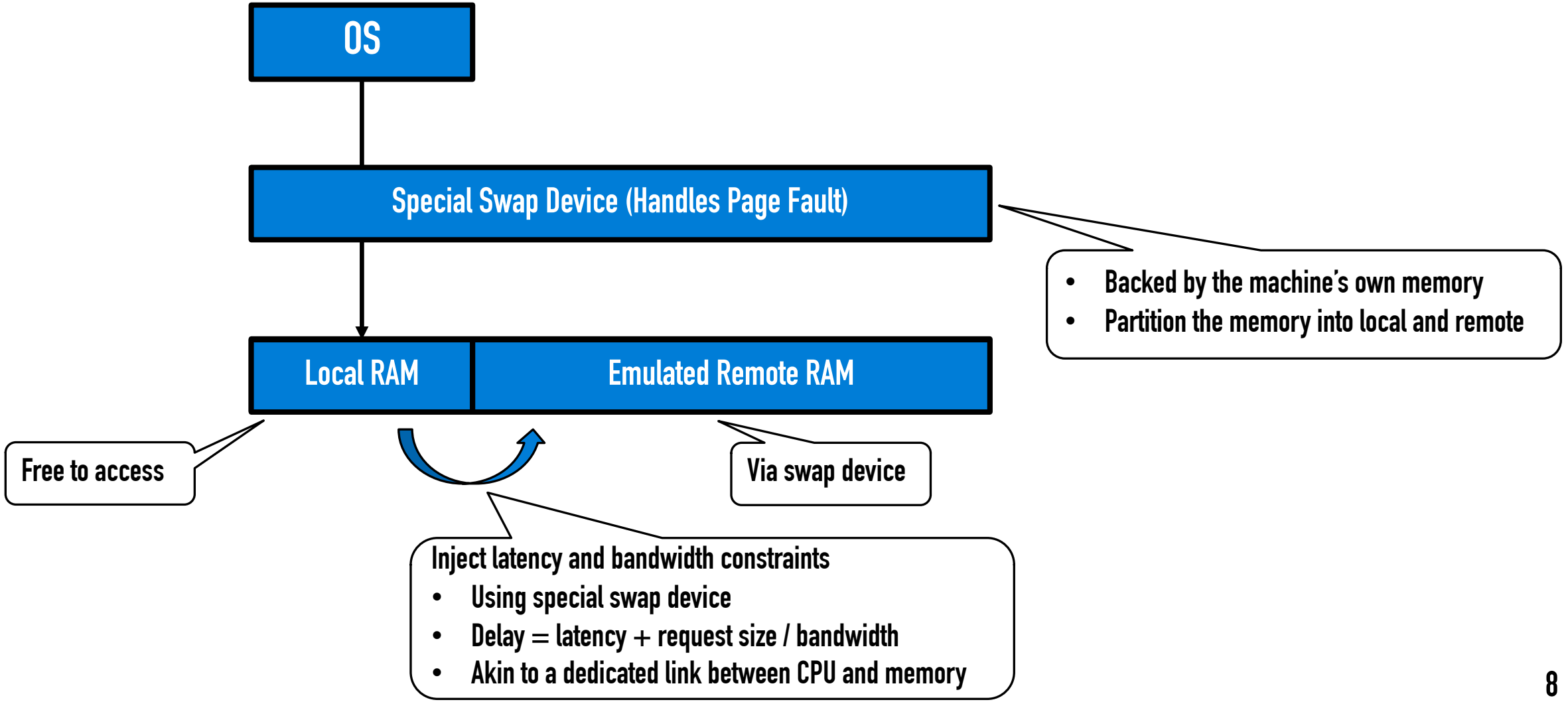
Methodology: Workload Driven

- 10 workloads on 8 applications
- ~ 125 GB input data

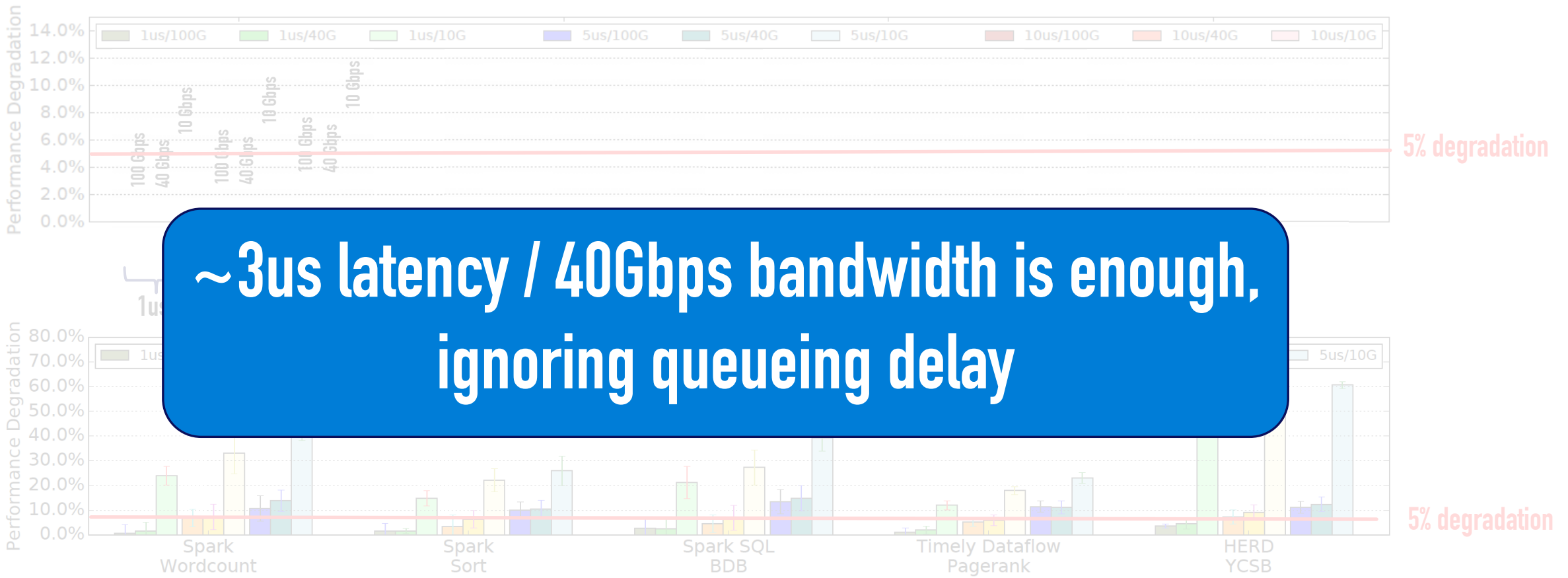
- 5 m3.2xlarge EC2 nodes
- Virtual Private Cloud enabled



Disaggregated Datacenter Emulator



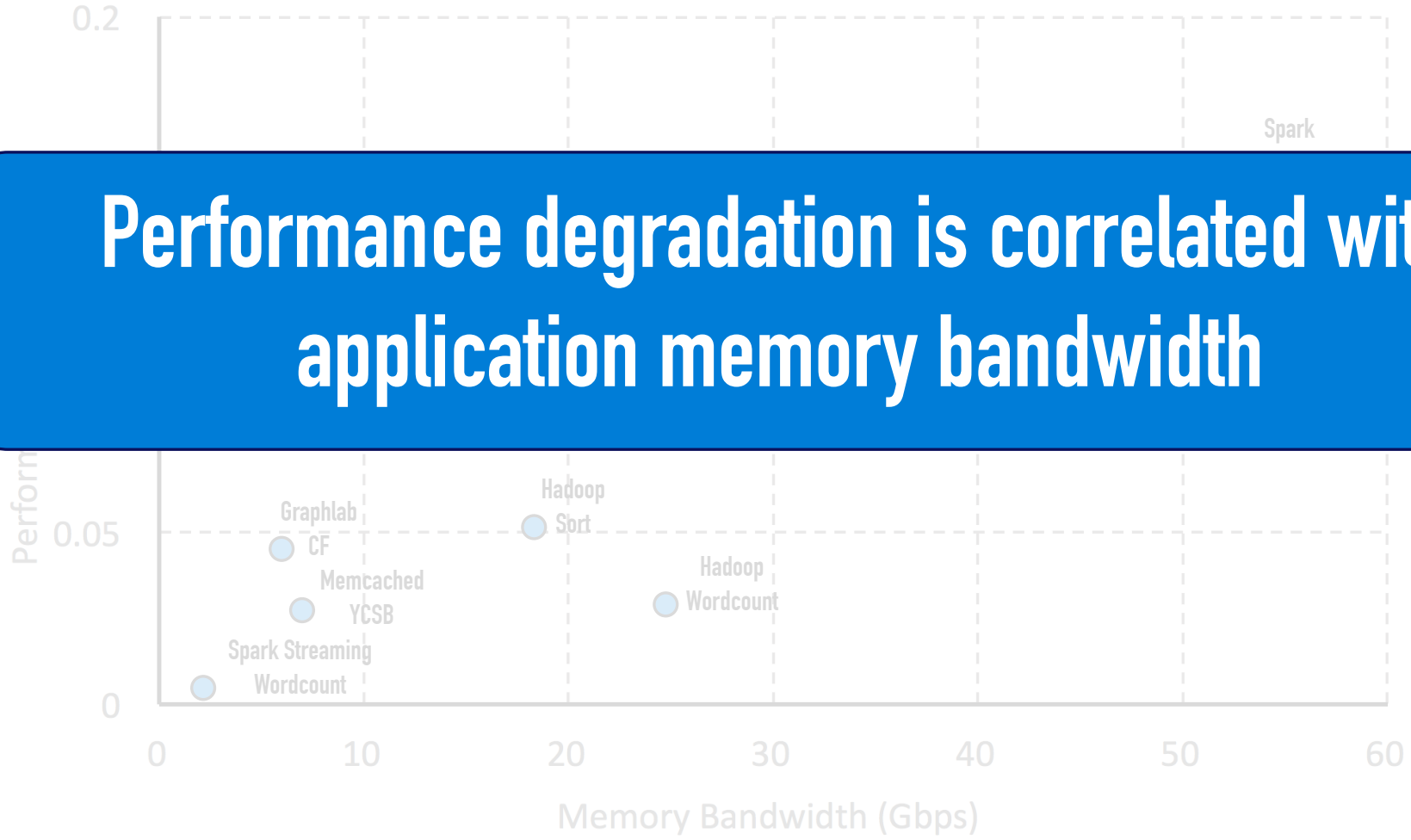
Latency and Bandwidth Requirement



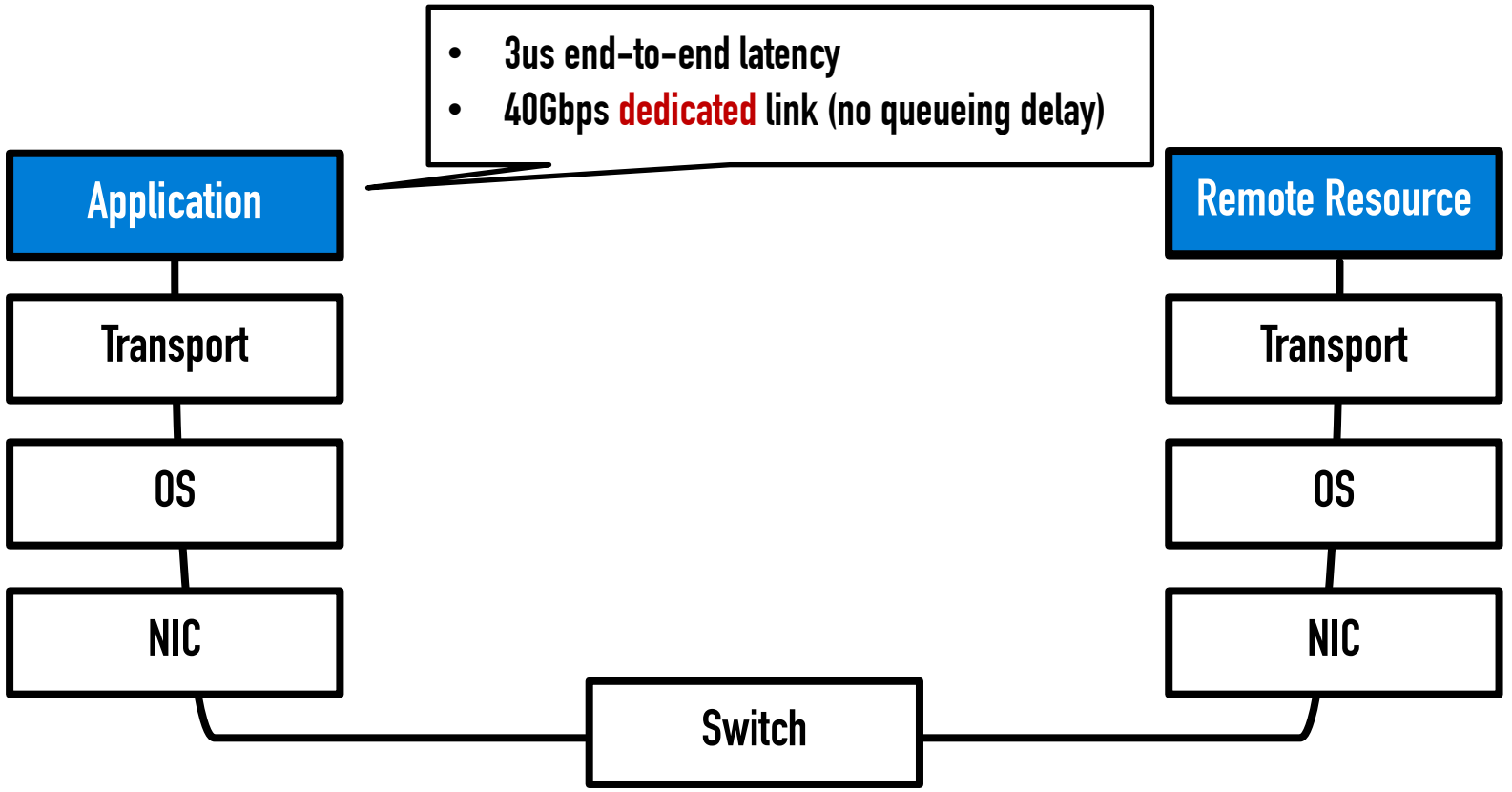
~3us latency / 40Gbps bandwidth is enough, ignoring queueing delay

*Note: Delay = latency + request size / bandwidth

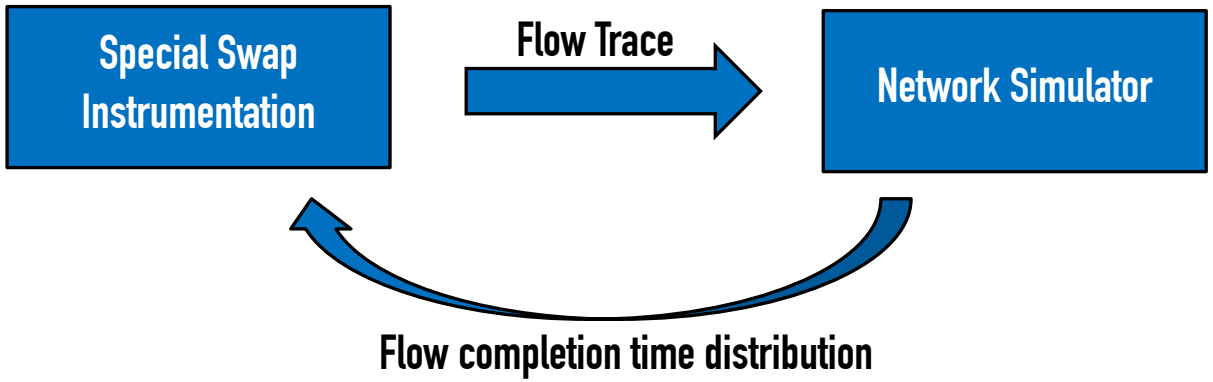
Understanding Performance Degradation

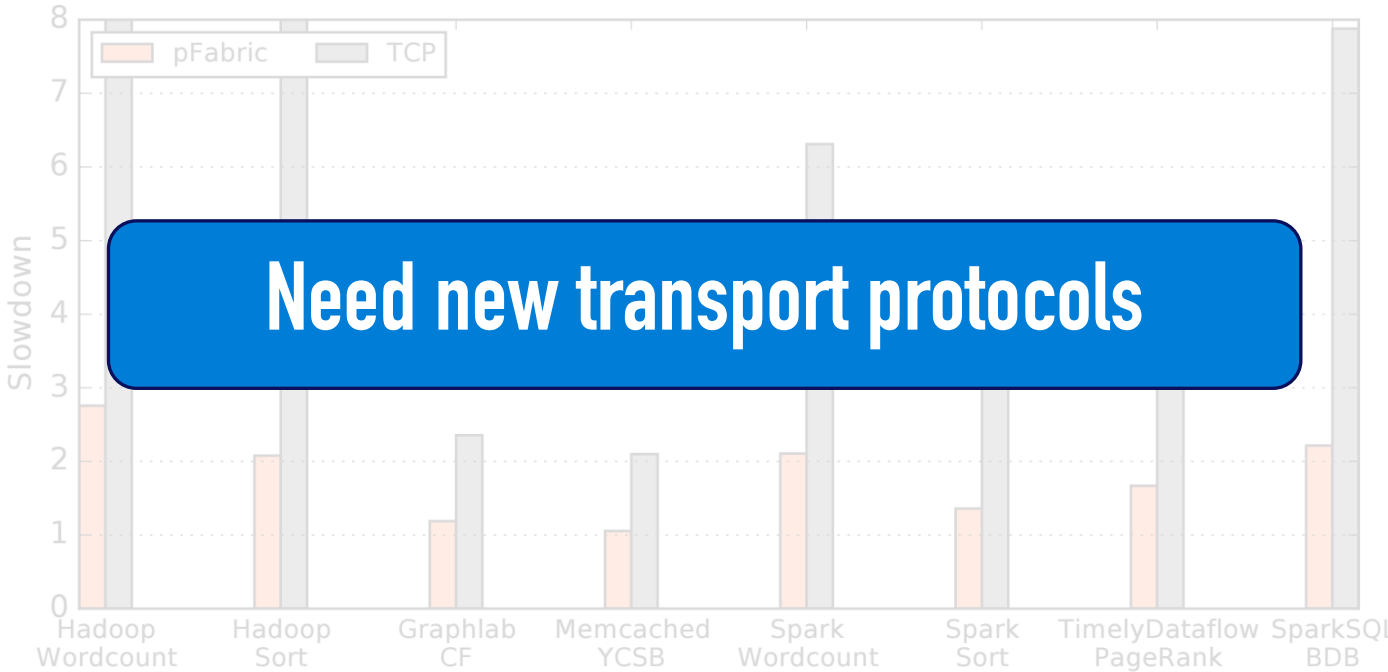


Performance degradation is correlated with application memory bandwidth


• 3us end-to-end latency
• 40Gbps dedicated link (no queueing delay)
Application
Transport
OS
NIC
Switch
Remote Resource
Transport
OS
NIC

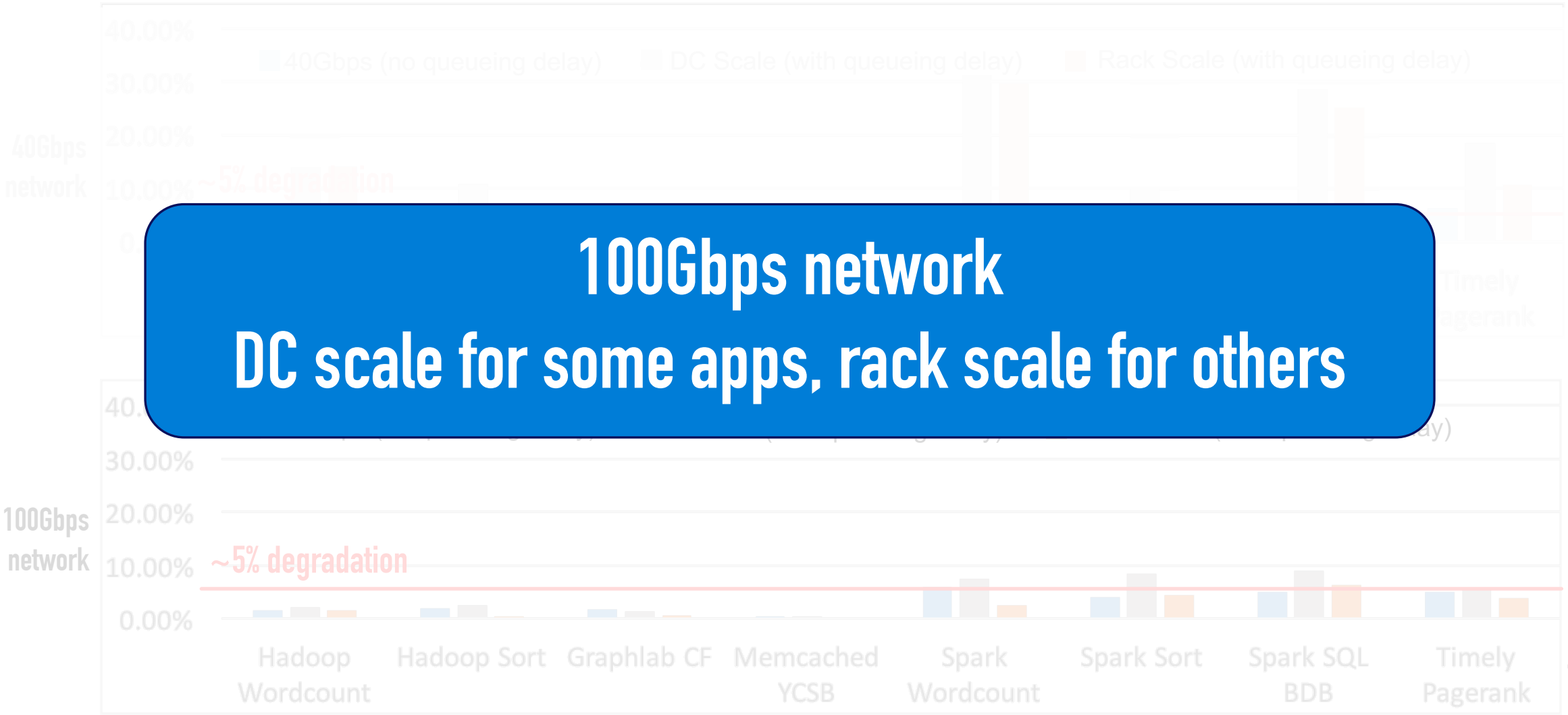
Transport Simulation Setting



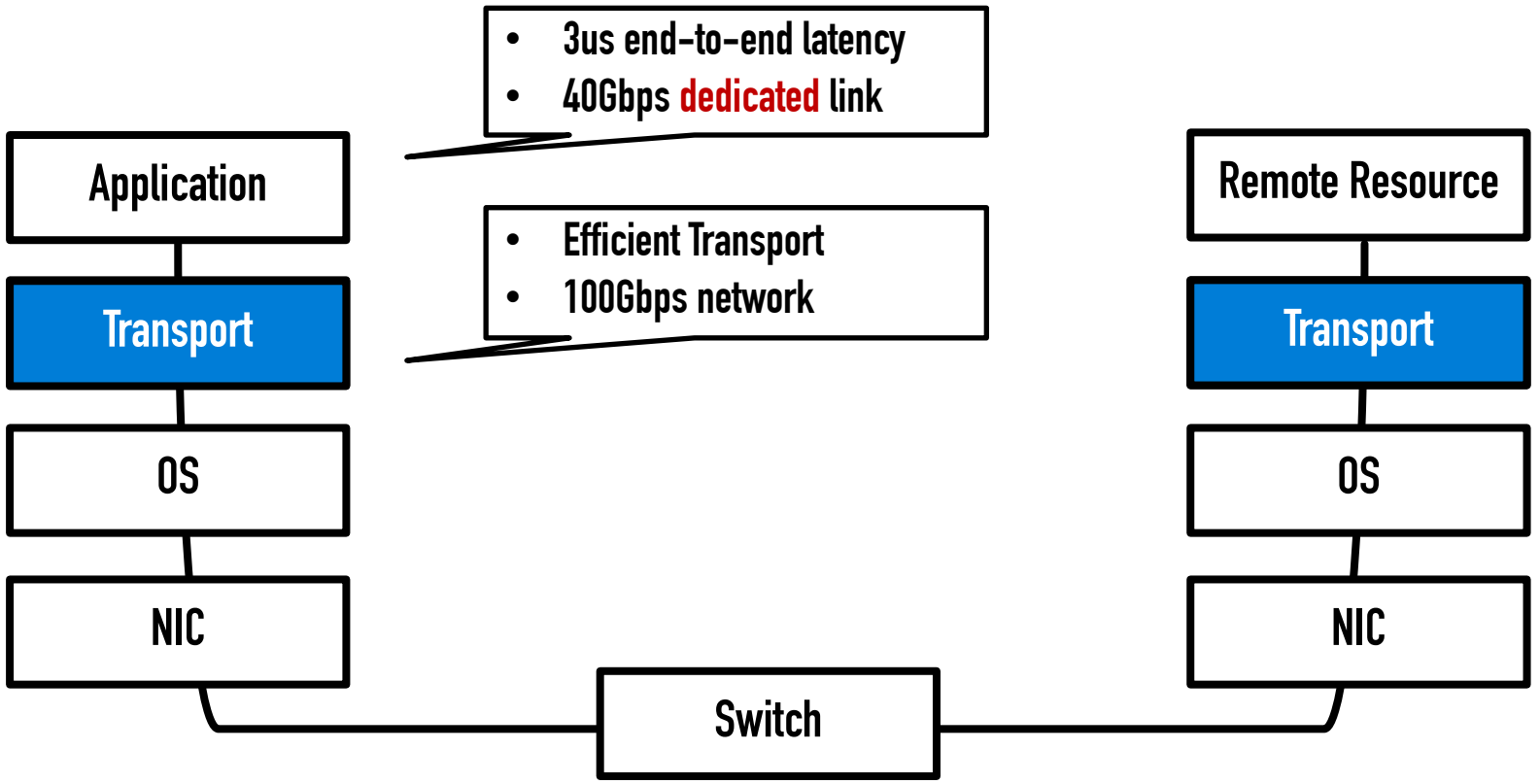


Need new transport protocols

Application Performance Degradation

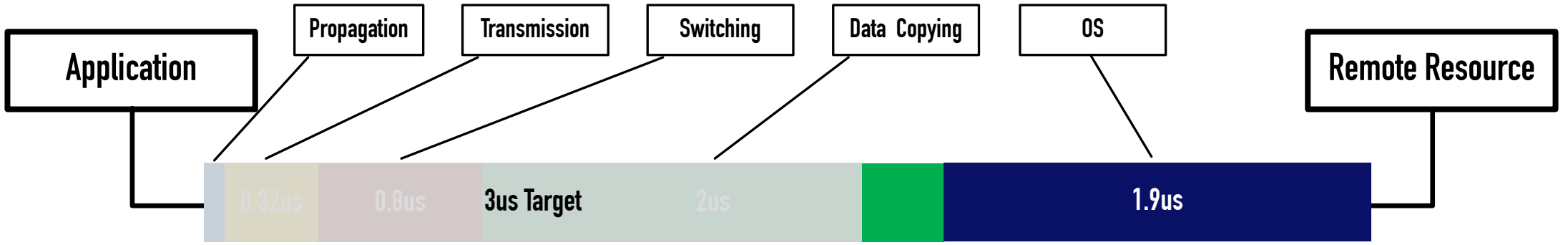


100Gbps network
DC scale for some apps, rack scale for others


• 3us end-to-end latency
• 40Gbps dedicated link
• Efficient Transport
• 100Gbps network
Application
Transport
OS
NIC
Switch
Remote Resource
Transport
OS
NIC

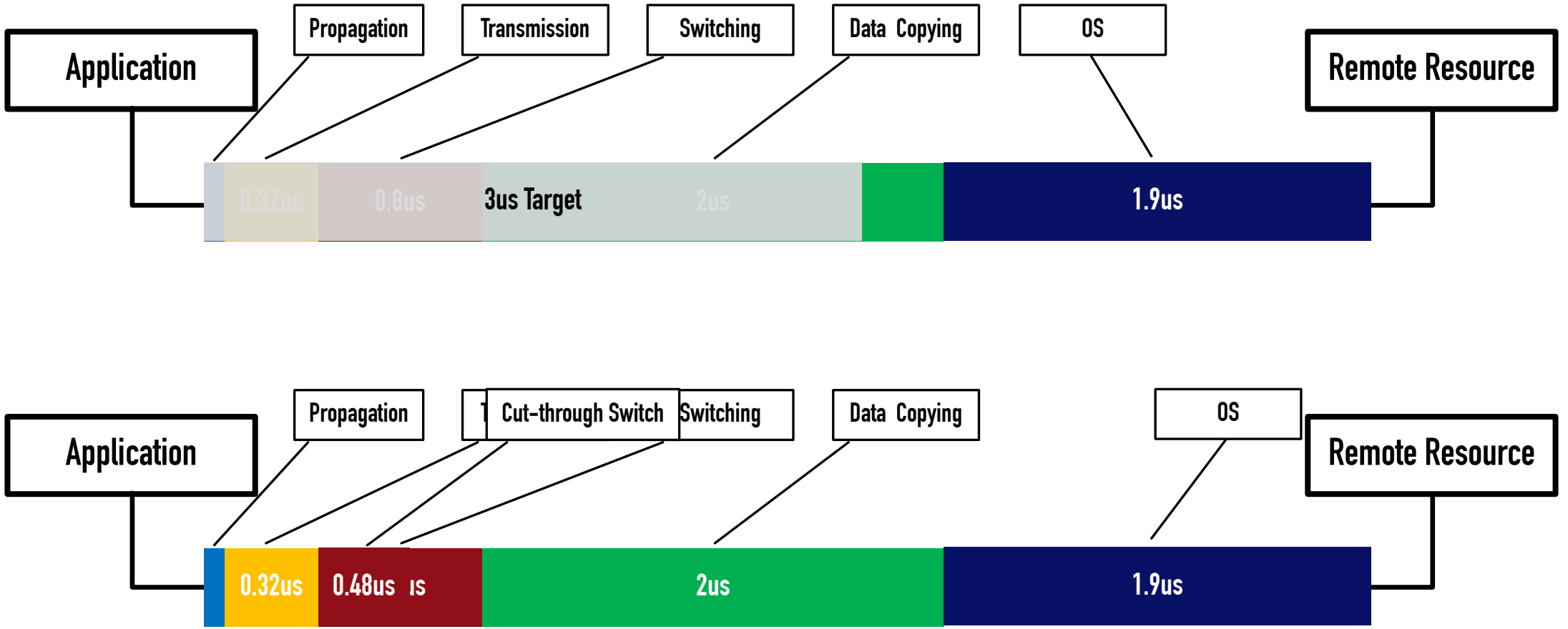
Is 100Gbps/3us achievable?

Feasibility of end-to-end latency within a rack



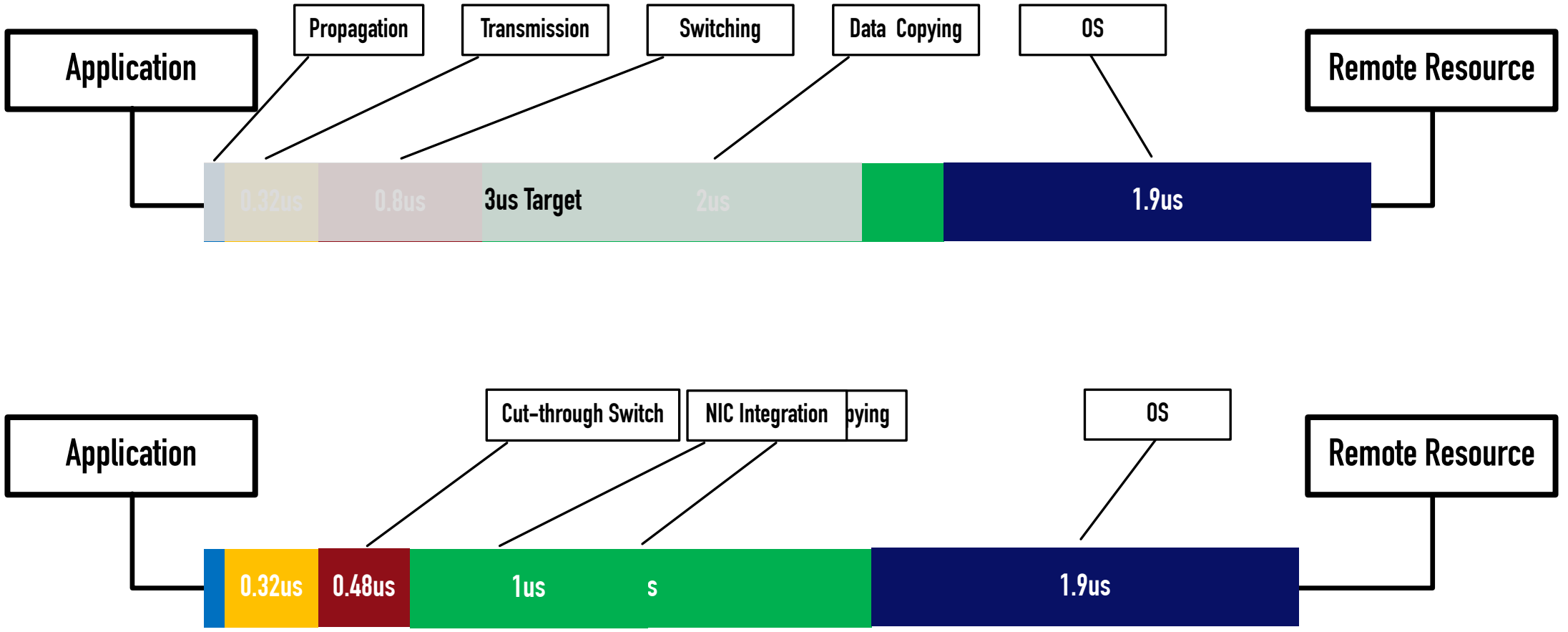
*Numbers estimated optimistically based on existing hardware

Feasibility of end-to-end latency within a rack



*Numbers estimated optimistically based on existing hardware

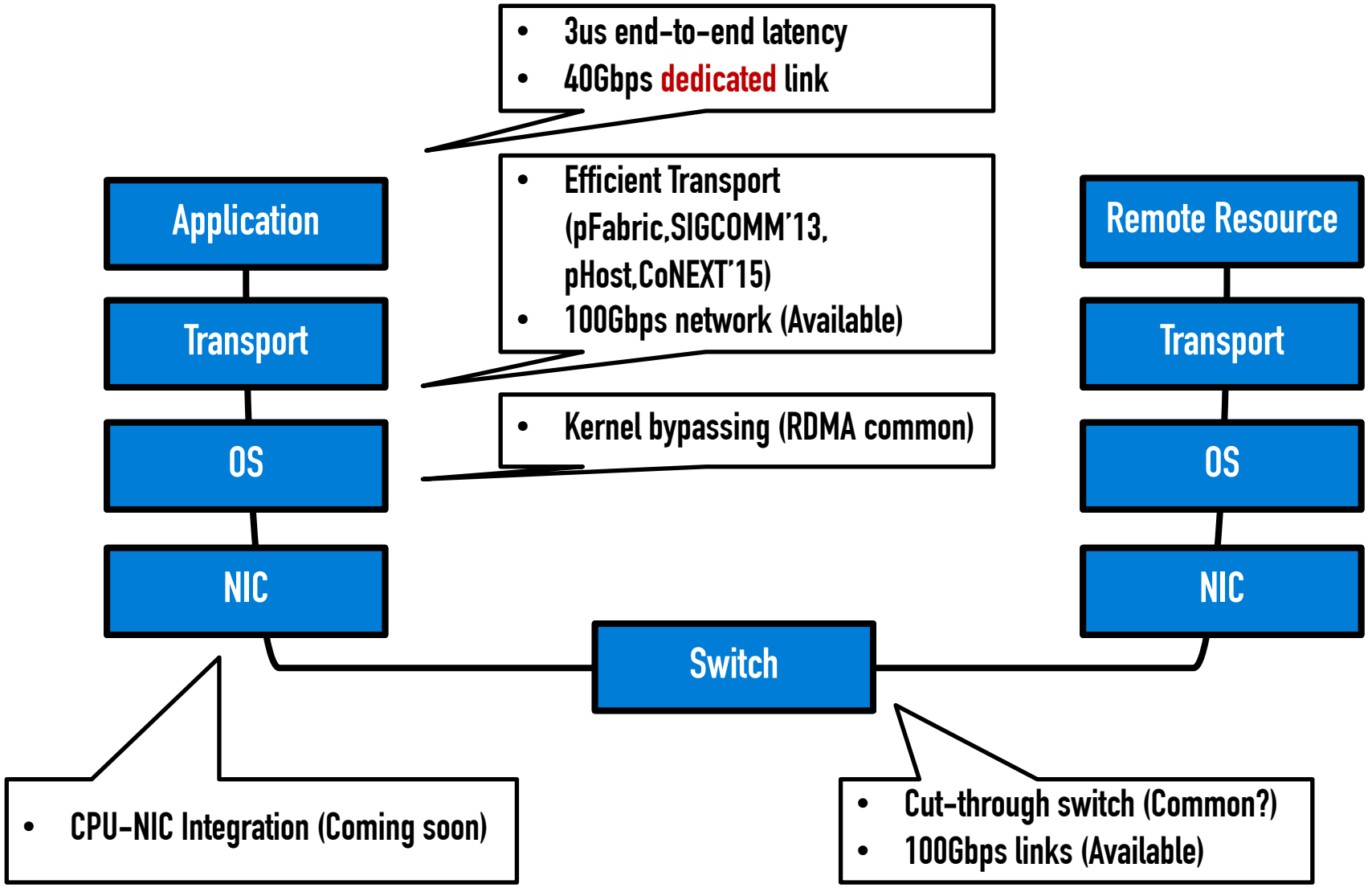
Feasibility of end-to-end latency within a rack



*Numbers estimated optimistically based on existing hardware

Feasibility of end-to-end latency within a rack

*Numbers estimated optimistically based on existing hardware


• 3us end-to-end latency
• 40Gbps dedicated link
• Efficient Transport (pFabric,SIGCOMM'13, pHost,CoNEXT'15)
• 100Gbps network (Available)
• Kernel bypassing (RDMA common)
Application
Transport
OS
NIC
Remote Resource
Transport
OS
NIC
Switch
• CPU-NIC Integration (Coming soon)
• Cut-through switch (Common?)
• 100Gbps links (Available)

What's next?

Application Design

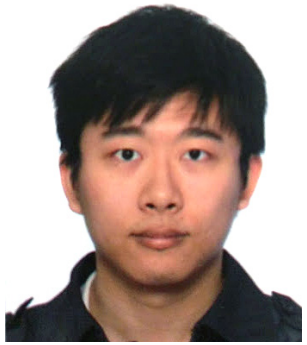
Rethinking OS Stack

- Storage
- Network Stack
- Failure Models

Network Fabric Design

Please refer our paper for evaluations on improving application performance in disaggregated datacenters

Thank You!



Peter X. Gao

Akshay Narayan

Sagar Karandikar

Joao Carreira

Sangjin Han

Rachit Agarwal

Sylvia Ratnasamy

Scott Shenker