SDC
STORAGE DEVELOPER CONFERENCE
SNIA SANTA CLARA, 2015

PCIe Non-Transparent Bridging for RDMA

Roland Dreier
Pure Storage

<roland@purestorage.com>
@rolanddreier

Modern storage pushes interconnects

- Resilience and scale require multiple controllers
- Flash raises performance demands

RDMA is the answer

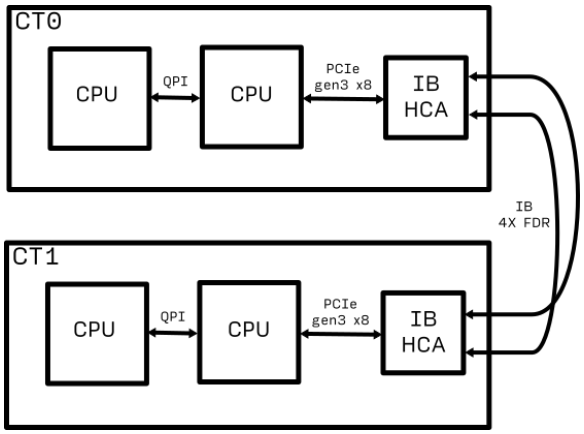
- Highest throughput, lowest latency
- Offloaded data movement frees CPU for storage services
- Kernel bypass avoids context switch overhead

InfiniBand checks all the boxes

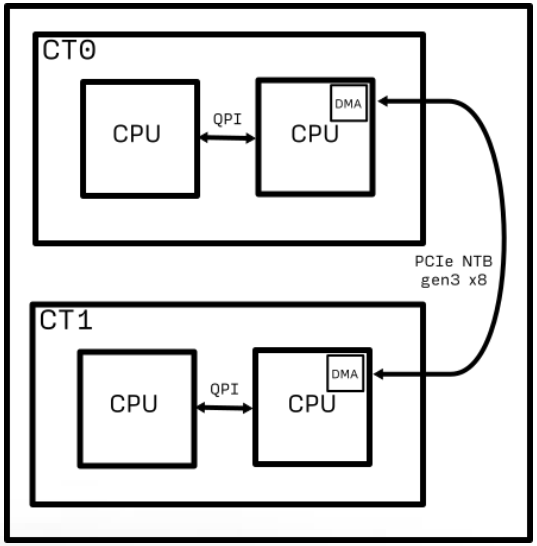
- IB is *the* standard for RDMA in HPC
- However: features and complexity required to scale to 1000s of nodes are a disadvantage in storage systems

PCIe NTB as an alternative

- Today's CPUs integrate PCIe
- Non-transparent bridging allows direct links between independent CPUs



vs.



PCIe NTB as an alternative

- Today's CPUs integrate DMA engines
- Offload data movement to PCI

PCIe NTB as an alternative

- Today's CPUs integrate IO virtualization (IOMMU)
- Linux vfio and similar allow kernel bypass

PCIe NTB as an alternative

- High throughput, low latency
- Offloaded data movement
- Kernel bypass

Experiences with PCIe NTB

- Performance can match and exceed IB
- Not widely used yet, may need to work around HW errata
- No RDMA stack available yet, need to write register-level code

Thanks!

Questions?