

S9709

Dynamic Sharing of GPUs and IO in a PCIe Network

Håkon Kvale Stensland

Senior Research Scientist / Associate Professor
Simula Research Laboratory / University of Oslo

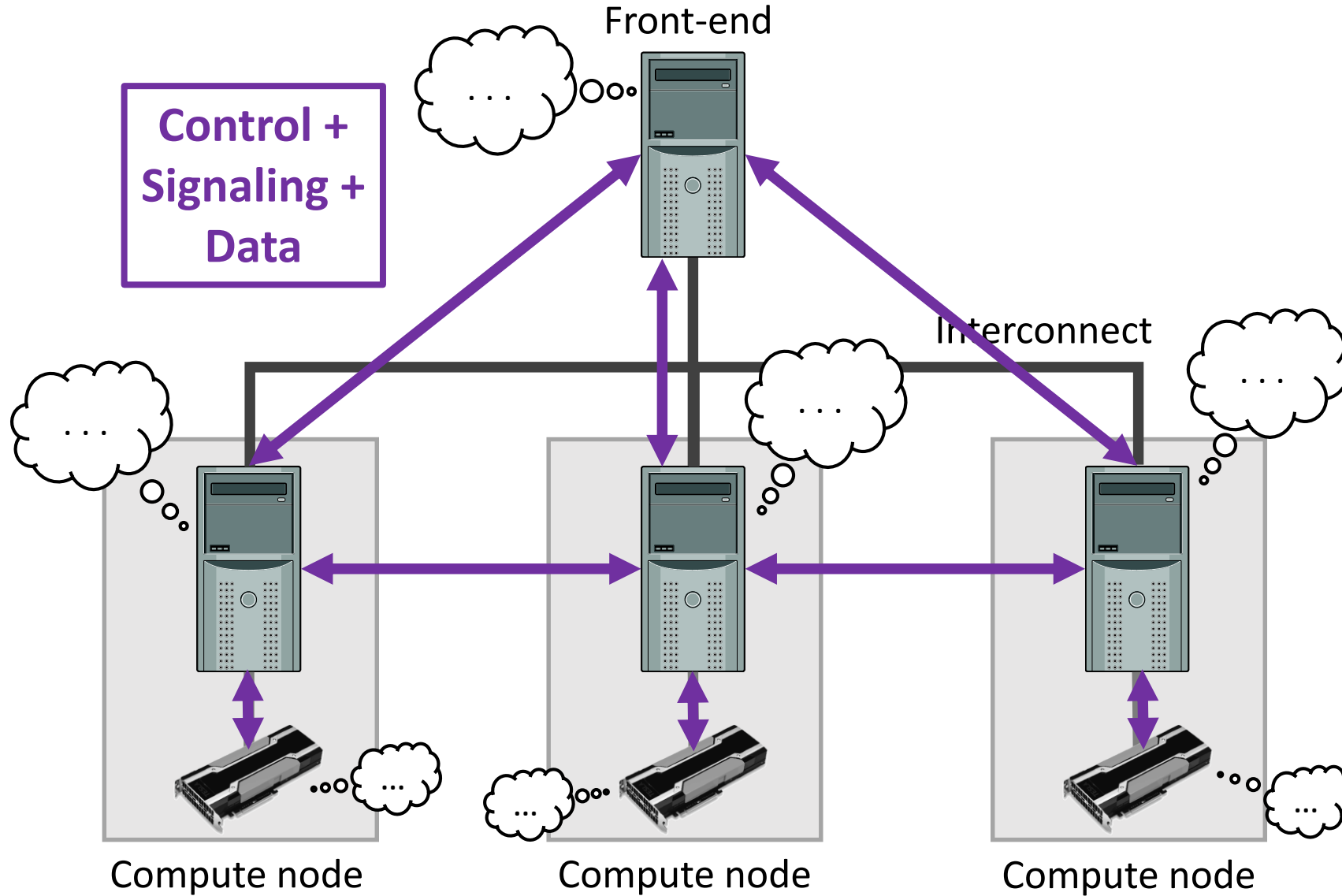


Outline

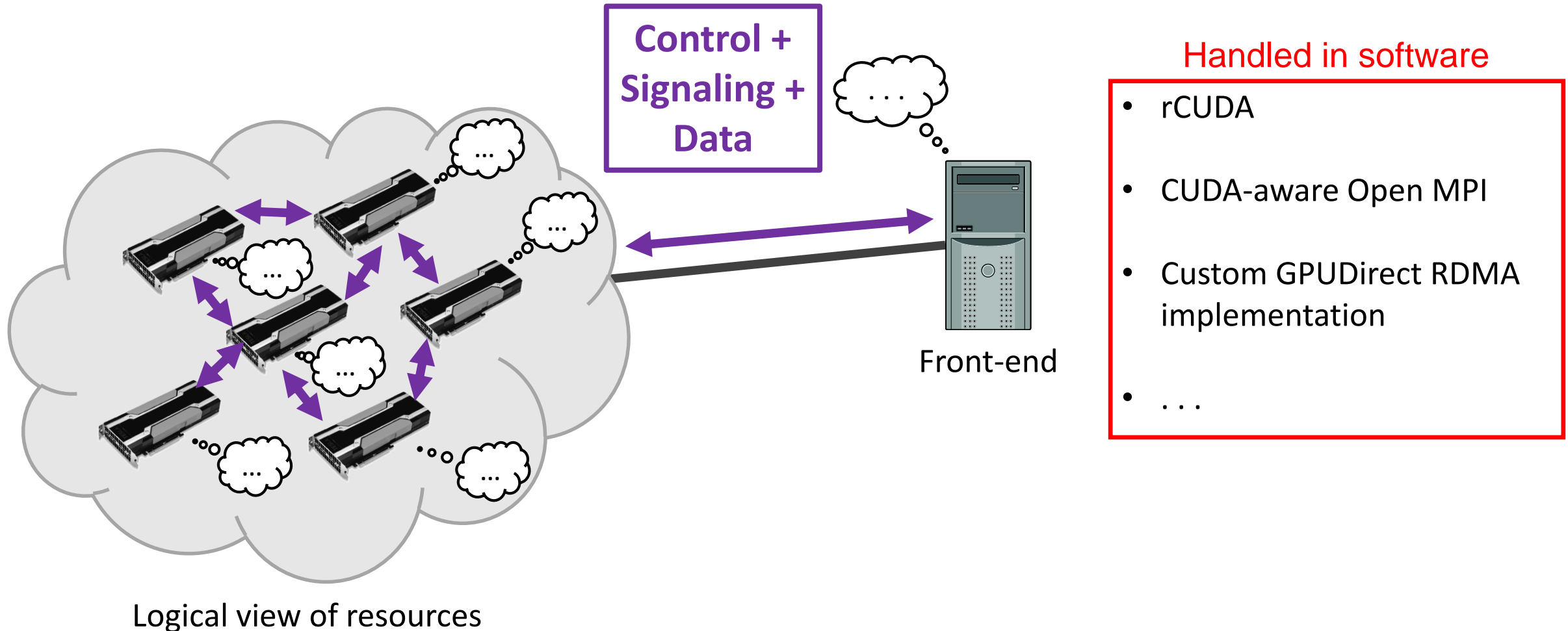
- Motivation
- PCIe Overview
- Non-Transparent Bridges
- Dolphin SmartIO
 - Example Application
 - NVMe sharing
- SmartIO in Virtual Machines



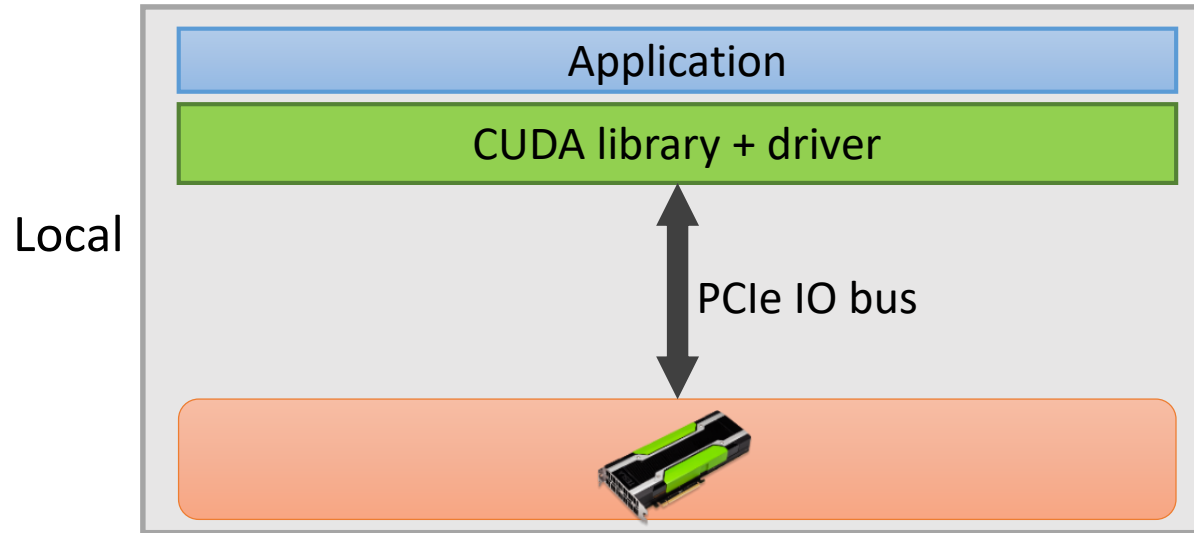
Distributed applications may need to access and use IO resources that are physically located inside remote hosts



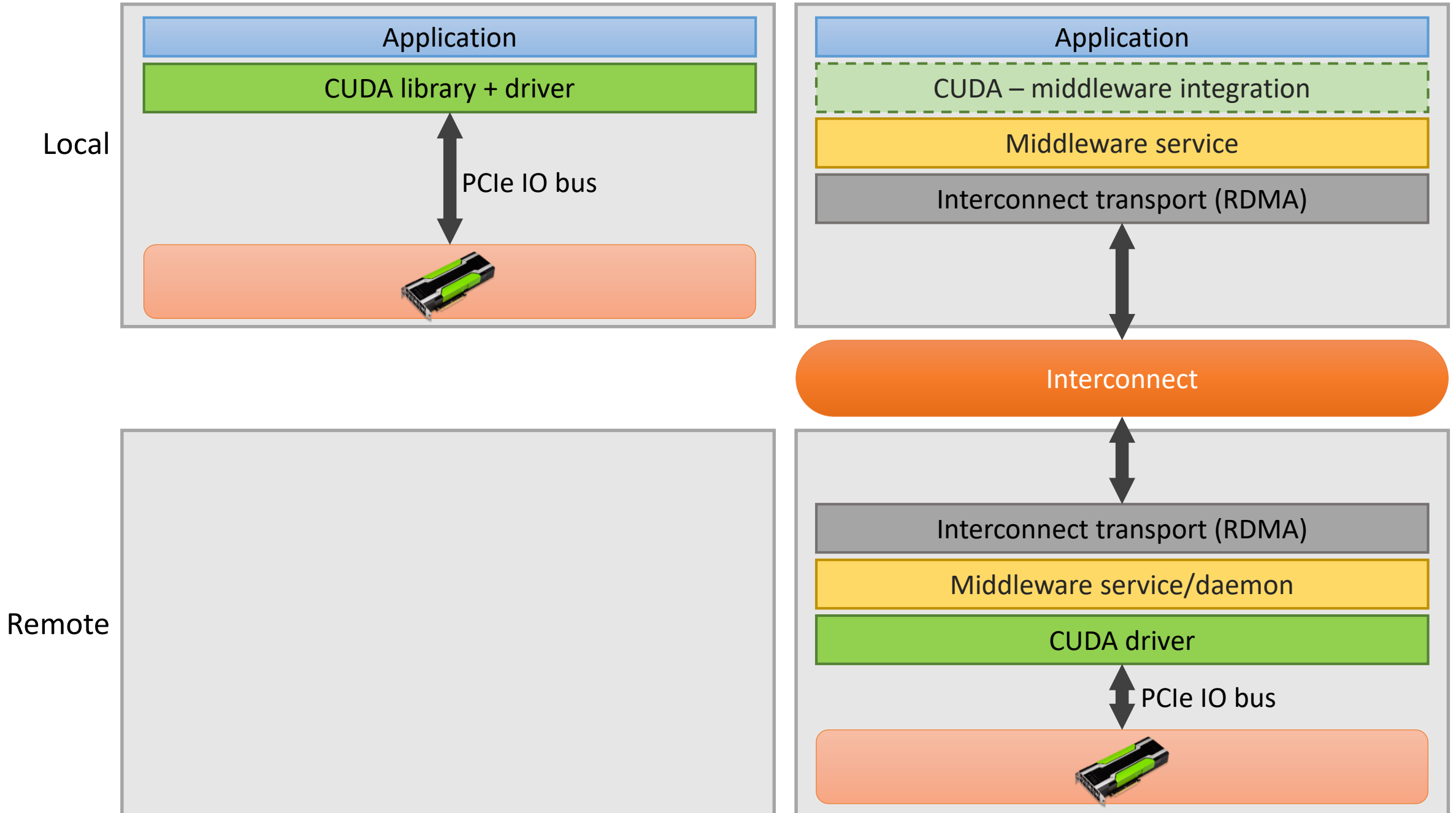
Software abstractions simplify the use and allocation of resources in a cluster and facilitate development of distributed applications



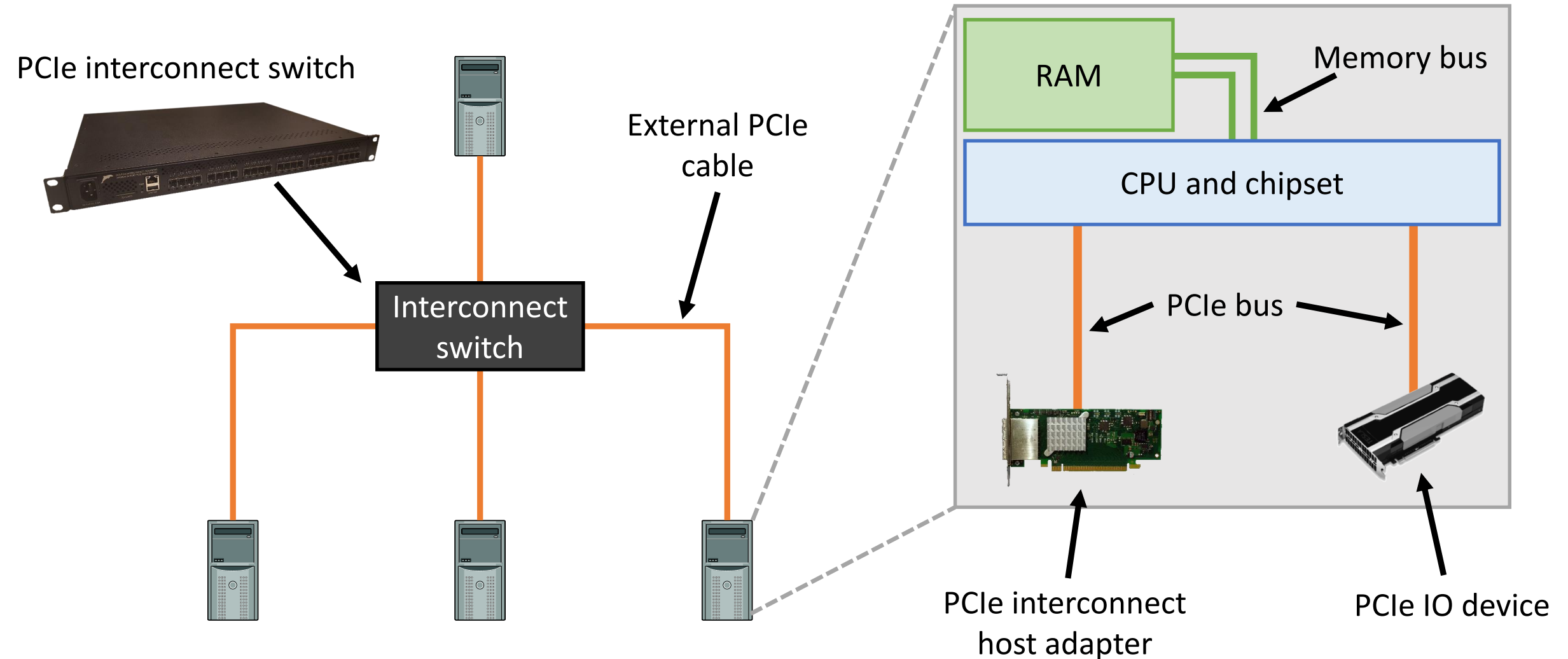
Local resource



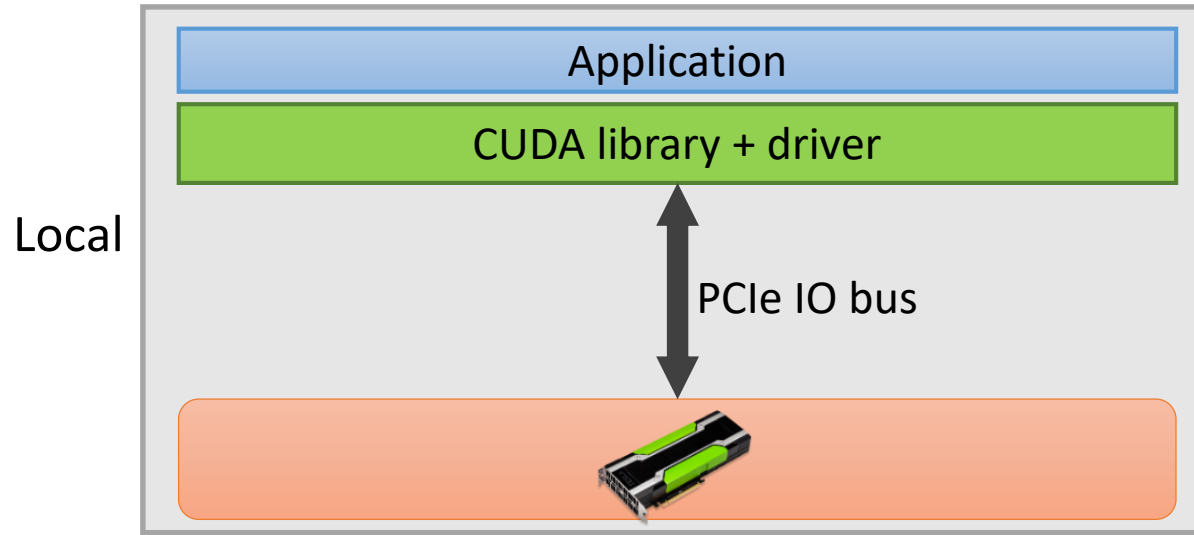
Remote resource using **middleware**



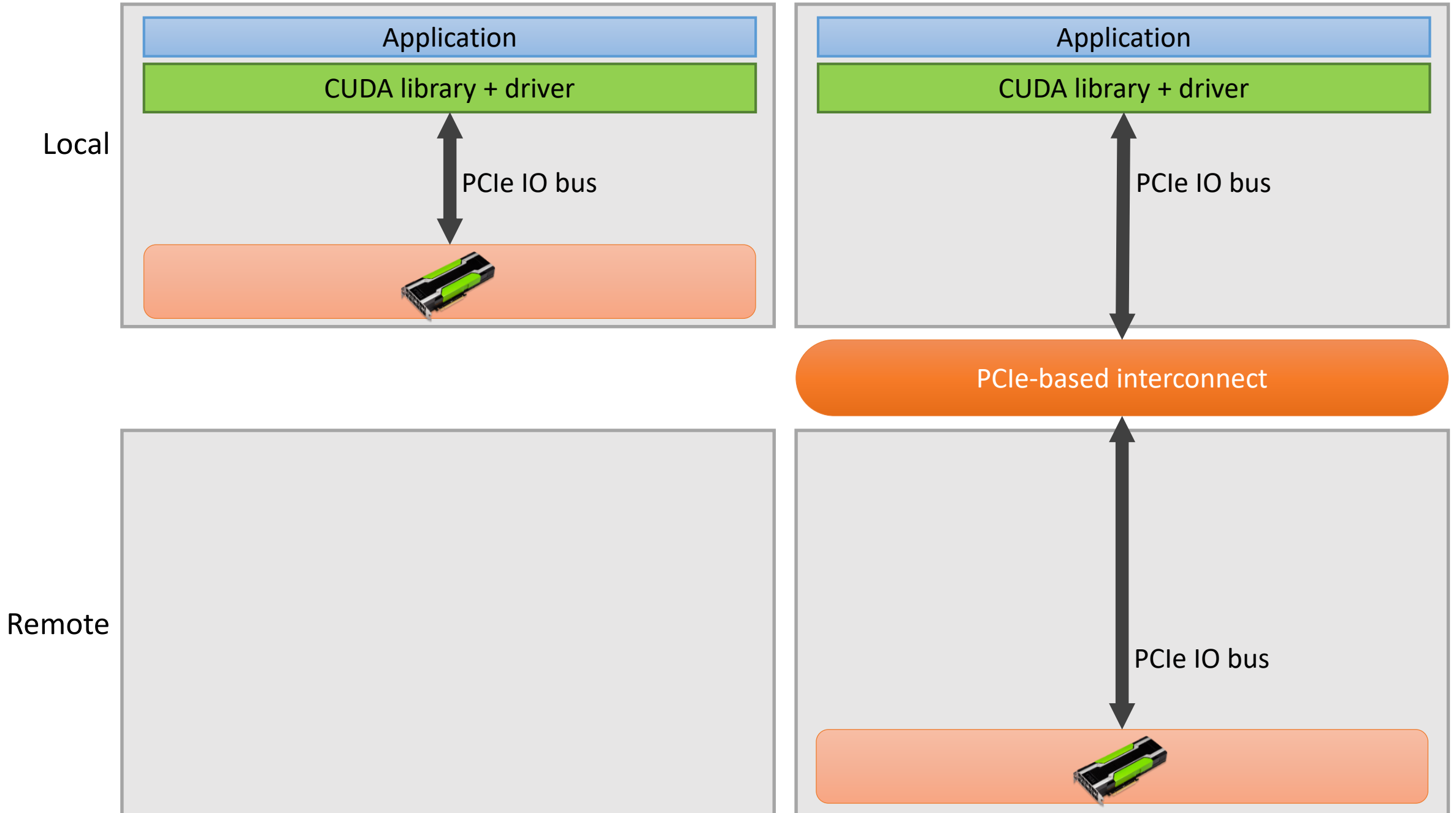
In PCIe clusters, the same fabric is used both as local IO bus within a single node and as the interconnect between separate nodes



Local resource

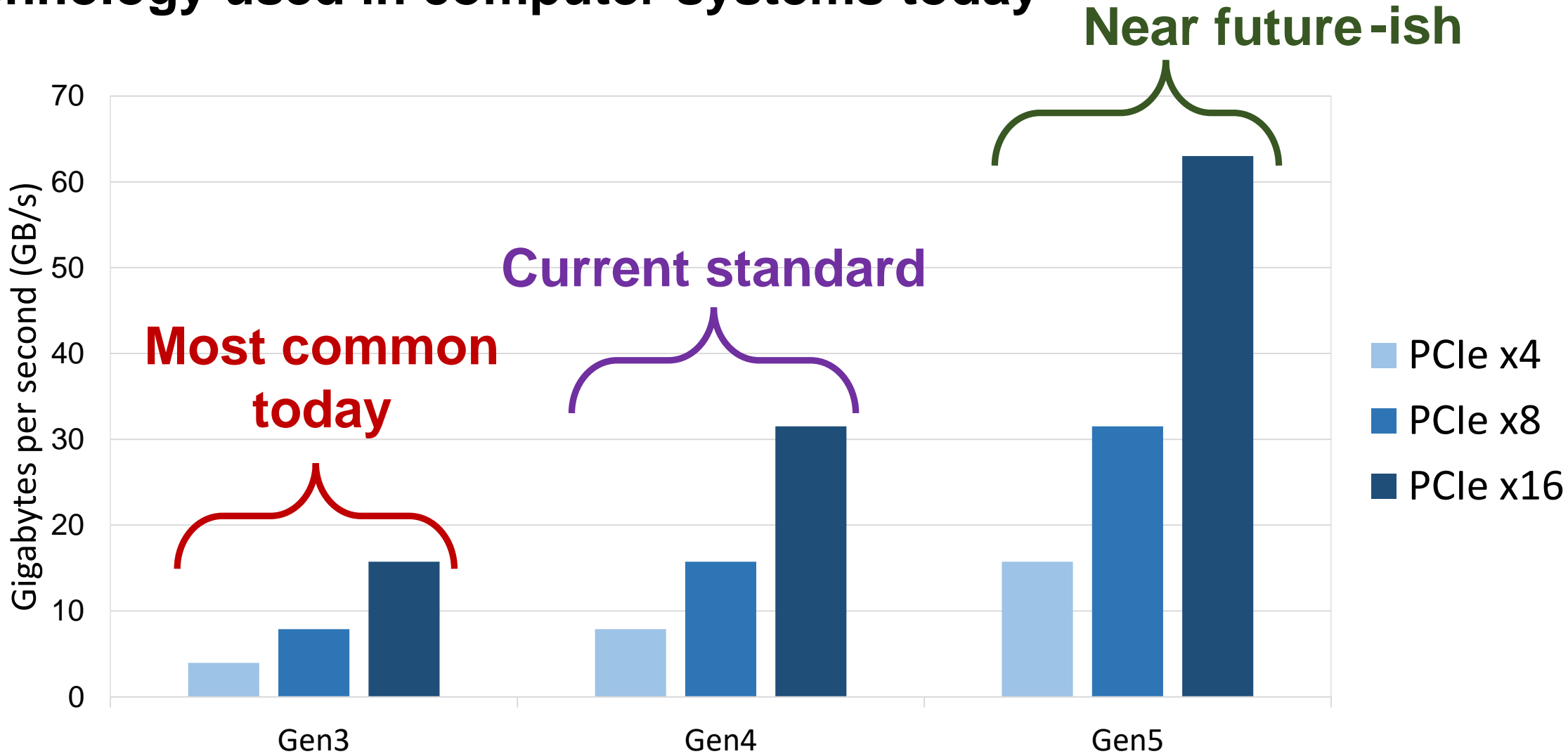


Remote resource over **native fabric**

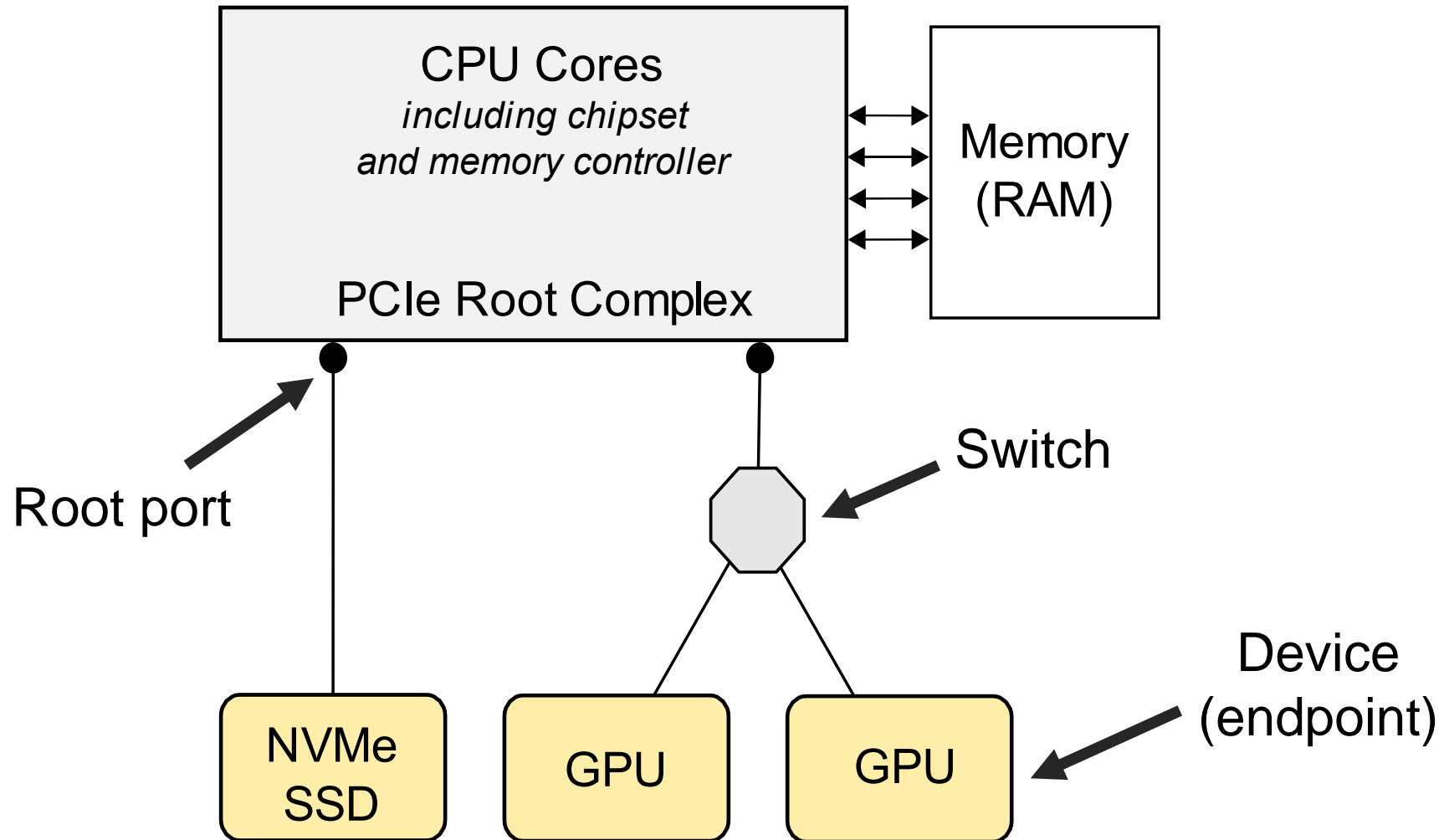


PCIe Overview

PCI Express (PCIe) is the most widely adopted I/O interconnection technology used in computer systems today

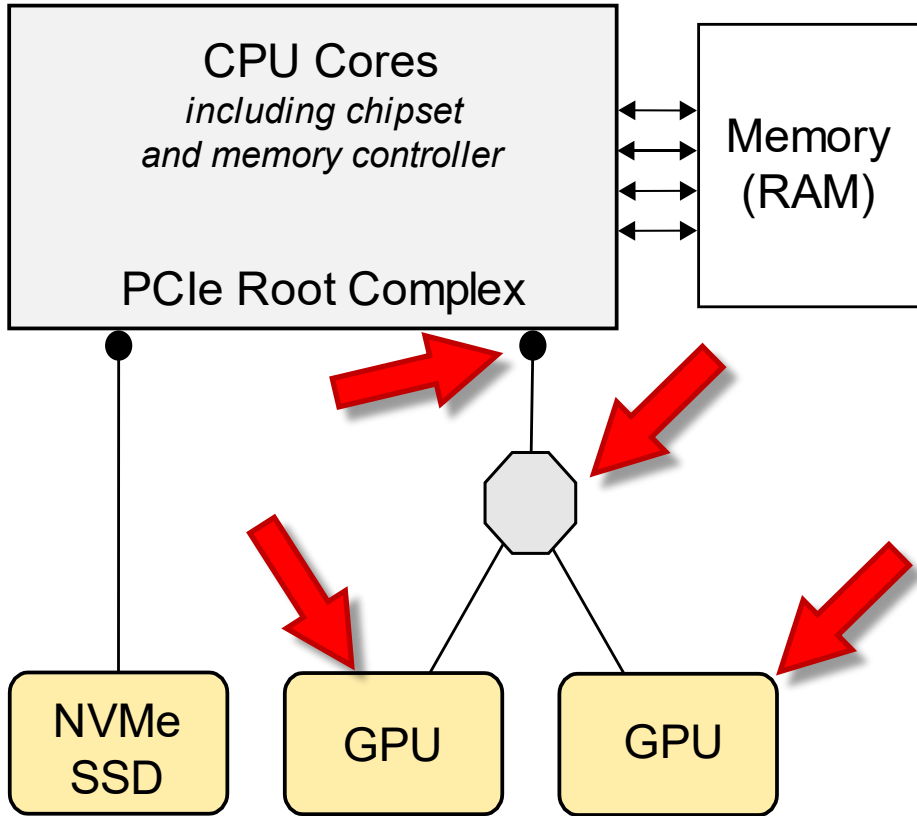




The PCIe fabric is structured as a tree, where devices form the leaf nodes (endpoints) and the CPU is on top of the root



The PCIe fabric is structured as a tree, where devices form the leaf nodes (endpoints) and the CPU is on top of the root

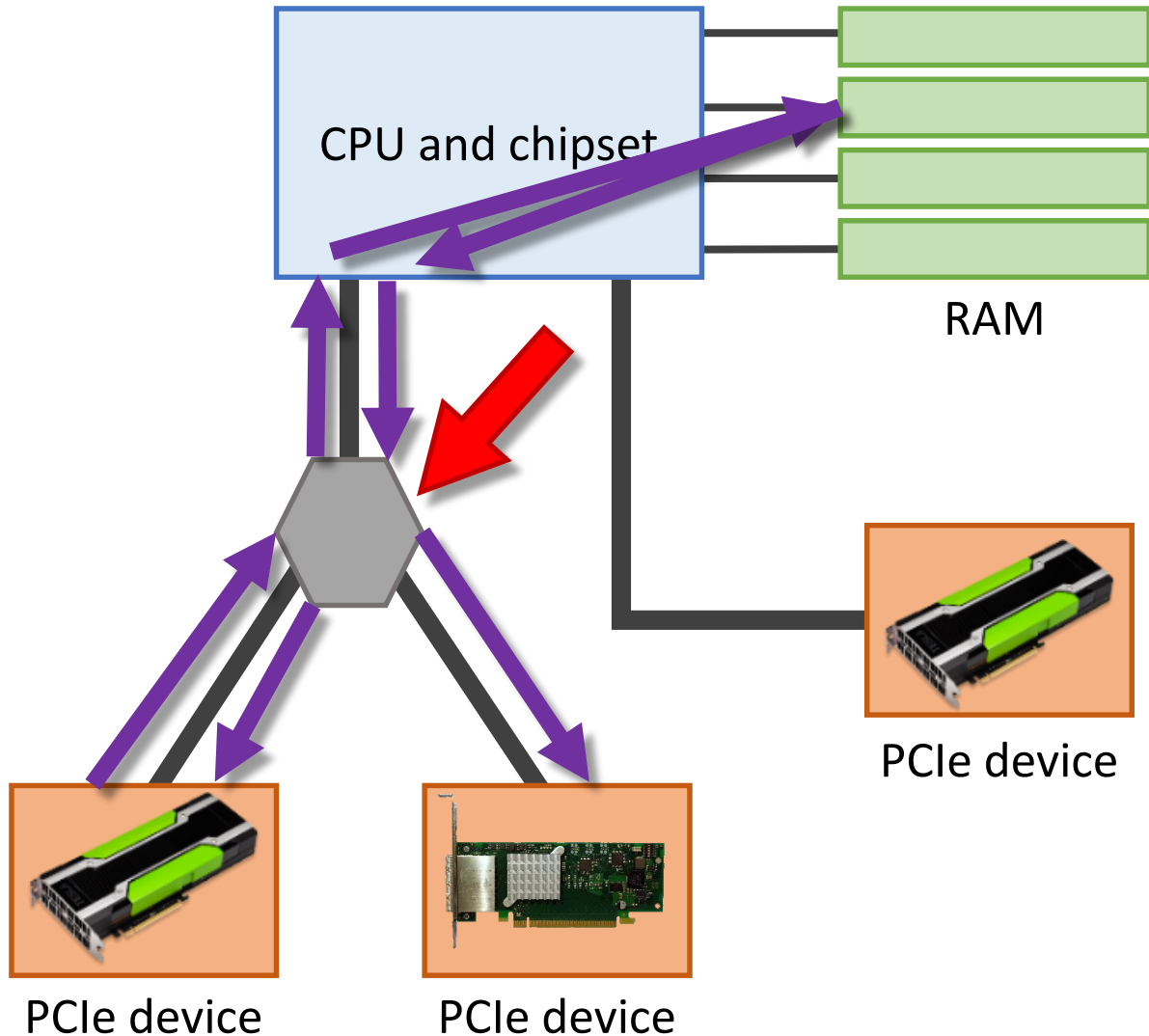
```
$ lspci -tv
```



● Root port — Link  Switch  Endpoint

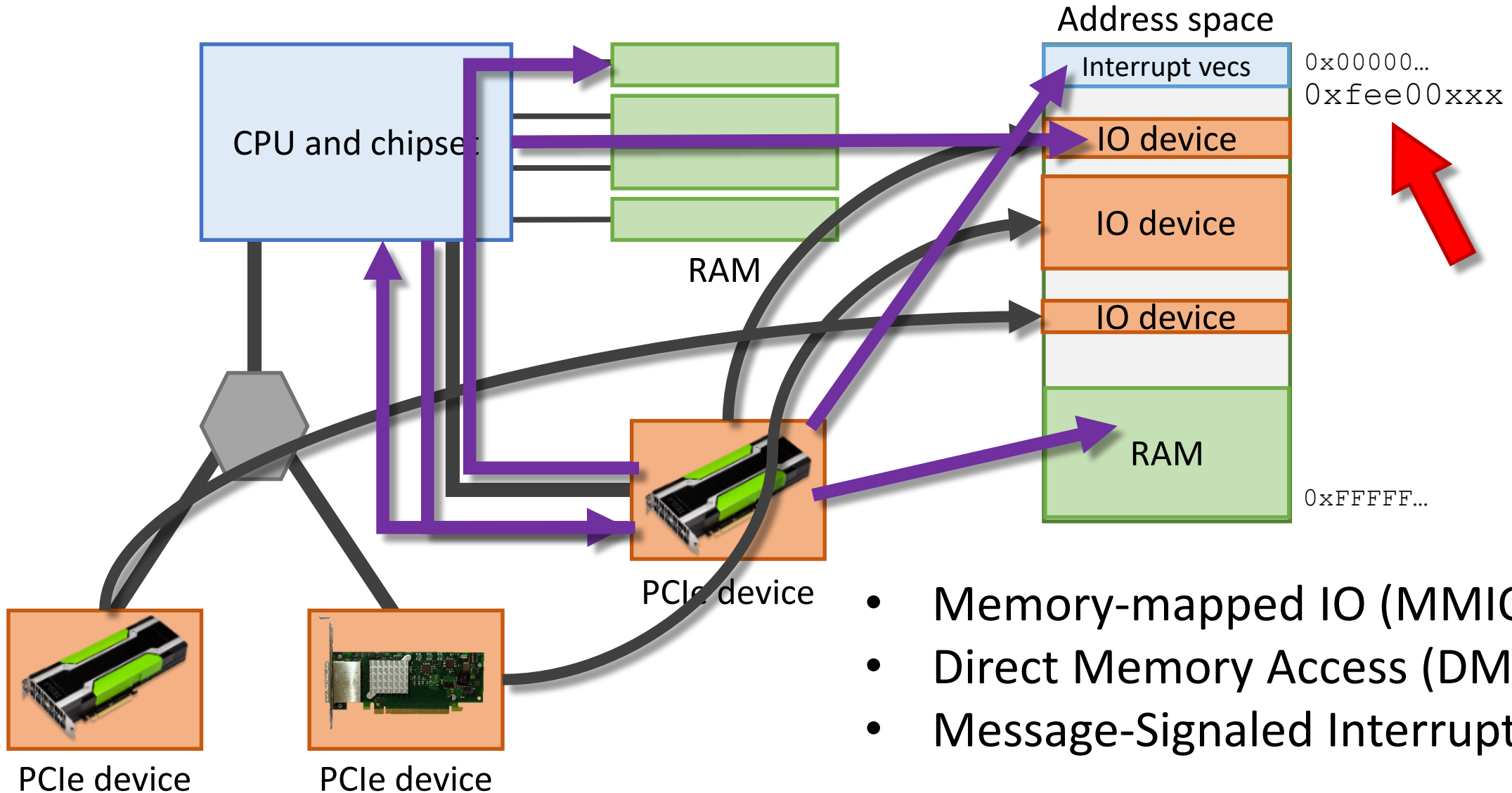
[illegible]

Memory reads and writes are handled by PCIe as transactions that are packet-switched through the fabric depending on the address



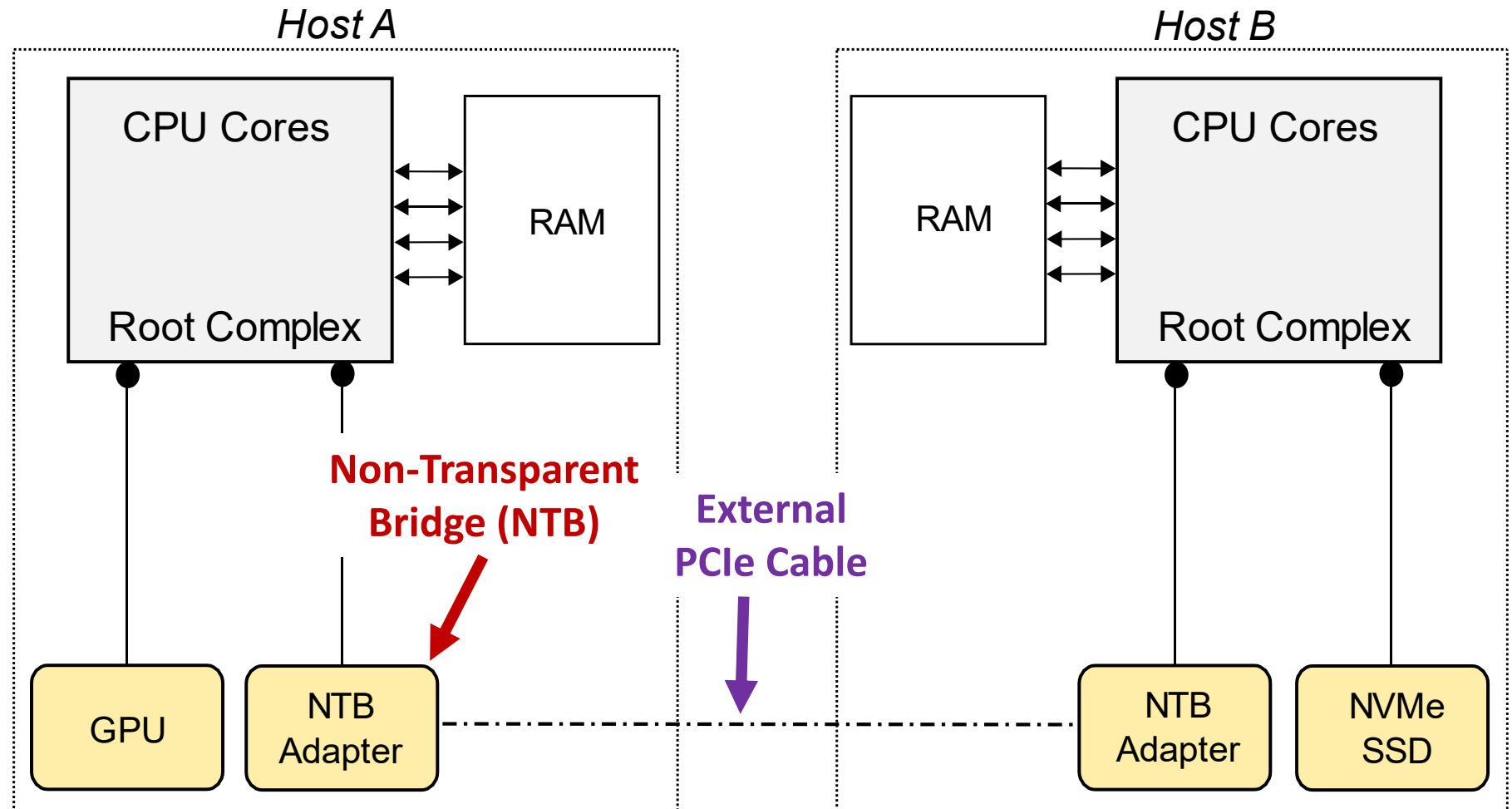
- Upstream
- Downstream
- Peer-to-peer (shortest path)

IO devices and the CPU share the same physical address space, allowing devices to access system memory and other devices

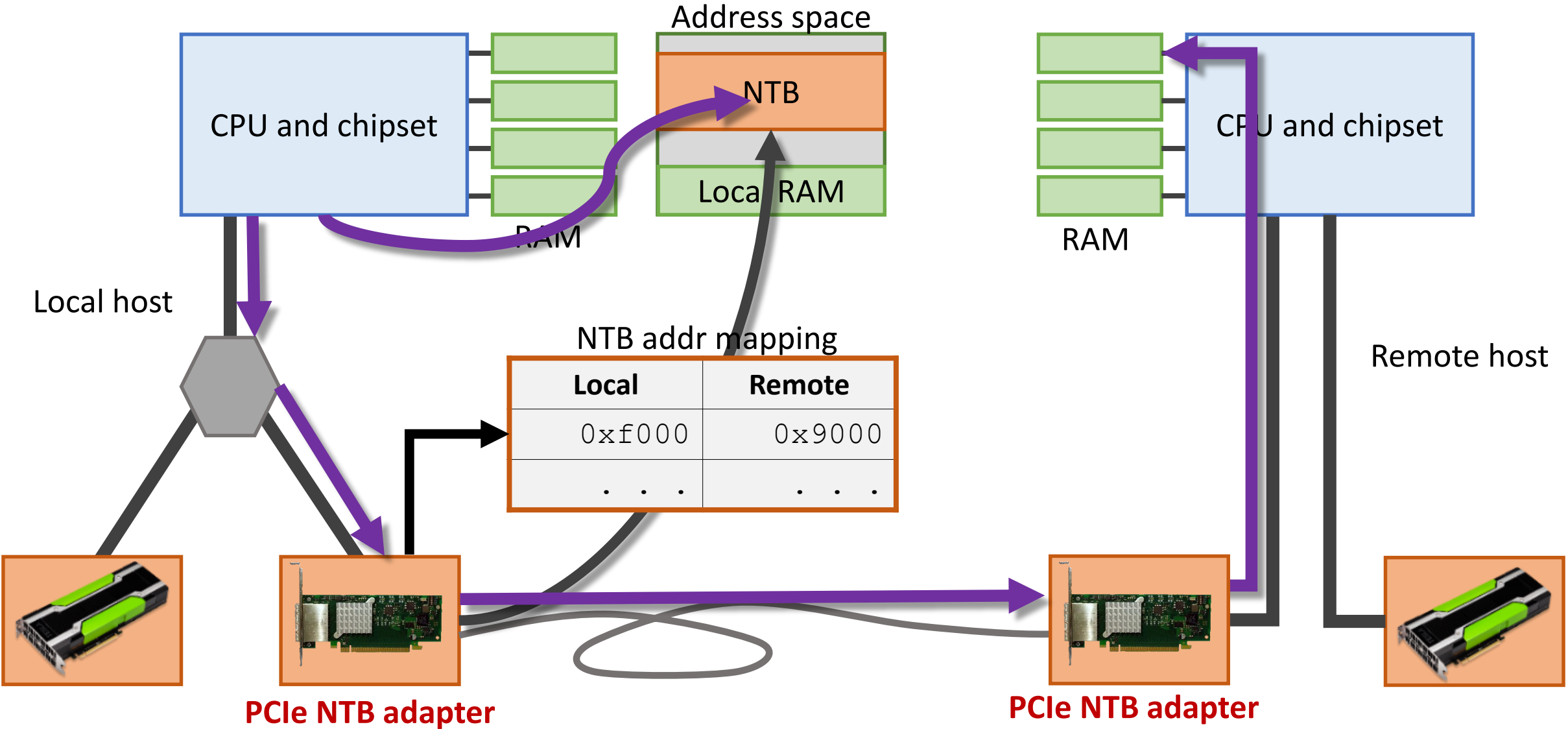


Non-Transparent Bridges

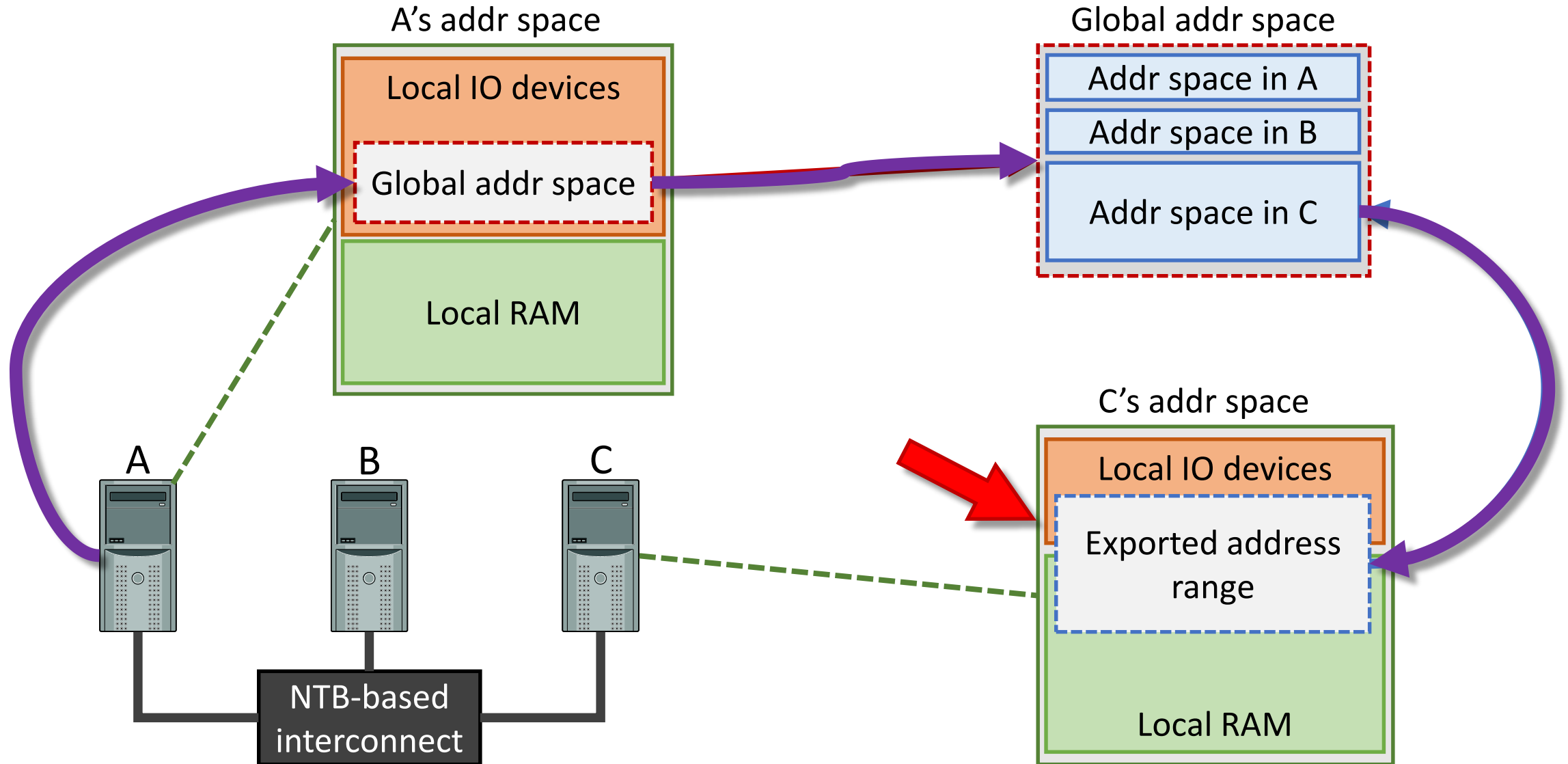
We can interconnect separate PCIe root complexes and translate addresses between them using a non-transparent bridge (NTB)



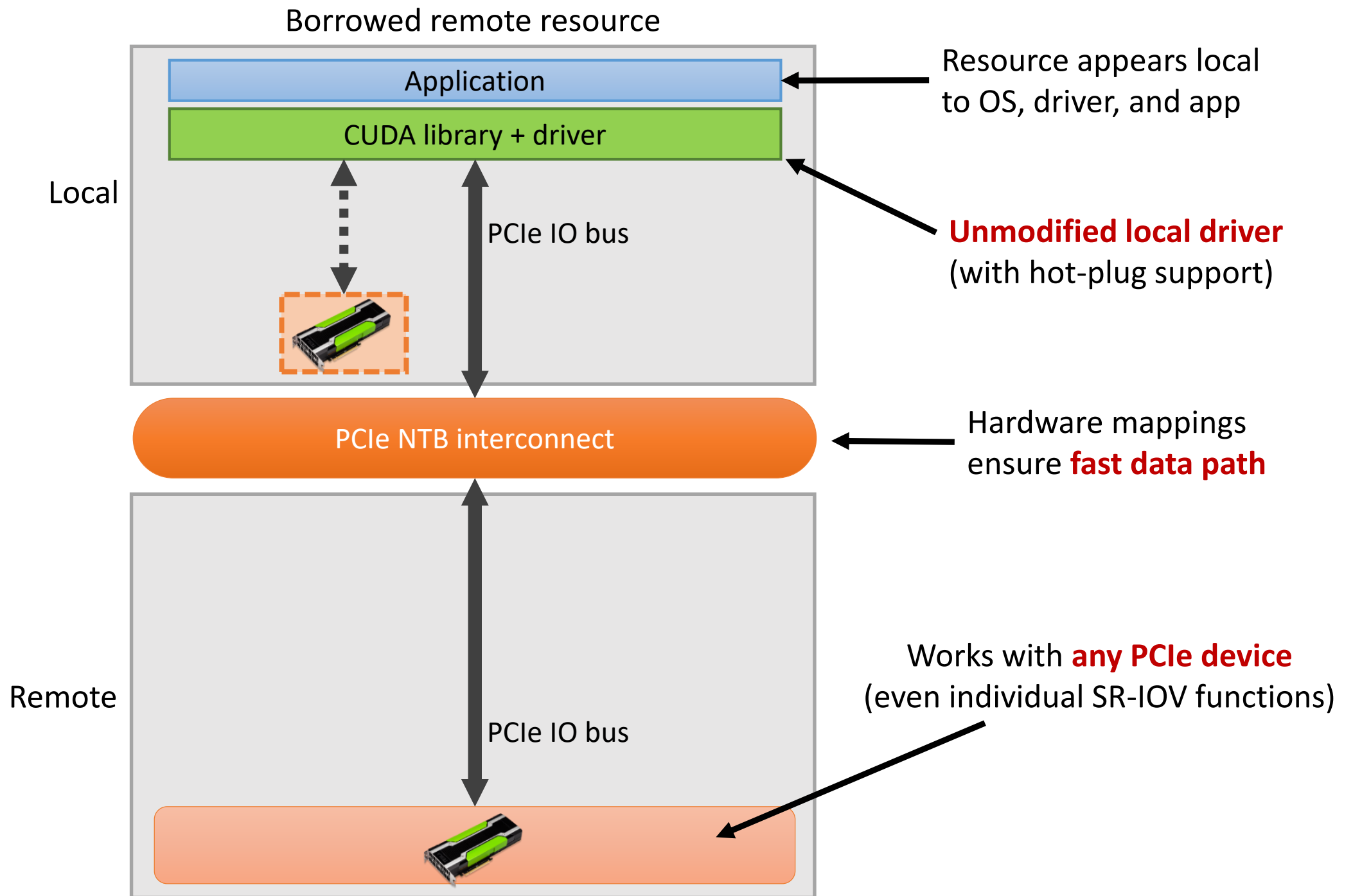
Remote address space can be mapped into local address space by using PCIe Non-Transparent Bridges (NTBs)



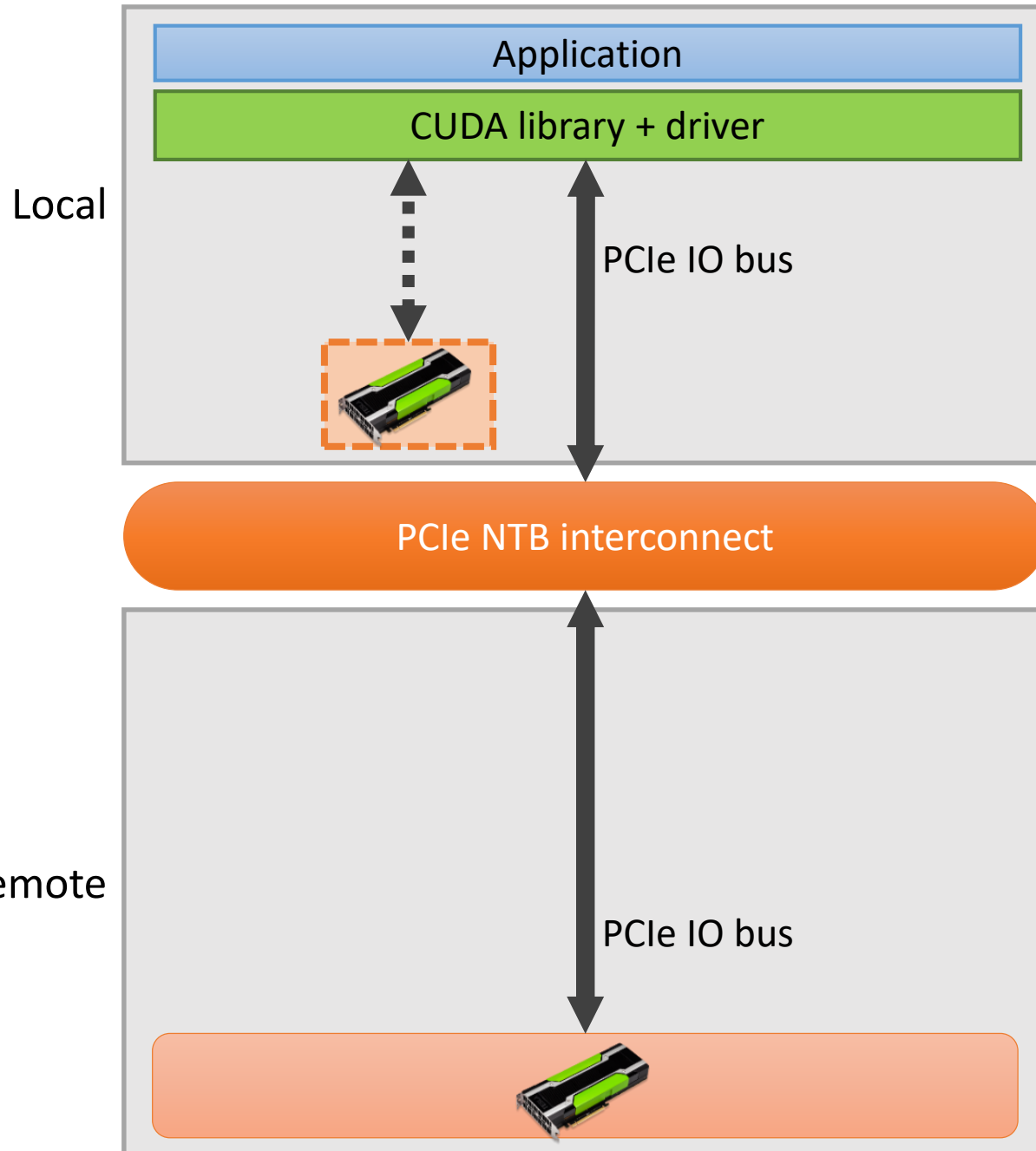
Using NTBs, each node in the cluster take part in a shared address space and have their own “window” into the global address space



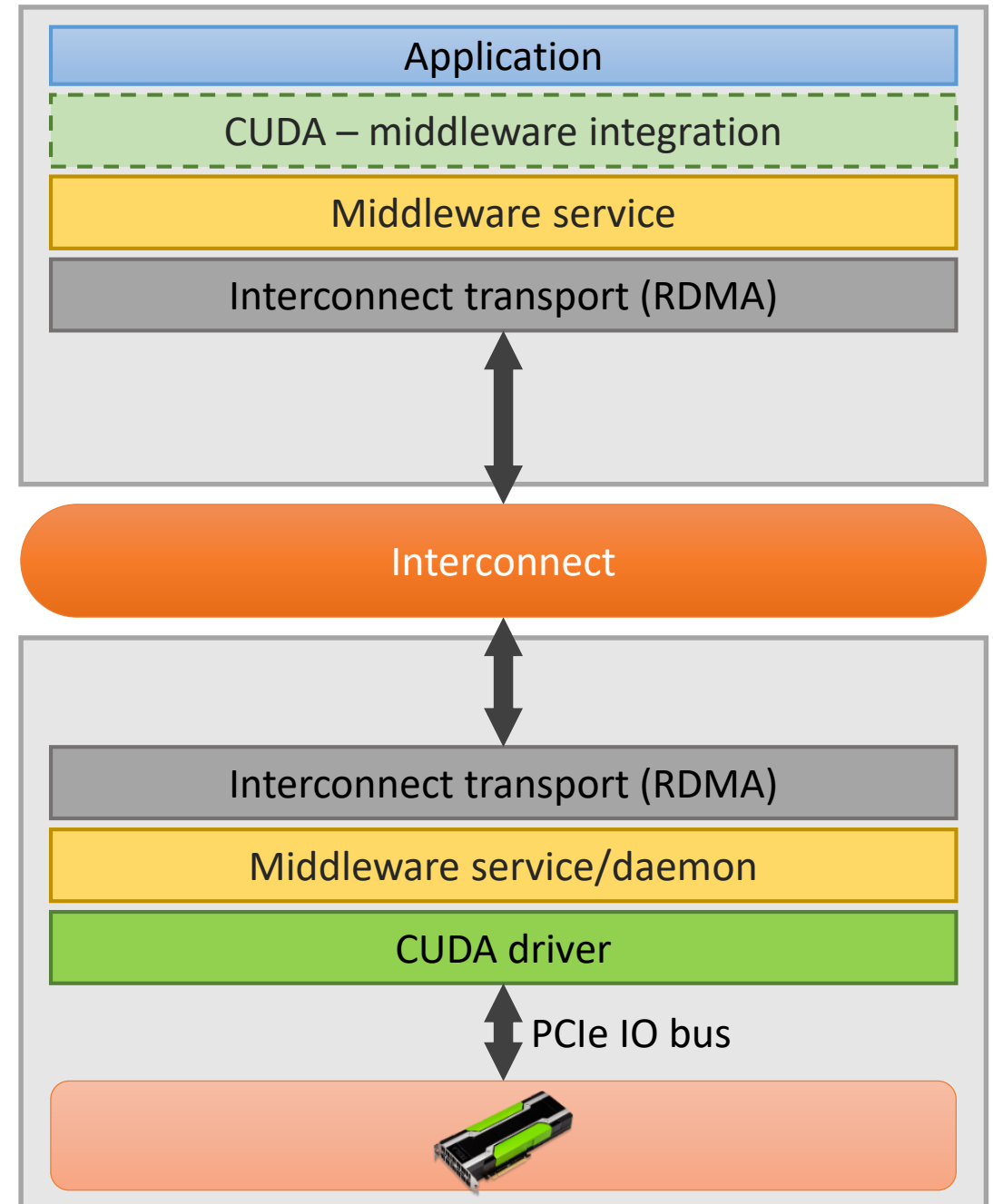
SmartIO



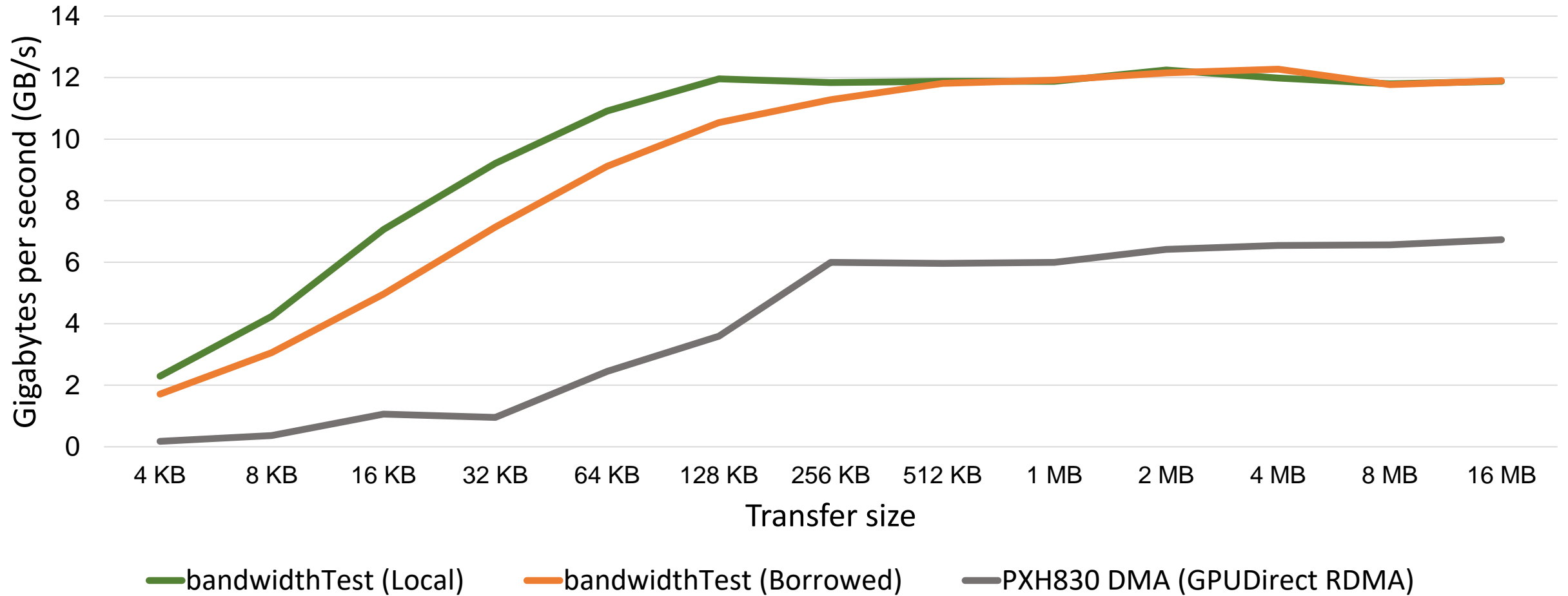
Borrowed remote resource



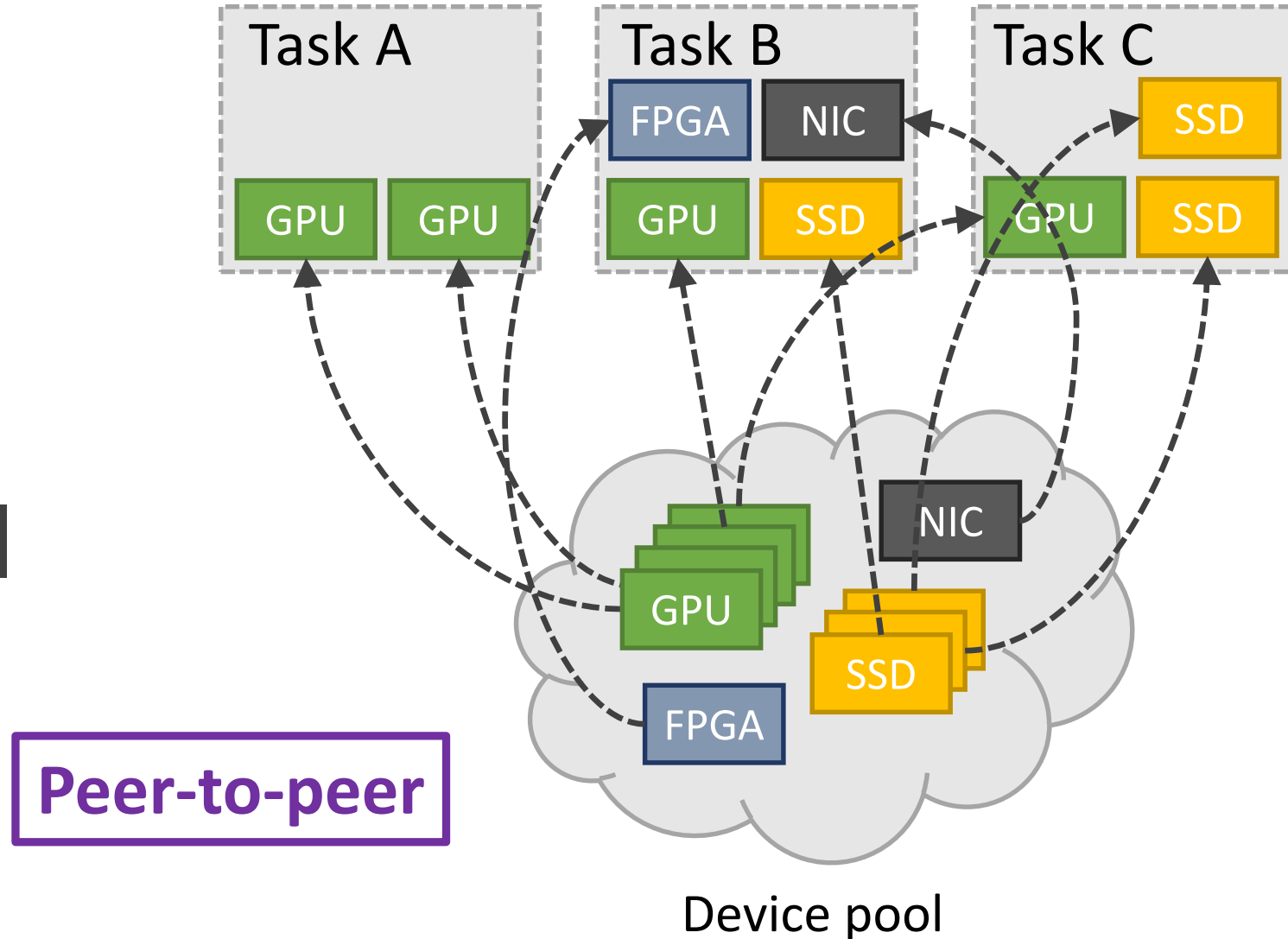
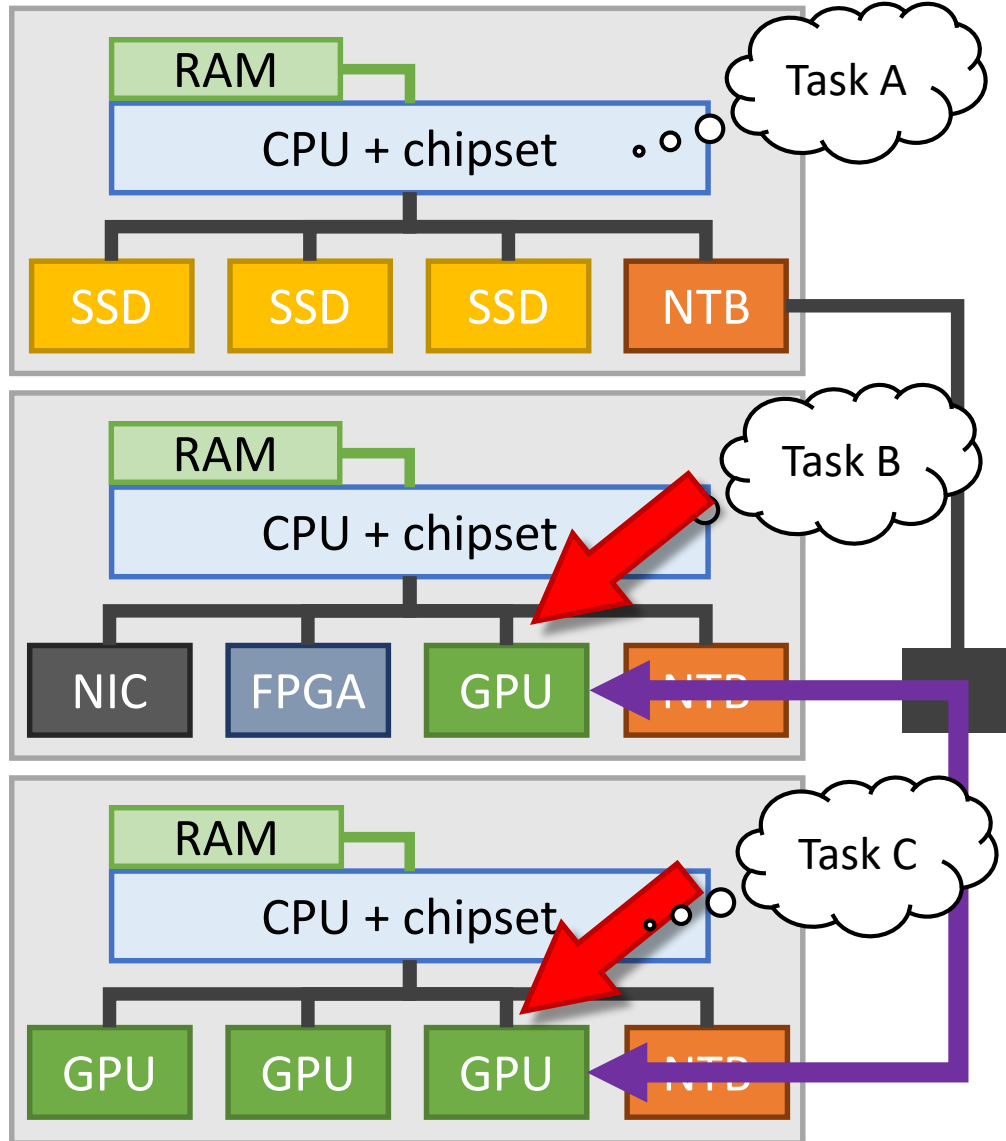
Remote resource using middleware



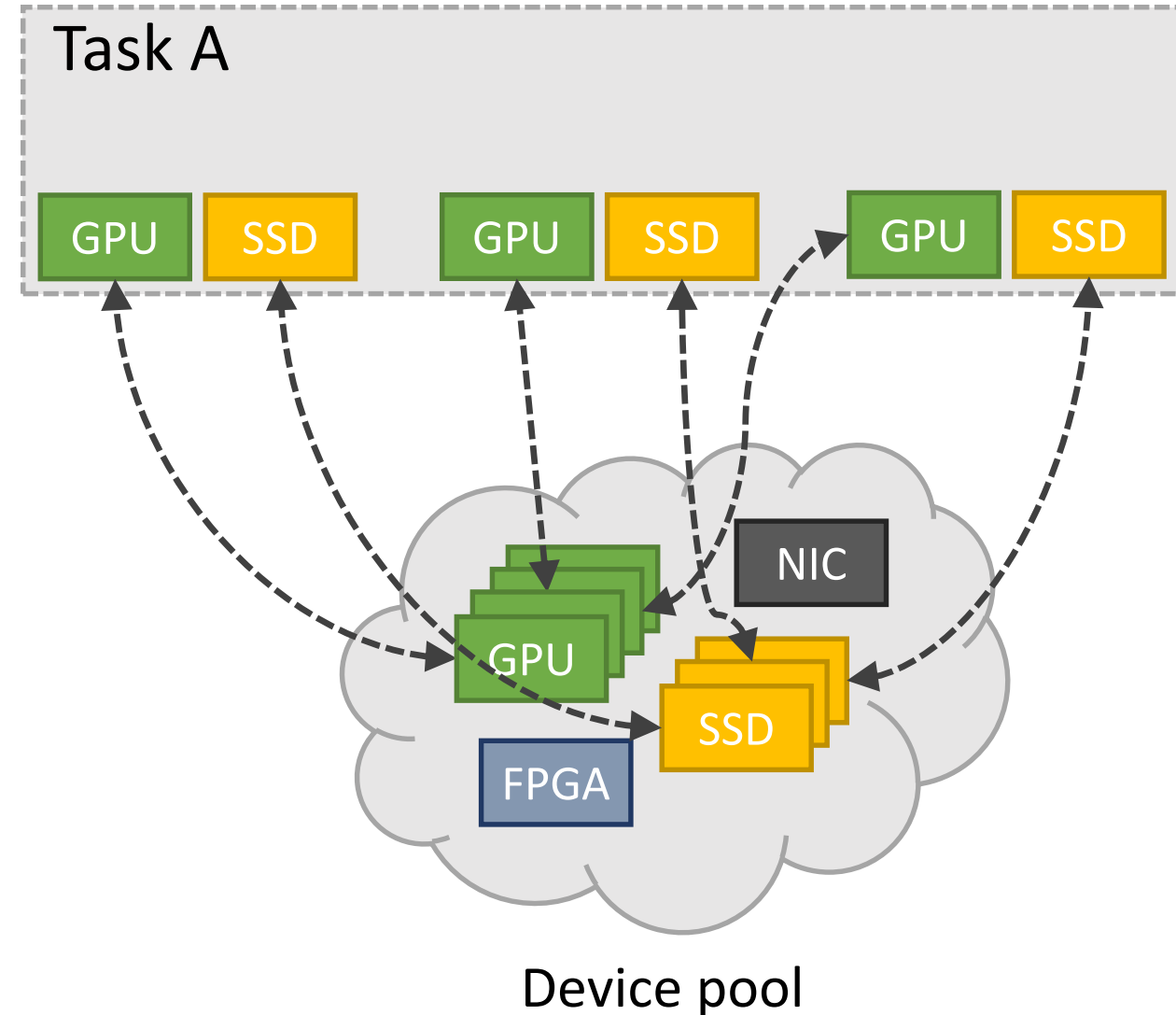
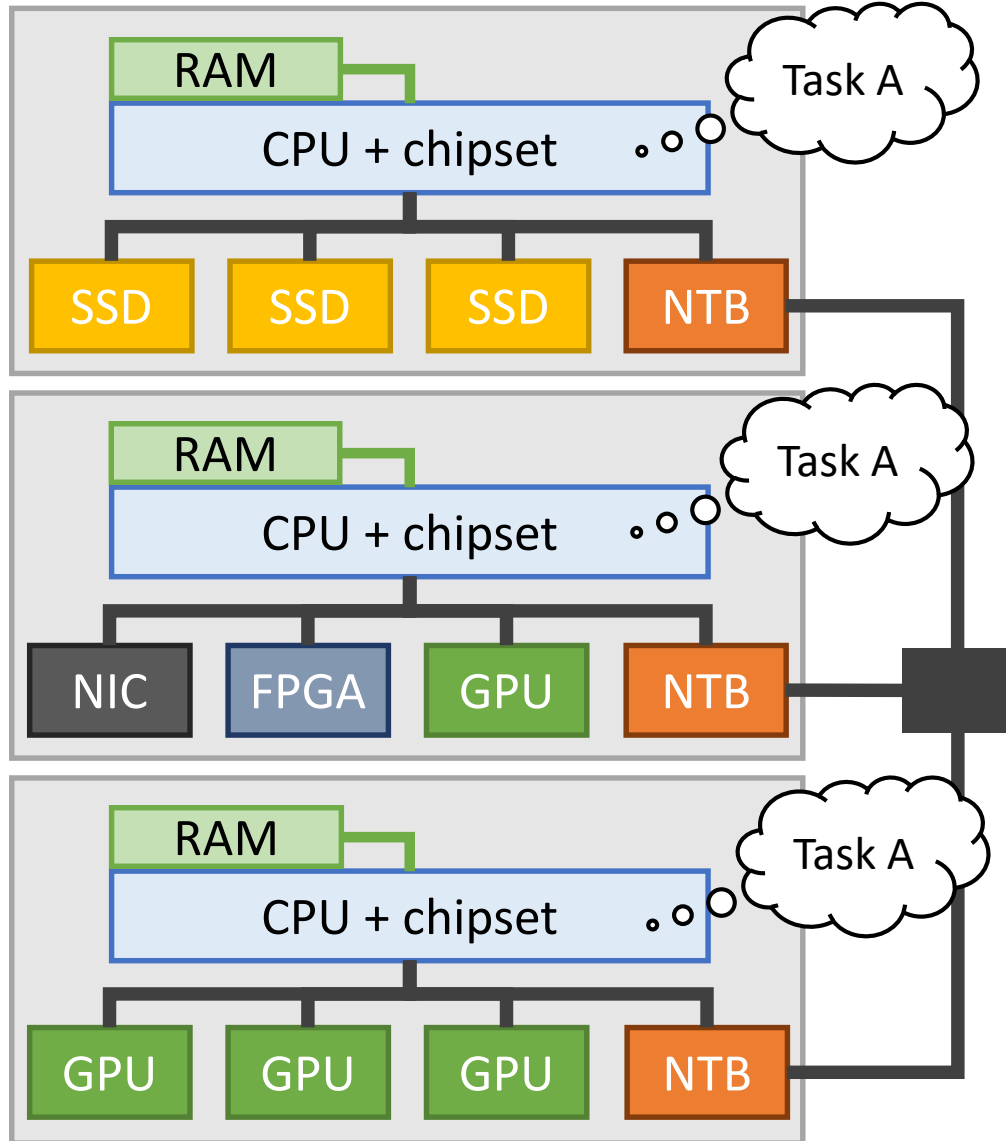
Device to host transfers: Comparing local to borrowed GPU



Using Device Lending, nodes in a PCIe cluster can share resources through a process of borrowing and giving back devices



Using Device Lending, nodes in a PCIe cluster can share resources through a process of borrowing and giving back devices

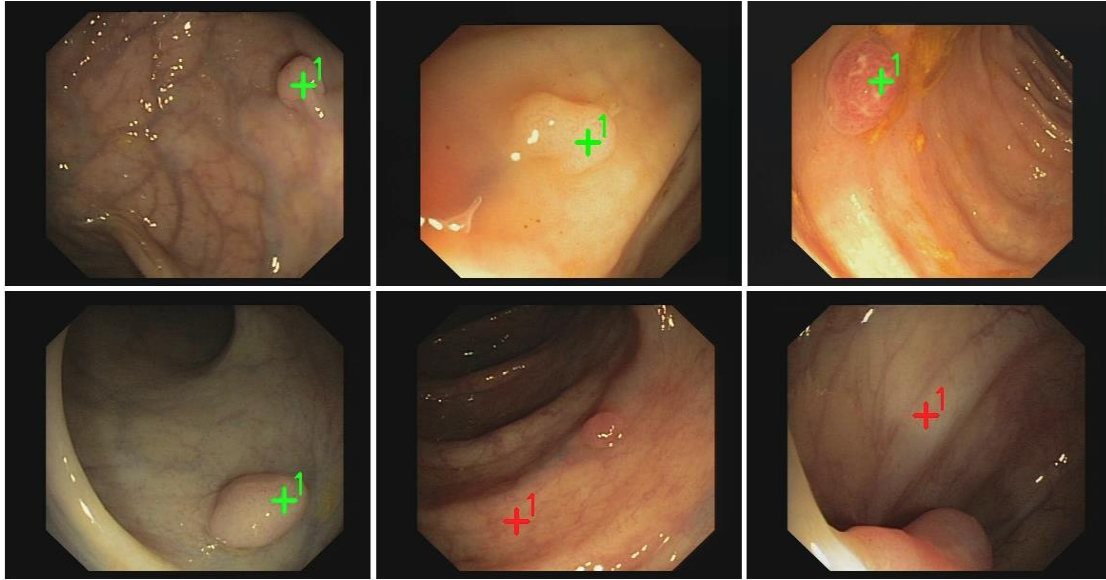


Example Application

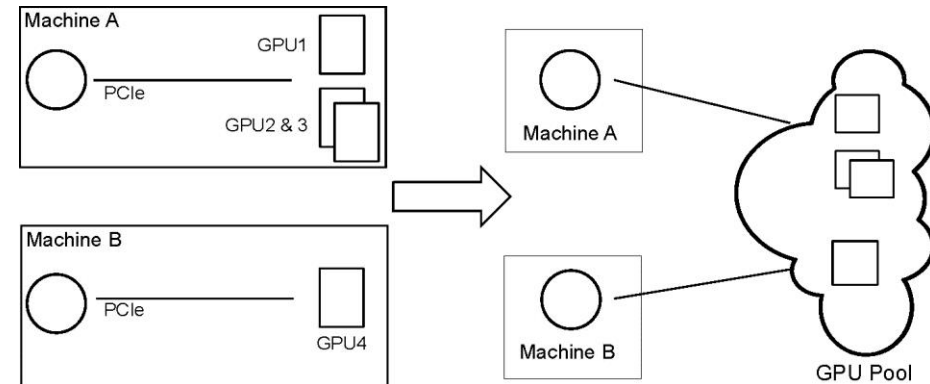
Processing of Medical Videos

P9258 - Efficient Processing of Medical Videos in a Multi-auditory Environment Using GPU Lending

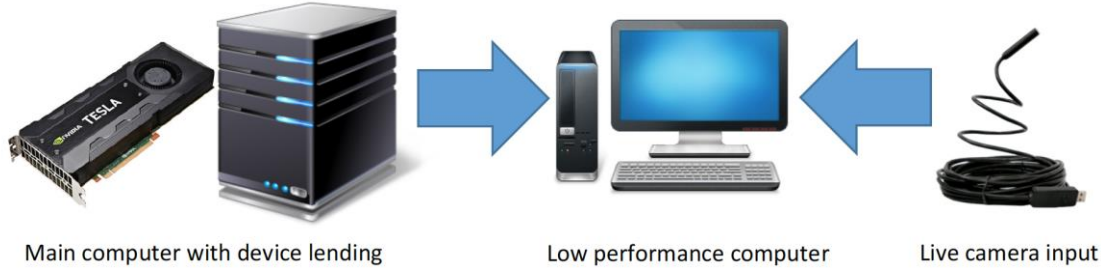
Scenario: Real-time computer-aided polyp detection



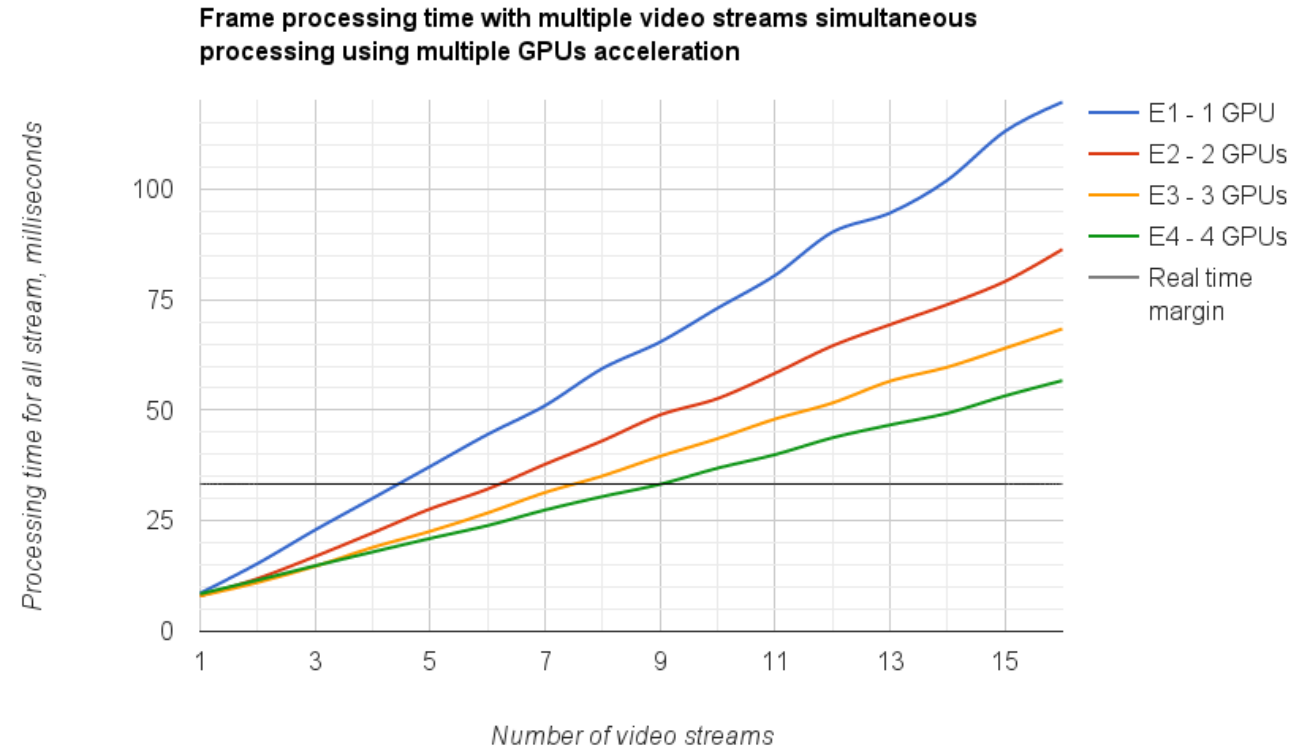
- PCIe fiber cables can be up to 100 meters.
- Enable "thin clients" to use GPUs in remote machine room



Flexible sharing of GPU resources between multiple examination rooms



- System uses a combination of classic computer vision algorithms and machine learning.
- Research prototype since 2016.



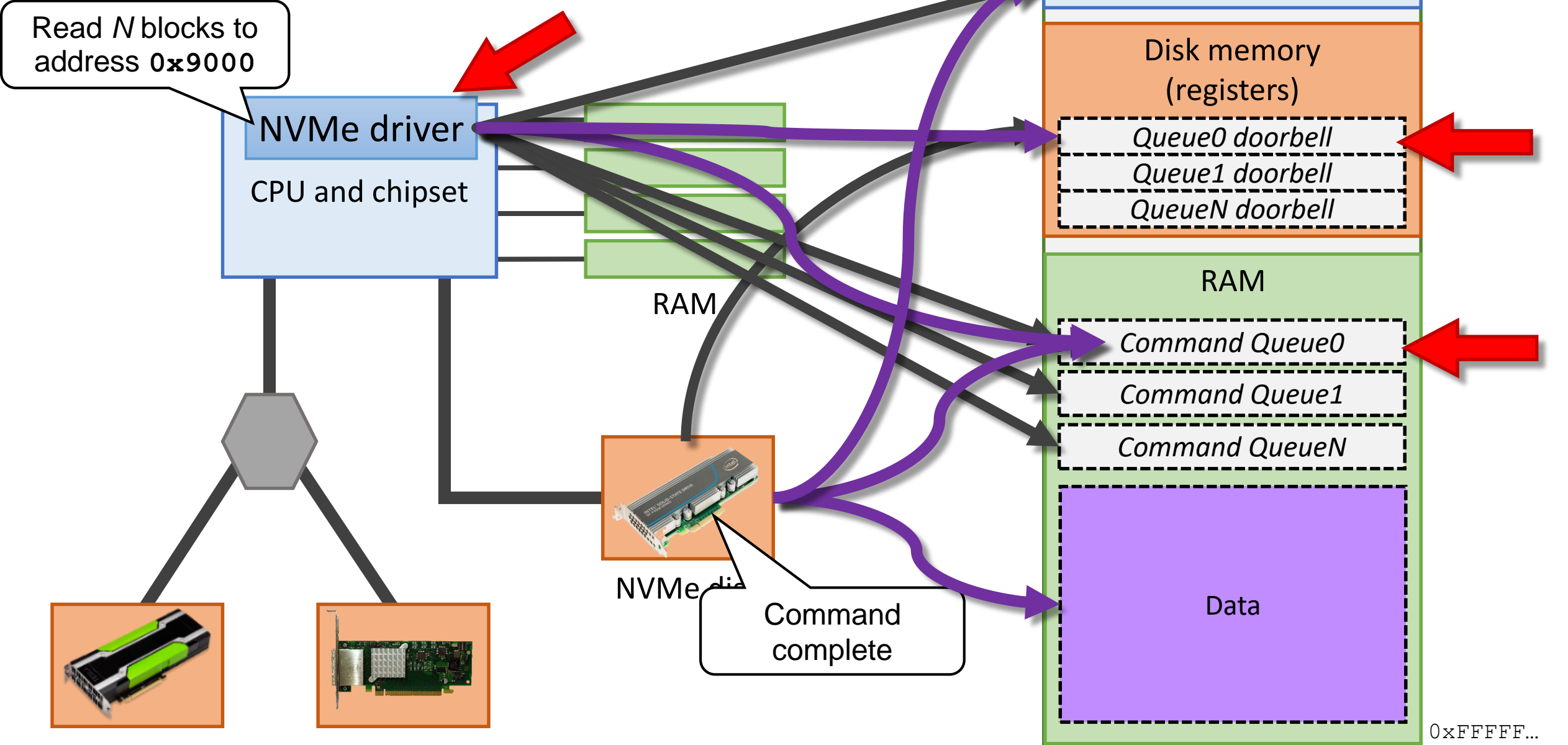
Sharing of NVMe drives

For more details:

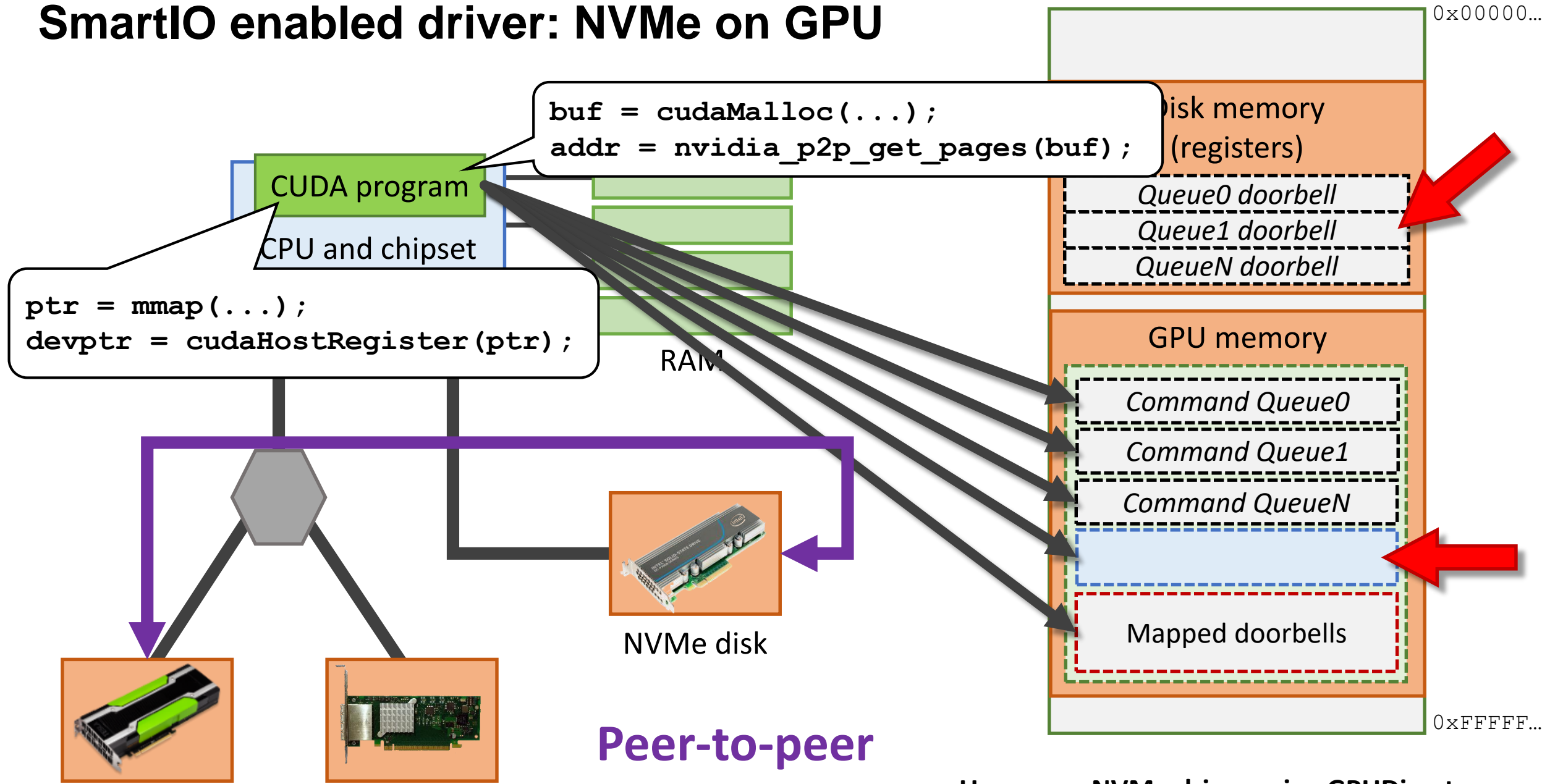
S9563 - Efficient Distributed Storage I/O using NVMe and GPU Direct in a PCIe Network
or

Visit Dolphin Interconnect Solutions in booth 1520

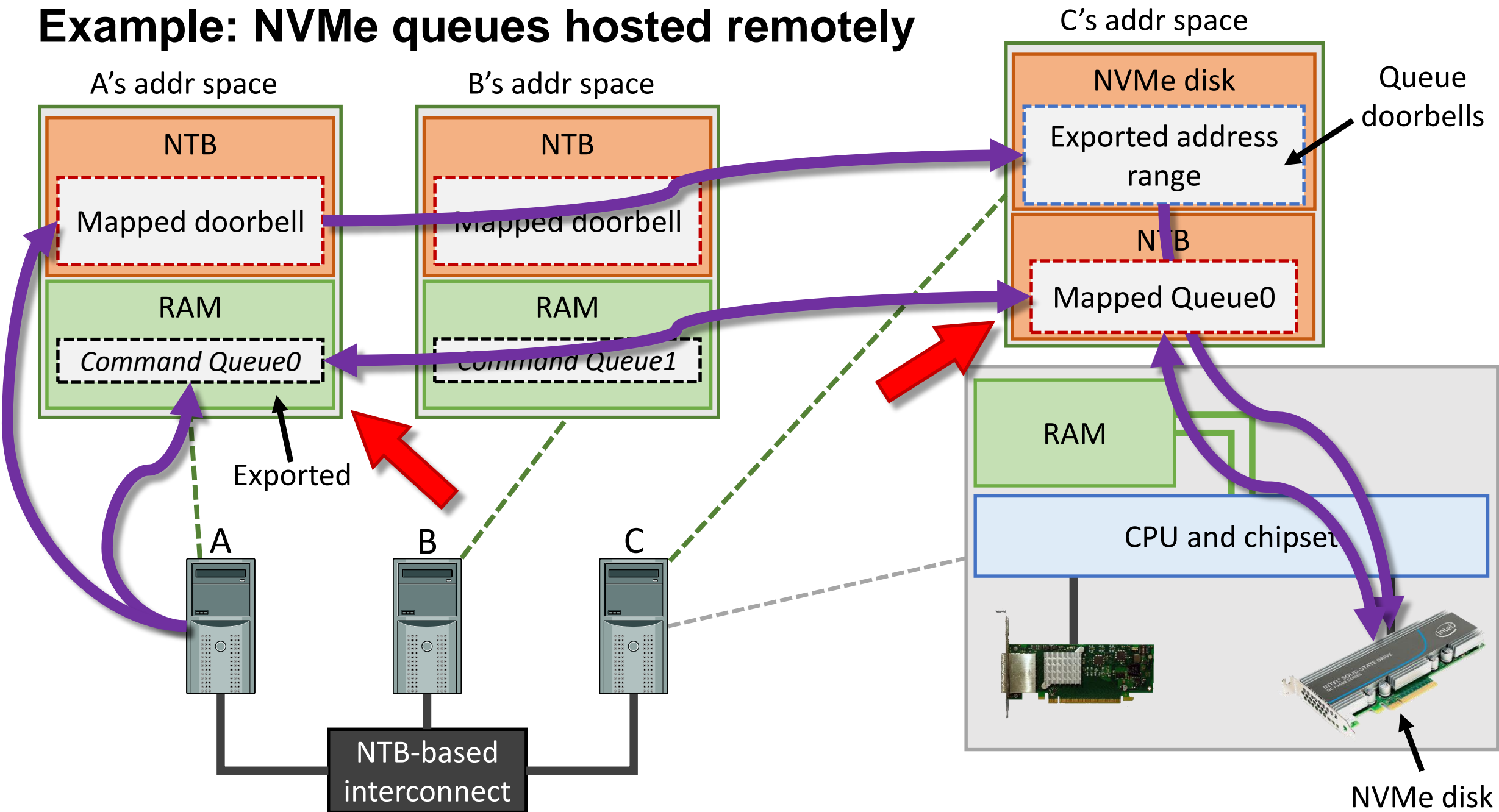
Example: NVMe disk operation (simplified)



SmartIO enabled driver: NVMe on GPU

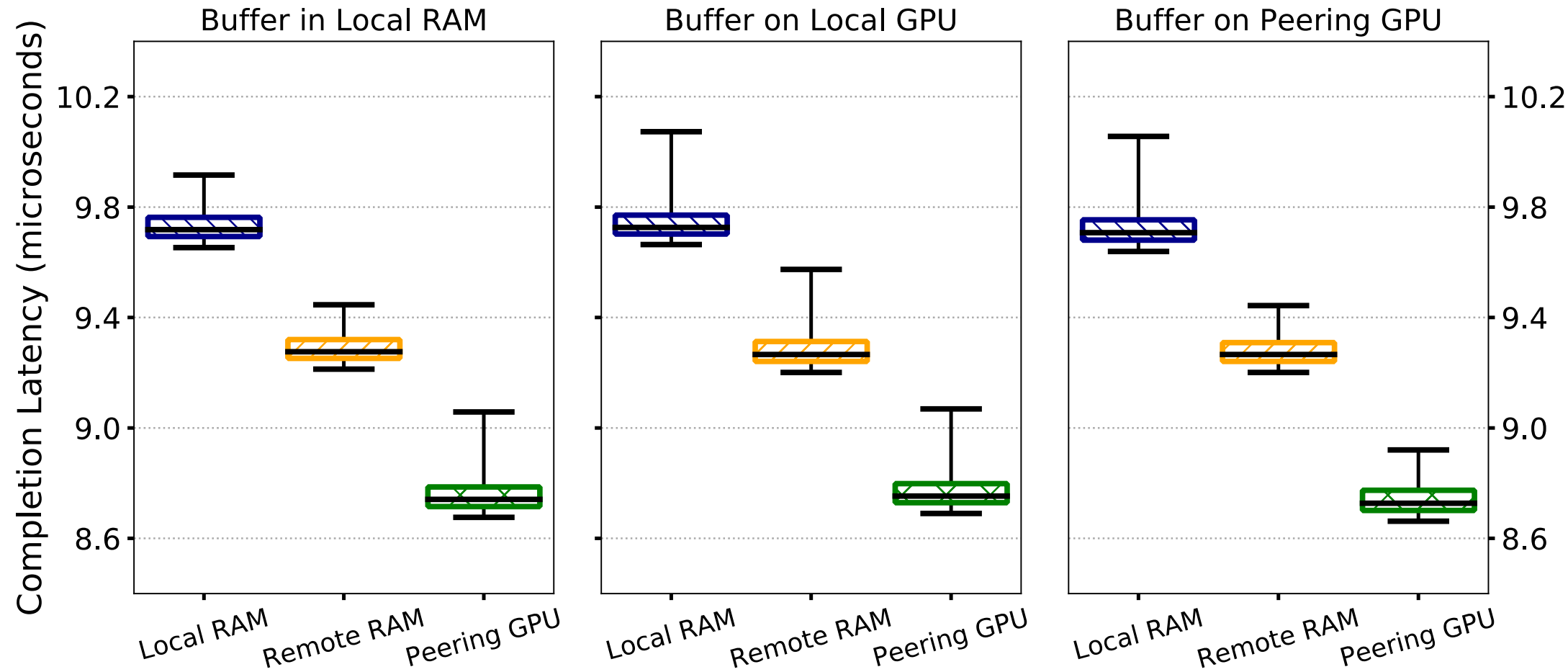


Example: NVMe queues hosted remotely



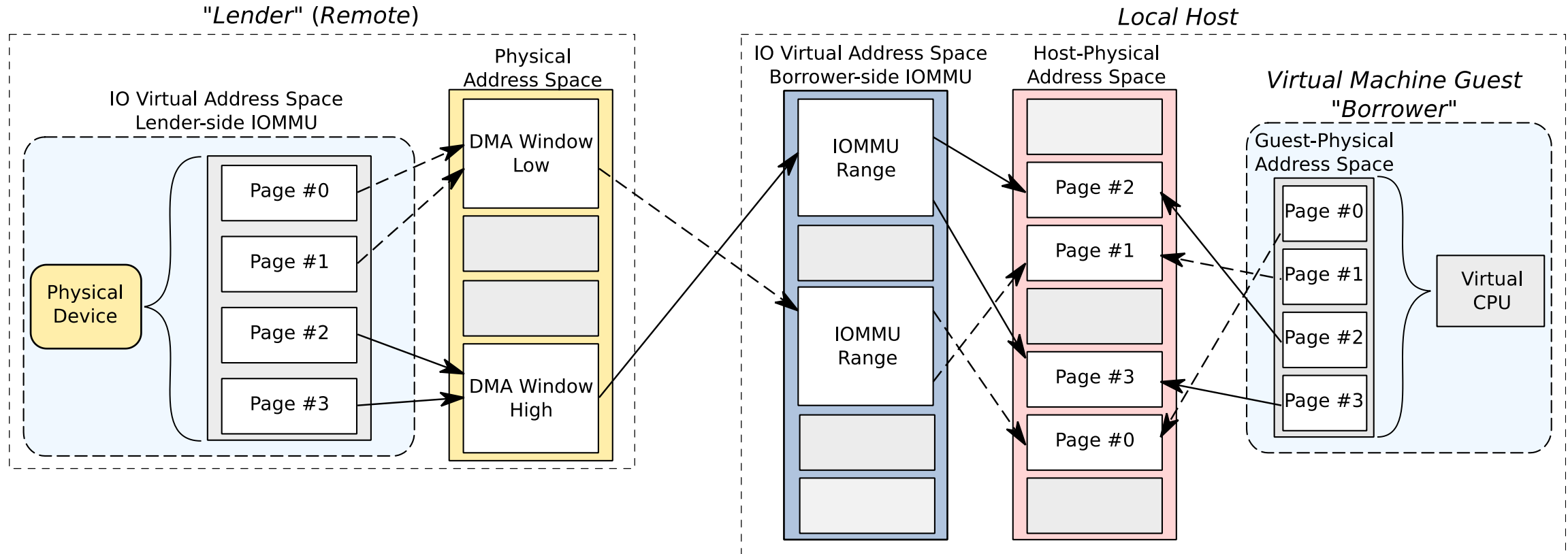
Read latency for reading blocks from a NVMe disk into a GPU: Local versus borrowed disk

Command Submission Latency
Random Reads, Queue Depth=1, PRPs per Command=1

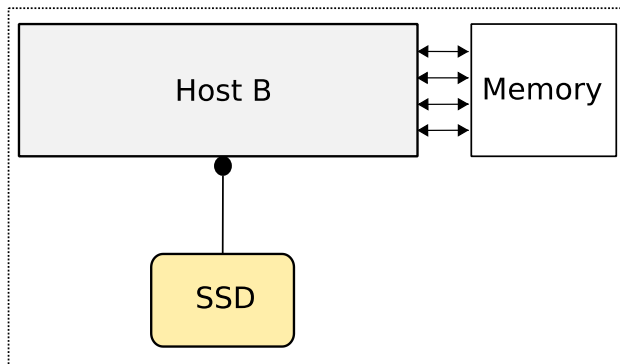
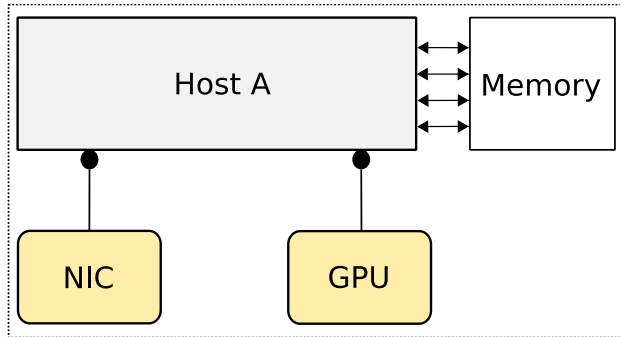


SmartIO in Virtual Machines

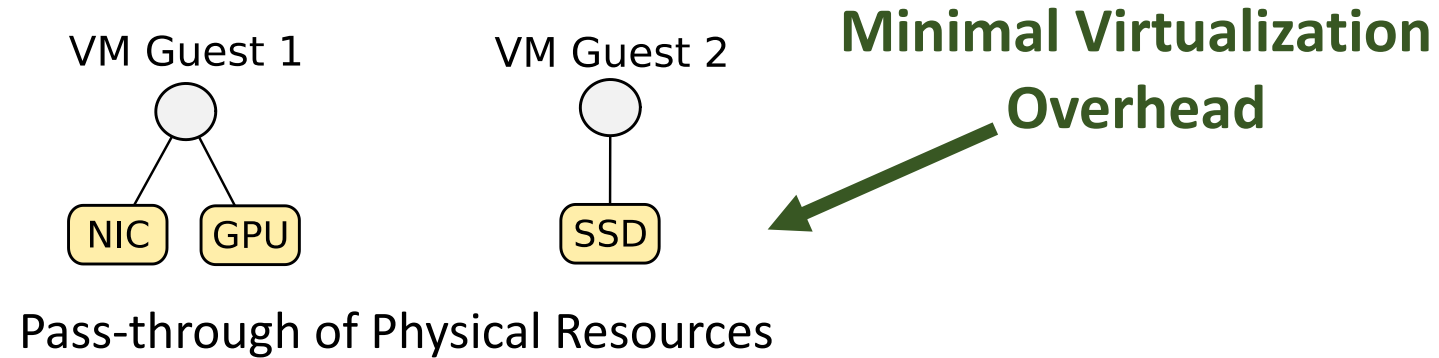
SmartIO fully supports to lend devices to virtual machines running in Linux KVM using Virtual Function IO API (VFIO)



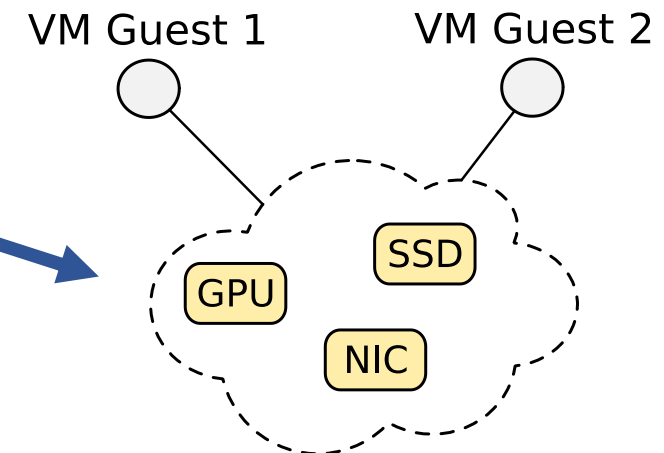
Pass-through allows physical devices to be used by VMs with minimal overhead, but is not as flexible as resource virtualization



Physical View

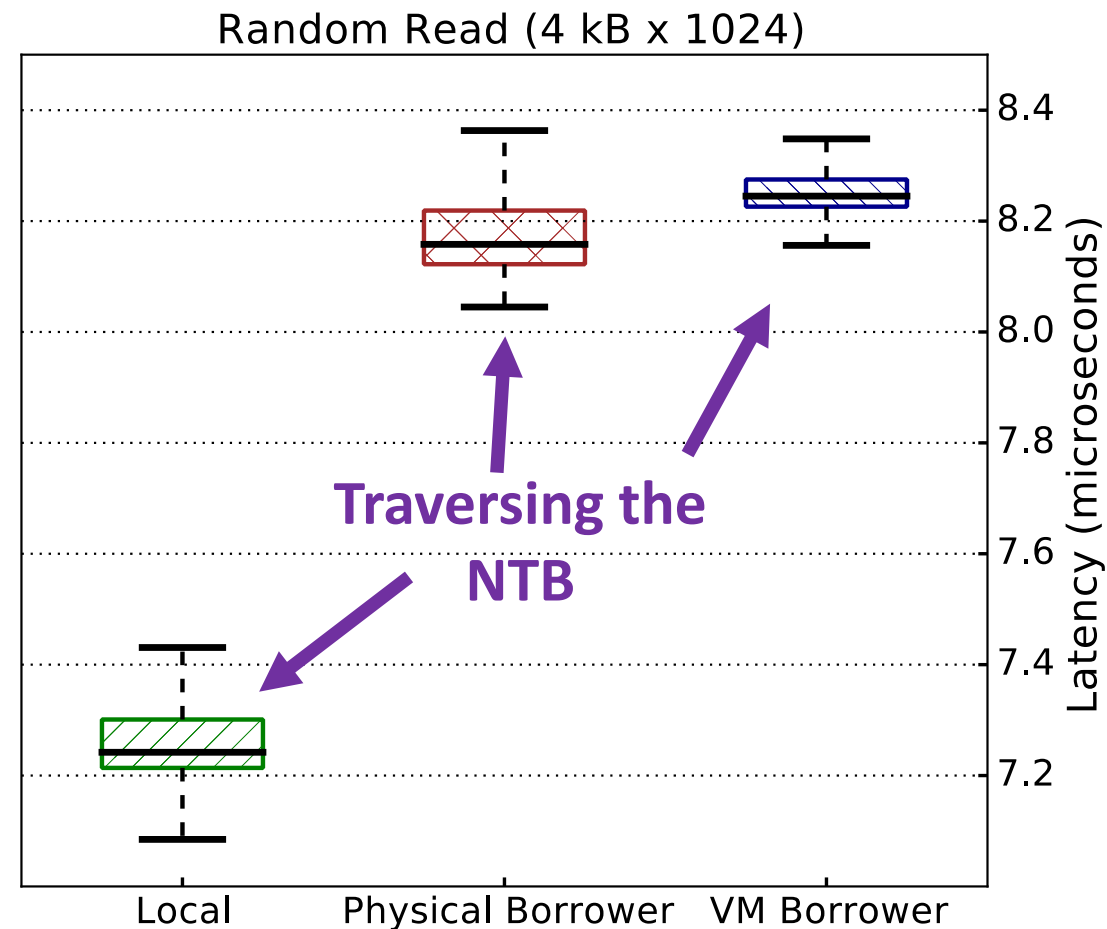
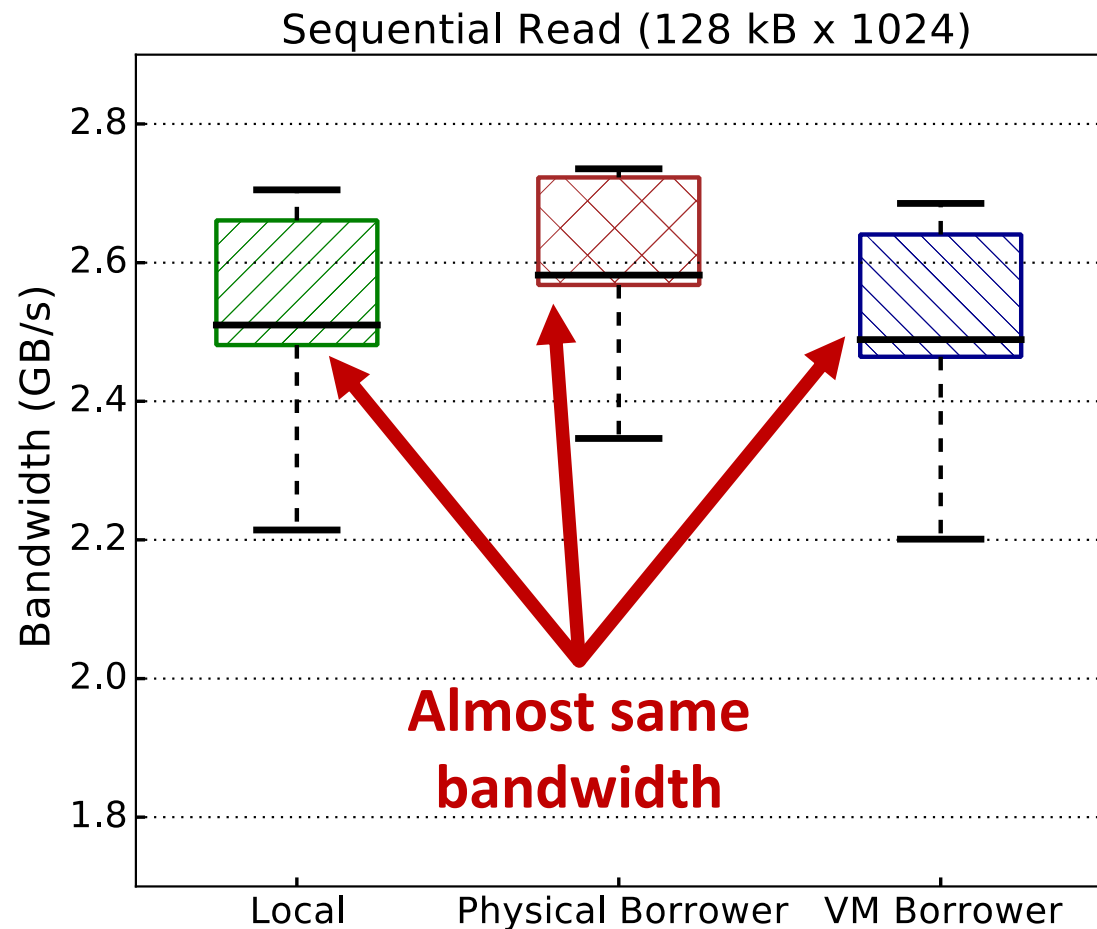


Dynamic Provisioning & Flexible Composition



Virtual or Paravirtualized Resources

Passing through a remote NVMe disk to a VM only adds the latency of traversing the NTB and is comparable to a physical borrower



Guest OS: Ubuntu 17.04, Host OS: CentOS 7
VM: Qemu 2.17 using KVM
NVMe Disk: Intel 900P Optane (PCIe x4 Gen3)

Thank you!

Selected
publications

“Device Lending in PCI Express Networks”
ACM NOSSDAV 2016

“Efficient Processing of Video in a Multi Auditory Environment using Device Lending of GPUs”
ACM Multimedia Systems 2016 (MMSys'16)

“Flexible Device Sharing in PCIe Clusters using Device Lending”, International Conference on Parallel Processing Companion (ICPP'18 Comp)

haakonks@simula.no

SmartIO & Device Lending demo with GPUs, NVMe and more

Visit Dolphin in the exhibition area (booth 1520)

