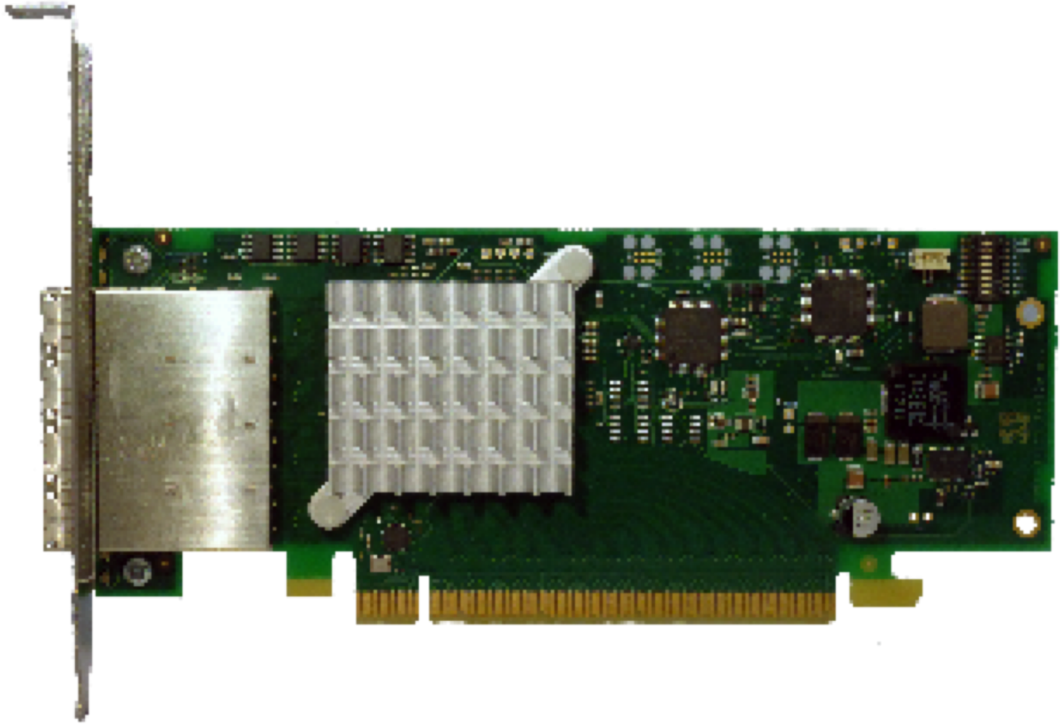
S8511
SmartIO: Dynamic Sharing of GPUs and IO in a PCIe Cluster

Håkon Kvale Stensland
Research Scientist / Associate Professor
Simula Research Laboratory / University of Oslo

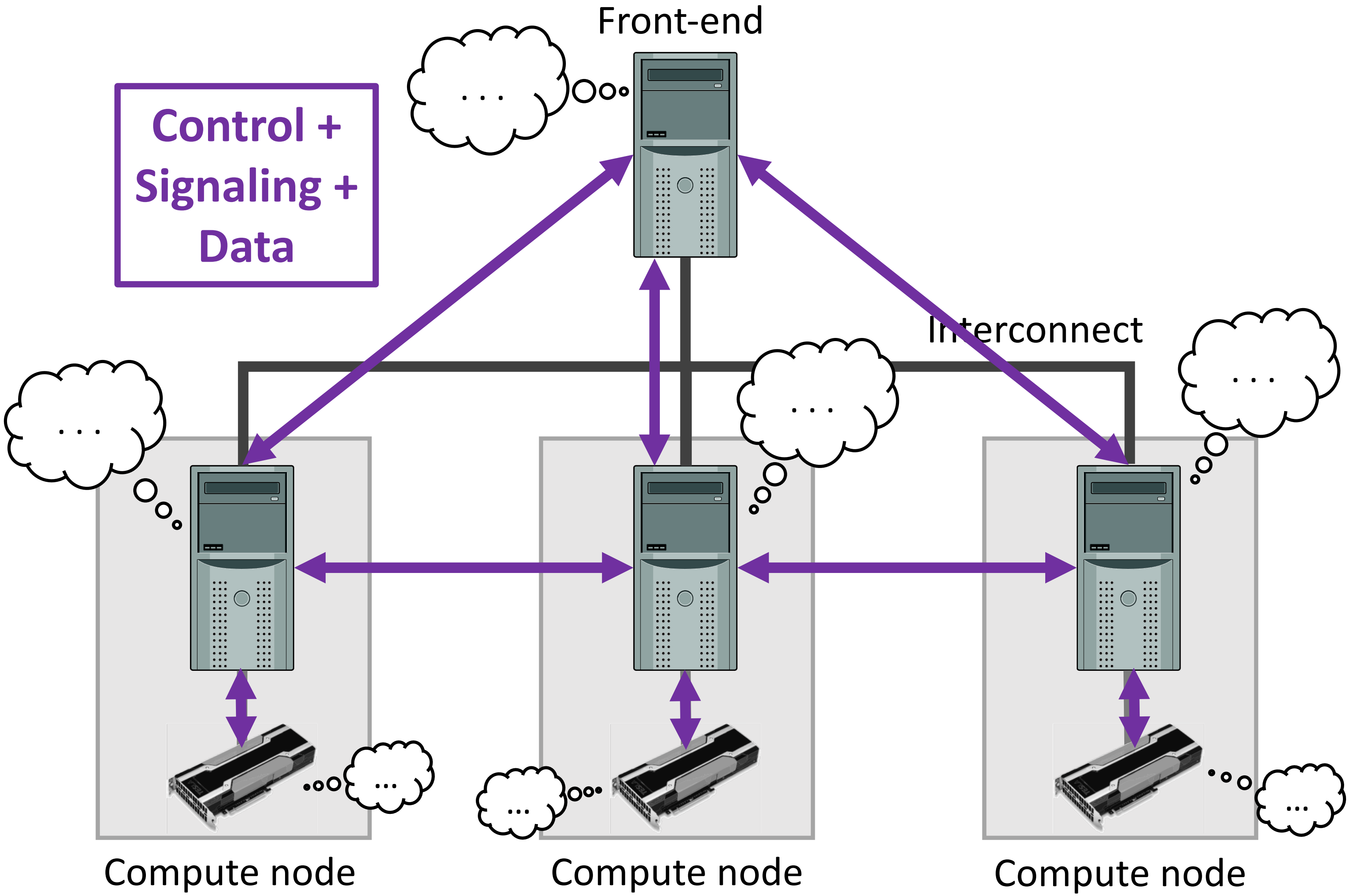


simula Dolphin INTERCONNECT SOLUTIONS UNIVERSITAS OSLOENSIS MDCCCXI

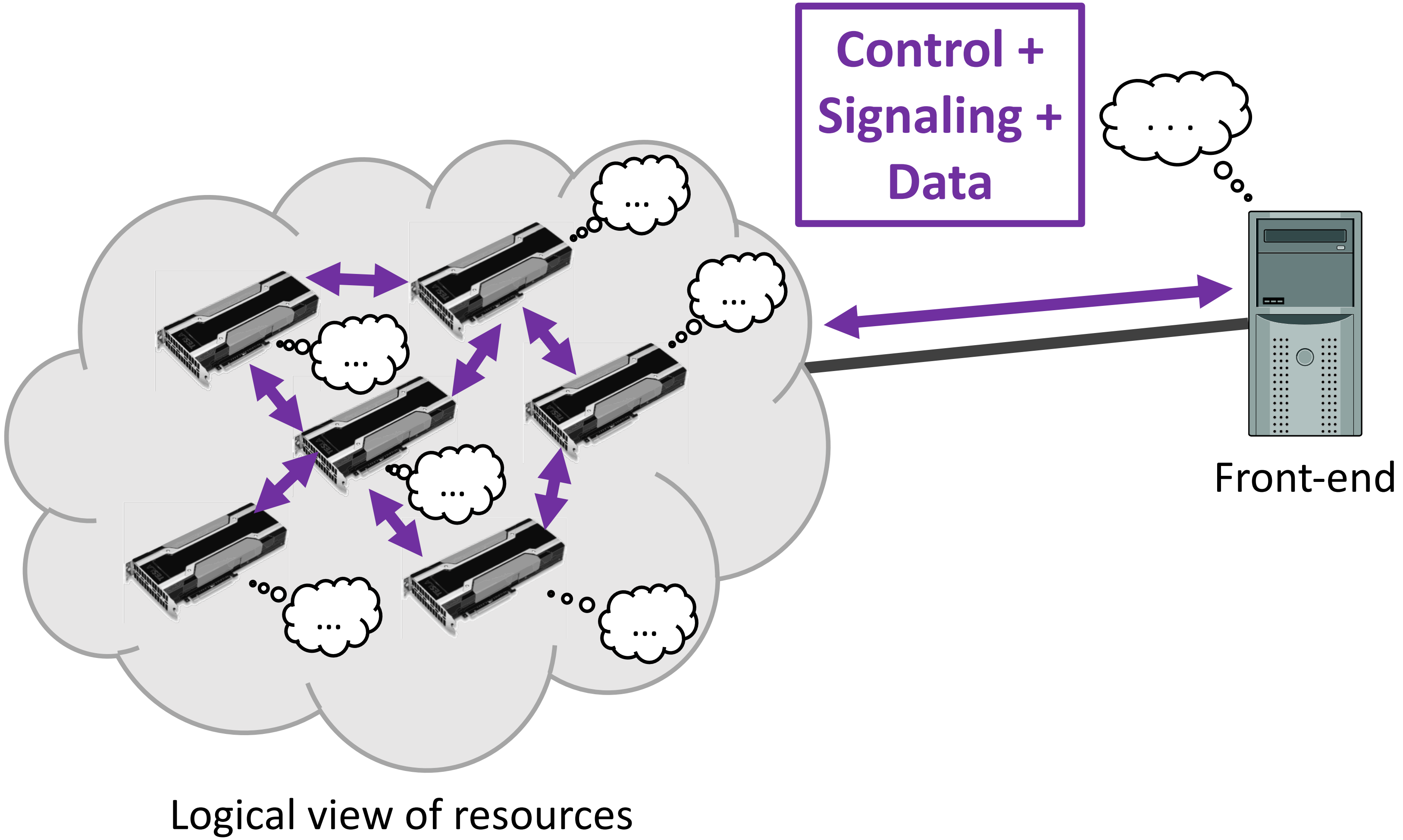
Outline

- Motivation
- PCIe Overview
- Non-Transparent Bridges
- Dolphin SmartIO

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

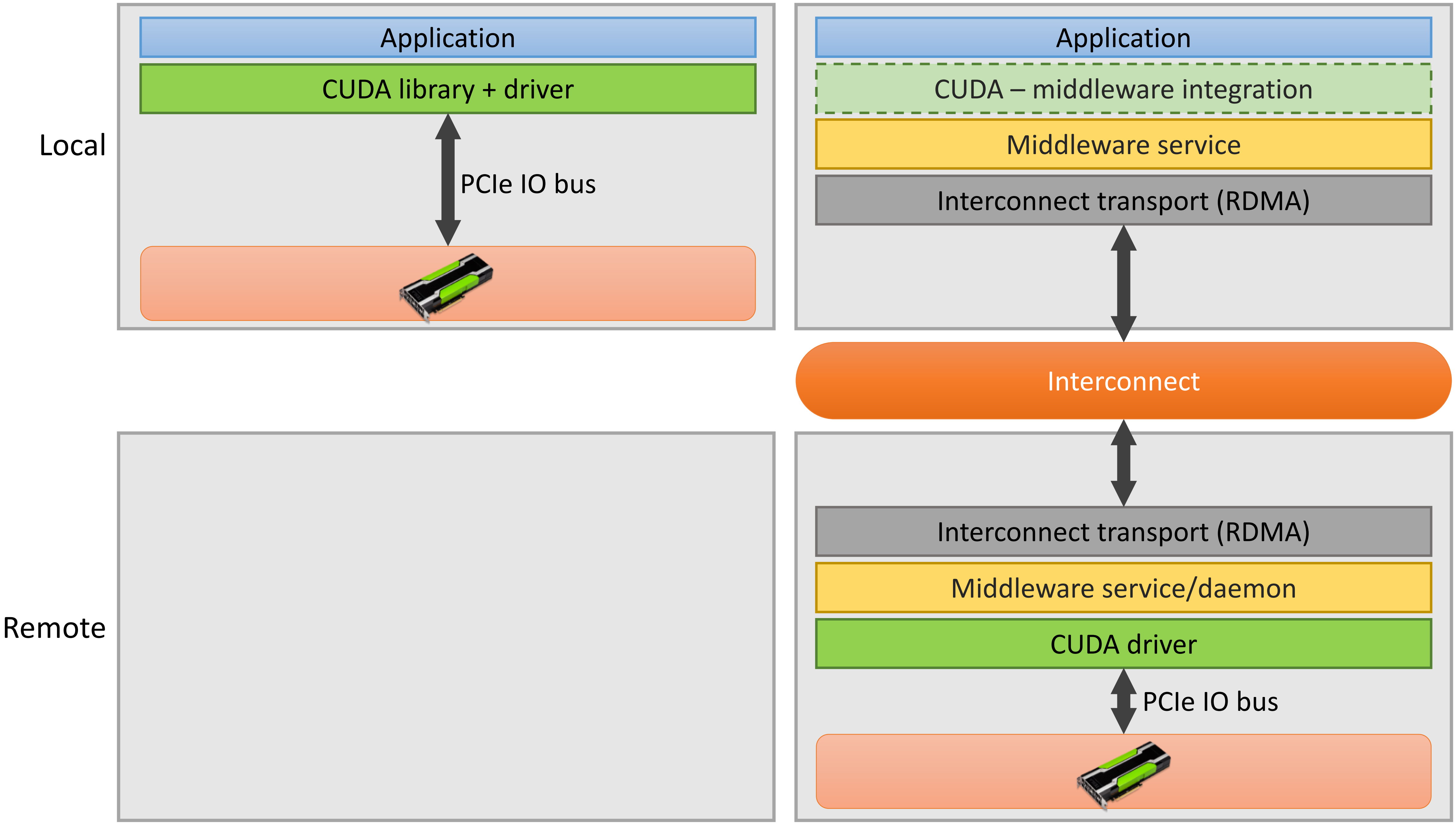


Software abstractions simplify the use and allocation of resources in a cluster and facilitate development of distributed applications



Handled in software

- rCUDA
- CUDA-aware Open MPI
- Custom GPUDirect RDMA implementation
- . . .



Local resource

Remote resource using **middleware**

Local

Application

CUDA library + driver

PCIe IO bus

Application

CUDA – middleware integration

Middleware service

Interconnect transport (RDMA)

Interconnect

Remote

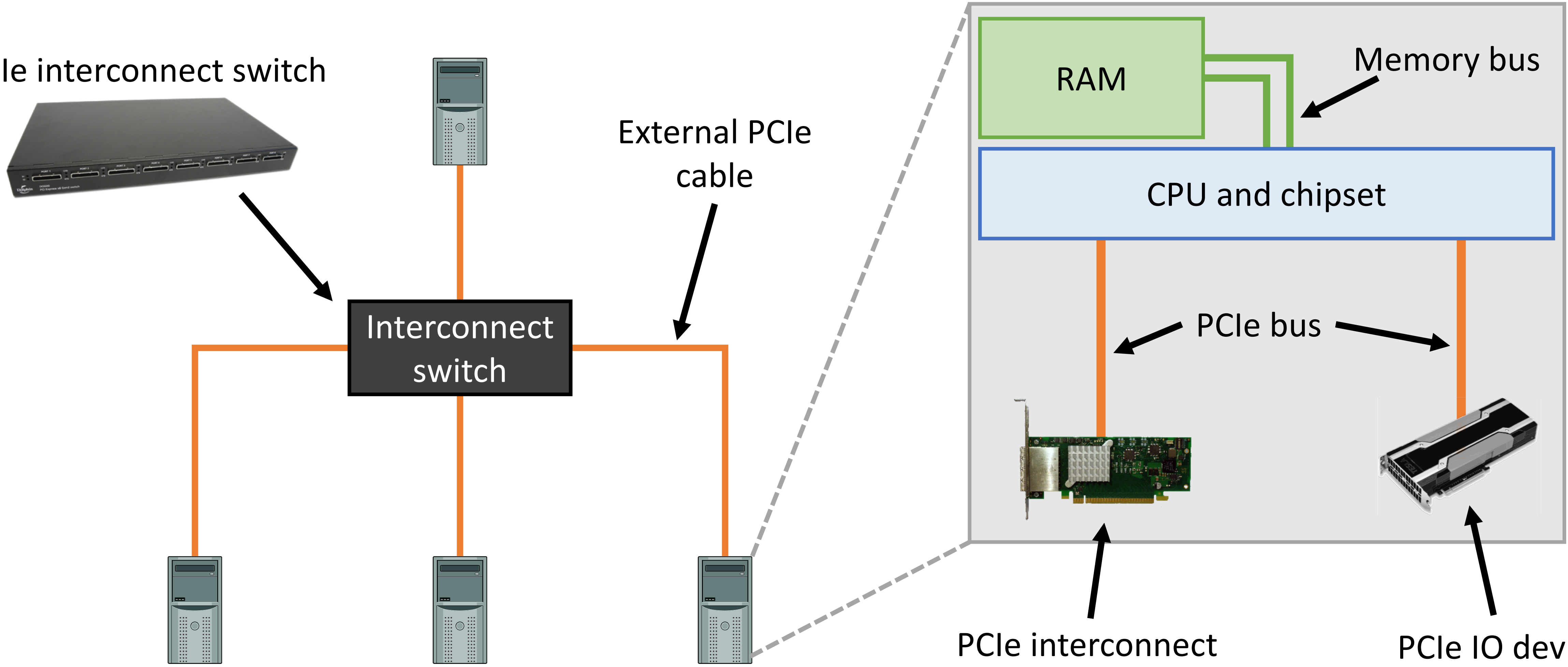
Interconnect transport (RDMA)

Middleware service/daemon

CUDA driver

PCIe IO bus

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



PCIe interconnect switch

Interconnect switch

External PCIe cable

RAM

Memory bus

CPU and chipset

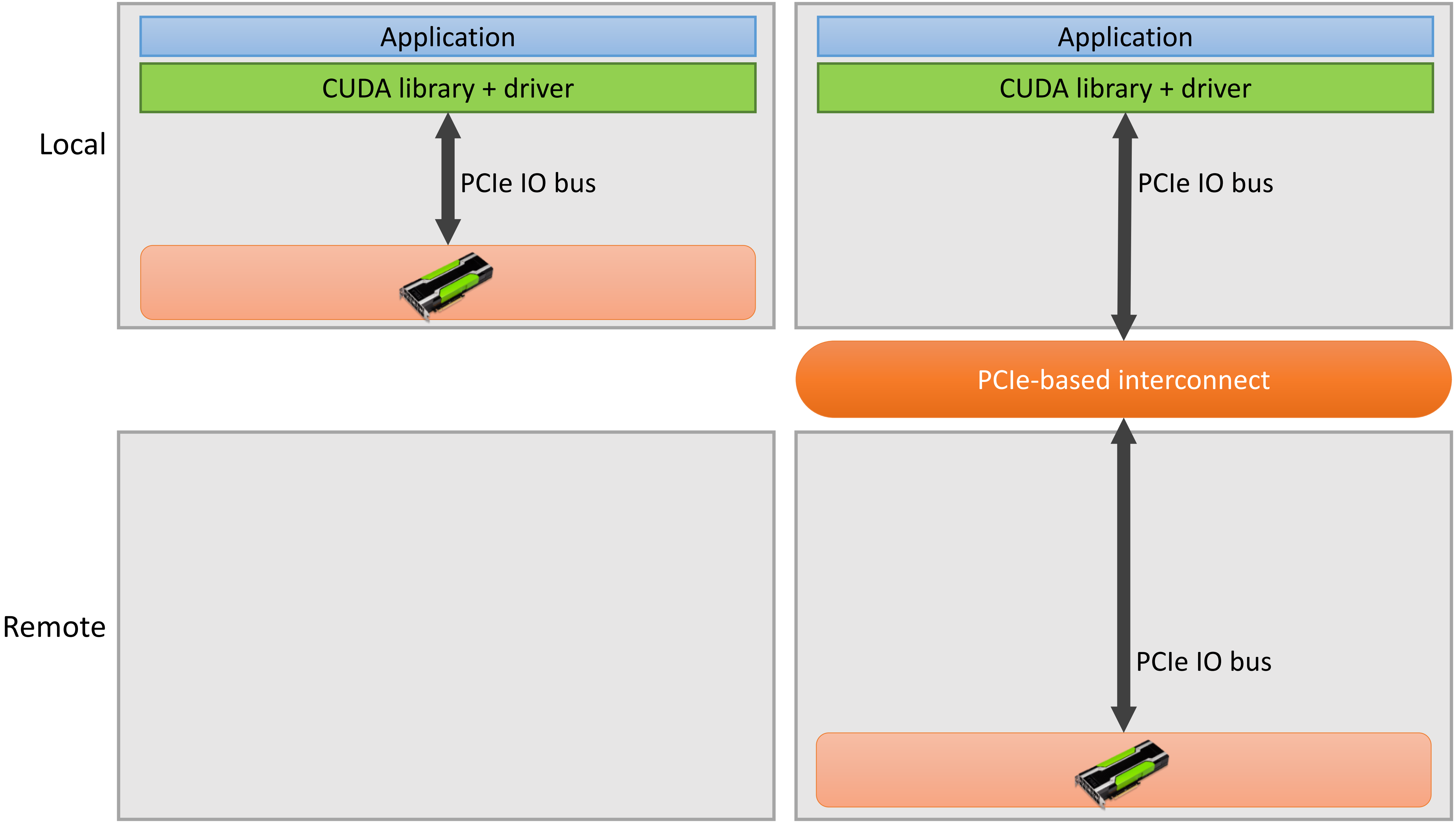
PCIe bus

PCIe interconnect host adapter

PCIe IO device

Local resource

Remote resource over **native fabric**



Local

Application

CUDA library + driver

PCIe IO bus

Application

CUDA library + driver

PCIe IO bus

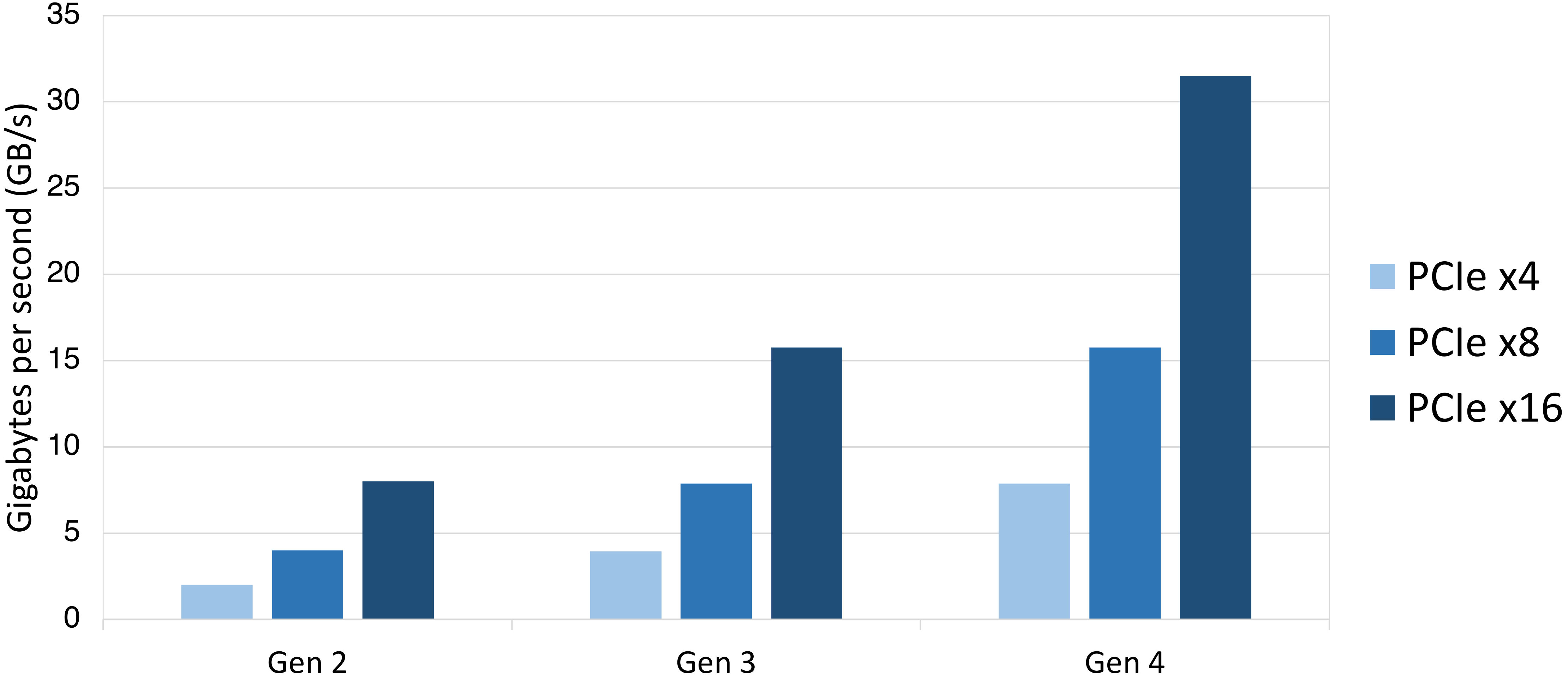
PCIe-based interconnect

Remote

PCIe IO bus

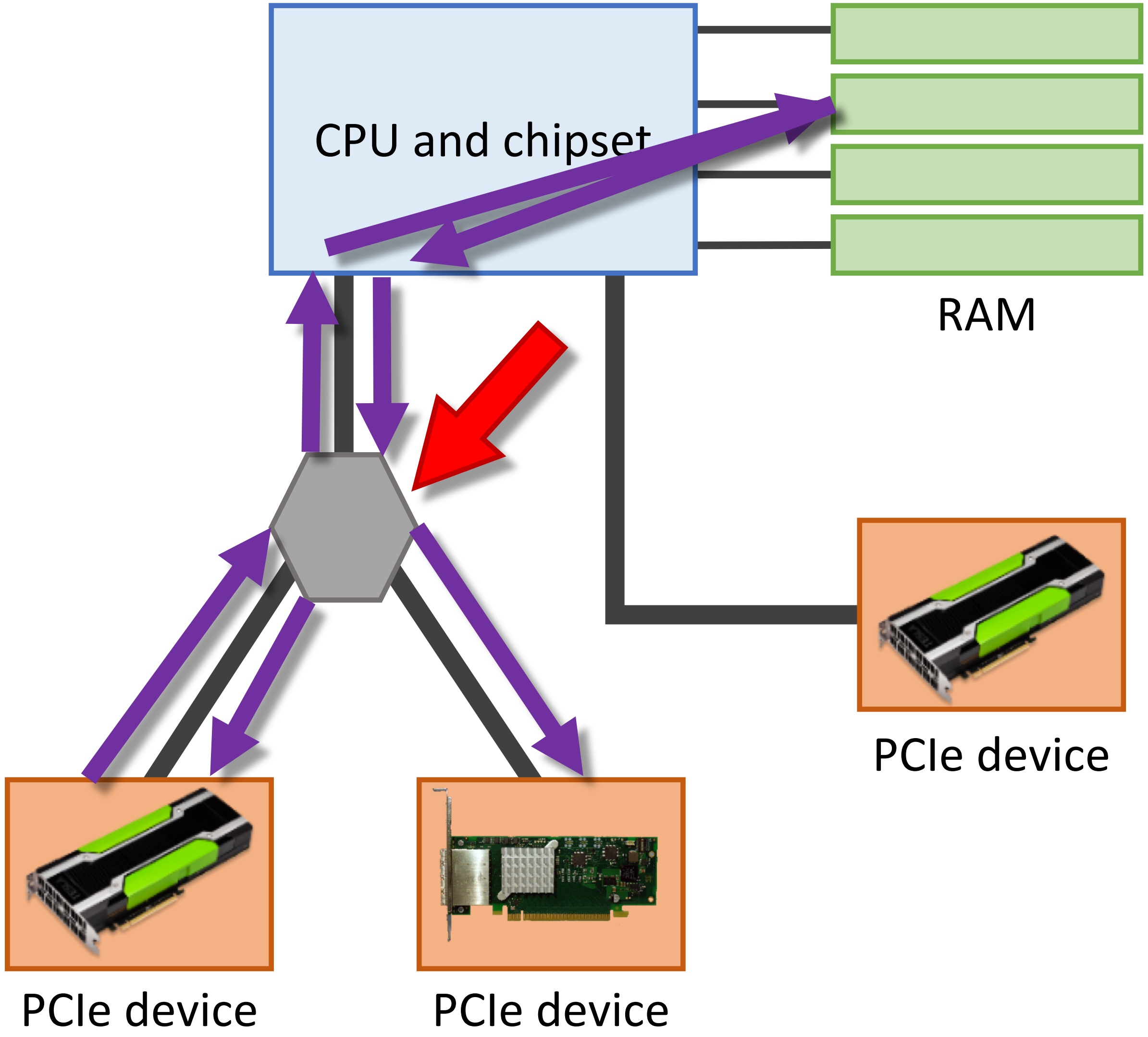
PCIe Overview

PCIe is the dominant IO bus technology in computers today, and can also be used as a high-bandwidth low-latency interconnect



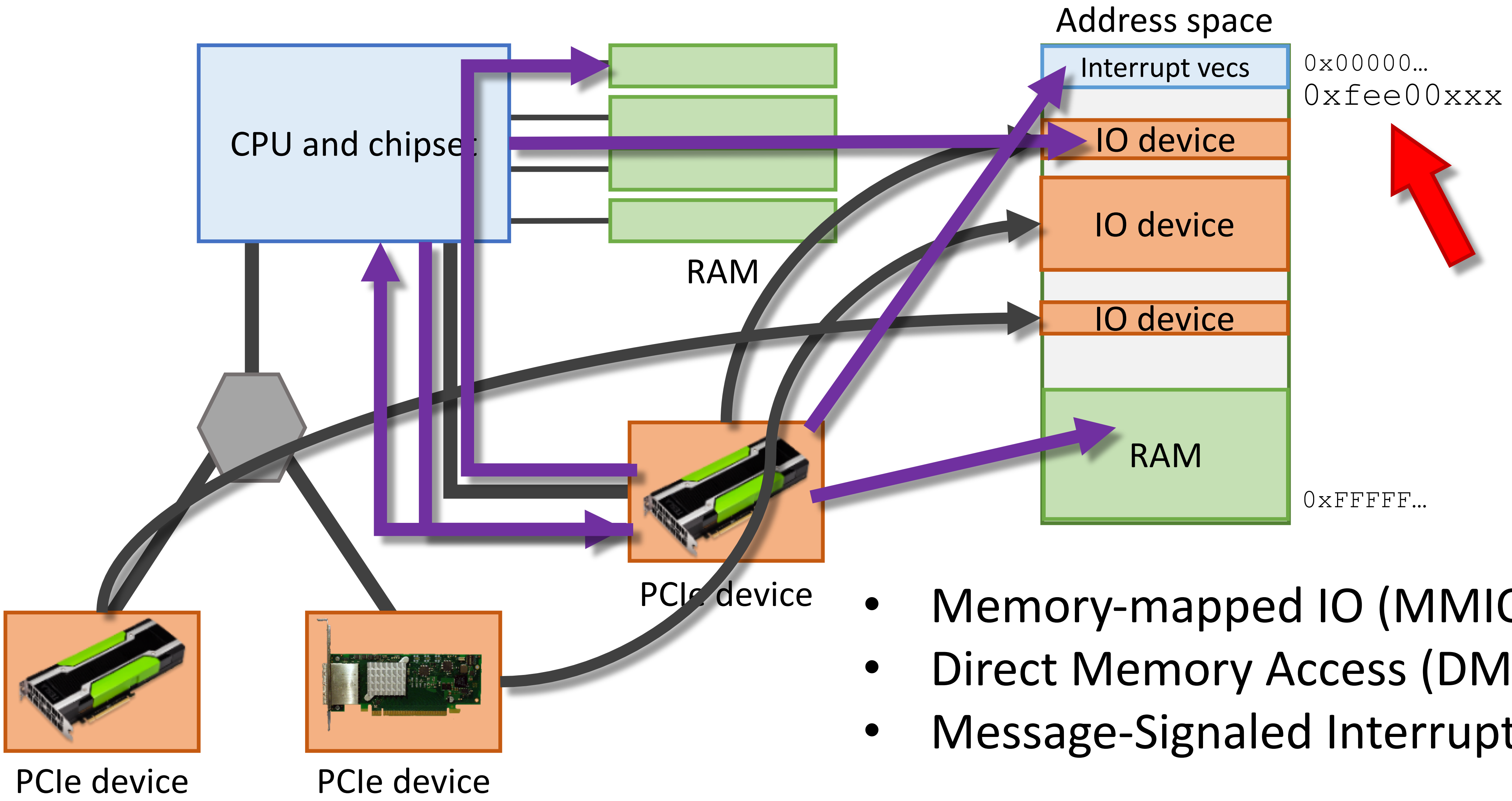
PCI-SIG. PCI Express 3.1 Base Specification, 2010.
http://www.eetimes.com/document.asp?doc_id=1259778

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)

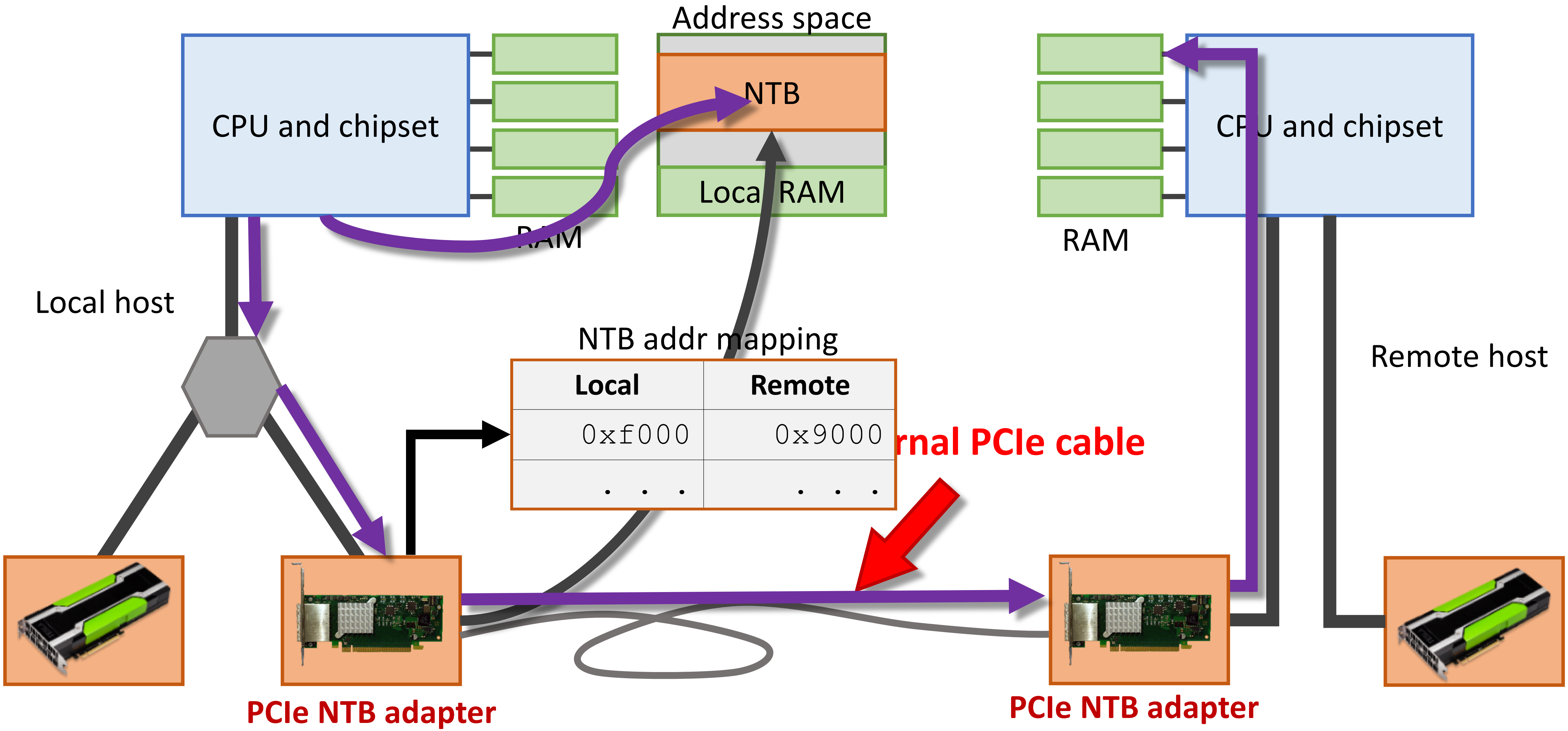
IO devices and the CPU share the same physical address space, allowing devices to access system memory and other devices



- Memory-mapped IO (MMIO / PIO)
- Direct Memory Access (DMA)
- Message-Signaled Interrupts (MSI-X)

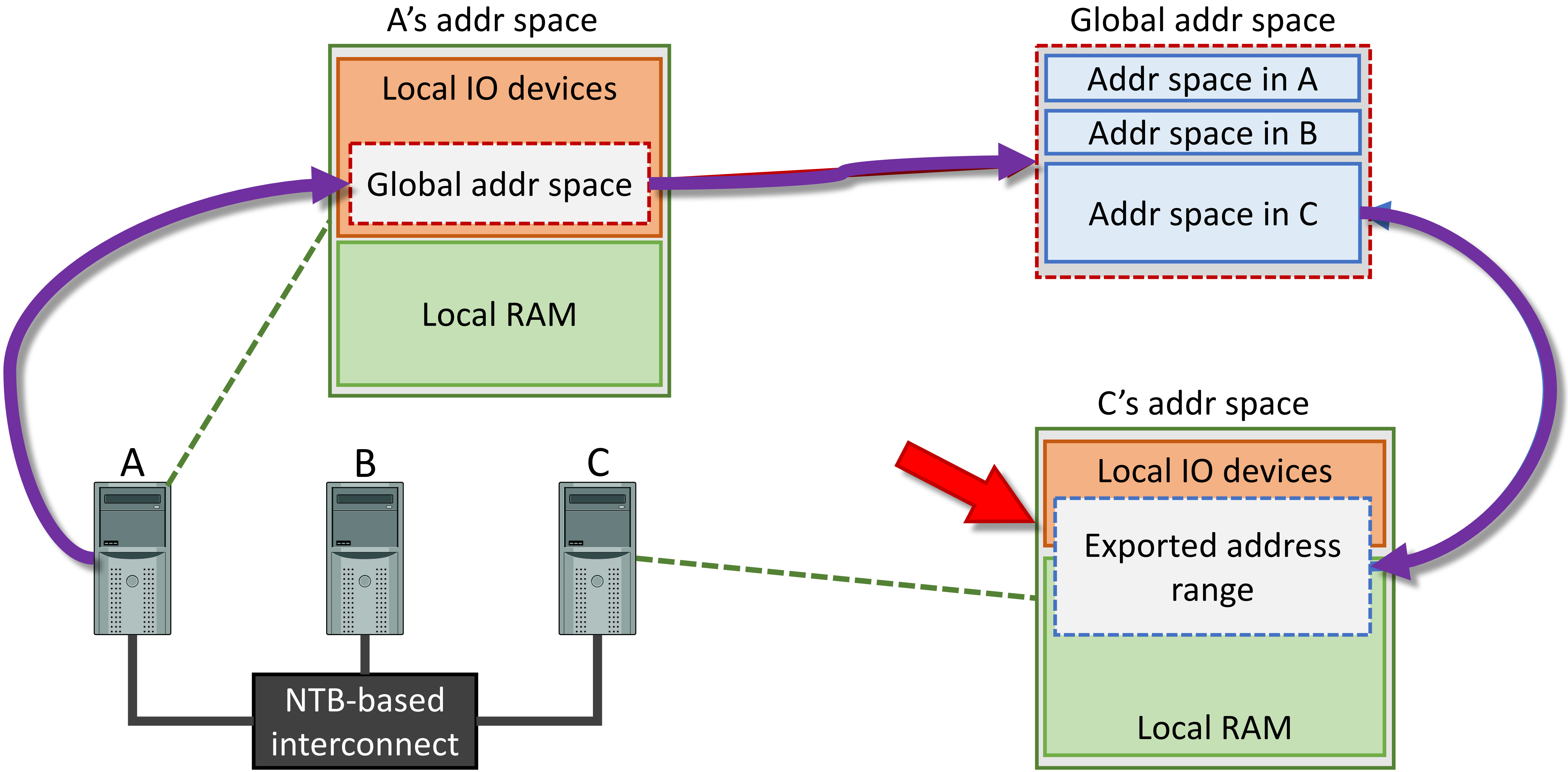
Non-Transparent Bridges

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

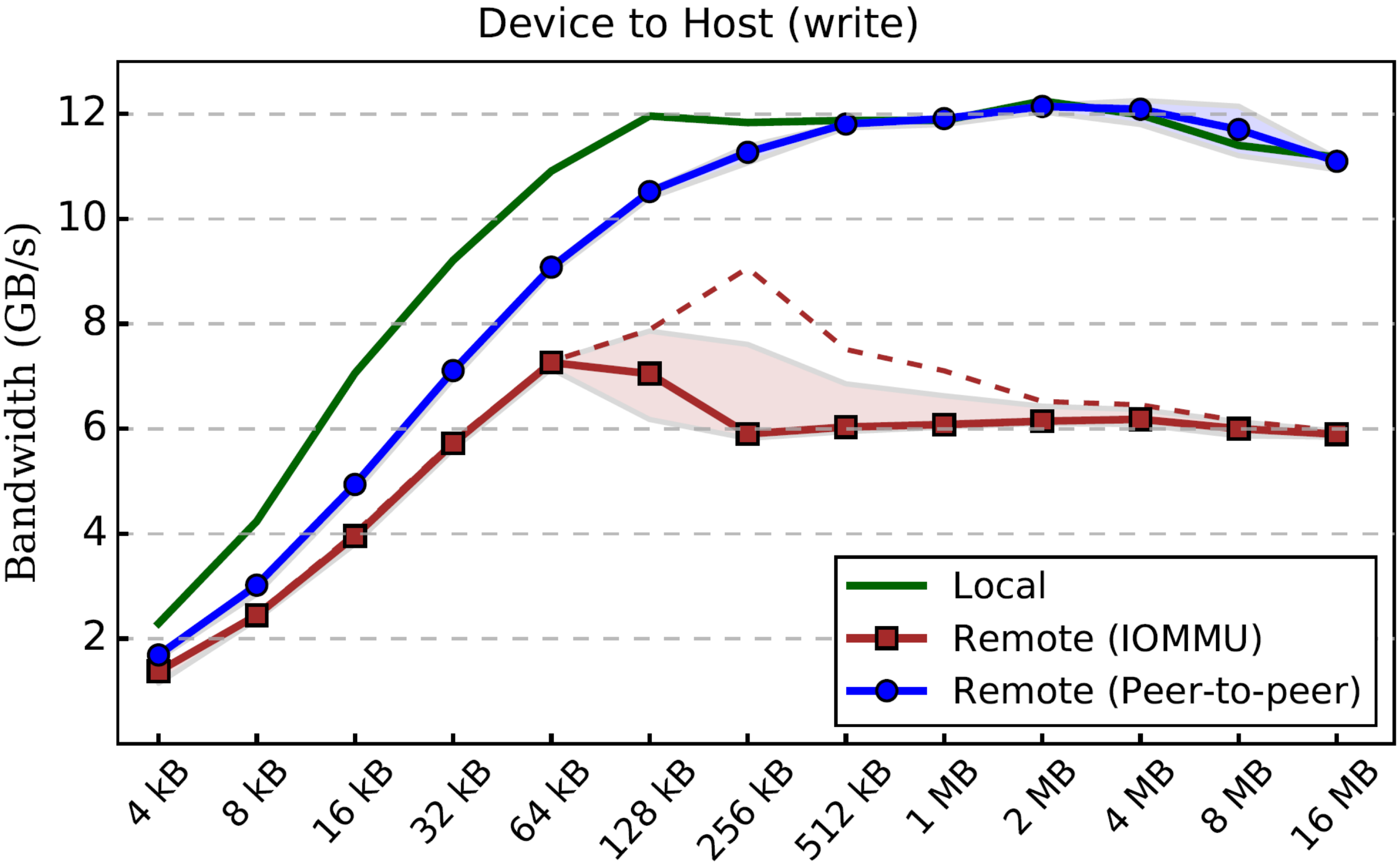


Local	Remote
0xf000	0x9000
. . .	. . .

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

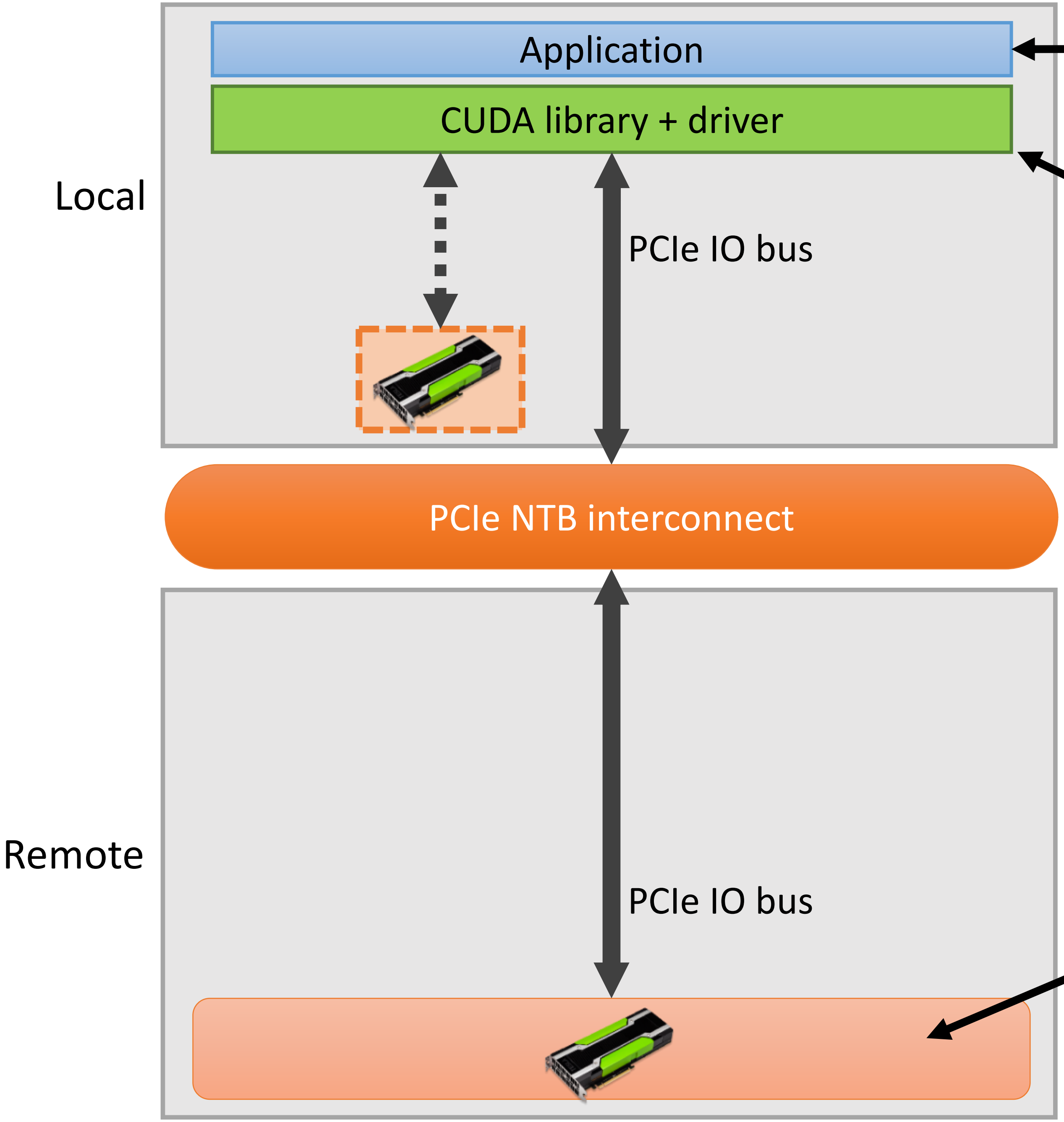


Device to host transfers: Comparing local to borrowed GPU



SmartIO

Borrowed remote resource

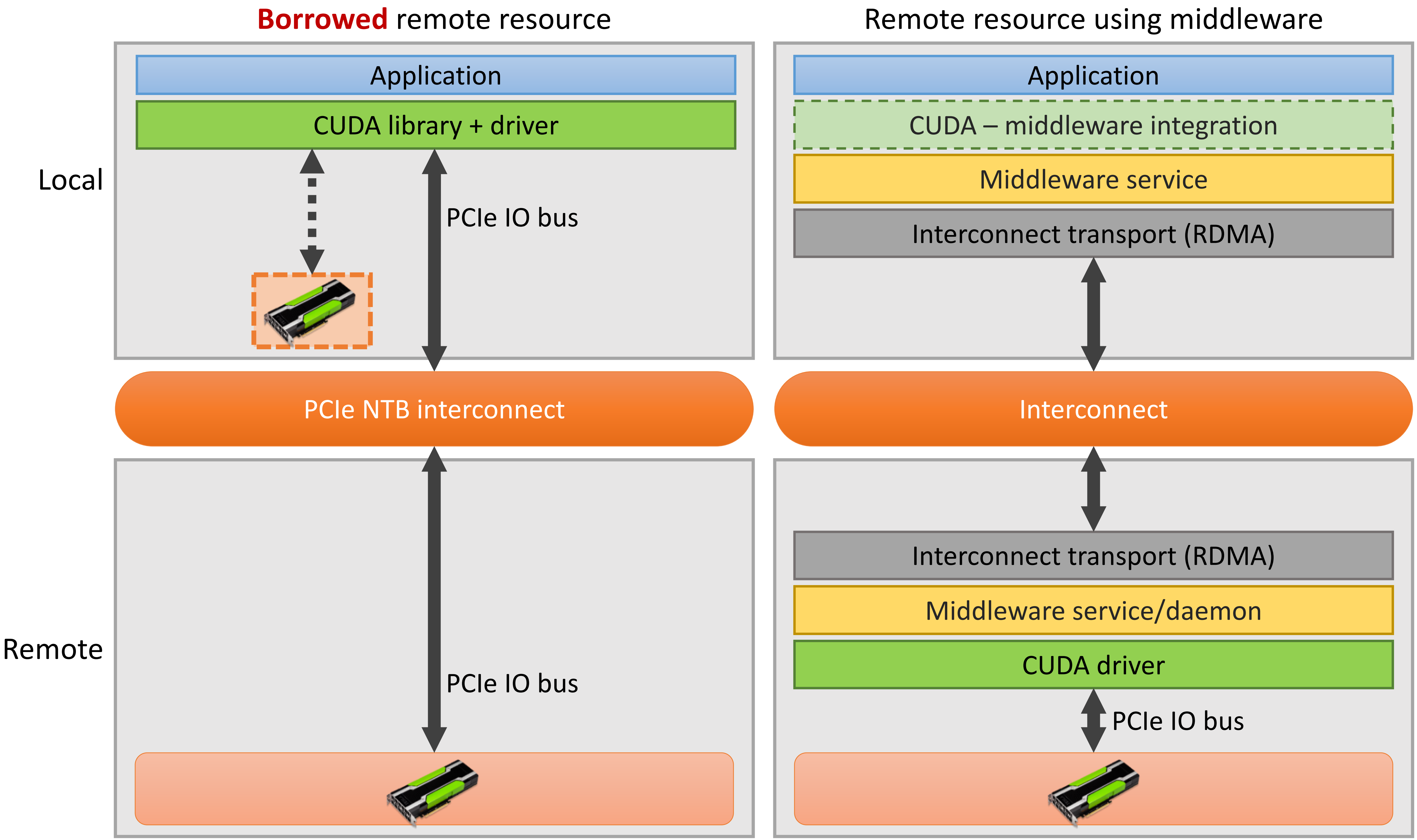


Resource appears local to OS, driver, and app

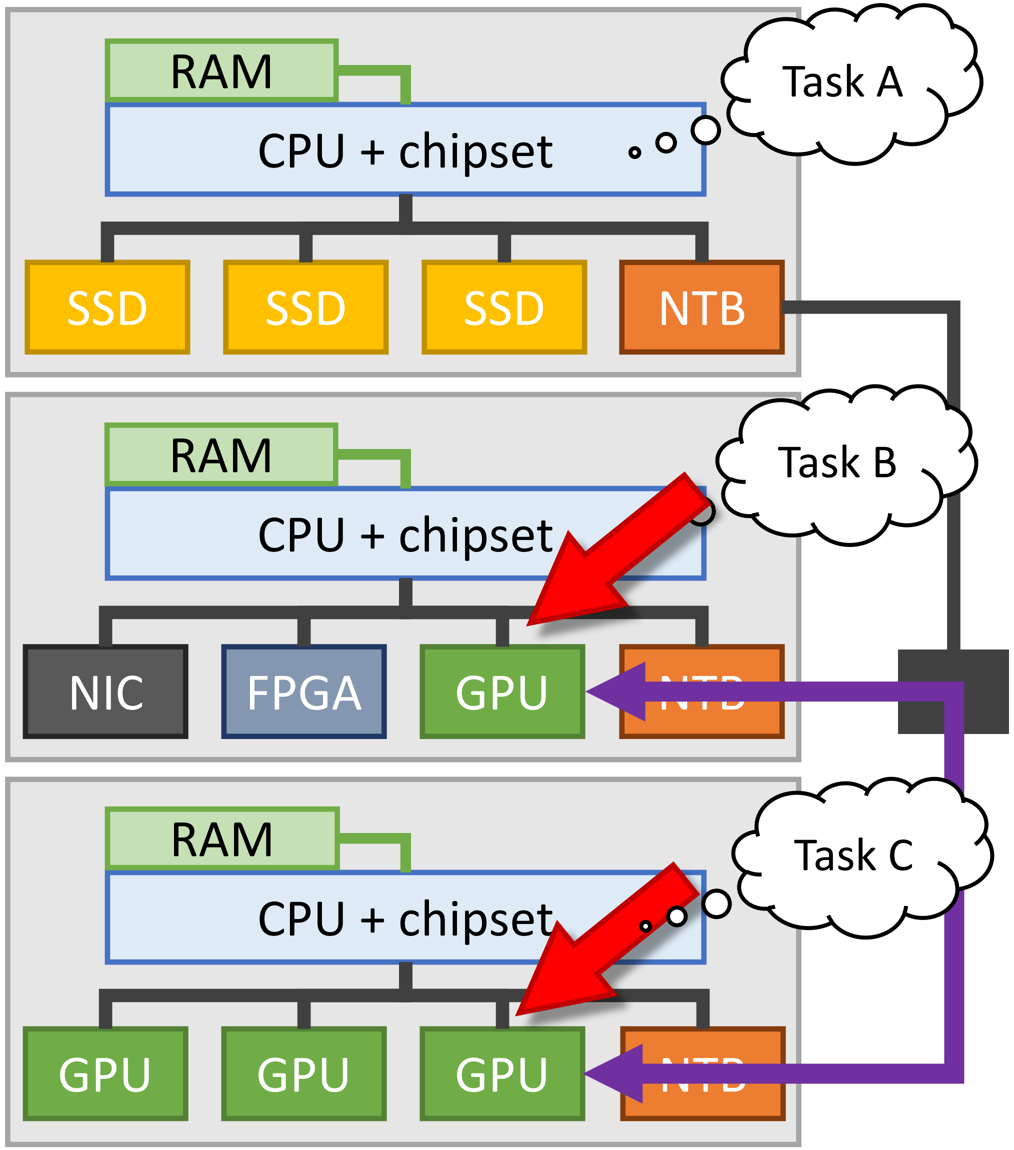
Unmodified local driver (with hot-plug support)

Hardware mappings ensure **fast data path**

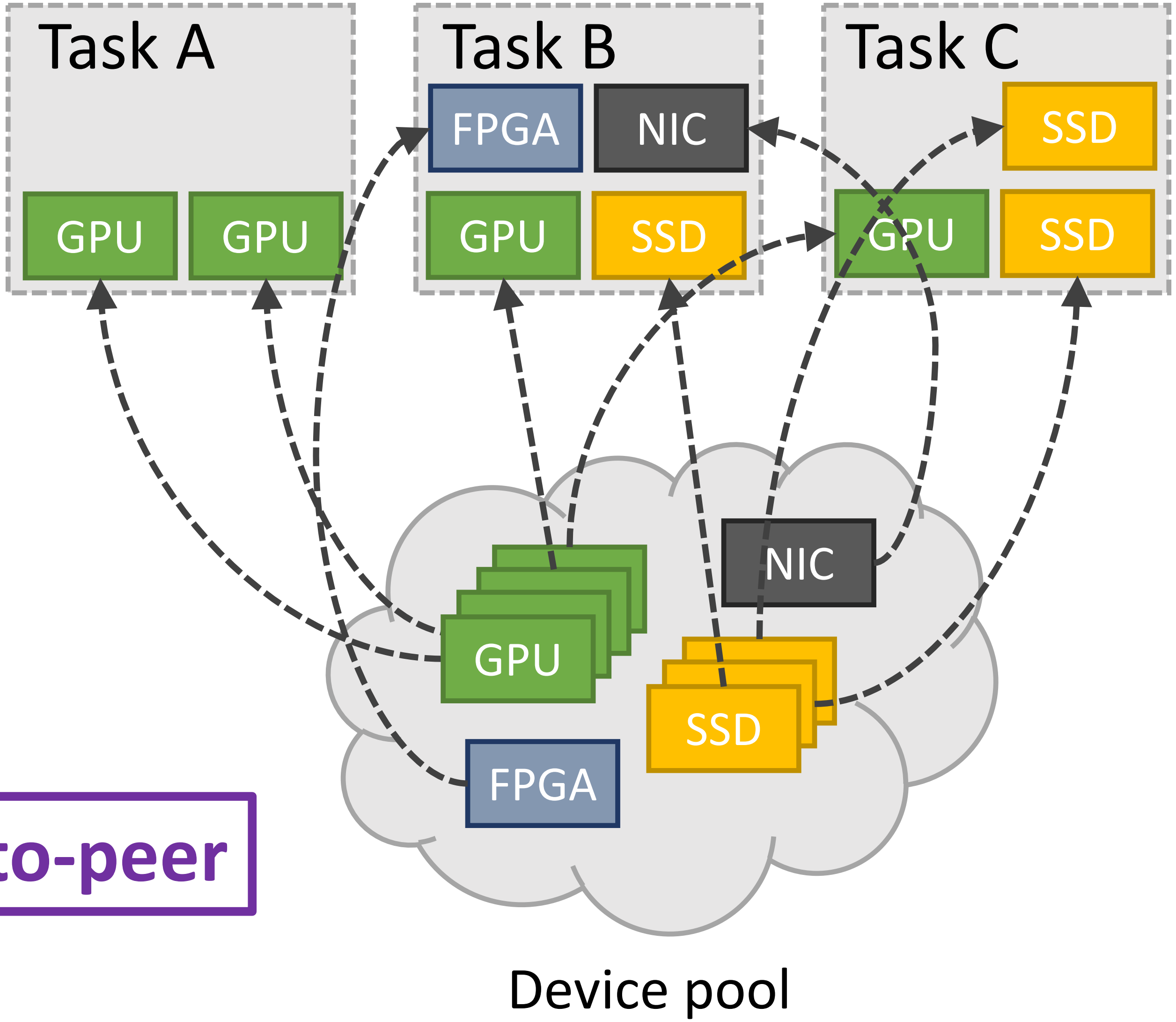
Works with **any PCIe device** (even individual SR-IOV functions)


Borrowed remote resource
Remote resource using middleware
Local
Remote
Application
CUDA library + driver
PCIe IO bus
PCIe NTB interconnect
PCIe IO bus
Application
CUDA – middleware integration
Middleware service
Interconnect transport (RDMA)
Interconnect
Interconnect transport (RDMA)
Middleware service/daemon
CUDA driver
PCIe IO bus

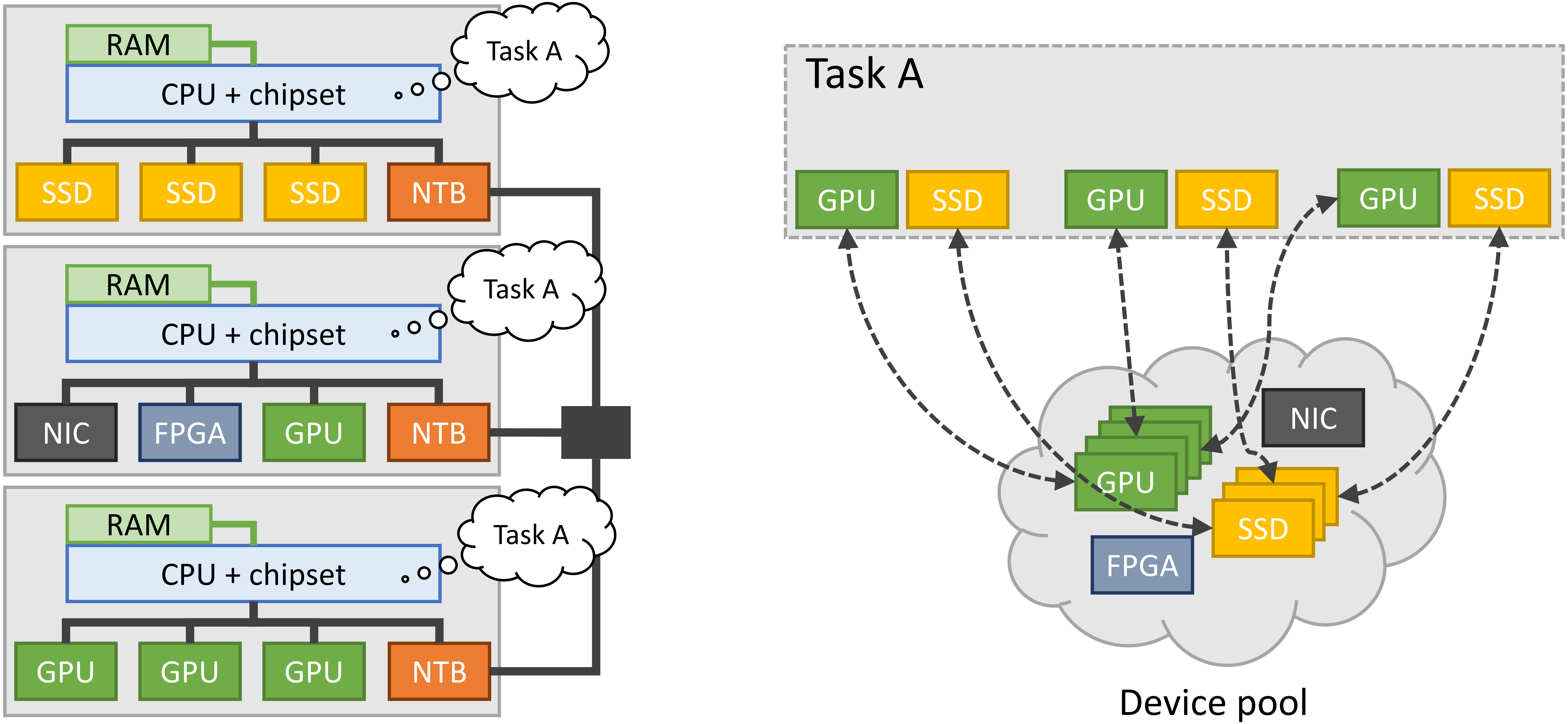
Using Device Lending, nodes in a PCIe cluster can share resources through a process of borrowing and giving back devices



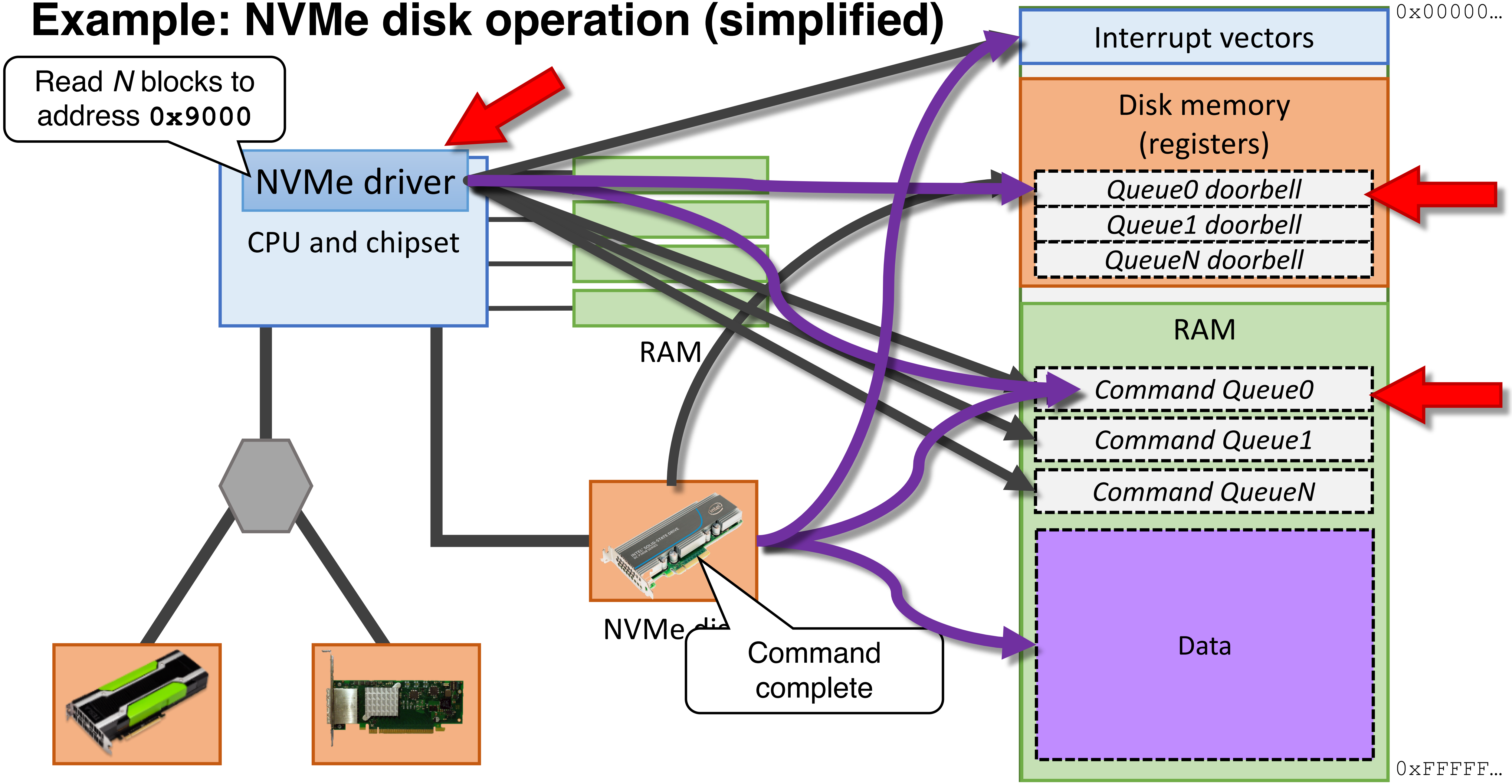
Peer-to-peer



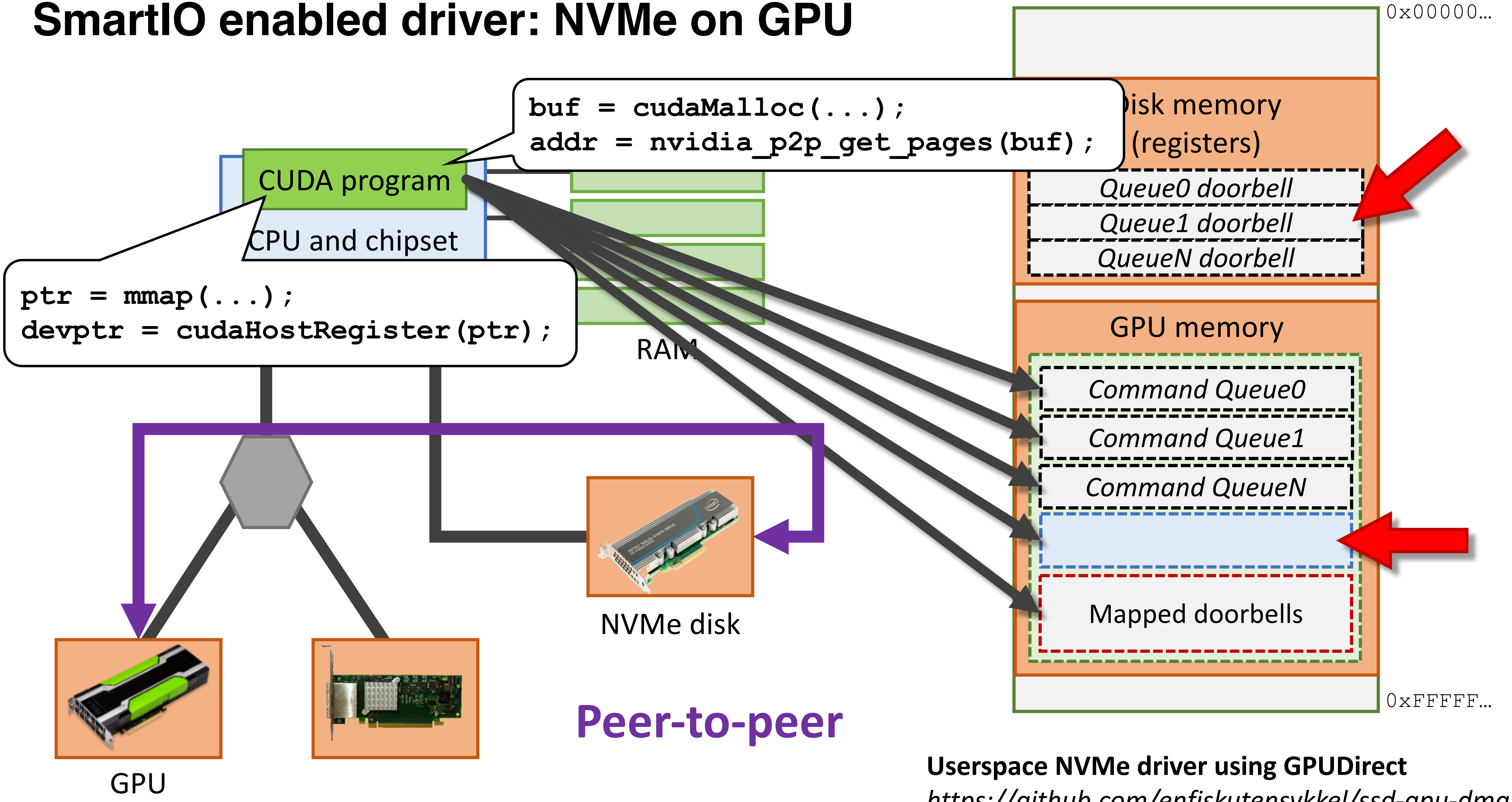
Using Device Lending, nodes in a PCIe cluster can share resources through a process of borrowing and giving back devices



Example: NVMe disk operation (simplified)



SmartIO enabled driver: NVMe on GPU



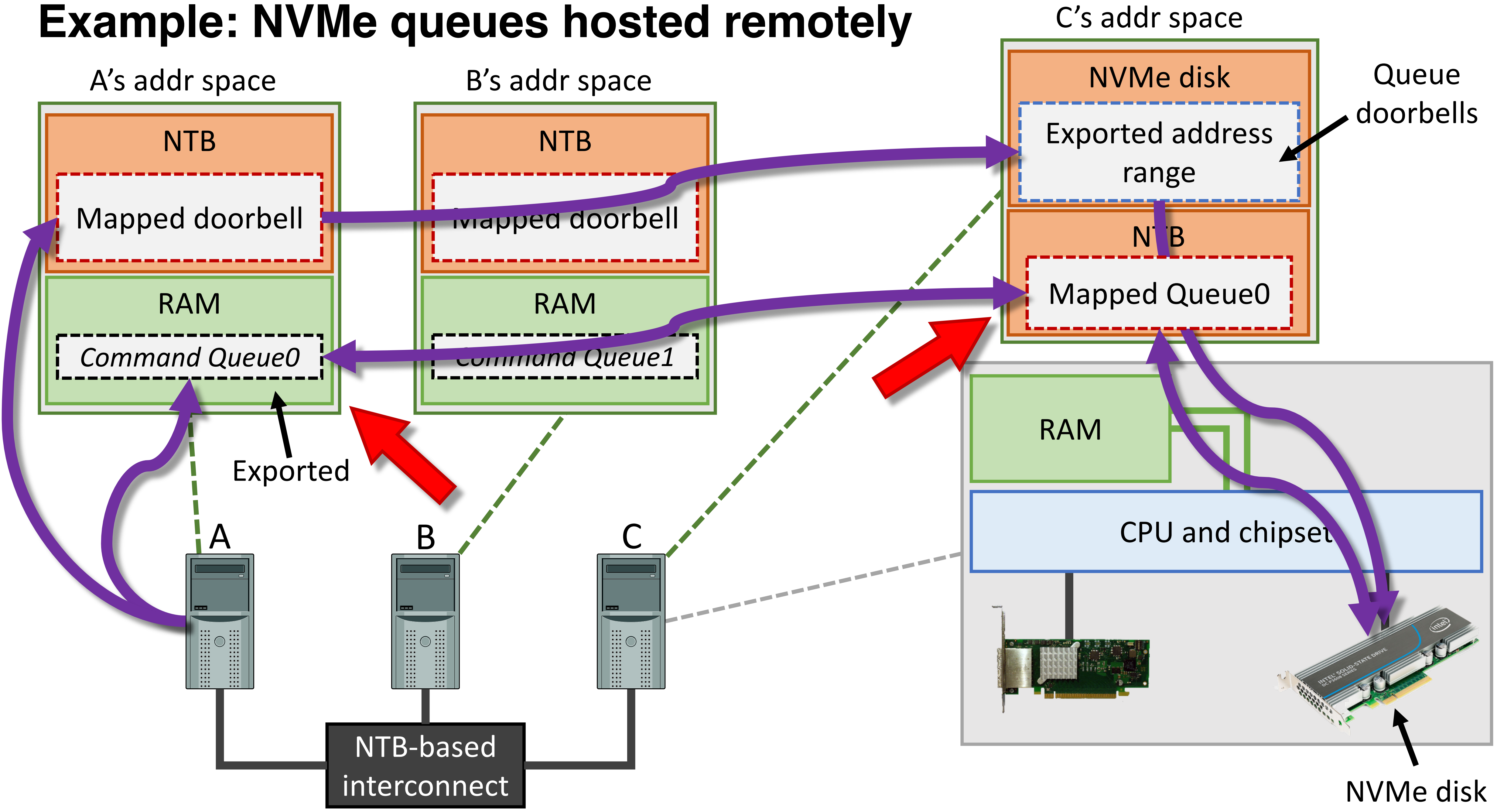
```
buf = cudaMalloc(...);
addr = nvidia_p2p_get_pages(buf);
```

```
ptr = mmap(...);
devptr = cudaHostRegister(ptr);
```

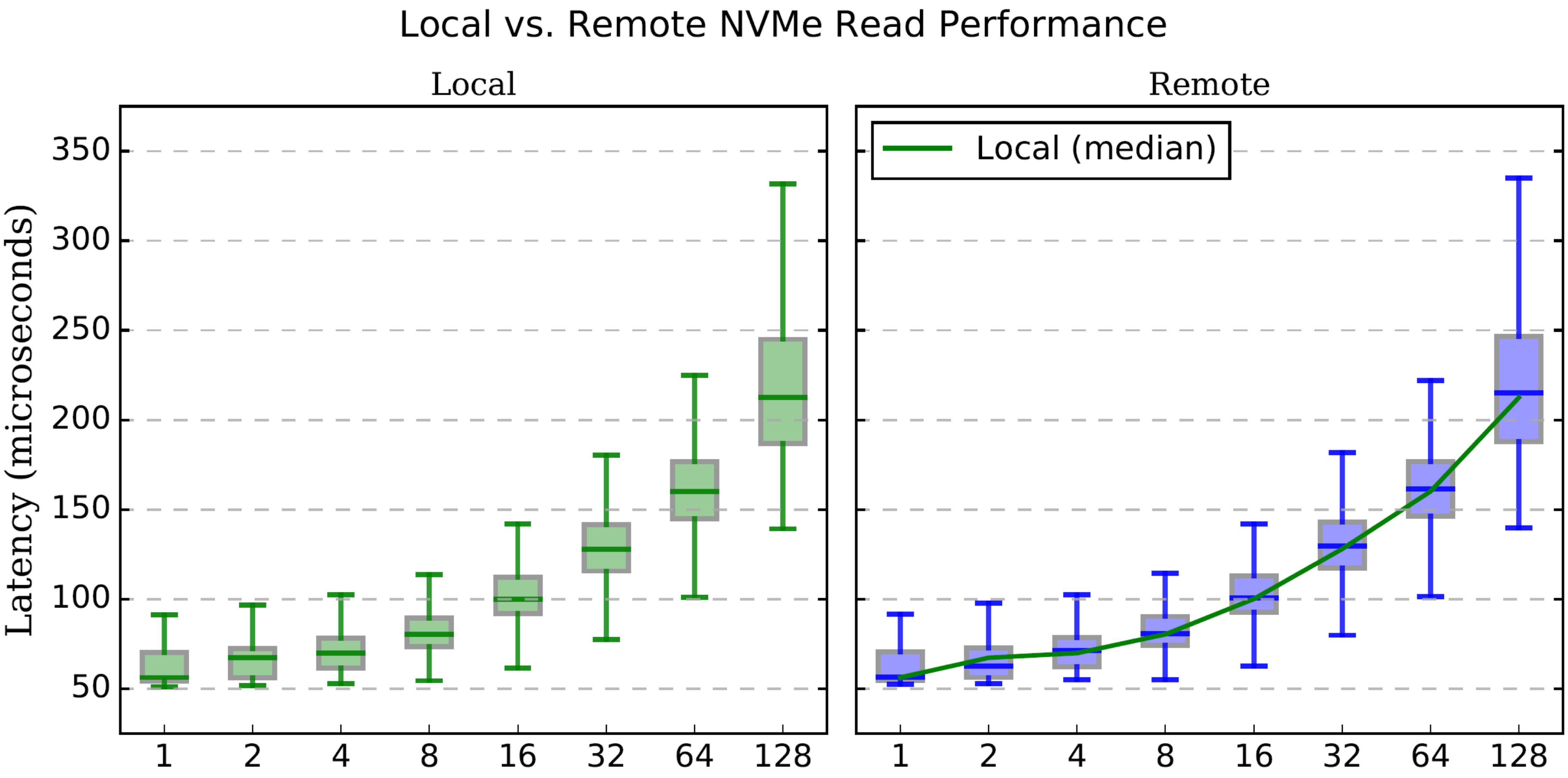
Userspace NVMe driver using GPUDirect

https://github.com/enfiskutensykkel/ssd-gpu-dma

Example: NVMe queues hosted remotely



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



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)

"PCIe Device Lending"
University of Oslo 2015

My email address

haakonks@simula.no

SmartIO & Device Lending demo and more

Visit Dolphin in the exhibition area (booth 523)

simula Dolphin INTERCONNECT SOLUTIONS UNIVERSITAS OSLOENSIS MDCCCXI