Flash Memory Summit

Accelerating NVMe to GPU Transfer Rates Using PCIe Fabrics and Logical Volumes

Vincent Haché

Technical Staff, Applications Engineering

Accelerating NVMe to GPU Transfer Rates Using PCIe Fabrics and Logical Volumes

- A critical performance metric in every AI/ML workload is the rate at which training or inference data can be transferred to GPU memory
- Systems need the right infrastructure and storage technologies to ensure workloads aren’t limited by storage performance

Accelerating NVMe to GPU Transfer Rates Using PCIe Fabrics and Logical Volumes

- Logical storage volumes (RAID) deliver maximum instantaneous bandwidth to NVMe storage
- PCIe fabrics enable optimal routing for peer-to-peer transfers, decreasing root port congestion and surpassing CPU performance bottlenecks

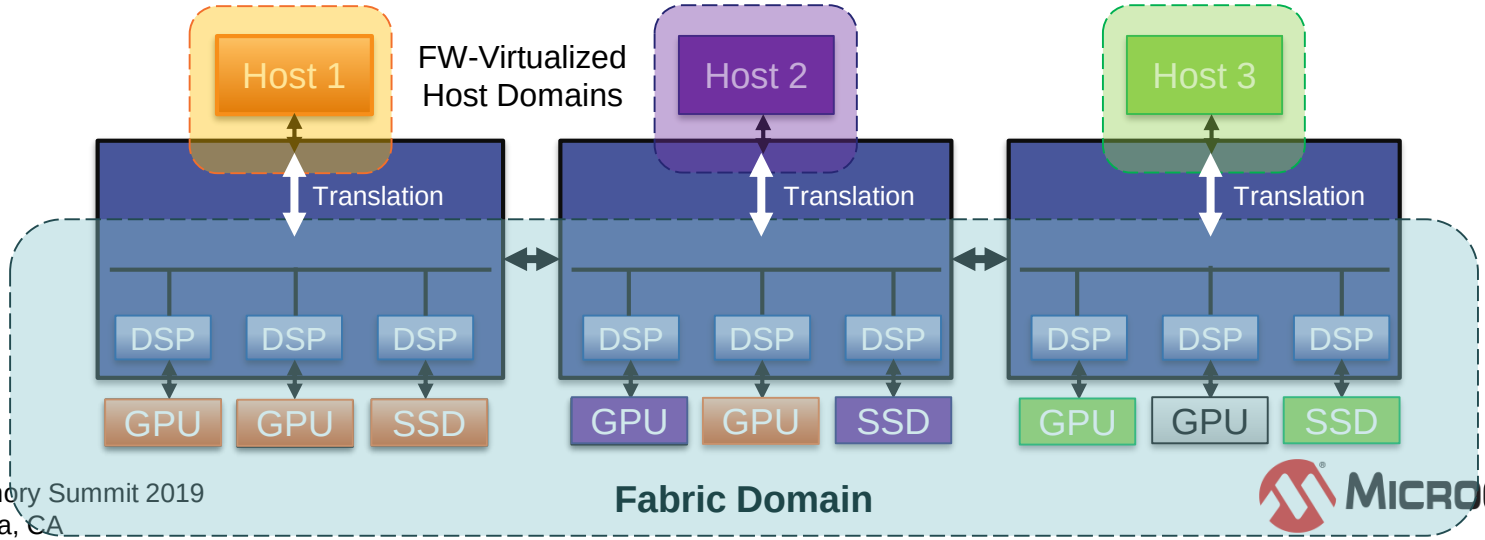
Accelerating NVMe to GPU Transfer Rates Using PCIe Fabrics and Logical Volumes

PCIe Fabrics + High-Performance NVMe Logical Volumes:

- Combined these technologies provide considerable improvements to GPU file accesses

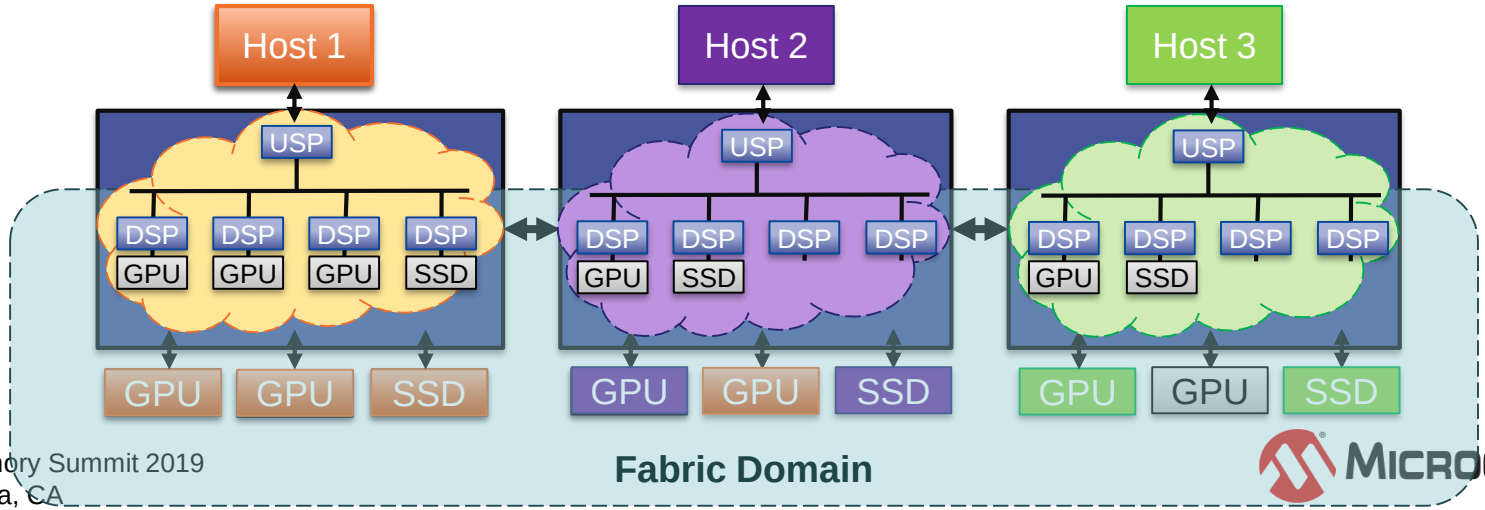
PCIe Fabrics Overview

- PCIe Fabrics span multiple switches and EPs
- Hosts are kept in separate virtual domains



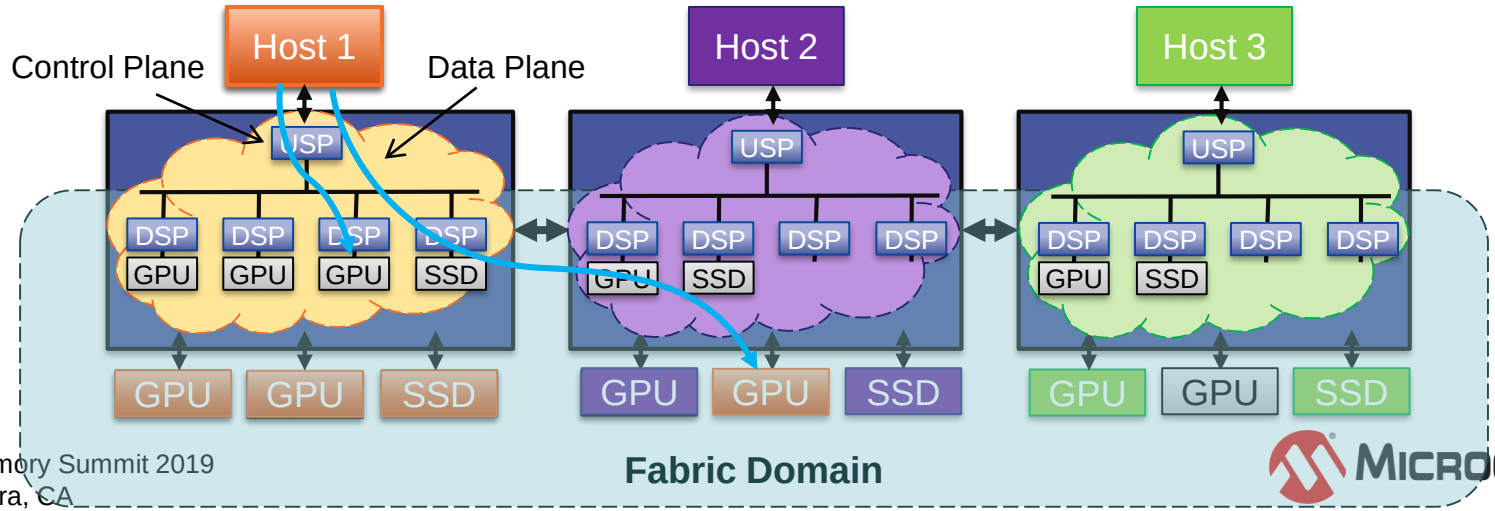
PCIe Fabrics Overview

- FW running on embedded CPU virtualizes simple, PCIe spec-compliant switch



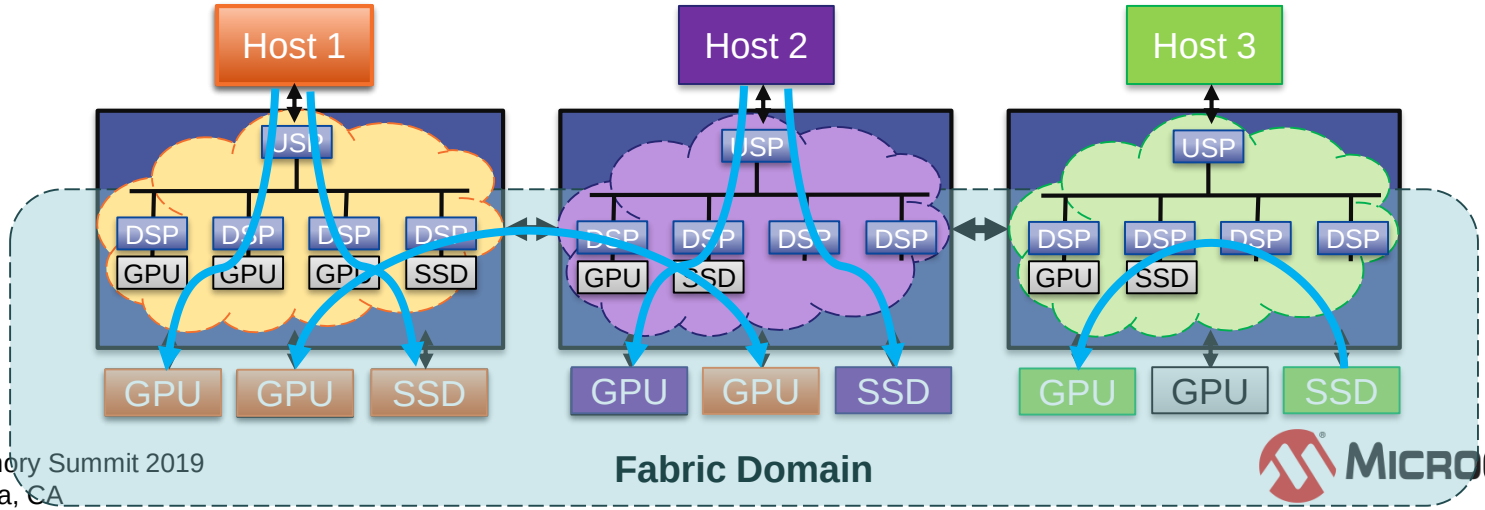
PCIe Fabrics Overview

- Embedded CPU handles control plane, but data is routed directly by switch HW



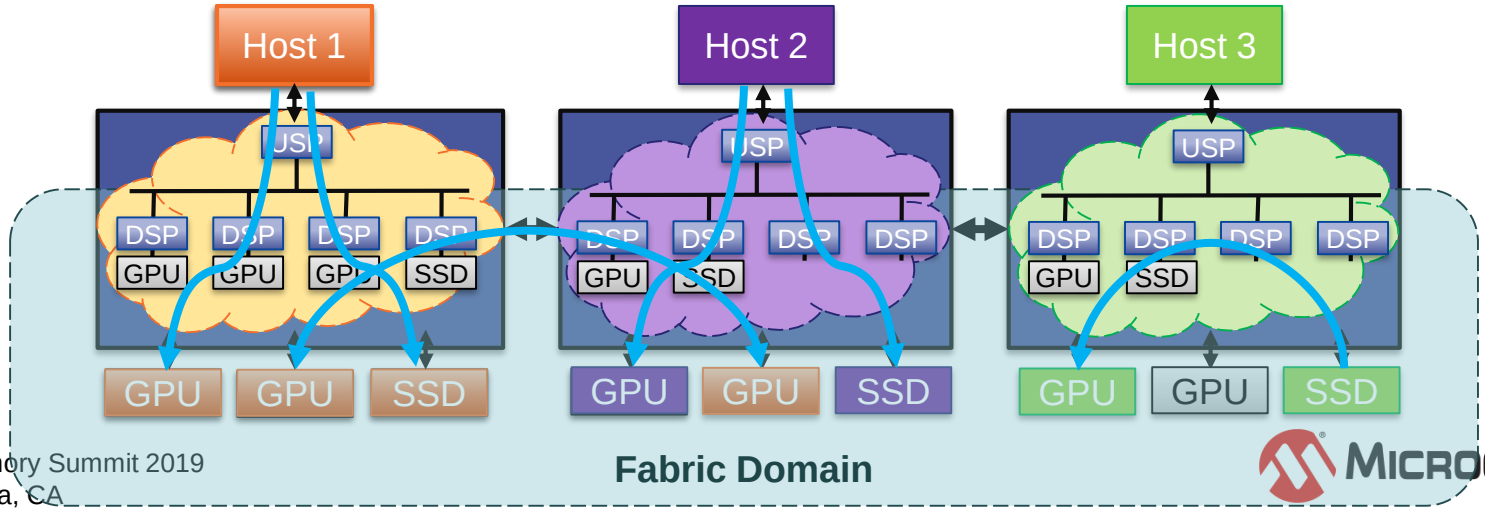
PCIe Fabrics Overview

- Fabric links are shared among hosts
- P2P transfers routed through fabric domain



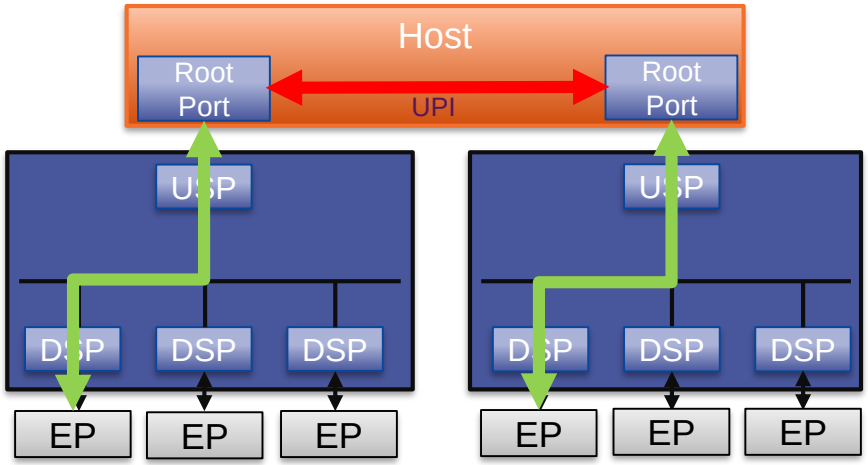
PCIe Fabrics Overview

- Fabric routing is proprietary, non-hierarchical
- No restrictions on fabric configuration



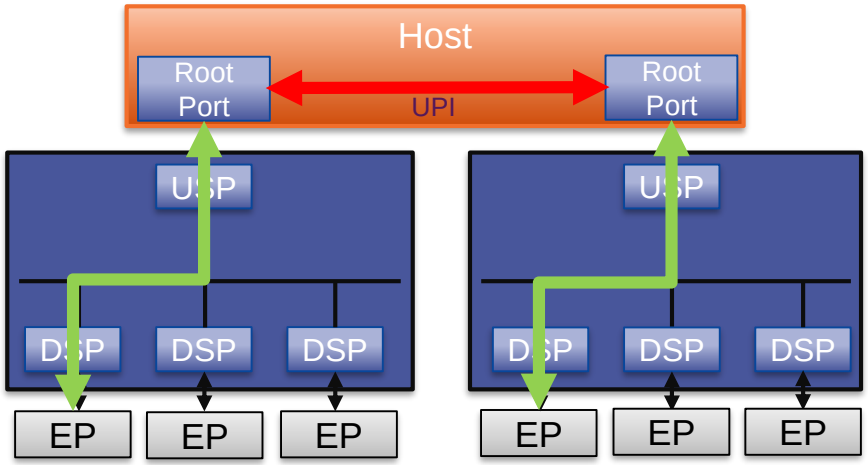
Standard PCIe Routing Restrictions

- No spec support for loops or redundant paths
- Spec-compliant routing passes through UPI



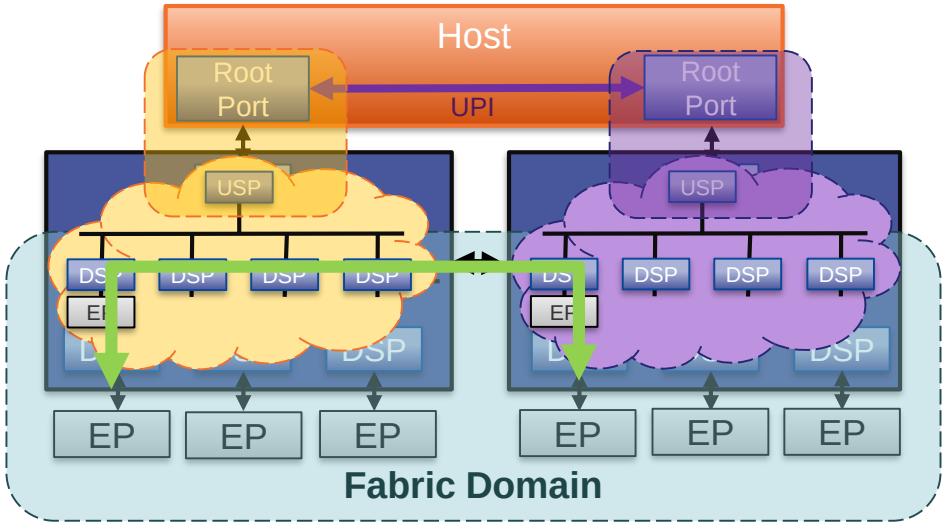
Standard PCIe Routing Restrictions

- P2P rates limited by UPI and RC packet sizes
- Root port congested with P2P traffic



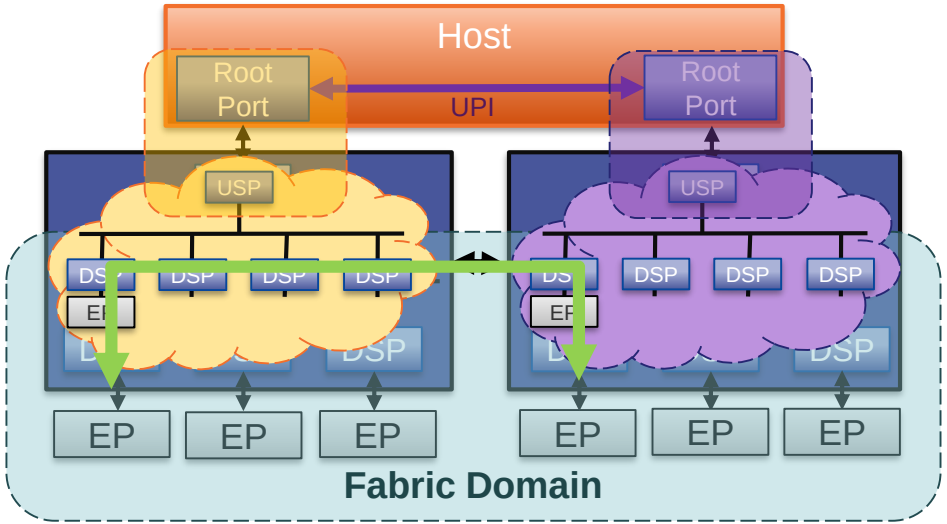
PCIe Fabrics for Performance

- PAX's "Cross-Domain P2P" feature bridges host domains for multi-root hosts

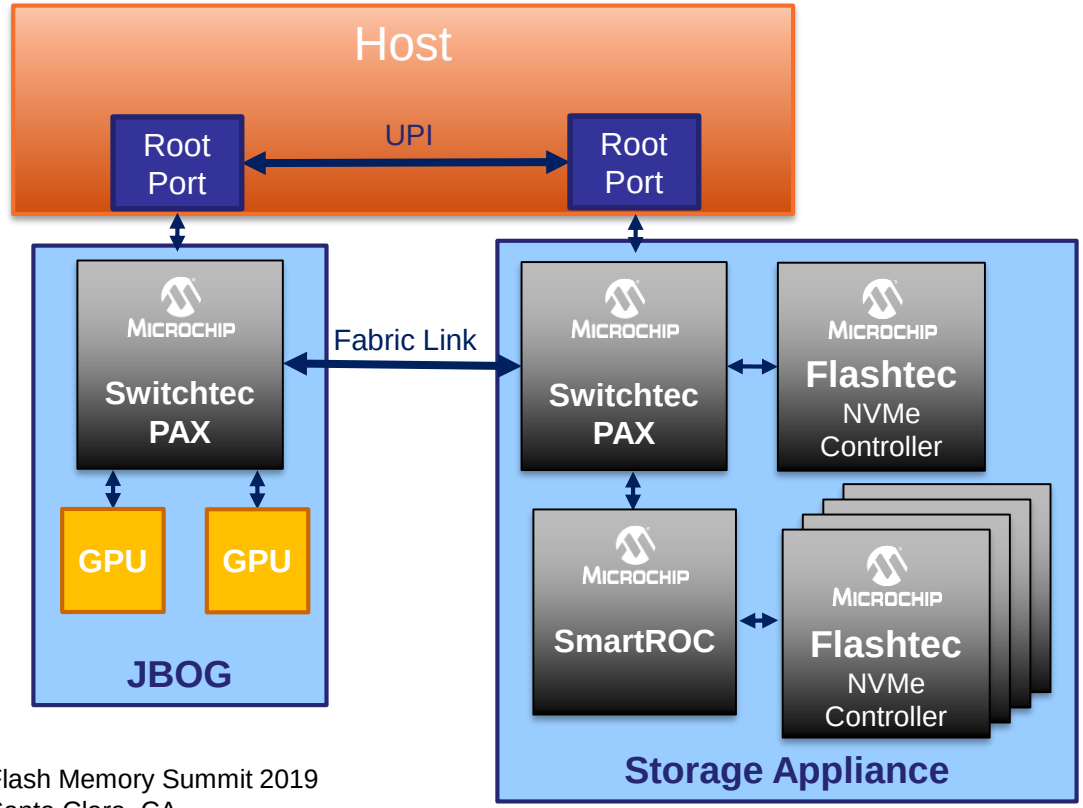


PCIe Fabrics for Performance

- P2P traffic is routed through the fabric, decreasing congestion and bypassing UPI



Demo – PAX Cross-Domain P2P



- Dual-Socket Intel Sky Lake host running Ubuntu Server 18.04
- 2 PAX Gen 4 switches
- 2 Nvidia Tesla T4 GPUs
- SmartROC Gen 4 Tri-Mode Controller
- 9 Gen 3 Flashtec-based NVMe SSDs

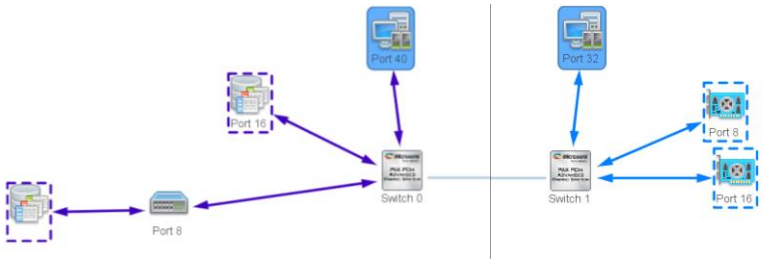
Demo – PAX Cross-Domain P2P

Each socket sees a simple PCIe switch

```
root@esdapps-p2p3:~# lspci -tv
-+-[0000:d7]-+-00.0-[d8]--
 |           +-05.0  Intel Corporation Device 2034
 |           +-05.2  Intel Corporation Sky Lake-E RAS Configuration Registers
 |           +-05.4  Intel Corporation Device 2036
 |           +-0e.0  Intel Corporation Device 2058
 |           +-0e.1  Intel Corporation Device 2059
 |           +-0f.0  Intel Corporation Device 2058
 |           +-0f.1  Intel Corporation Device 2059
 |           +-10.0  Intel Corporation Device 2058
 |           +-10.1  Intel Corporation Device 2059
 |           +-12.0  Intel Corporation Sky Lake-E M3KTI Registers
 |           +-12.1  Intel Corporation Sky Lake-E M3KTI Registers
 |           +-12.2  Intel Corporation Sky Lake-E M3KTI Registers
 |           +-12.4  Intel Corporation Sky Lake-E M3KTI Registers
 |           +-12.5  Intel Corporation Sky Lake-E M3KTI Registers
 |           +-15.0  Intel Corporation Sky Lake-E M2PCI Registers
 |           +-16.0  Intel Corporation Sky Lake-E M2PCI Registers
 |           +-16.4  Intel Corporation Sky Lake-E M2PCI Registers
 |           \-17.0  Intel Corporation Sky Lake-E M2PCI Registers
 +-[0000:ae]-+-00.0-[af-b5]----00.0-[b0-b5]--+-00.0-[b1]----00.0  Adaptec Smart Storage PQI 12G SAS/PCIe
 |           |                               +-01.0-[b2]----00.0  Micron Technology Inc Device 51b1
 |           |                               +-02.0-[b3]--
 |           |                               +-03.0-[b4]--
 |           |                               \-04.0-[b5]--
 |           +-05.0  Intel Corporation Device 2034
 |           +-05.2  Intel Corporation Sky Lake-E RAS Configuration Registers
```

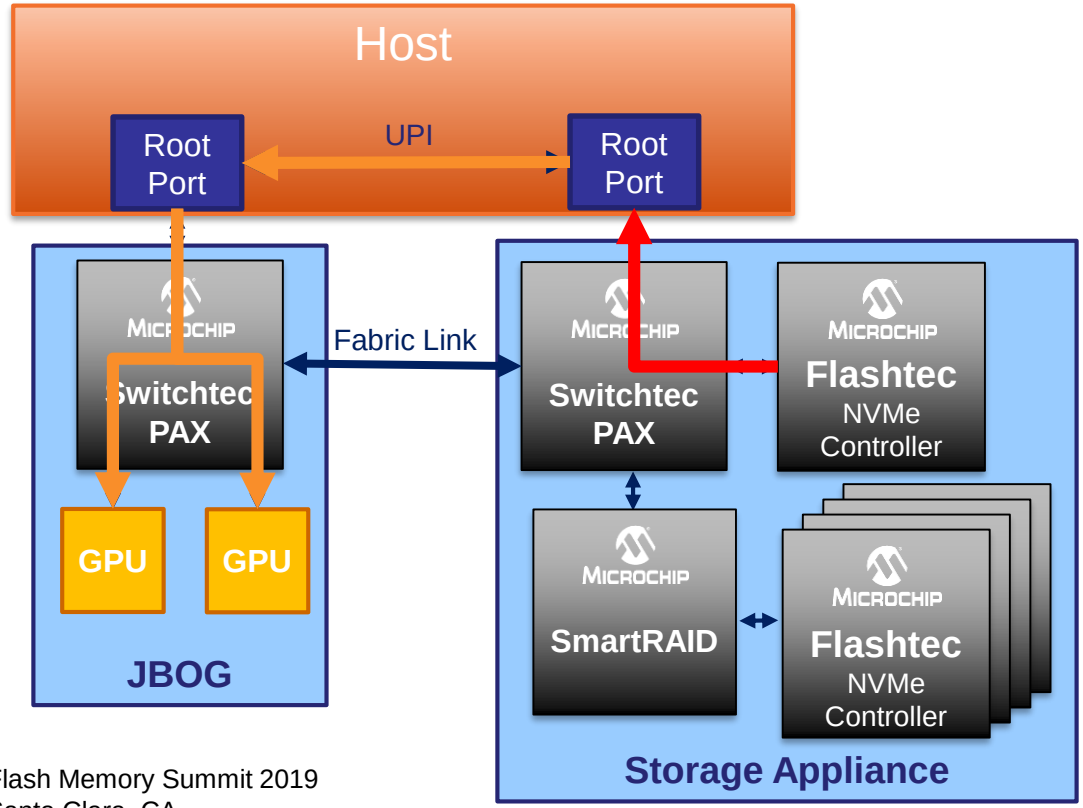
```
 |           +-12.5  Intel Corporation Sky Lake-E M3KTI Registers
 |           +-15.0  Intel Corporation Sky Lake-E M2PCI Registers
 |           +-16.0  Intel Corporation Sky Lake-E M2PCI Registers
 |           +-16.4  Intel Corporation Sky Lake-E M2PCI Registers
 |           \-17.0  Intel Corporation Sky Lake-E M2PCI Registers
 +-[0000:36]-+-00.0-[37-3d]--+-00.0-[38-3d]--+-00.0-[39]----00.0  NVIDIA Corporation Device 1eb8
 |           |               |               +-01.0-[3a]----00.0  NVIDIA Corporation Device 1eb8
 |           |               |               +-02.0-[3b]--
 |           |               |               +-03.0-[3c]--
 |           |               |               \-04.0-[3d]--
 |           |               \-00.1  PMC-Sierra Inc. Device 4100
 |           +-05.0  Intel Corporation Device 2034
```

GPUs bound under first socket, storage bound under second socket



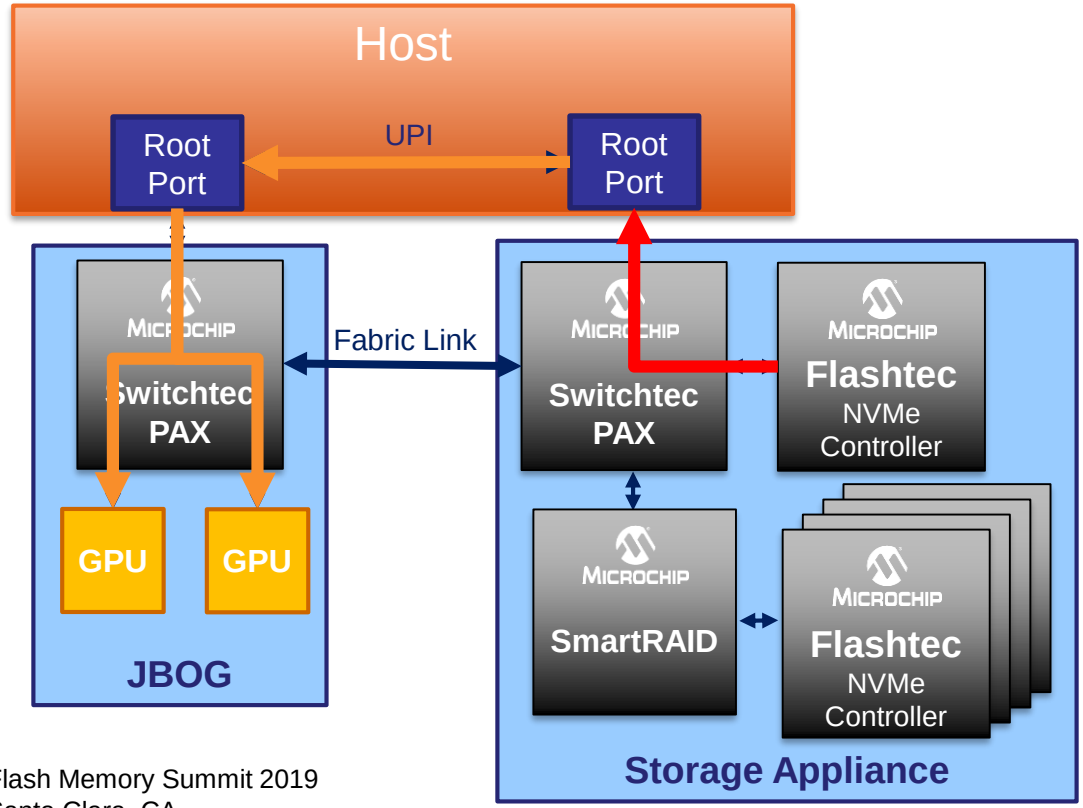
Only bound PCIe EPs are visible to host

Demo – PAX Cross-Domain P2P



- GPU running custom CUDA application to read directly from volume
- Transfer initiated from single NVMe SSD to GPU memory

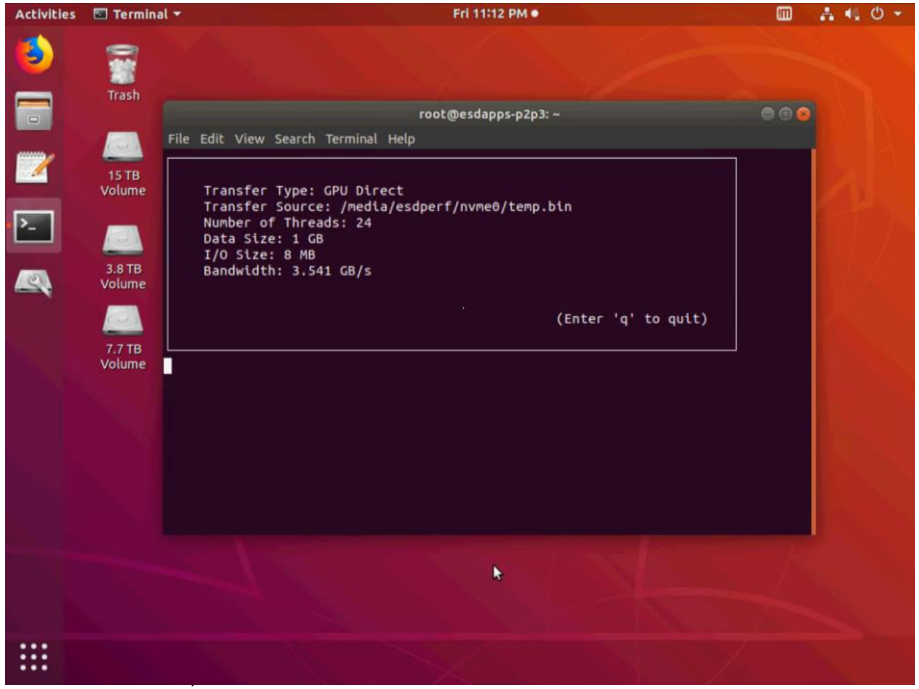
Demo – PAX Cross-Domain P2P



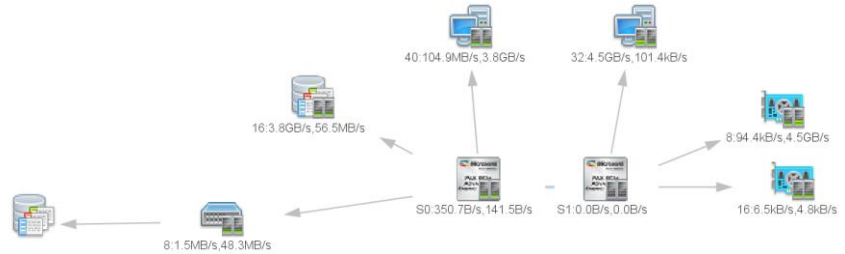
- Conventional NVMe usage limits file access to ***3.5 GB/s***
- Performance limited by single drive bandwidth

Demo – PAX Cross-Domain P2P

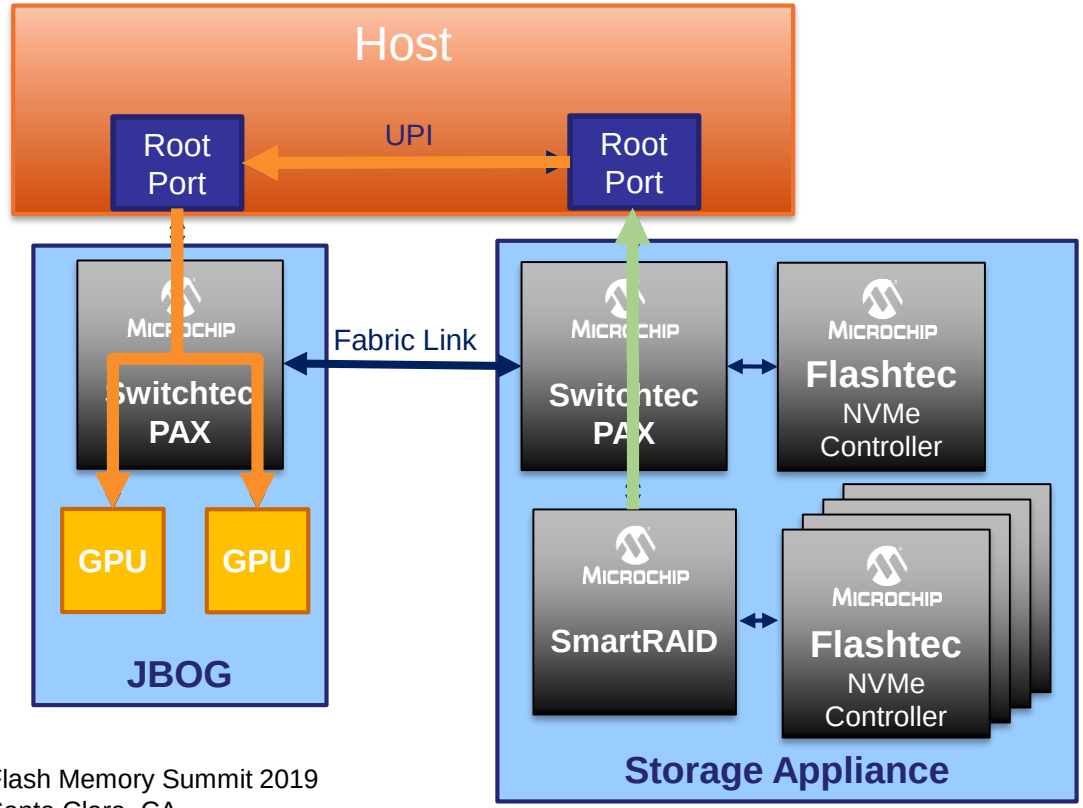
GPU reading data from NVMe SSD at **3.5 GB/s**



Traffic runs through root ports, **4.5 GB/s** raw bandwidth with added overhead

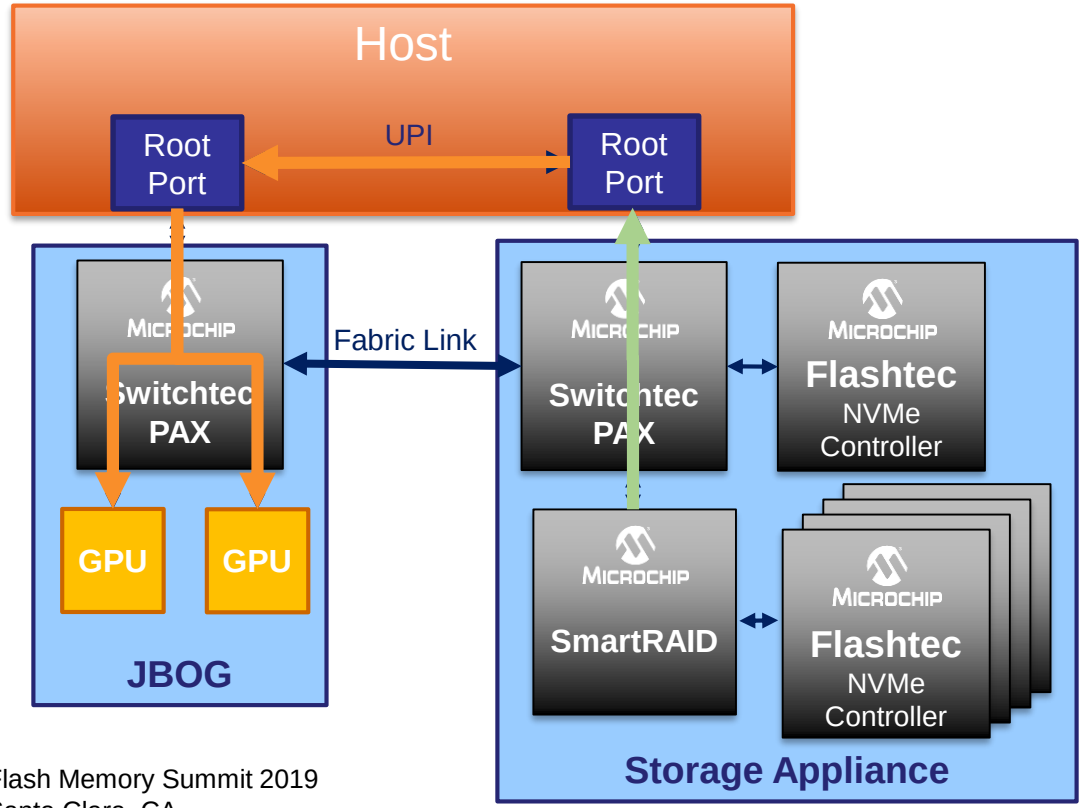


Demo – PAX Cross-Domain P2P



- SmartRAID creates 2 RAID volumes with 4 NVMe each
- With “Cross-Domain P2P” disabled, traffic runs over UPI

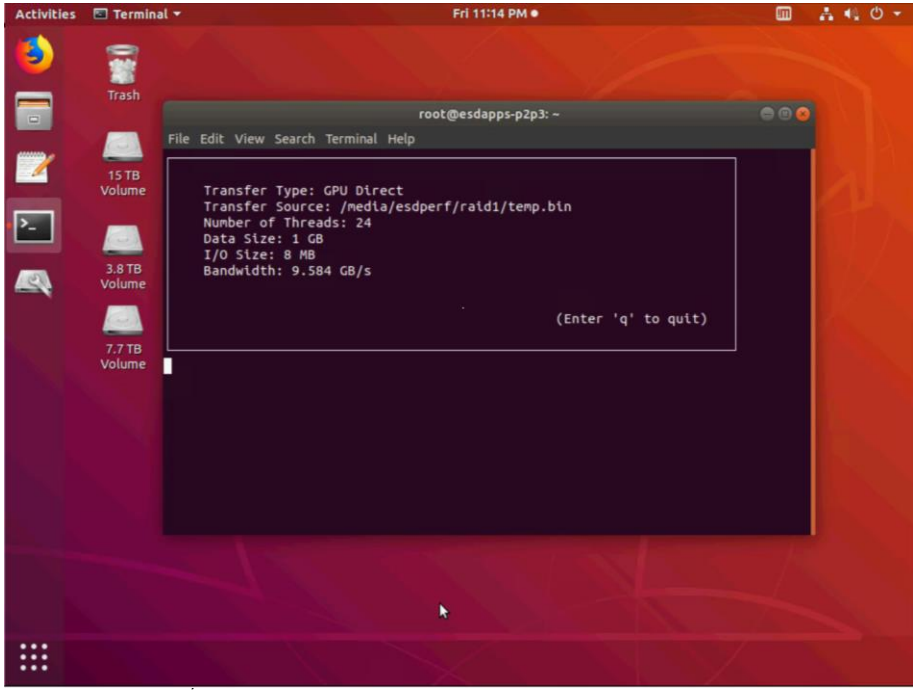
Demo – PAX Cross-Domain P2P



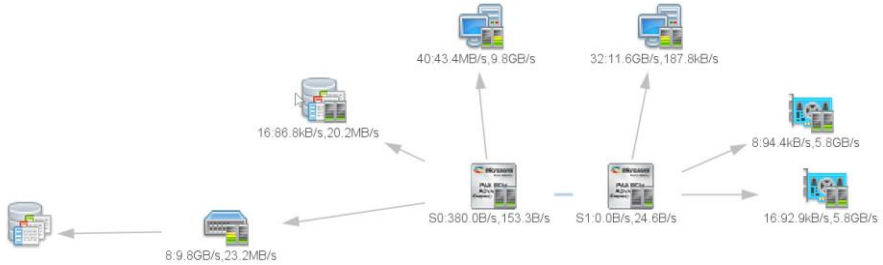
- RAID aggregates NVMe BW, but UPI limits multi-root P2P to ***9.5 GB/s***
- Root ports congested with P2P, TLPs broken into 64B payloads

Demo – PAX Cross-Domain P2P

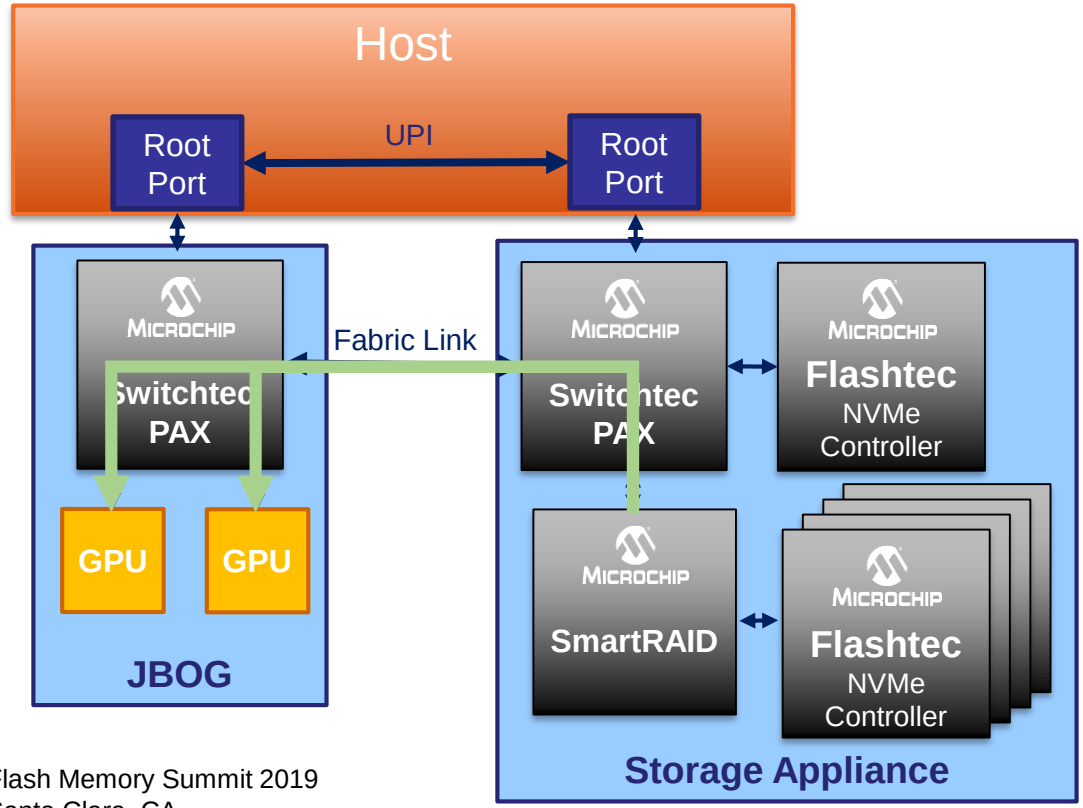
GPU reading data from RAID volume at **9.5 GB/s** aggregate BW



Traffic runs through root ports, **11.5 GB/s** raw bandwidth with added overhead



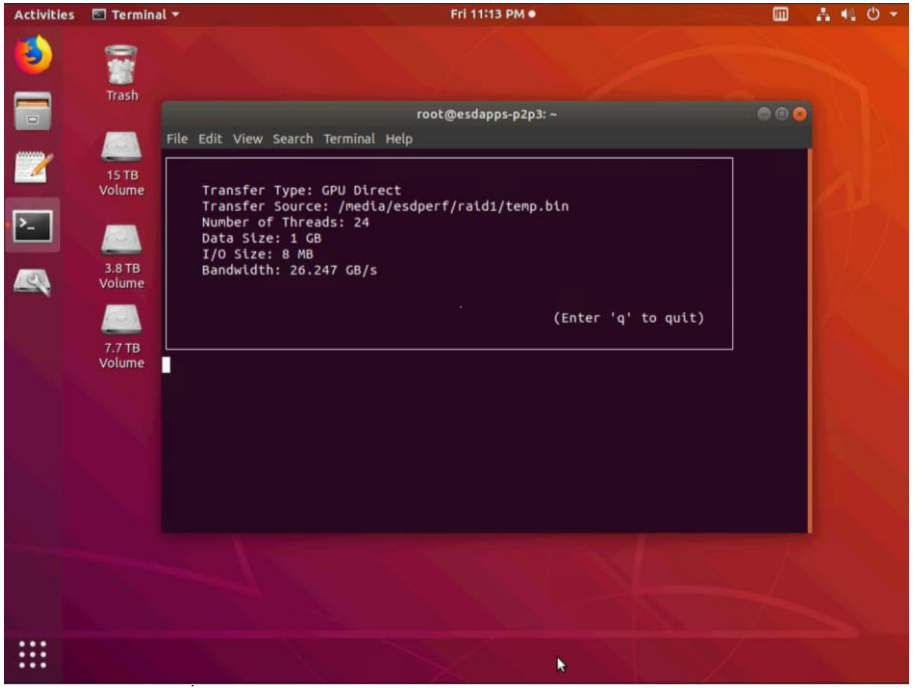
Demo – PAX Cross-Domain P2P



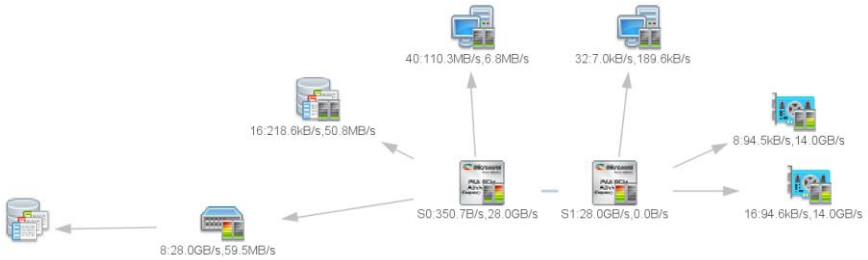
- "Cross-Domain P2P" is enabled
- P2P traffic routed to fabric link, easing root port congestion, surpassing UPI at ***26 GB/s***

Demo – PAX Cross-Domain P2P

GPU now reading data from RAID volume at **26 GB/s** aggregate BW



Root port traffic drops to just GPU control commands



Fabric link passing **28 GB/s** raw bandwidth

Demo – PAX Cross-Domain P2P

7x

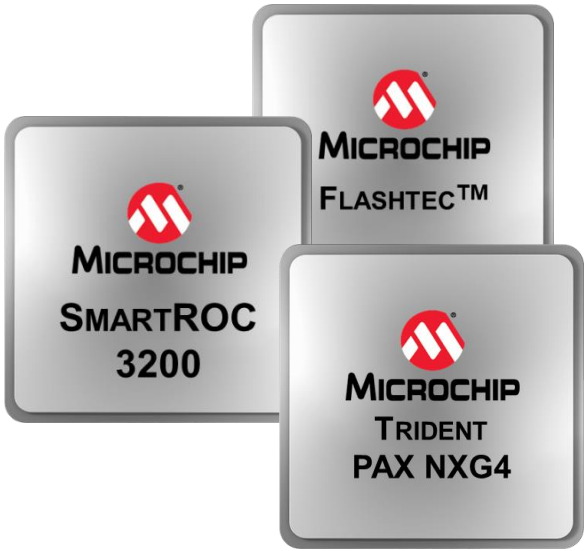
Increase in Transfer Rates

- Aggregate NVMe performance and protect vital training data with High Bandwidth, Low Latency RAID
- Increase PCIe P2P rates and decrease Root Port Congestion with PCIe Fabrics

Summary

Accelerating NVMe to GPU transfers:

- Logical Volumes aggregate NVMe performance to dramatically improve storage bandwidth
- PCIe fabrics enable new architectures and peer-to-peer routing paths for next-gen server designs



Live Demo at Booth #601