Multi-Host Sharing of NVMe Drives and GPUs Using PCIe Fabrics

Vincent Haché
Principal Applications Engineer,
Microsemi Corporation
Introduction

- Increase in use of GPUs and NVM in DC
- System designers need:
  - Efficient resource deployment
  - High-BW, low-latency interconnect
  - Flexible, composable architectures
- There are restrictions in standard PCIe that present challenges for system design
PCIe Hierarchy Restriction

- PCIe hierarchy is restrictive, making scale out challenging
PCIe Hierarchy Restriction: Host 1
PCle Hierarchy Restriction: Host 2

[Diagram showing a hierarchical structure with Host 2 at the top, followed by USP, DSP, and EP levels branching out below.]
PCle Hierarchy Restriction: Host 3
The multiple links required for transparent scale out complicates design and decreases efficiency.
PCIe Single Domain Restriction

- PCIe is single domain
  - Unused EPs are stranded
  - Complicated, non-standard NT drivers required for sharing
PCIe Fabrics for Scaling

- Fabric routing is proprietary, non-hierarchical
- Fabric links are shared among hosts
FW running on embedded CPU virtualizes a simple switch compliant with the PCIe spec

- Fabric Domain
PCIe Fabrics for Scaling (continued)

- Embedded CPU handles the control plane, but data is routed directly by switch HW.
Unused devices can be dynamically assigned (no longer stranded)

Low-latency, high-BW P2P within the rack

Standard drivers to simplify system development
Multi-host Sharing of SSDs

- Fabric resources assigned by function
- SR-IOV: EP appears as multiple functions
Demo: Multi-host Sharing of NVMe and GPUs

- Dynamic partitioning of GPUs and multi-host sharing of SR-IOV SSDs in real time
  - Standard host drivers in Windows Server 2016 and Ubuntu Server 16.04 LTS
- GPU P2P transfers across the fabric
  - CUDA P2P BW test and TensorFlow cifar10 image classification multi-GPU training algorithm
Demo: Multi-host Sharing of NVMe and GPUs (continued)

All GPUs are assigned to Host 1 to increase performance.

Multi-host sharing of SR-IOV NVM device.

Fabric managed through UART.

Flash Memory Summit 2018
Santa Clara, CA
Demo: Multi-host Sharing of NVMe and GPUs (continued)

Domain Virtualized as a Spec-compliant PCIe Switch

NVM VF Appears as a Standard NVM Device
Demo: Multi-host Sharing of NVMe and GPUs (continued)

CUDA P2P Bandwidth

Running Tensorflow Model
Demo: Multi-host Sharing of NVMe and GPUs (continued)

Host 1 workload completes and GPUs are released back into fabric pool

Host 1

PAX  

PAX  

PAX  

PAX  

Host 2

SSD

VF  

VF

GPU

GPU

GPU

GPU
Demo: Multi-host Sharing of NVMe and GPUs (continued)

Spare GPUs are assigned to Host 2
Demo: Multi-host Sharing of NVMe and GPUs

(continued)

Domain Virtualized as a Spec-compliant PCIe Switch

NVM VF Appears as a Standard NVM Device

Windows Host Still Running During Dynamic Reassignment
Demo – Multi-host Sharing of NVMe and GPUs

- PCIe spec-compliant host domain
- Simple management
- Standard drivers
- Dynamic reassignment appears as spec-compliant surprise-plug
Summary

Microsemi’s Switchtec PAX Switches enable new architectures for next-gen solutions

Benefits of PCIe fabrics with PAX:

- Scalable, low-latency, cost-effective
- Simple Management (PCIe, UART, TWI, Ethernet)
- Multi-host sharing of SR-IOV NVMe devices
- Standard host drivers

Live Demo at Booth #213