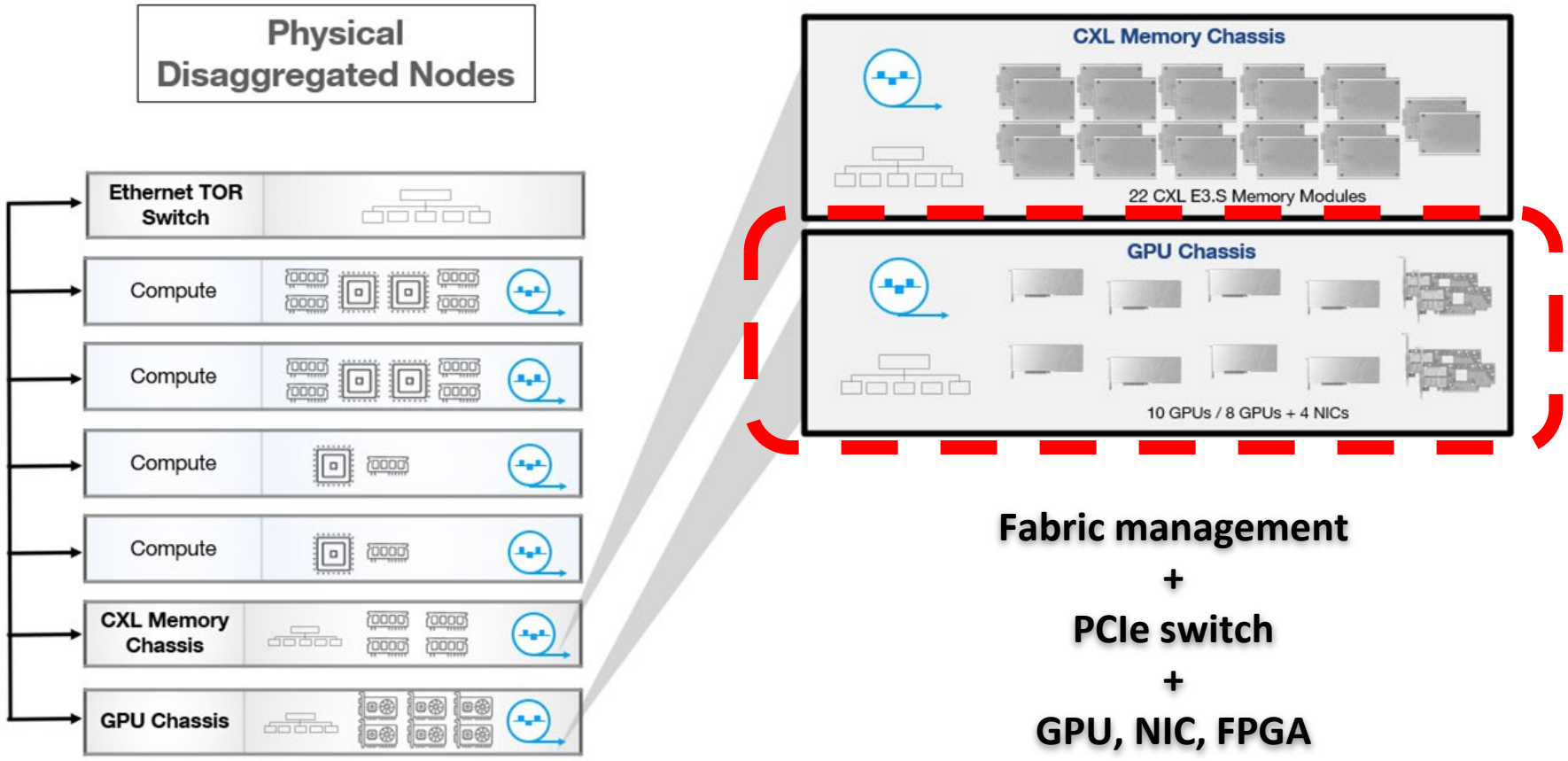
PCIe Gen5 Fabric and Management

Brian Pan
H3 Platform

FMS the Future of Memory and Storage

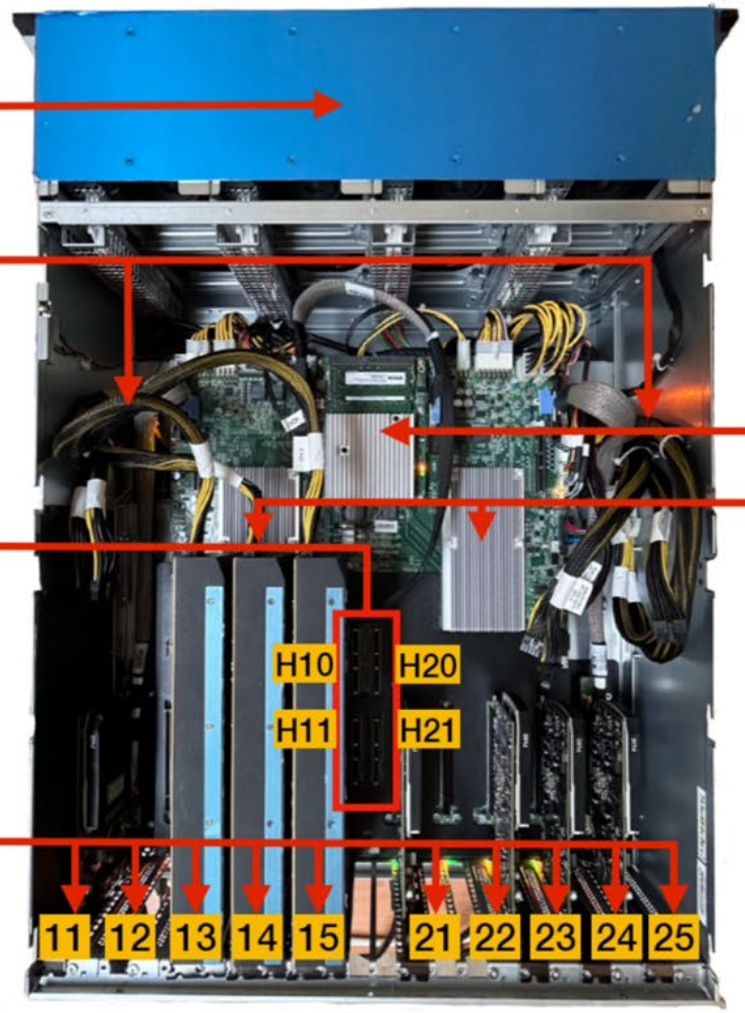
Table of Contents

Architecture: Composable PCIe System



Fabric management
+
PCIe switch
+
GPU, NIC, FPGA

Chassis Top View



8 System Fans

GPU Power Cables

BMC and mCPU
Chassis management
and Device management

PCIe Switches
2 Broadcom PCIe 5.0 switches

Host Ports
4 CDFP Connectors

Device slots
- **Configuration A:**
  8 dual slots for GPUs +
  4 slots for NICs
- **Configuration B:**
  10 dual slots for GPUs

Key Software Specification

Features

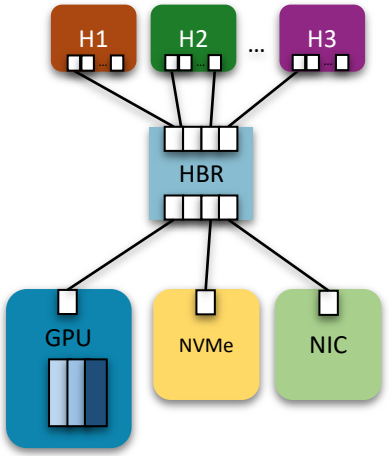
- GPU composability
- Device surprise add and remove
- Device peer-to-peer (GPU P2P)
- PCIe port configuration (Host or Device)
- Performance and error monitoring

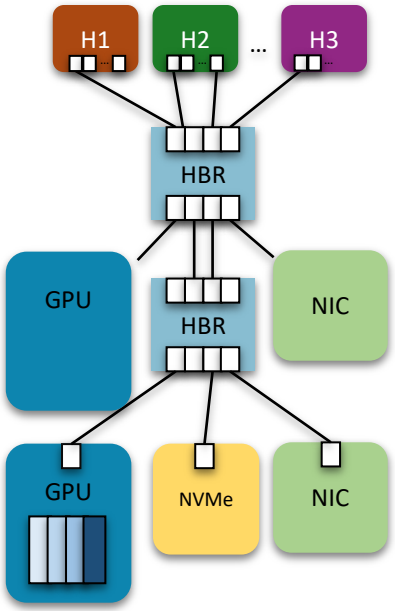
Management Interface

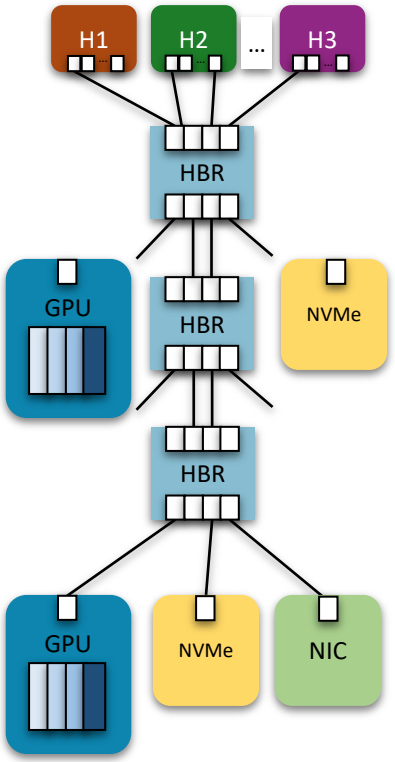
- Redfish®, RESTful API, GUI

Fabric Topology: PCIe Hierarchical Switches

The host can be connected to the first-layer switch ONLY.



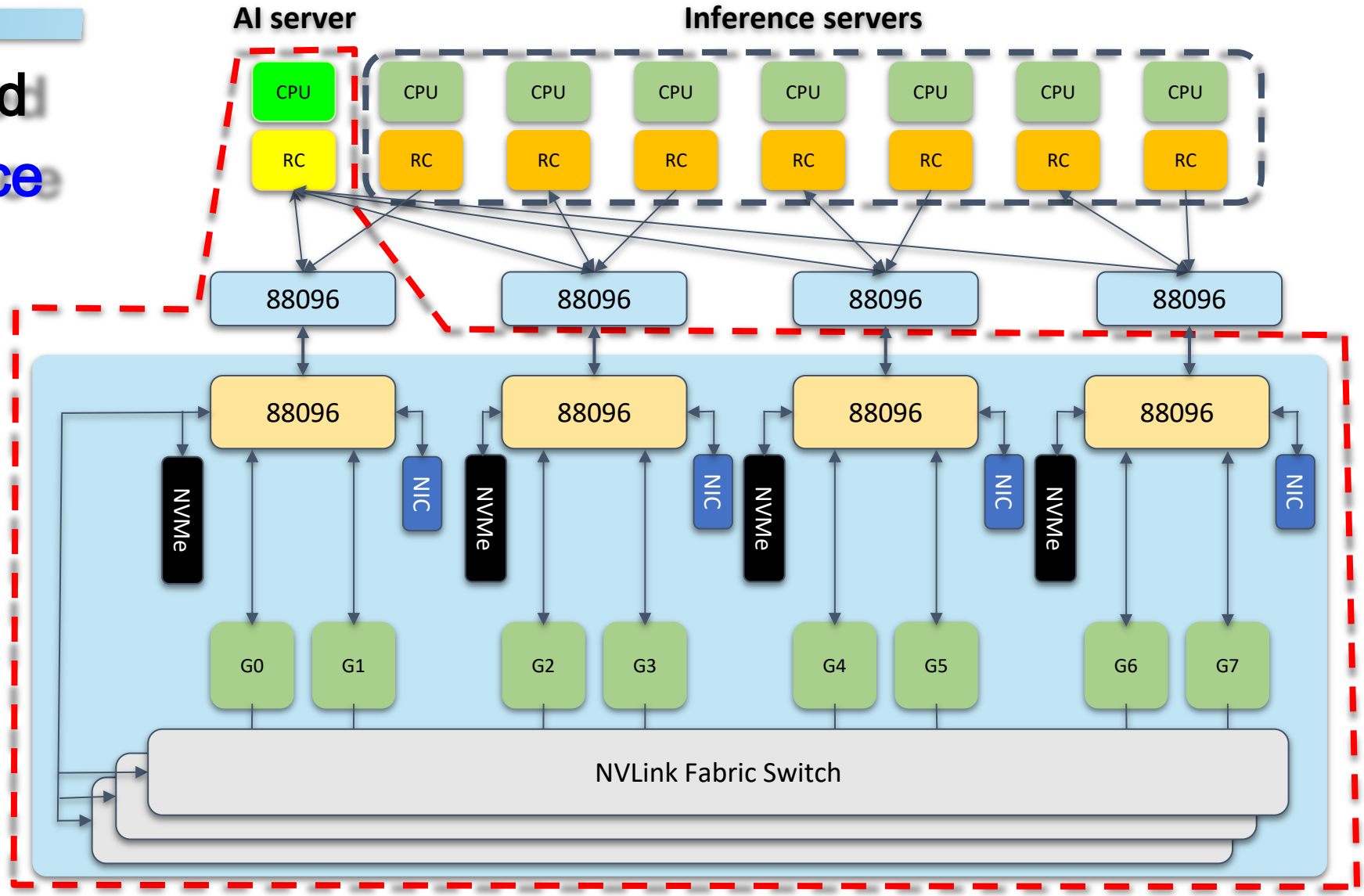




- **HBR**: Hierarchy-Based Routing Switch
- **GPU**: Graphic Processor Unit
- **NVMe**: NVMe SSD
- **Network card**: Ethernet/ Infiniband network card

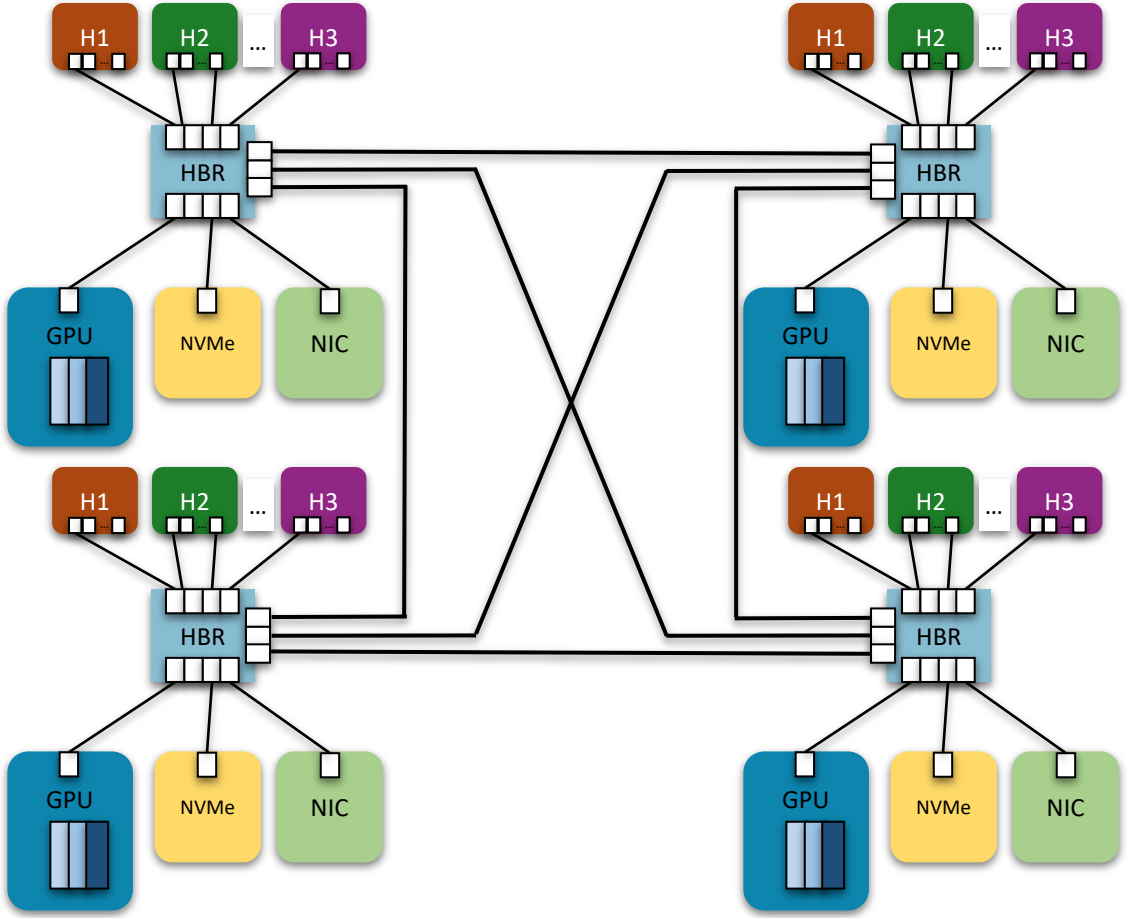
Usage Case of PCIe Hierarchical Switch: GPU for AI and Inference

The GPUs are used for AI and Inference in different time period of the day



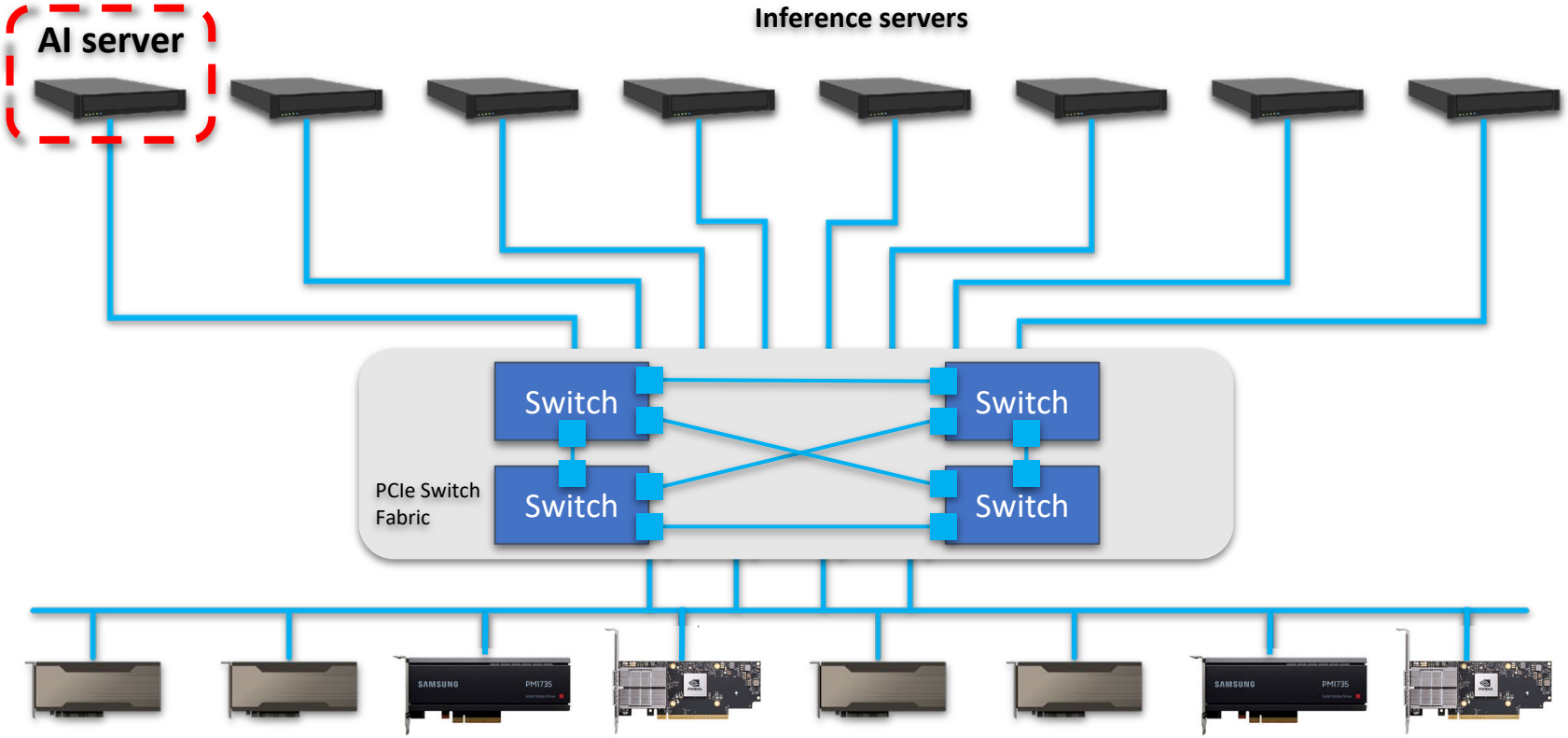
Fabric Topology: PCIe Hierarchical Switches with Fabric Capability

The host can be connected to any switch. The switches are in mesh topology.



- **PCIe Fabric Switch** : Hierarchy-Based Routing Switch with Fabric Capability

Usage Case of PCIe Hierarchical Switch with Fabric: Scale Up GPUs or PCIe Device Hub

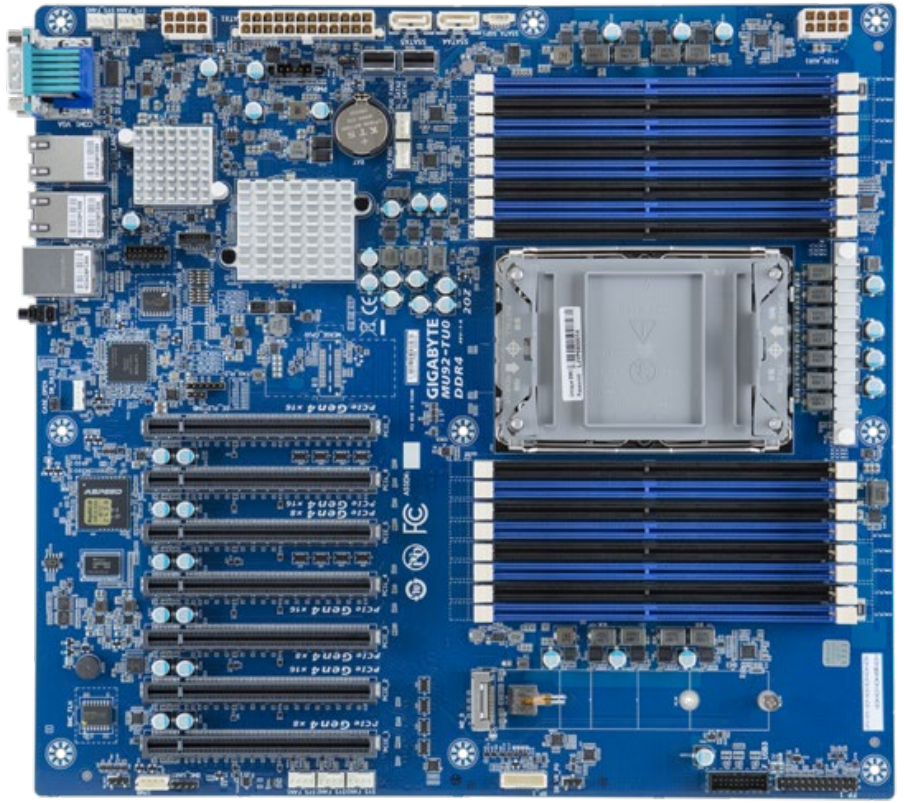


All the PCIe devices in the PCIe fabric can be assigned to any host.

Lessons Learned: Server

Server bus number and memory address

- Not enough bus number and memory addressin server slots
  a. BIOs bus allocation on each PCIe CEM slot
  b. BAR0, BAR1, BAR2 memory address



Lessons Learned: Re-timer

Signal integrity, bifurcation, reset, thermal

- Re-timer should test against with server slots for signal integrity
- Bifurcation, clock, and reset design
- High speed re-timer need extra cooling

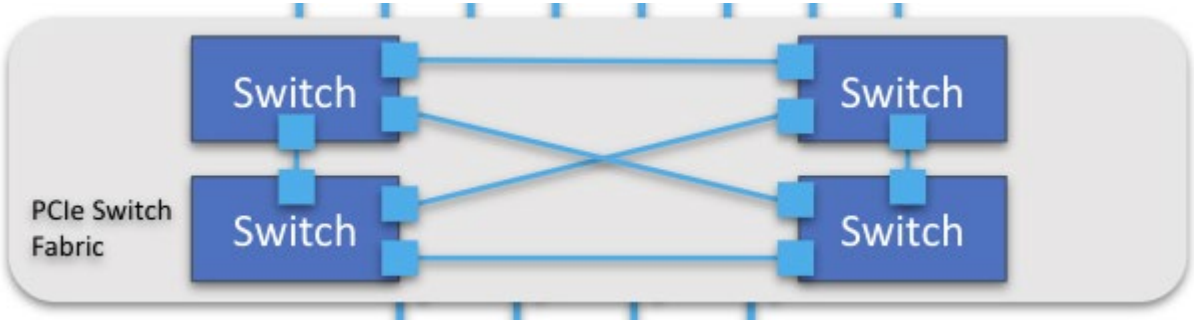
Lessons Learned: PCIe fabric

GPU reset when hotplug

- When re-provisioning the GPU, the GPU should be reset through out-of-band or inband secondary bus reset

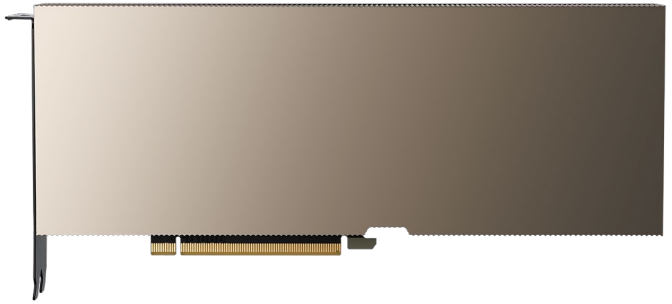
GPU peer to peer setting when re-provisioned

- GPU P2P setup in the PCIe switch fabric
  - Single GPU device
  - GPU with internal PCIe bridge
  - NVMe and NICs



Lessons Learned: Devices

PCIe device in one card

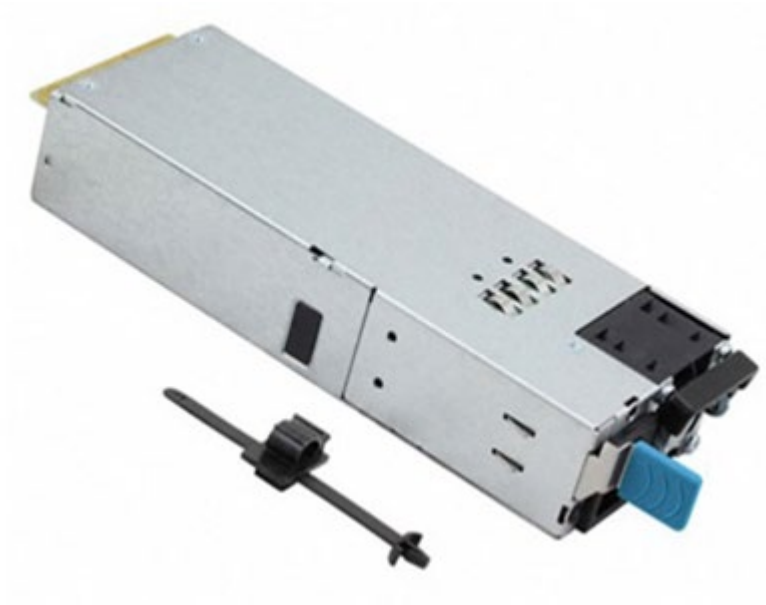


- Single GPU
- Multiple GPUs in a single device (GPU and PCIe switch)
- Multiple devices in a single device (GPU, PCIe switch, and NIC)

Lessons Learned: Devices

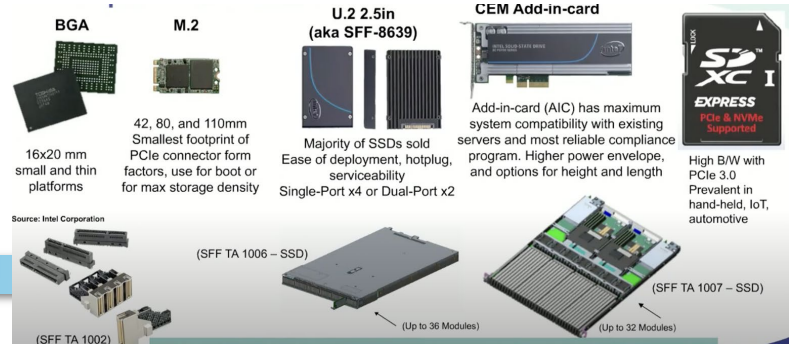
Device power and cooling

- GPU up to 1KW and the NVMe is only 20W



Form factor of PCIe device

- FHFL, HHFL, HHHL, U.2, E3.S..

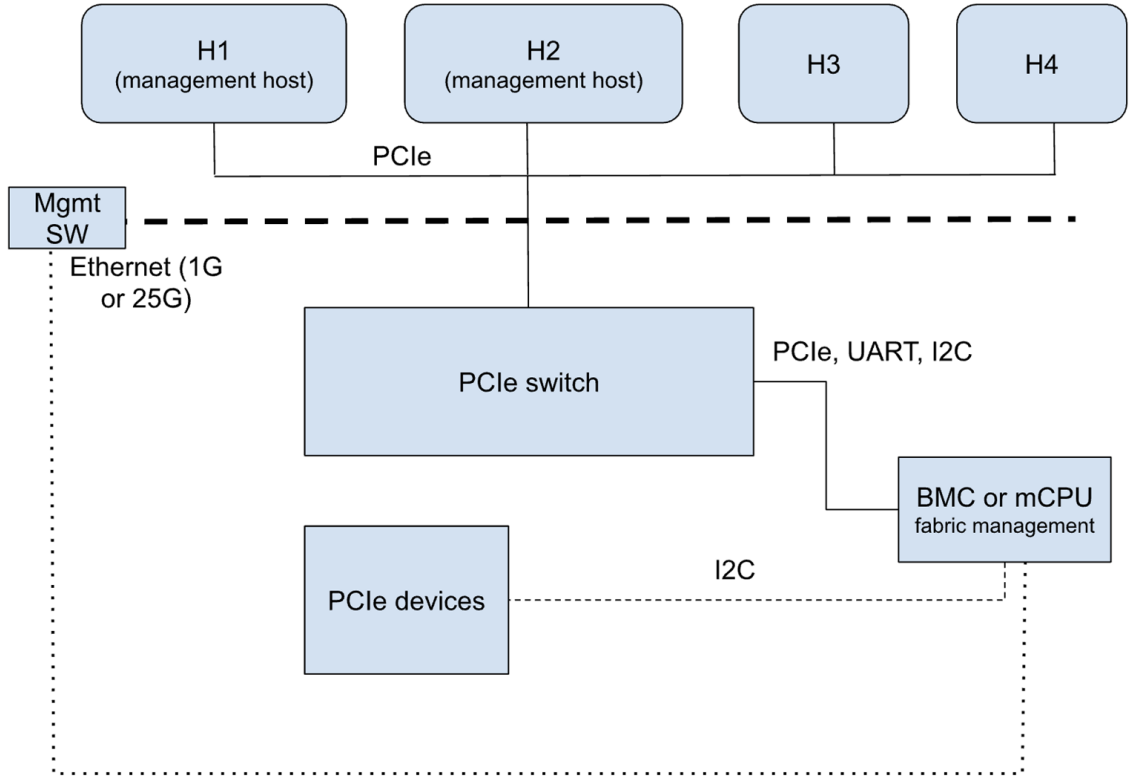


Lessons Learned: Management and API

Management path

PCIe device management path

- Ethernet (Data or Management path)
- PCIe management
- I2C out-of-band



Lessons Learned: Management and API

Orchestration and API

Many consortiums are working on the standard orchestration and API for the composable solutions