CXL is Coming to OCP

Siamak Tavallaei, Chief Systems Architect, Google
Jeff Dodson, Hardware Architect, Broadcom

EMPOWERING OPEN.

OCP GLOBAL SUMMIT
OCTOBER 18-20, 2022
SAN JOSE, CA

CXL is Coming to OCP

Siamak Tavallaei, Chief Systems Architect, Google
Jeff Dodson, Hardware Architect, Broadcom

Abstract

This Talk will explore CXL (Compute Express Link) opportunities for servers as CXL is getting ready for prime time! It will cover a brief set of use-cases along with HW and SW considerations based on the well-known PCIe plus new use cases which drive new aspects such as module Security, RAS, CXL Fabric Manager. This Talk will offer a glimpse into getting ready for Compute Disaggregation with CXL.

Outline

CXL: industry-standard, multi-host protocol that expands PCIe and enables cache semantics for memory transfers

- Takes advantage of PCIe investment: Same PCIe electricals, programming, and DMA
- Can start out with just PCIe
- Brief summary of PCIe TLPs, UIO, .cache, and .mem messages

Option to scale to multiple CXL Switches of connectivity

- Various topologies (include multi-headed CXL Device connected to multiple CXL Switches)
- PBR ID overview
- PCIe tree baseline

Use cases

- Memory expansion (covered during earlier sessions)
- Peer-to-peer (directly via CXL Switch)
- Host-to-Host via Shared Memory

Larger Systems

- In this session, we are presenting the larger systems based on multi-host-capable CXL Switches and the notion of CXL Fabric
- Based on the soon-to-be-available CXL memory controllers and Devices
- Multi-ported CXL Devices may connect to multiple CXL Switches

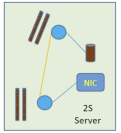
Challenges

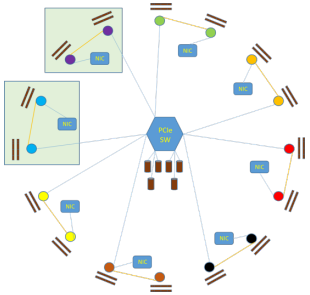
- Dive into what it will take for a system designer to think about the challenges of building a large system! (a multi-Host, multi-Device system, ... within one Chassis, and beyond... to a Rack)
- CXL Fabric Manager and Dynamic Capacity Device (DCD) concepts have been covered ahead of this Talk
- End-to-end Security, Fabric Management, etc.
- The interconnect (backplane, cables, ...)
- Topology
- Fault-tolerance and High-availability (HA) requirements (First, do no harm!)

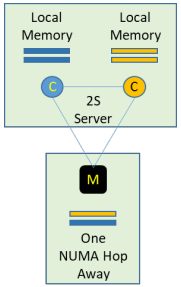
CXL Enables Extensible Solutions

Topology

- Point-to-point
- Multi-port
- Switched





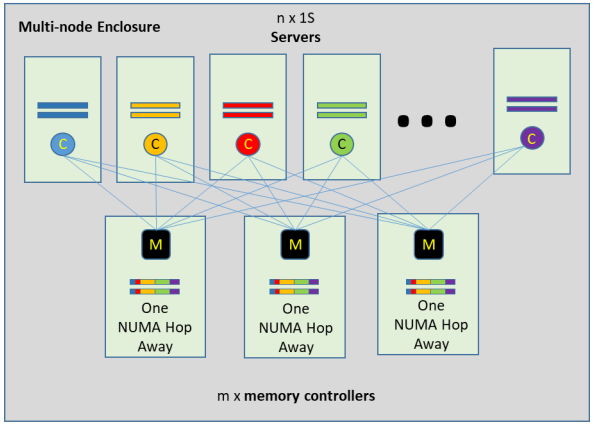


Density (multi-port)

- Dense packaging of (n x m) multi-ported Devices
- Liquid cooling

Reach (SERDES)

- Longer Links *(to all devices including memory!)*
- Modular Enclosures
- Cabled Solutions
- Photonics



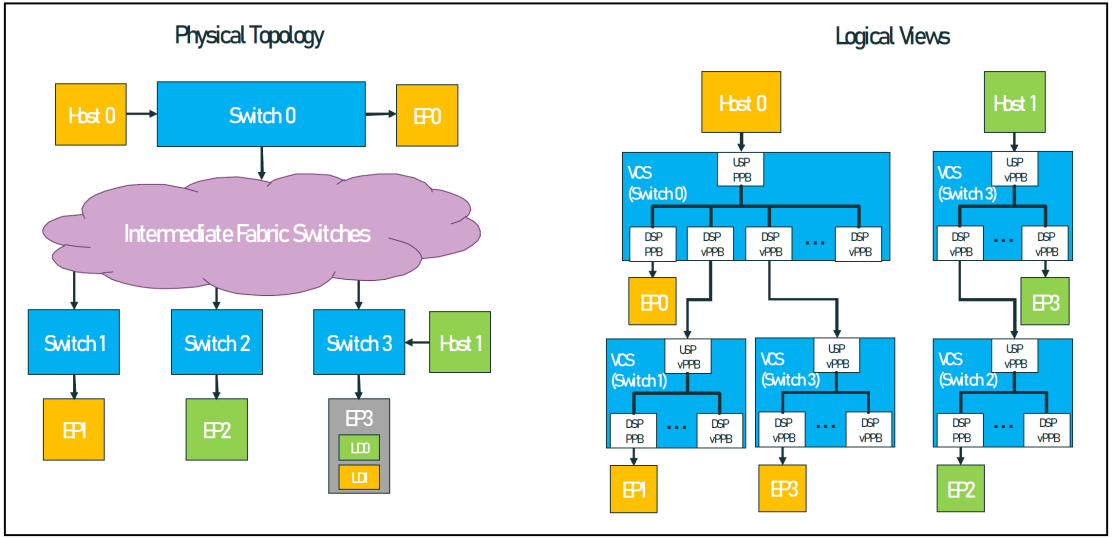
Capacity-matched
Bandwidth-matched

Extensibility (heterogeneous)

- Compute (xPU), Memory, Storage, Networking

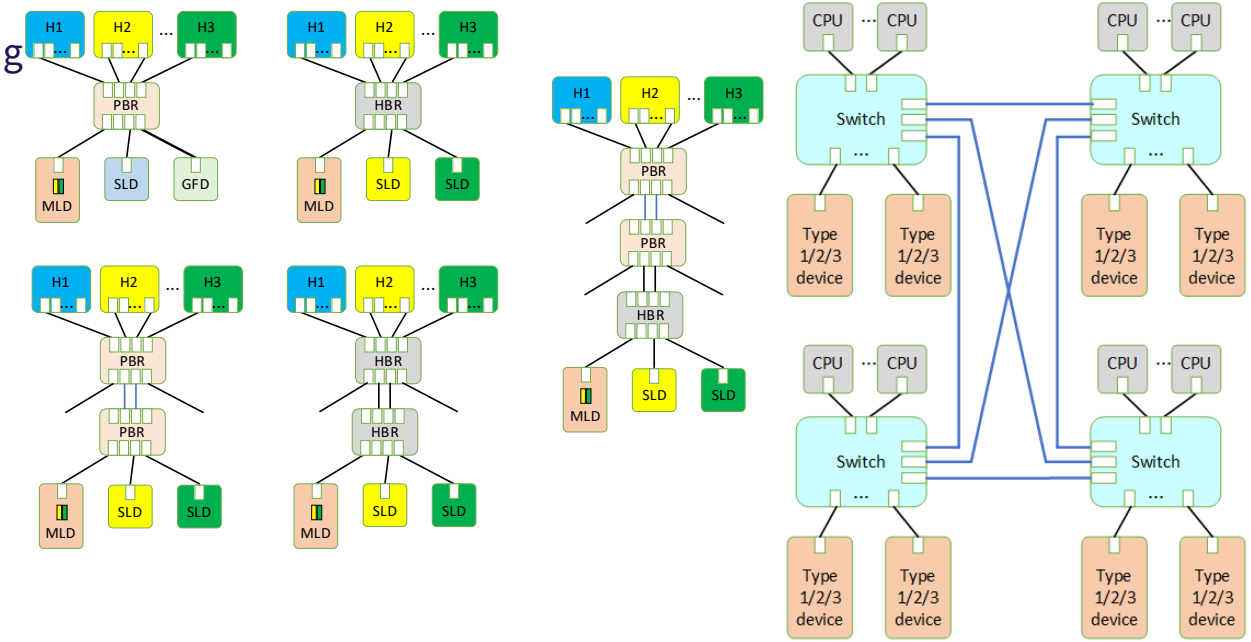
System View of a Switch

- Physical Topology
  - What you build
  - Intermediate switches only for connectivity
- Logic view, **per host**
  - No intermediate switch
  - Simple PCIe tree
  - Discovery, enumeration all PCIe compliant
  - CXL capability enables CXL memory and cache
- Links can be PCIe or CXL
  - CXL link is a ‘flex bus’ – runs both PCIe and CXL.cache, CXL.mem



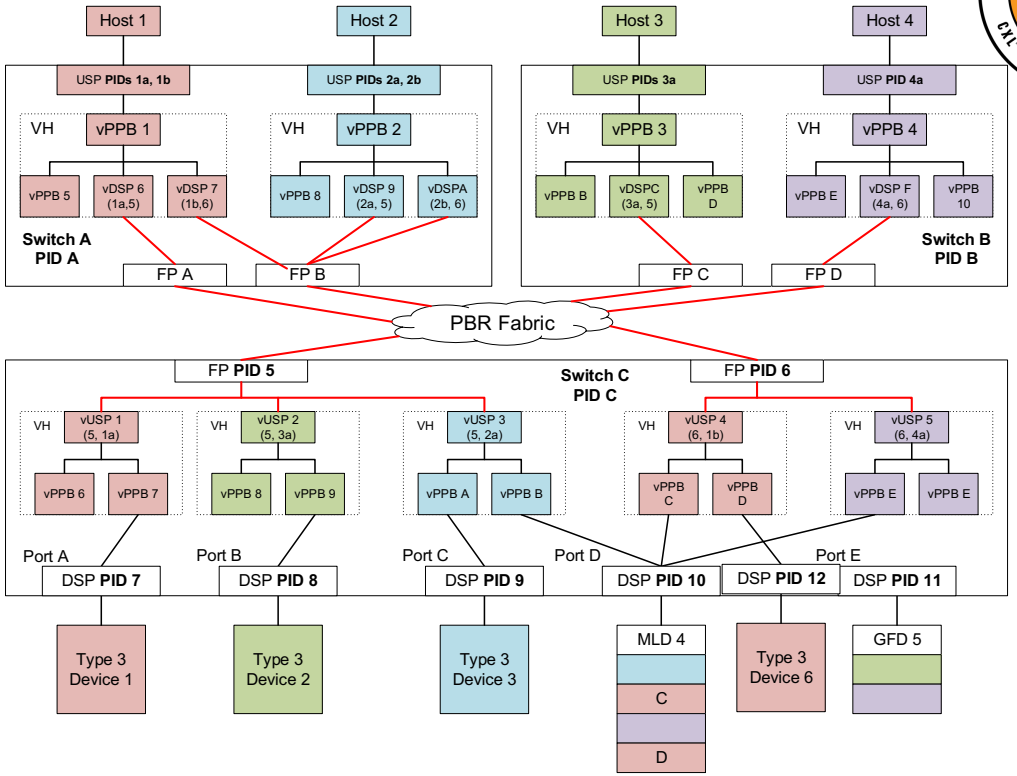
Switch configurations

- HBR = Hierarchy-Based Routing
  - Switches based on CXL2.0 specification
- PBR = Port-Based Routing
- SLD = Single logical device
- MLD = Multi logical device
- GFD = Global Fabric Attached Memory Device



PBR ID Overview

- PBR IDs assigned to:
  - Switch host ports
  - Switch downstream ports
  - Switches
- Route by Destination PBR ID
  - Tables in switch pick a path
  - Multi-paths supported
- Requests needing a response provide a Source PBR ID
  - Response uses the source PBR ID as a Destination PBR ID
- PBR IDs applied to PCIe (cxl.io) TLPs as well as CXL.cache and CXL.mem messages



Things to Consider

- System discovery, enumeration, and authentication
- PBR ID assignment, routing, multipath routing, and failover
- Composability: hot-add, hot-remove of a device (not hot-add/remove of a Switch)
- Latency, traffic isolation per VH, throughput metrics, QoS
- Mix between IO and CM traffic

Call to Action

- Join the effort, provide feedback, introduce new use cases, help solve the problems, develop new products
- OCP Programs currently targeting these issues:
  - DC-MHS
  - Multi-node Systems
  - OAI

Thank you!

EMPOWERING OPEN.

OCP GLOBAL SUMMIT

OCTOBER 18-20, 2022
SAN JOSE, CA

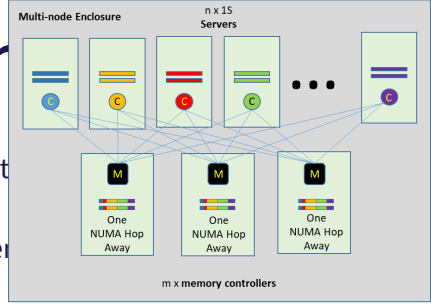
Open Discussion

EMPOWERING OPEN.

OCP GLOBAL SUMMIT

OCTOBER 18-20, 2022
SAN JOSE, CA

Compute Disaggregation Taxonom



Capacity-matched
Bandwidth-matched

- **Pooling** (dividing a resource to multiple non-overlapping logical units and assigning them to different servers/host
- **Sharing**
  - Serialized Sharing (a device may be fully mapped to a server at one time and to a different server at a differe time)
  - Concurrent Sharing (multiple servers/hosts are assigned to the same portion of a device at the same time; coherence and access ordering may be enforced by hardware or software)
- **Borrowing** (as part of its own separate coherence domain, a server may get permission to access a portion of a second server's resource. This resource will leave the second server's coherence domain.)
- **Fabric** (a mechanism for dynamically interconnecting heterogeneous elements to form computing systems)
- **Physical Disaggregation** (interconnected chassis: Server Head-node + Expansion Chassis such as a JBOD, JBOF, JBOG, …)
- **Logical Disaggregation** (composing several servers via a ***Fabric*** to provide access to shared or pooled sets of resources)
- **Local Disaggregation**
  - Multiple servers **in one Chassis** accessing a shared or pooled set of resources
  - Using a multi-host capable *Switch* or via multi-ported *End Devices*
- **Extend** memory via increased memory capacity and range of the same type of medium (e.g., **DRAM** with <3x mem latency)
- **Expand** memory via managing **multiple tiers** of memory (e.g., NAND Flash backing DRAM) & swapping/paging techniques
- Coherence (enforced by HW or maintained by SW via access ordering sequences and appropriate flush mechanisms)
- Scale-up (single host, homogeneous computing, scale via the same type of interconnect protocol)
- Scale-out (networked-based or via changing interconnect protocol, heterogeneous or distributed multi-server computing)

Enablers (Software and Firmware Ingredients)

CXL Fabric Manager

- Secure composability, allocation, on-lining/off-lining

Pre-boot Environment

- Discovery, Enumeration

CXL Bus Driver

- Configuration, Resource allocation

CXL Memory Device Driver

- Interactions with Bus Driver, Fabric Manager, and VMM
- RAS, Security, Fault-isolation, On-lining, Off-lining, …

  ECN: Error Isolation on CXL.mem and CXL.cache (Enabled by the Root Port; requires Software Stack to recover from faults)

OS-specific Software

- VMM, Hypervisor
- VM Allocation, Orchestration, Fault-isolation & Recovery

Thank you!

EMPOWERING OPEN.

OCP GLOBAL SUMMIT

OCTOBER 18-20, 2022
SAN JOSE, CA