CXL for Storage

REGIONAL SDC 24
BY Developers FOR Developers
APRIL 24, AUSTIN, TX

Kevin Marks
Dell Technologies

A SNIA Event

Agenda

- About Me
- CXL Overview and its Evolution
- Possible CXL Storage Use Cases

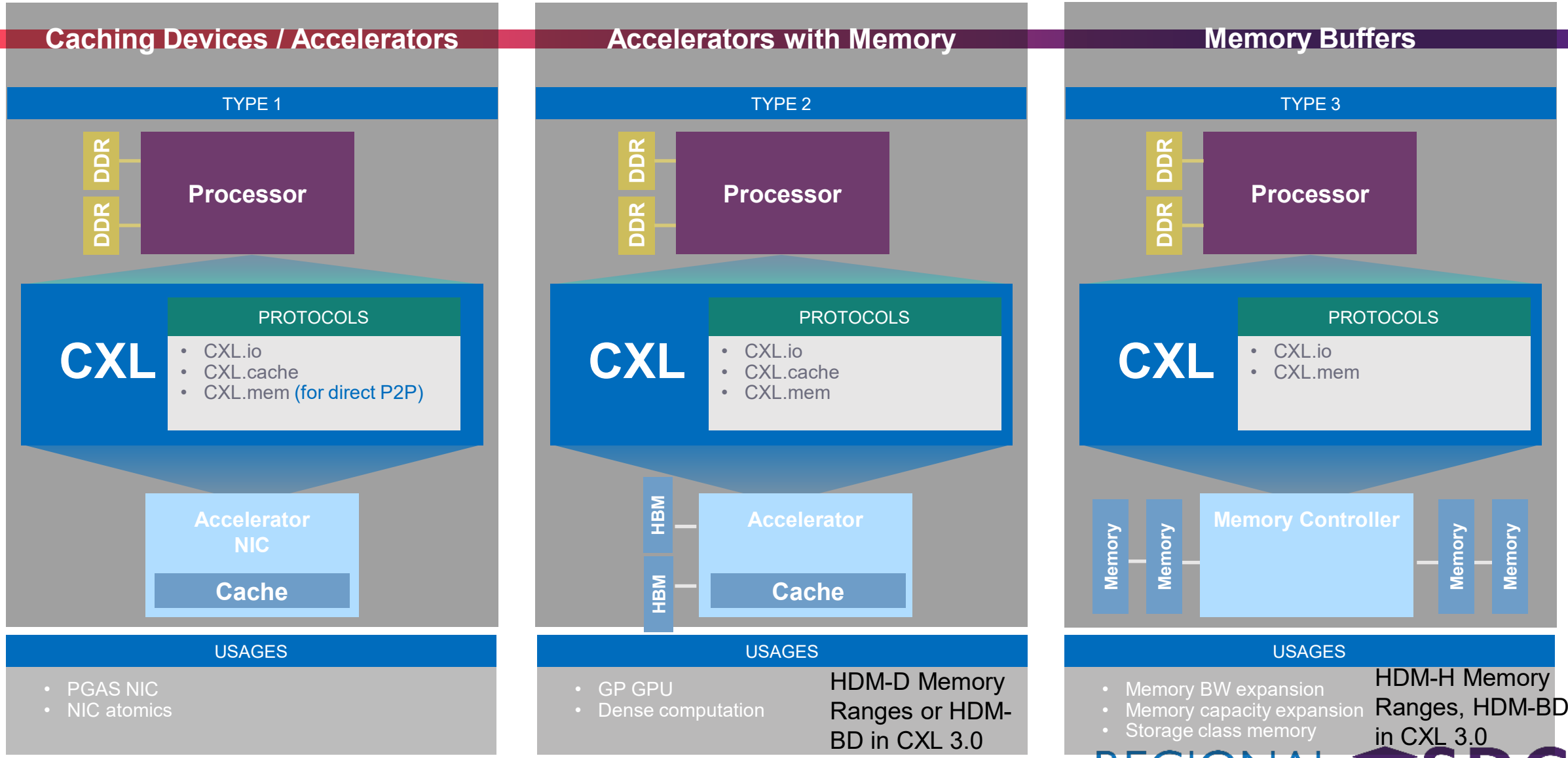
About me

- Technologist and Distinguished Engineer
- 25+ years at Dell Technologies
- 20+ years of standards participation
  - T10, T13, SATA-IO, NVMe, TCG, CXL, SNIA TWGs
- CXL Protocol WG Co-Chair
- Current focus:
  - GenAI solutions enablement
  - CXL standards
- 1st year Empty Nester

CXL Basics Starting at CXL 1.1

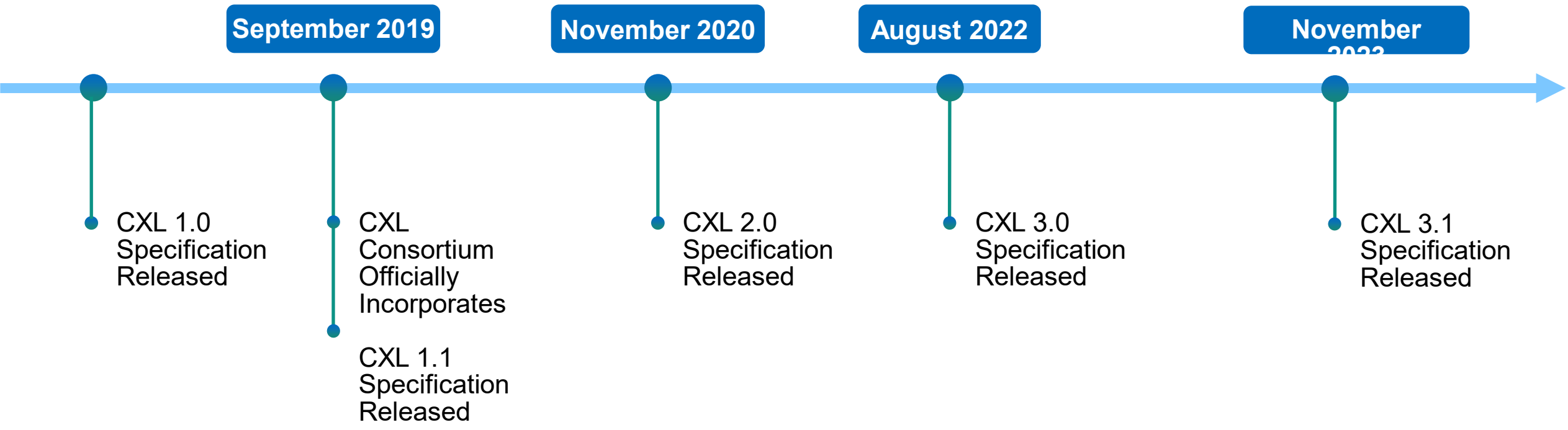
- PCIe Gen5 electricals (using Alternate Protocol Negotiation)
- Communicates on links using FLITs which contain slots
- Slots have defined fixed formats based on FLIT headers
- Slots carry protocol request, response, data and control
- 3 protocols (muxed on the link), well muxed between .cache/.mem vs .io
  - .mem (host to device) - Read/write device exposed host device memory (HDM)
  - .cache (device to host) - Allows device to coherently cache host memory
  - .io (PCIe over CXL) - Required for discovery/configuration
- CXL 1/1.1 is direct attached only

Representative CXL Usages



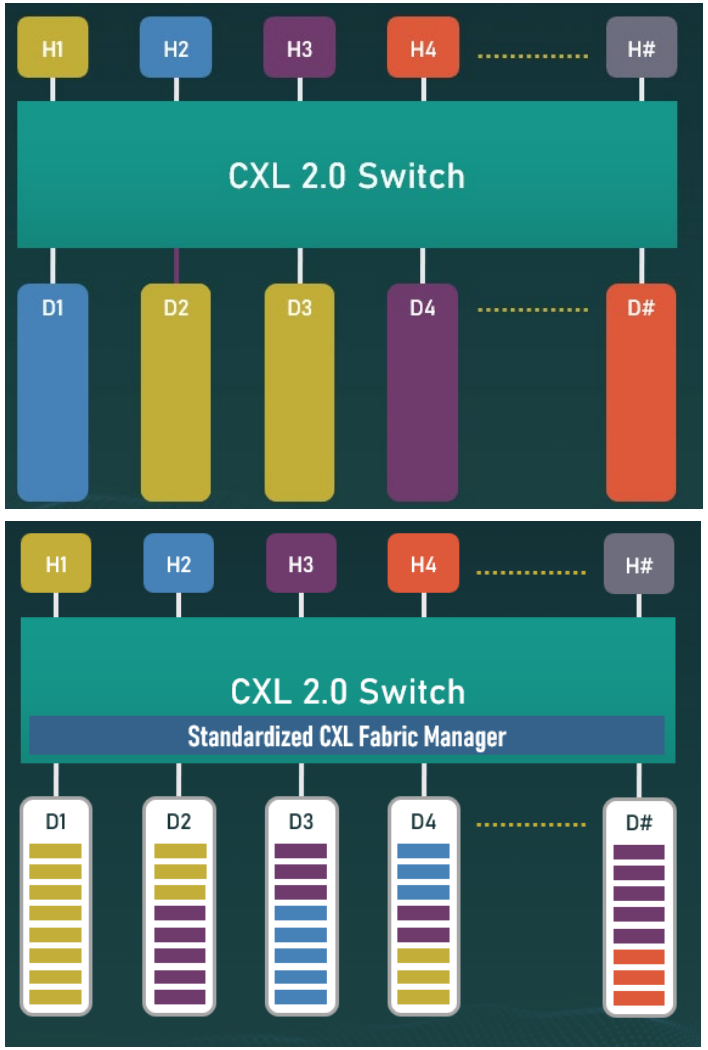
*Original Image credit to CXL™ Consortium

CXL Overview/History



CXL 2.0 (High level additions)

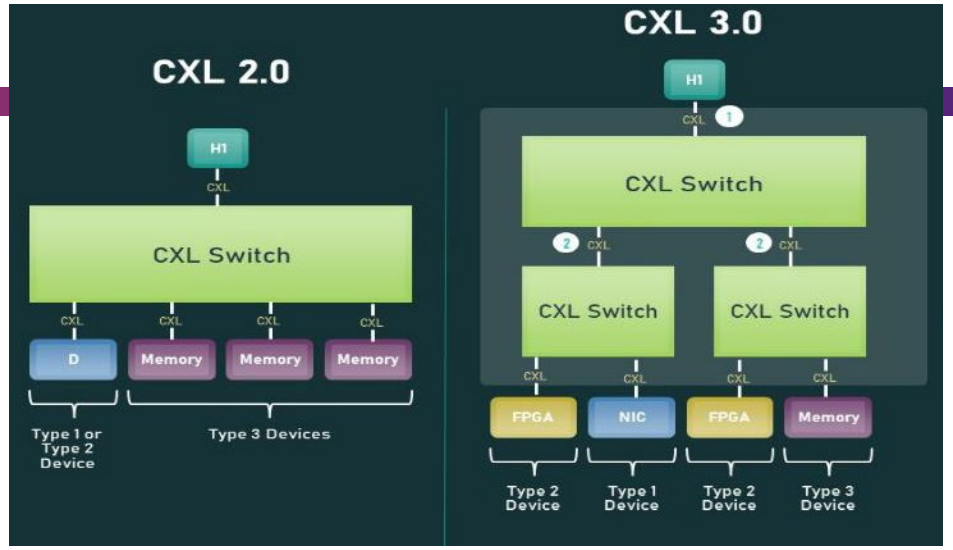
- Single level switching allowing for T3 memory fanout and virtual switching/virtual hierarchies (e.g., zoning)
  - Only 1 T1/T2 .cache device per VH
- T3 MLD device sharing (not memory sharing)
  - Multi-Logical Device (up to 16 LDs)
- Global Persistence Flush – 2 stage system flush mechanism for persistent memory
- IDE (Integrity and Data Encryption) Link encryption
- Defined CXL Fabric Manager API

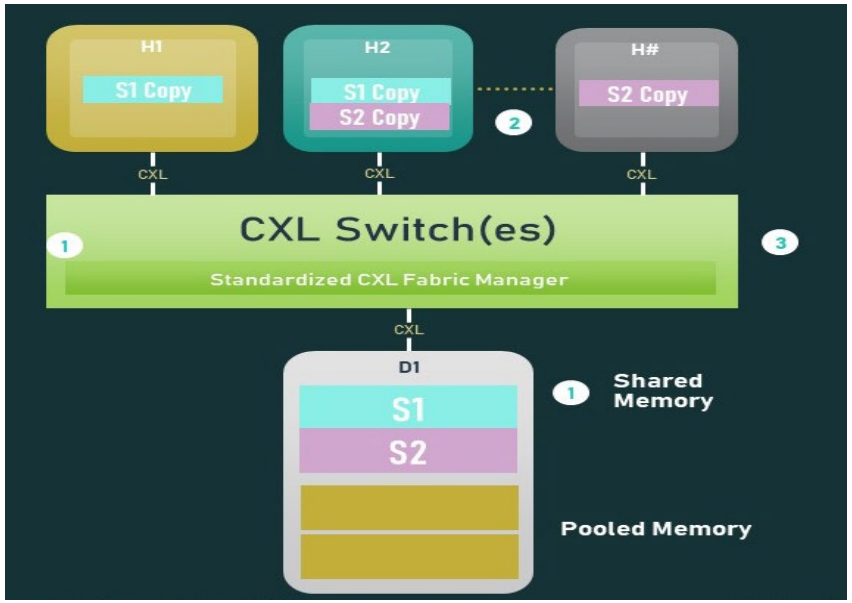


*Original Image credit to CXL™ Consortium

CXL 3.0 (High level additions)

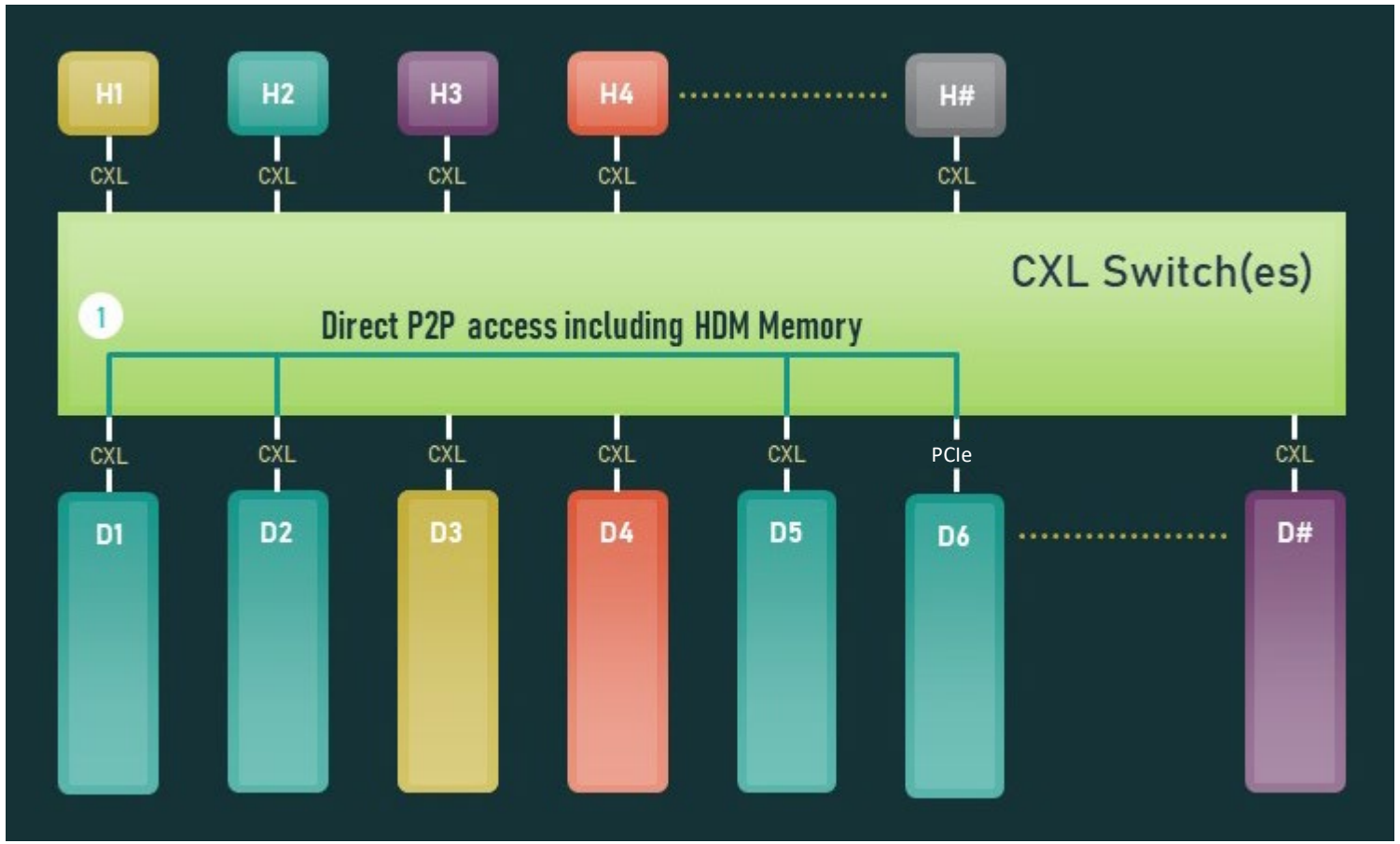
- Speed bump to Gen 6 PCIe electricals (64 GT/s)
- Multi-level switching
  - Memory address-based routing and port-based routing*
  - Up to 16 T1/T2 devices per VH
- Enhanced coherency (256 byte flit only)
  - Back Invalidate / Snooping (HDM-DB)
- T3 MLD memory sharing (256 byte flit)
  - Software or hardware-based coherency
- Direct memory access for peer-to-peer
  - Reversed m2s channel (device initiated)
- Fabric capabilities (256 byte flit)
- PCIe UIO support in devices and switches
  - Allows UIO requests to access HDM ranges
- GFAM (Global Fabric Attached Memory)





*Original Image credit to CXL™ Consortium

UIO and BI for Device to Device Connectivity

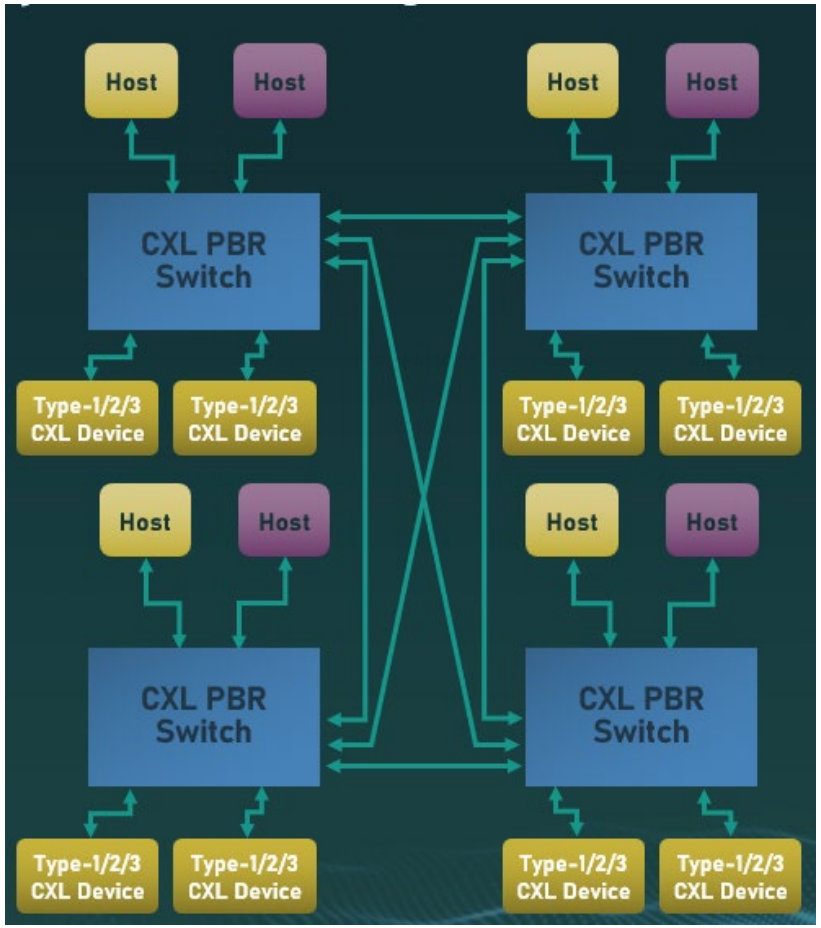


- Unordered I/O (UIO) defined in PCI SIG was enhanced to support P2P in CXL allowing an IO coherent P2P targeting .mem address ranges over .io (including switch changes)
- Depending on the source or target of the P2P, Back-Invalidation flows (CXL.Mem) may come into play to resolve coherency.
- Also allows PCIe devices to access CXL memory in a P2P fashion

*Original Image credit to CXL™ Consortium

CXL 3.1 (High level additions)

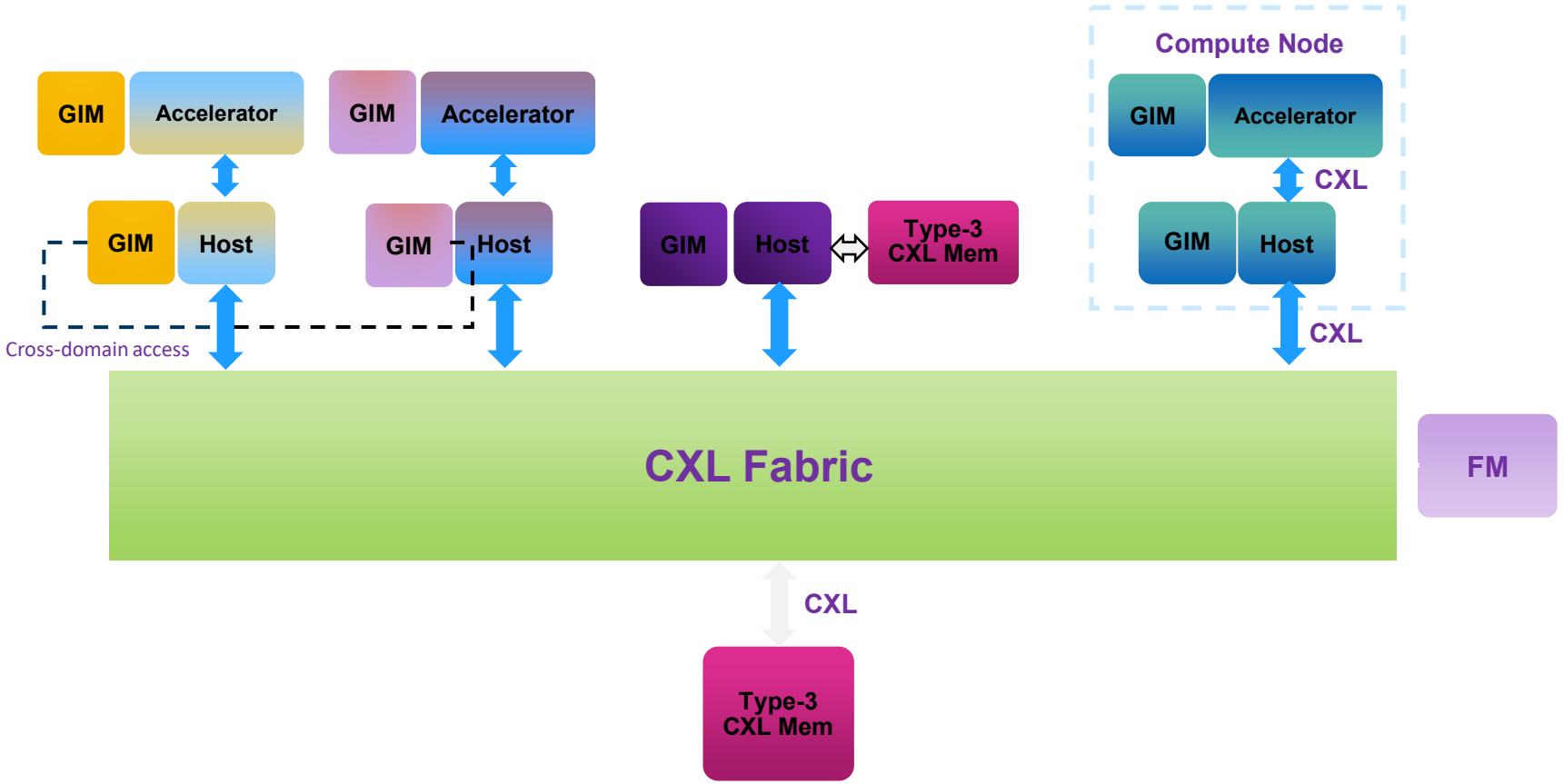
- Fabric Manager API definition for PBR based switching
- Trusted-Execution-Environment (TEE) / Trusted Security Protocol (TSP)
  - Direct attached only and HDM-H only
- Host-to-Host communication with Global Integrated Memory (GIM) concept
  - Sharing host memory
- Memory device enhancements (up to 32-bit of meta data, RAS capability enhancements)
- Direct Peer-to-peer using PBR and MLDs



*Original Image credit to CXL™ Consortium

CXL 3.1: Fabric Enhancement Features

Host-to-Host communication with Global Integrated Memory (GIM) concept



Host to Fabric-Attached Memory (FAM) communication with Global Integrated Memory (GIM)

- Multiple Hosts mapping CXL Fabric-Attached Memory devices
- Hosts and FAM devices can initiate cross-domain accesses to GIM
- Uses UIO and PBR

*Original Image credit to CXL™ Consortium

CXL Evolution

Features	CXL 1.0 / 1.1	CXL 2.0	CXL 3.0	CXL 3.1
Release date	2019	2020	August 2022	November 2023
Max link rate	32GTs	32GTs	64GTs	64GTs
Flit 68 byte (up to 32 GTs)	✓	✓	✓	✓
Flit 256 byte (up to 64 GTs)			✓	✓
Type 1, Type 2 and Type 3 Devices	✓	✓	✓	✓
Memory Pooling w/ MLDs		✓	✓	✓
Global Persistent Flush		✓	✓	✓
CXL IDE		✓	✓	✓
Switching (Single-level)		✓	✓	✓
Switching (Multi-level)			✓	✓
Direct memory access for peer-to-peer			✓	✓
Enhanced coherency (256 byte flit)			✓	✓
Memory sharing (256 byte flit)			✓	✓
Multiple Type 1/Type 2 devices per root port			✓	✓
Fabric capabilities (256 byte flit)			✓	✓
Fabric Manager API definition for PBR Switch				✓
Host-to-Host communication with Global Integrated Memory (GIM) concept				✓
Trusted-Execution-Environment (TEE) Security Protocol				✓
Memory expander enhancements (up to 32-bit of meta data, RAS capability enhancements)				✓

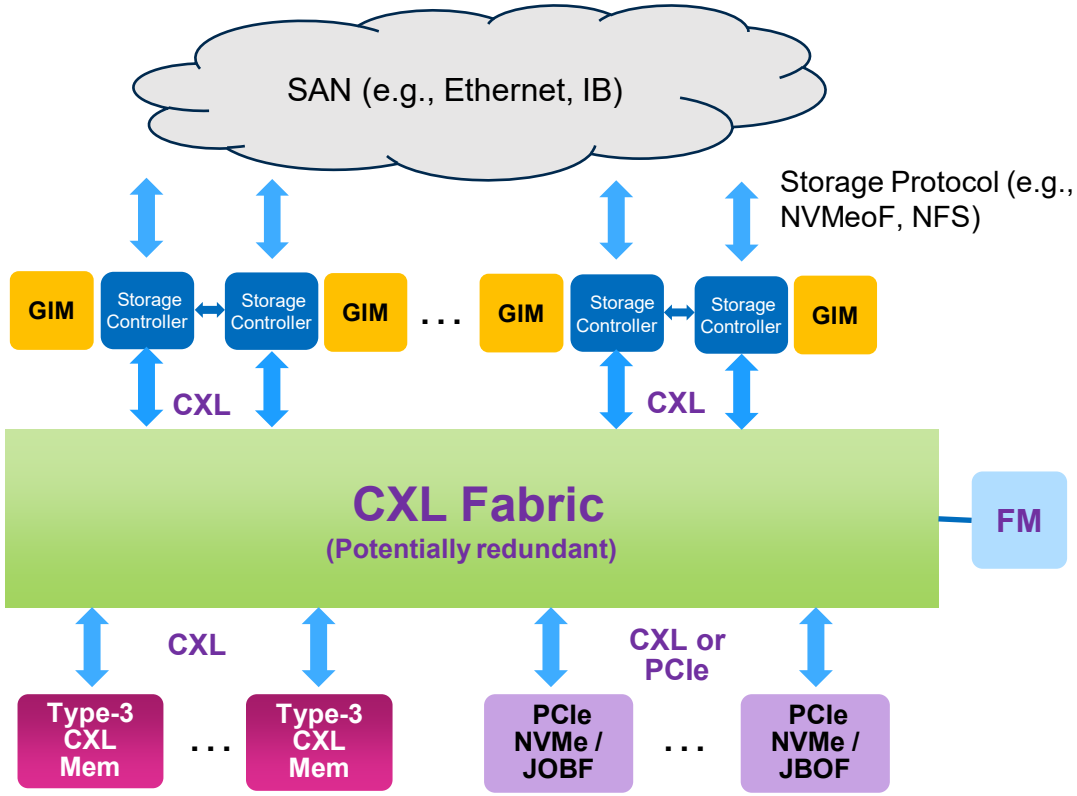
*Original Image credit to CXL™ Consortium

CXL Storage Use Cases (NVMe)

- Standard NVMe over CXL, Why?
  - Would just use CXL.io, and possibly be a little slower due to FLIT overhead
- But what about:
  - CMB/PMR concepts as HDM ranges (.mem) vs MMIO access (MRd/MRw)
    - Data source for commands (RDS/WDS)
    - Queues (CQS/SQS), not sure the value vs posted writes
  - Mapping a namespace or part of it to an HDM range
    - Would allow partial reads of a logical block and potentially faster access
    - Would allow coherent access via NVMe path and HDM path with HDM-DB range
    - Seeing some vendors speaking about these hybrid devices as of late
  - Computational Storage memory access
    - Jason Molgaard will cover later today
- What about adding UIO to NVMe device
  - Allows NVMe device commands to source or target CXL memory using P2P
  - Above assumes that the CXL T3 memory and CXL switches support UIO

CXL Storage Use Cases (External storage)

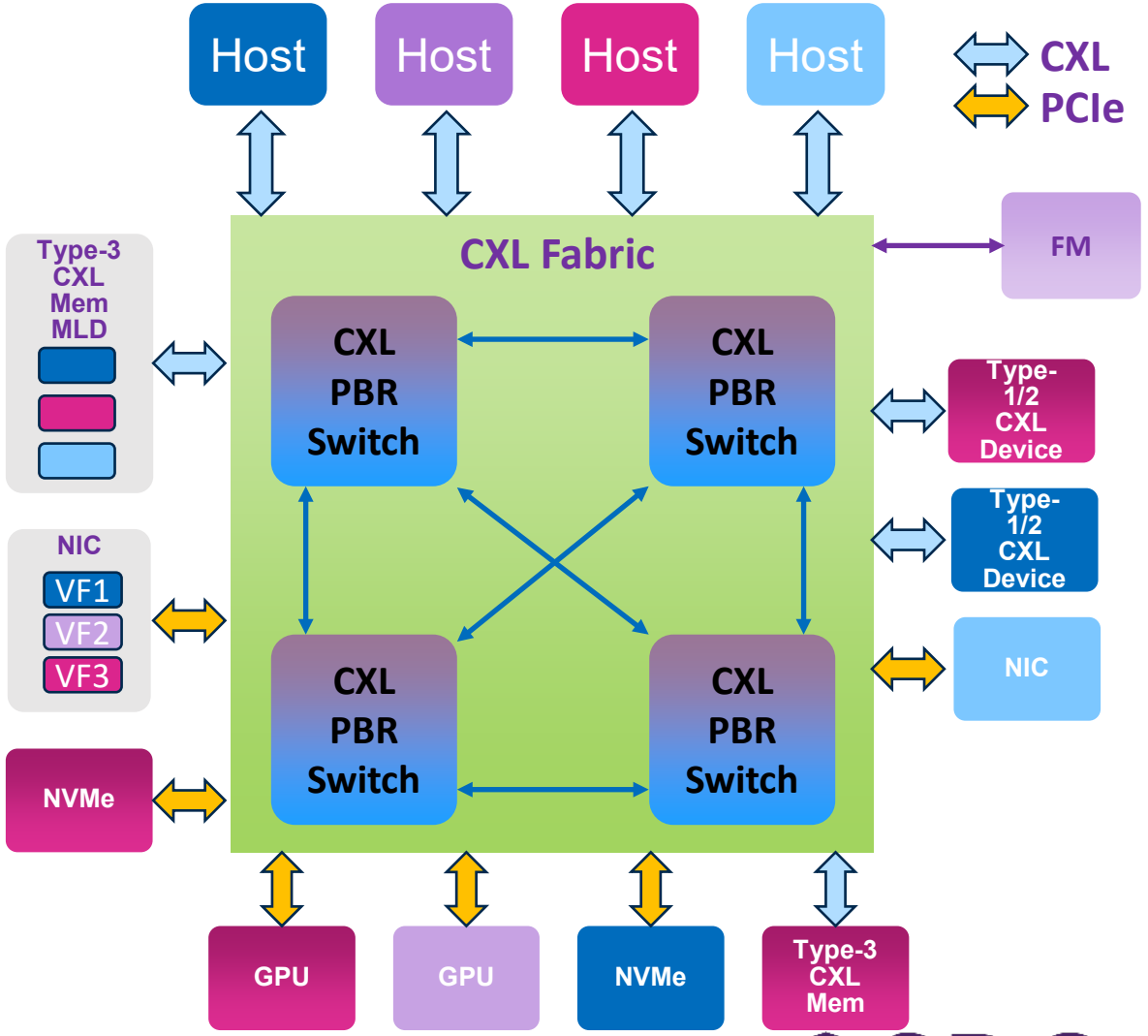
- Scalable Multi-node Storage system
  - Multiple storage controllers make up the storage systems for both performance scale and HA redundancy
- Front side – Traditional storage protocols
  - NVMeoF, NFS, FC
- Back side – CXL Fabric
  - NVMe or NVMe JBOF attachment for block storage
  - Type 3 CXL memory with HDM-DB for 2 Tier cache
  - GIM memory could be DRAM or storage controller local CXL T3 memory
  - GIM memory would be used to implement a global cache allowing IO coherent access from other controllers using UIO
- Inhibitors
  - Processor UIO support + GIM, PBR support, aspects of HA, robust error handling, etc.



REGIONAL SDC 24

CXL Storage Use Cases (Composable)

- With MR-IOV and vendor specific switch features, some of the composability story was possible. But ...
- With CXL 3.1, a more complete solution emerges with the combination of CXL PBR switches, MLD devices, PCIe support and the FM API.
- The FM API is “THE” enabler for a standardized Fabric Management for composability.
- And CXL Fabric features expected to grow with every release
  - Think more comprehensive QoS, multipathing, Error handling, etc.



THANK YOU

Please take a moment to rate this session.

REGIONAL SDC 24

BY Developers FOR Developers