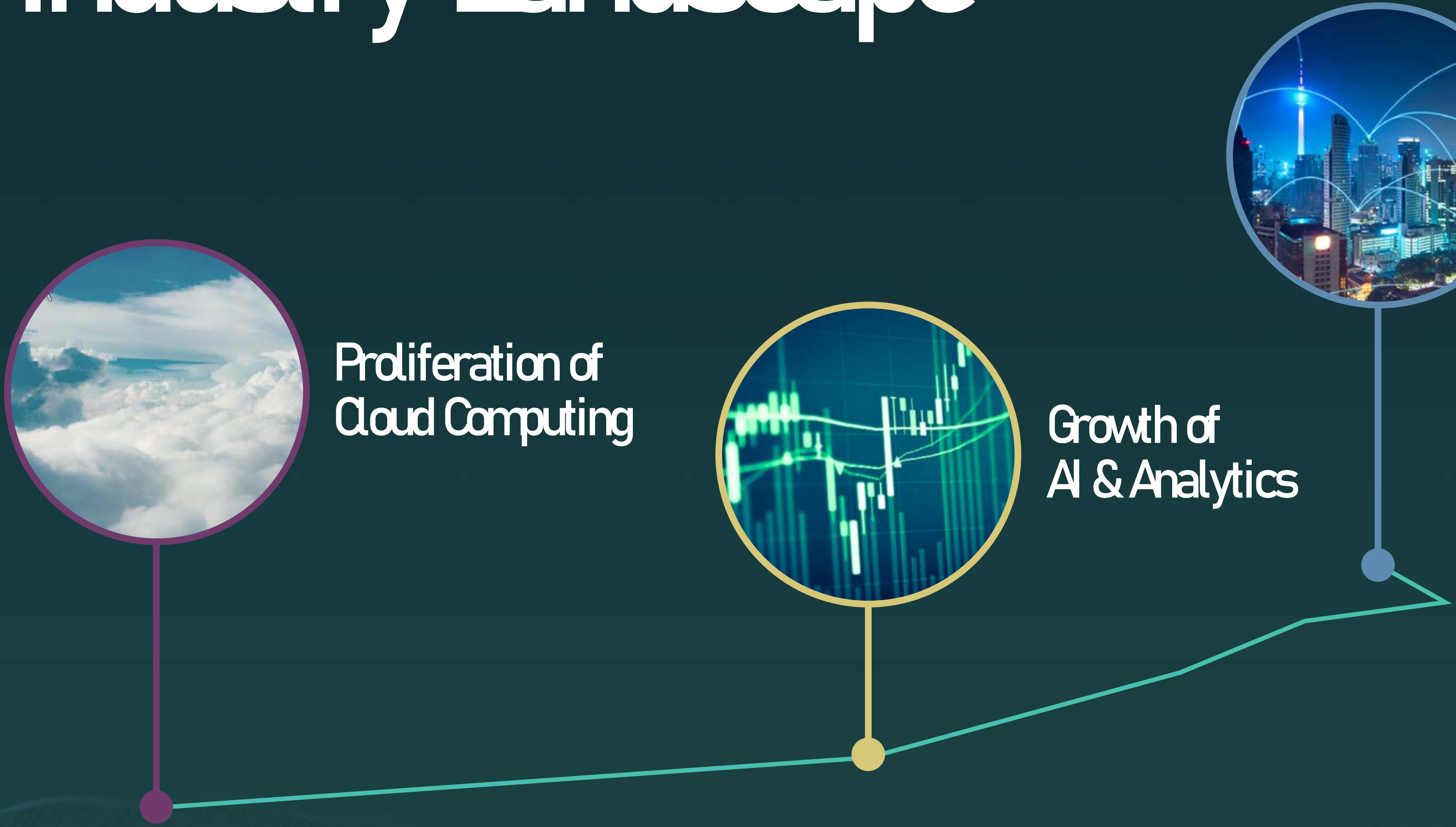
Compute Express Link™ (CXL™): Supporting Persistent Memory

Mahesh Natu, Datacenter Platform Architect, Intel Corporation

Thomas Won Ha Choi, Director, DRAM Product Planning & Enabling, SK hynix

Industry Landscape



Proliferation of Cloud Computing

Growth of AI & Analytics

Cloudification of the Network & Edge

Alibaba Group   AMD   arm   cisco   DELL EMC   facebook

Hewlett Packard Enterprise   HUAWEI   Google   IBM   intel

MICROCHIP   Microsoft   NVIDIA   XILINX

CXL Board of Directors

CXL Compute Express Link™

Industry Open Standard for High Speed Communications

150+ Member Companies

CXL Delivers the Right Features & Architecture

Challenges

Industry trends driving demand for faster data processing and next-gen data center performance

Increasing demand for heterogeneous computing and server disaggregation

Need for increased memory capacity and bandwidth

Lack of open industry standard to address next-gen interconnect challenges

CXL

An open industry-supported cache-coherent interconnect for processors, memory expansion and accelerators

Coherent Interface

Leverages PCIe® with 3 mix-and-match protocols

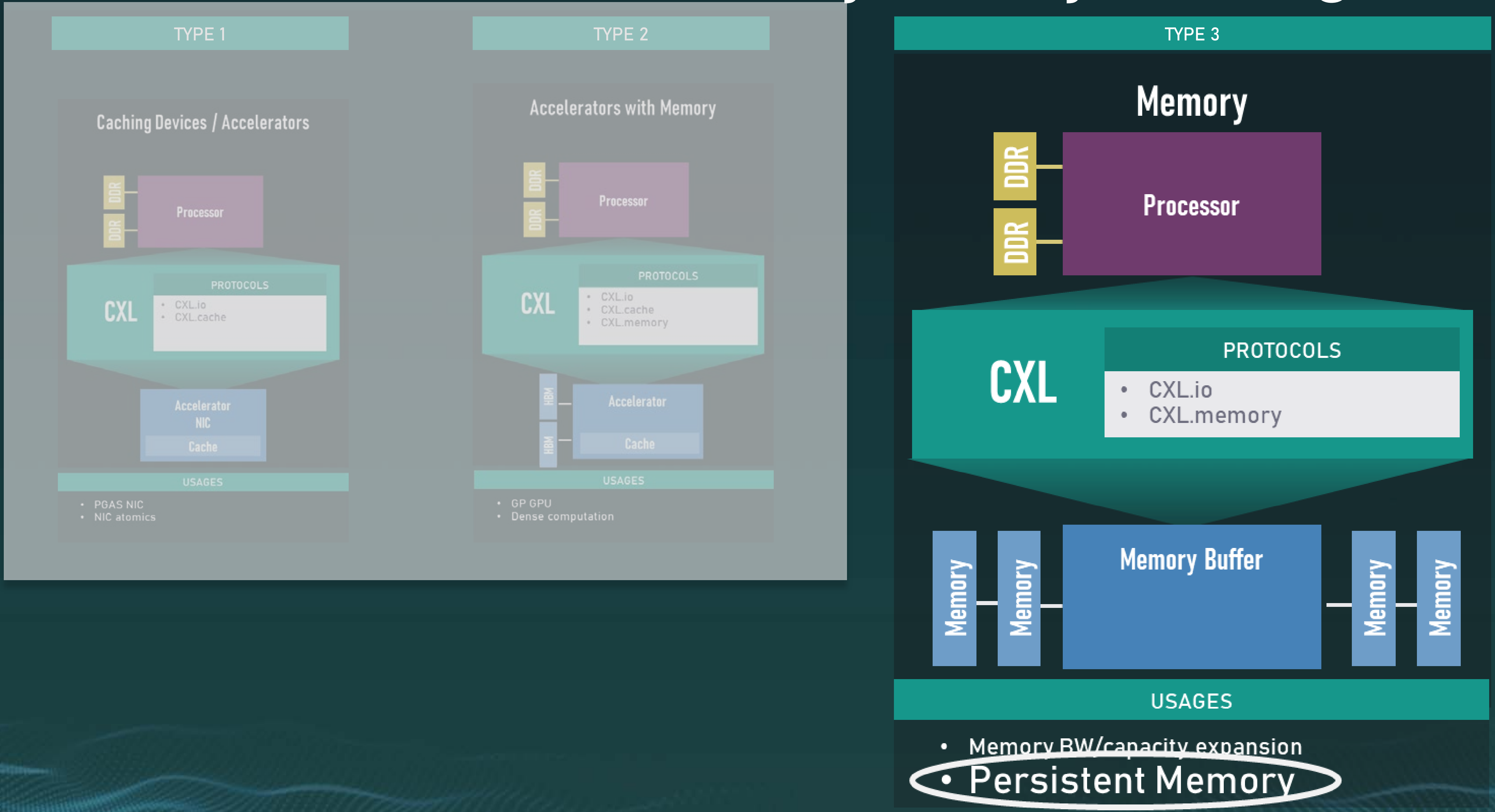
Low Latency

.Cache and .Memory targeted at near CPU cache coherent latency

Asymmetric Complexity

Eases burdens of cache coherent interface designs

Persistent Memory is a Key CXL Usage



TYPE 1

Caching Devices / Accelerators

DDR

DDR

Processor

CXL

PROTOCOLS

- CXL.io
- CXL.cache

Accelerator
NIC

Cache

USAGES

- PGAS NIC
- NIC atomics

TYPE 2

Accelerators with Memory

DDR

DDR

Processor

CXL

PROTOCOLS

- CXL.io
- CXL.cache
- CXL.memory

HBM

HBM

Accelerator

Cache

USAGES

- GP GPU
- Dense computation

TYPE 3

Memory

DDR

DDR

Processor

CXL

PROTOCOLS

- CXL.io
- CXL.memory

Memory

Memory

Memory Buffer

Memory

Memory

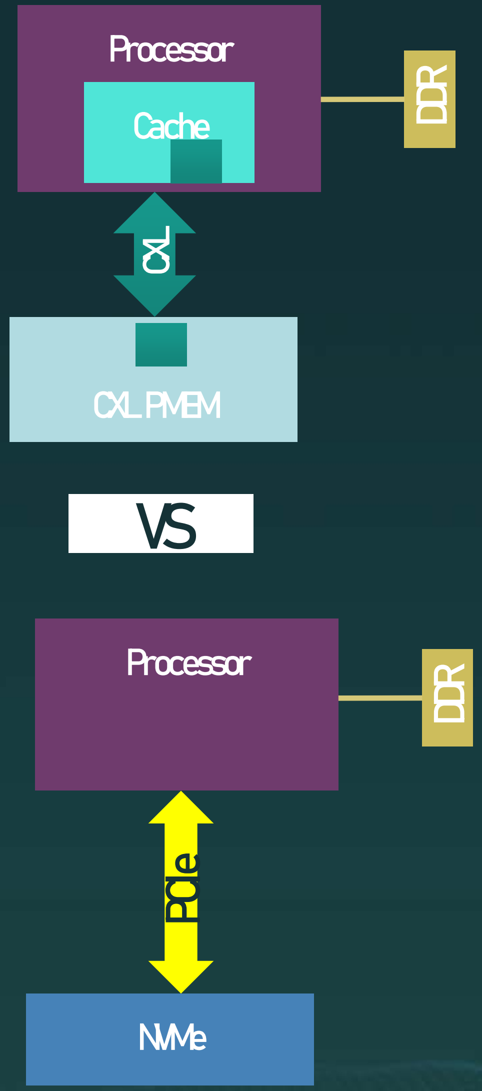
USAGES

- Memory BW/capacity expansion
- Persistent Memory

Persistent Memory

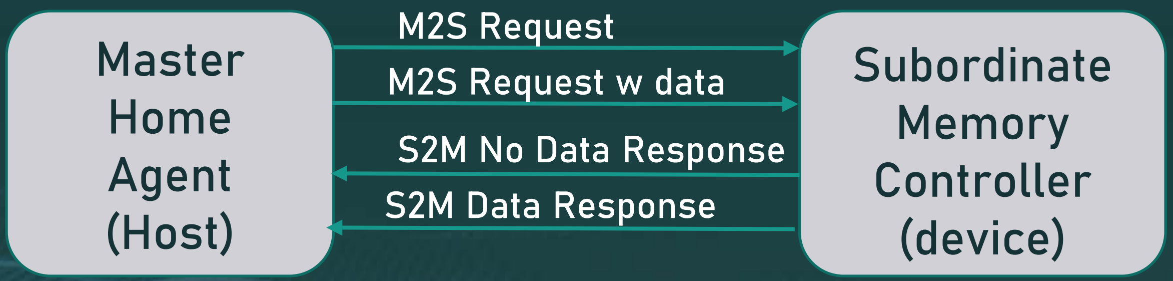
- Characteristics and benefits of Persistent Memory (PMEM)
    - Byte-addressable (vs. NVMe is block addressable)
    - Generally, lower latencies compared to SSD
    - Cacheable (vs. NVMe is uncached)
    - Data persists across power loss (vs. DRAM loses its content)
    - Generally, larger capacity (vs. DRAM)

- Many Workloads benefit from PMEM
    - Traditional Databases – Accelerated logging/journaling, instant recovery
    - Analytics/AI/ML – real time access to large datasets, faster checkpointing
    - Storage – caching, tiering, ..
    - HPC – Reduce checkpointing overhead
    - and more ..



CXL Protocol is Well Suited for PMEM

- CXL.mem protocol is transactional, see below
  - PMEM media may have longer latencies, or variable access latencies
  - Controller can hide longer and/or variable access latencies
- CXL.mem abstraction
  - Memory Controller and media are abstracted
  - Enables new and innovative media types
- CXL 2.0 introduces memory QoS
  - Device can synchronously report how loaded it is
  - Can prevent head of line blocking in heterogenous memory configuration (e.g. DRAM + PMEM)
- CXL 2.0 adds Memory Interleaving, a standardized register interface and Global Persistent Flush (GPF)



Master Home Agent (Host)

M2S Request

M2S Request w data

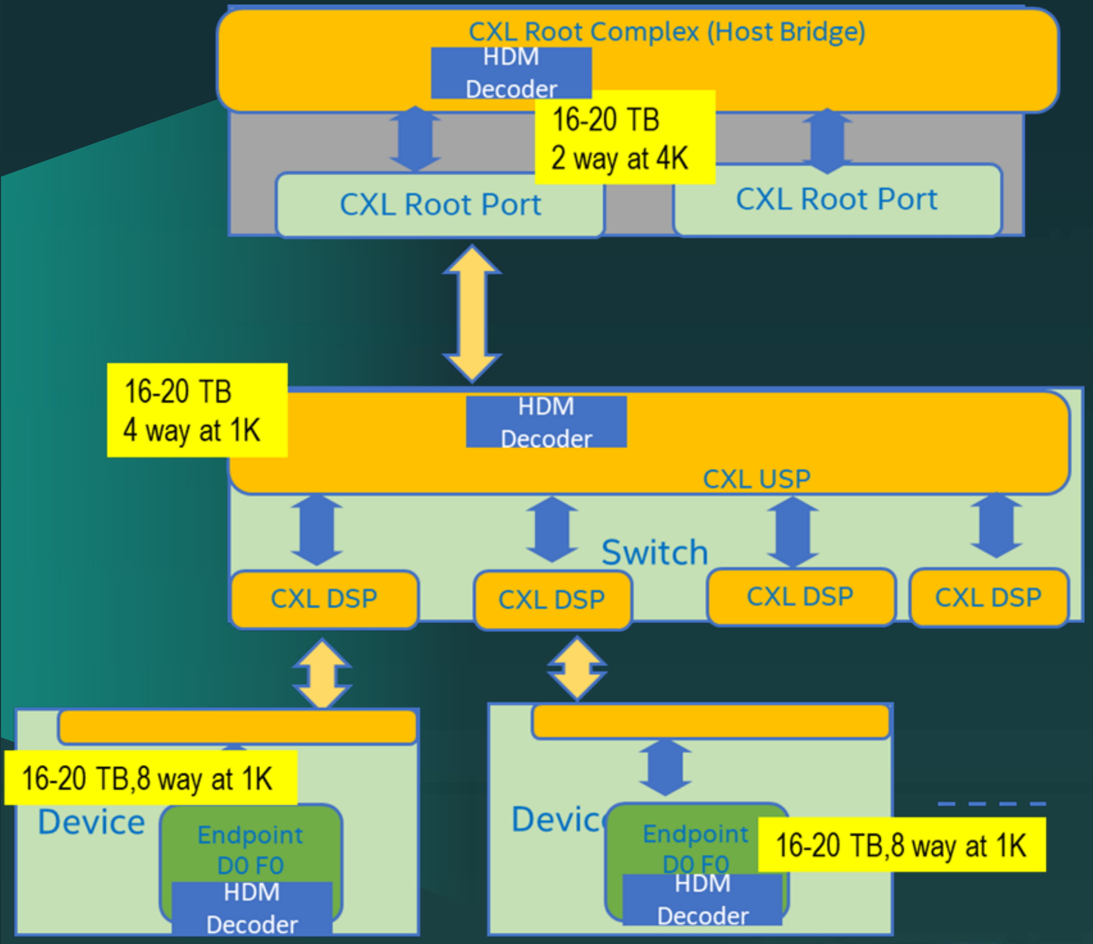
S2M No Data Response

S2M Data Response

Subordinate Memory Controller (device)

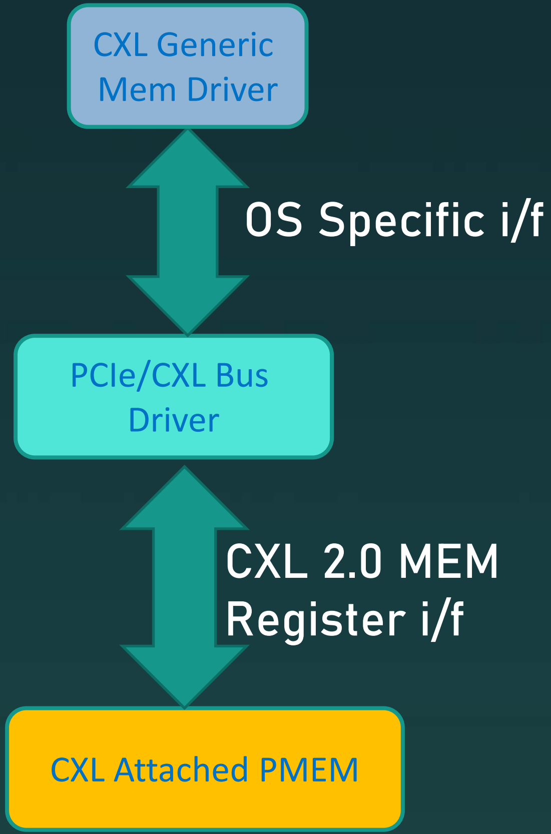
CXL Supports Interleaving of Memory Devices

- Interleaving is performance feature
- Example: 8-way interleaved device
- How it works
  - CXL Host Bridge HDM Decoders configured to select one of two Root Ports based on A[12]
  - HDM Decoders in every switch configured to select one of four DSPs based on A[11:10]
  - HDM Decoders in the device configured for 8 way interleave at 1K
  - Device removes A[12:10] from the Host Physical Address when computing the Device Physical Address



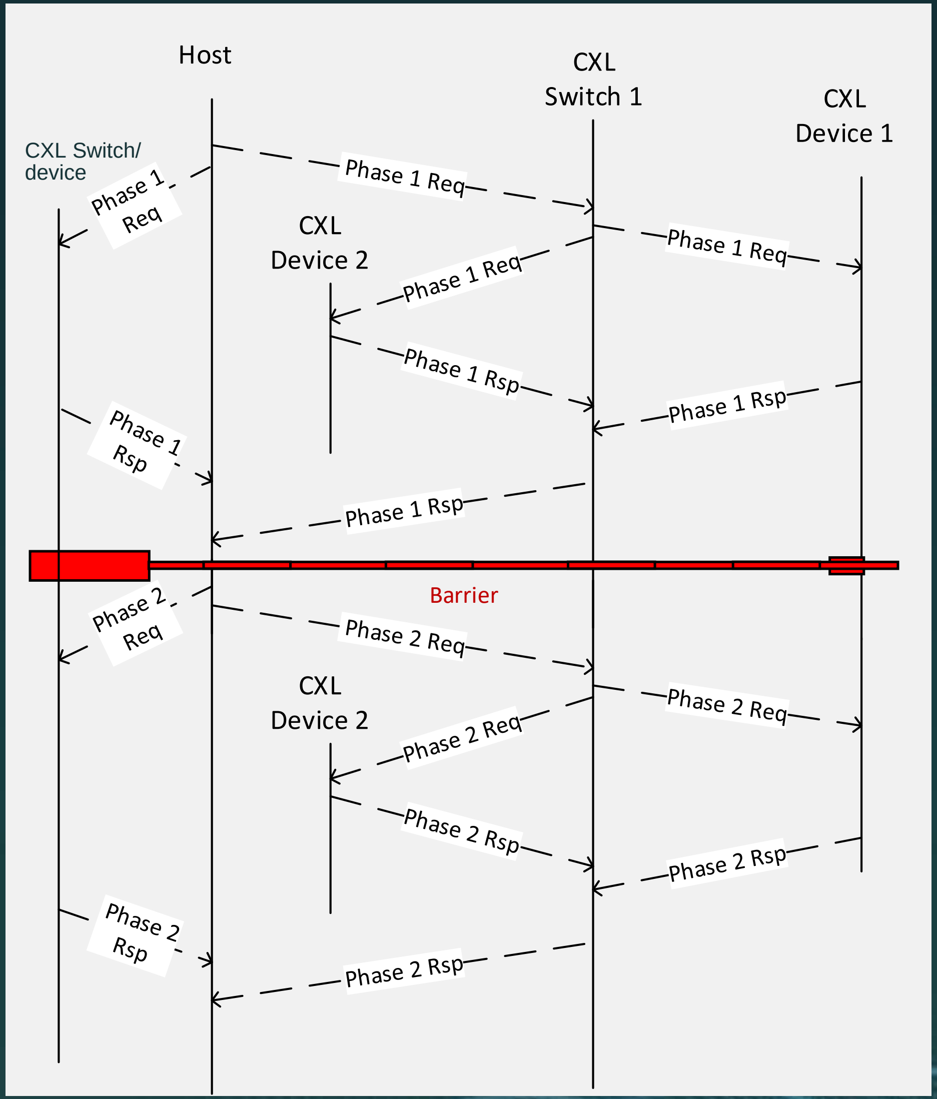
PMEM Configuration Interface

- Persistent memory devices rely on System Software for provisioning and management
- CXL 2.0 introduces a standard register interface for managing CXL attached memory devices including PMEM devices
- A generic memory device driver simplifies software enabling
- Architecture Elements
  - Defined as number of discoverable Capabilities
  - Capabilities includes Device Status and standard mailboxes, accessed via MMIO registers
  - Standardized mailbox commands that cover errors/health, alerts, partitioning, passphrases etc.
  - Allow Vendor specific extensions



Global Persistent Flush (GPF)

- PMEM aware applications expect that the completed writes are made persistent
- In reality, the write data may be held in Processor/CXL Device caches or Memory Device Write buffers for performance reasons
- Upon an event such as sudden power-loss, the system needs to push the data to Persistent domain in order to keep the promise.
- GPF is a Global event across cache coherency domain
- Controlled by host, enables coordination of flush activity between the host and the CXL domain
- Two phase CXL flow, with a barrier between the two phases
  1. Host ask each CXL device to stop injecting new traffic and flush its cache, device acks
  2. Host asks each CXL device to push data in local buffers to Persistent domain, device acks
- If error/timeout detected in phase 1, host propagates "error flag" to each device during Phase 2 so PMEM devices can log "dirty shutdown" event.



Failure Management: Dirty Shutdown Count (DSC)

GPF failure -> Dirty Shutdown triggered

- Shutdown State (device internal): set dirty when the GPF flow is not successful
- DSC: incremented when GPF failed or data untraceable of completion (i.e. shutdown state is dirty)
- DSC is exposed via Get Health Info (mailbox CMD), must account for DSC from other devices in the interleave set

@ Device Not in Use

```
<Wake up from reset>

if shutdown state is dirty {
  DSC++;
  shutdown state = clean;
}
else do nothing;
```

@ Device in Use

```
<Execute GPF flow>

<Set shutdown state: mailbox CMD>

if(GPF flow successful)
    shutdown state = clean;
else shutdown state = dirty;
```

@ Normal Shutdown

```
<Start normal shutdown flow>

<Set shutdown state: mailbox CMD>



shutdown state  = clean;
```

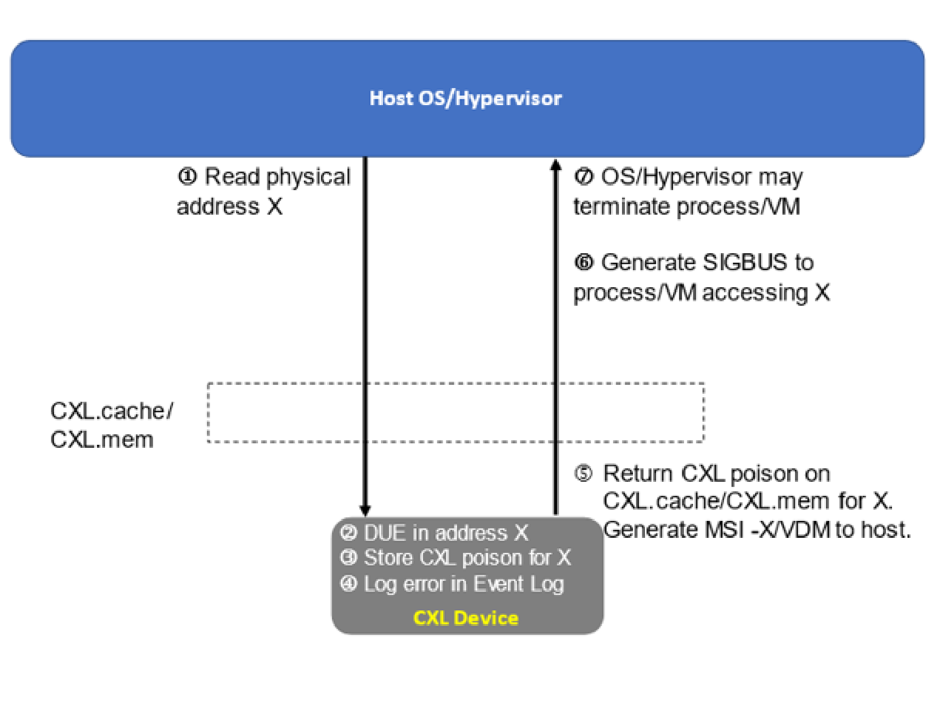
Internal Poison List Retrieval and Scan Media

Background

- Any non-fatal DUE indicated as poison to preserve RAS

1. Internal Poison List Retrieval (Get Poison List)
   - Obtains a complete list of poisoned locations on the memory device
   - Avoids host access to memory locations with faults (to avoid DUE)
   - Addition of new poisoned locations: notified via MSI or VDM notifications
   - Clearing a poisoned location: host issues "Clear Poison" command

2. Scan Media
   - Invoked when the poison list overflowed or complete scan is needed
   - Update of the scan outcome: notified via MSI or VDM notifications
   - Based on the outcome, the host addresses the poisoned media ranges and updates the poison list if DUE found
   - Slow background operation! May stop if mailbox is full



Host OS/Hypervisor

① Read physical address X

⑦ OS/Hypervisor may terminate process/VM

⑥ Generate SIGBUS to process/VM accessing X

CXL.cache/ CXL.mem

⑤ Return CXL poison on CXL.cache/CXL.mem for X. Generate MSI -X/VDM to host.

② DUE in address X

③ Store CXL poison for X

④ Log error in Event Log

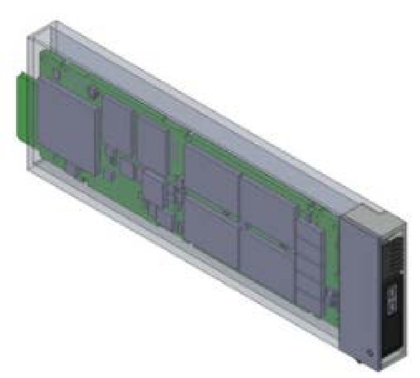
CXL Device

Poison Notification Example: Host Read, Device Response

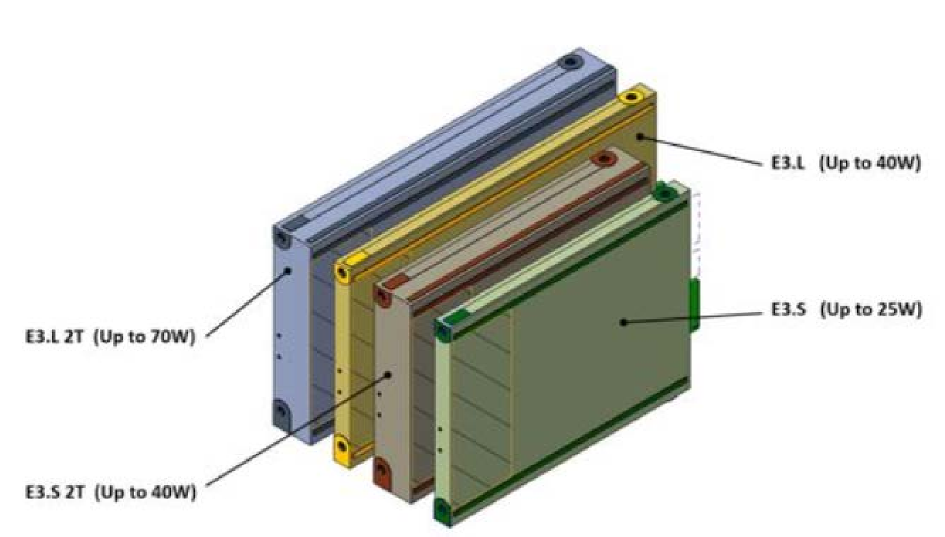
Form Factors for CXL Persistent Memory

All form factors supporting PCIe can be adopted for CXL persistent memory solutions.

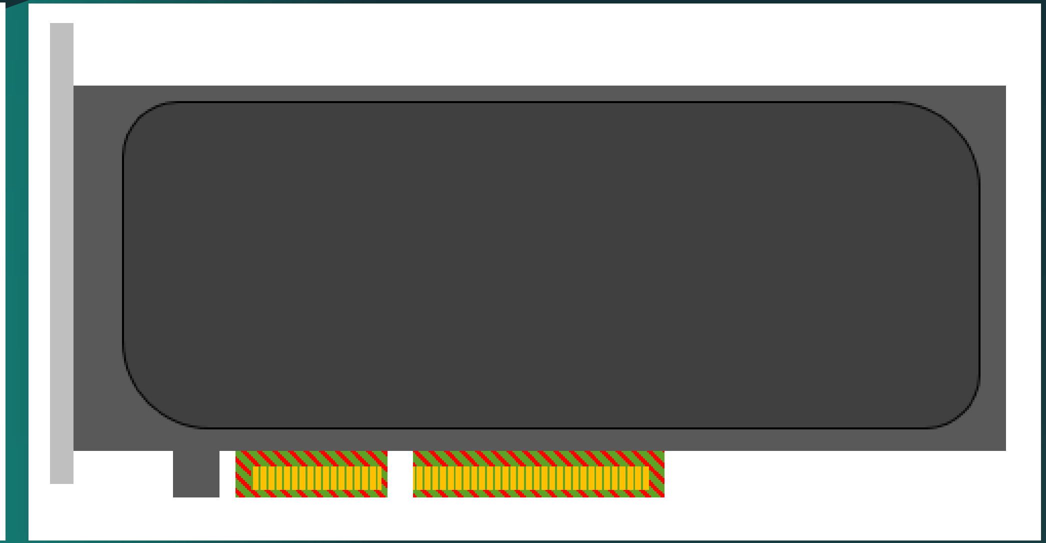
Ex #1: EDSFF E1.S



Ex #2: EDSFF E3.S / E3.L



Ex #3: Add-in Card (AIC)



Factors	EDSFF E1.S	EDSFF E3.S / E3.L	AIC
Area vs. DDRx Server DIMM	• Smaller	• Larger	• Larger (larger than E3.S/L)
Expected Max. Power Range	• 12 ~ 25W	• 25W~40W(1T), 40W~70W(2T)	• Similar range compared to E3.S/L

Reference: snia.org

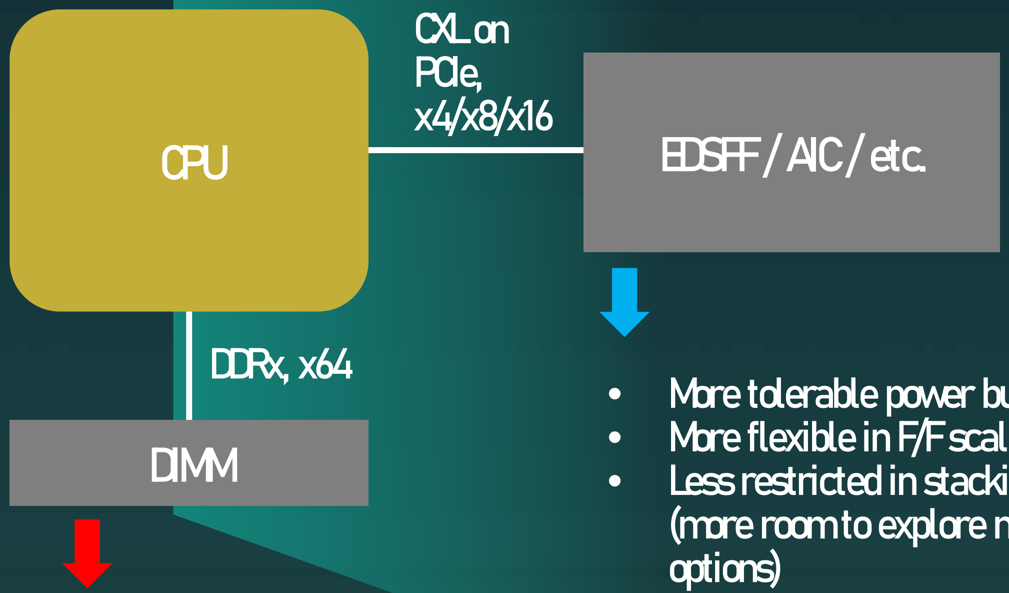
Advantages over DIMM Form Factors

CXL memory form factors allow better capacity scaling under separate expansion memory channel (i.e. PCIe).

DIMM Capacity Scaling Trend

DDRx Generation	Mainstream DIMM Speed	Max # of DIMMs per Channel
DDR3	1333 ~ 1866	3
DDR4	2133 ~ 3200	2
DDR5	4400 ~ 5600	2
	6400+	1

- 2 DIMMs per channel -> 1 DIMM per channel @ DDR5 era
- Hgh speed in DDRx restricts both capacity scaling and flexibility to allow persistent memory with relaxed BW



- More tolerable power budget (25W+)
- More flexible in F/F scaling up
- Less restricted in stacking options (more room to explore non-TSV options)
- Hot-plug more feasible

- Less tolerable power budget (15~18W)
- Less flexible in F/F scaling up
- More restricted in stacking options (harder to stack without TSVs)

Challenges in Enabling Persistent Memory

1. More experiences needed in enabling new features
   - Enabling persistent memory is still at the early stage
   - Some features need to reference the existing literature (DRAM-based), others need new paradigm
   - Example of new features applied to PMEM HW: power management, RAS, and security

2. Infrastructure readiness: HW development, SW infrastructure
   - HW development: throughput scaling is a big challenge considering power/thermal restrictions
   - SW infrastructure: ground works are done, but more exploration still needed for general purpose applications

3. User experience readiness: how to utilize the PMEM
   - Even through infrastructure is ready, still few more years needed for the users to learn how to utilize PMEM effectively!

In Summary

CXL Consortium momentum continues to grow

- 150+ members and growing
- Responding to industry needs and challenges

CXL is ideal for attaching Persistent Memory

- The protocol designed with PMEM in mind, media-agnostic
- Generic driver model eases SW enabling
- Robust RAS and reliability features
- Variety of Form factors enable innovative system designs

Call to action

- Join CXL Consortium
- Follow us on YouTube, Twitter and LinkedIn for more updates!

Compute Express Link™ and CXL™ Consortium are trademarks of the Compute Express Link Consortium.

CXL Compute Express Link ™

Thank You