STORAGE DEVELOPER CONFERENCE
SDC 21
BY Developers FOR Developers

Virtual Conference
September 28-29, 2021

A SNIA Event

Innovations in Load-Store I/O causing Profound Changes in Memory, Storage, and Compute landscape

Featured Keynote

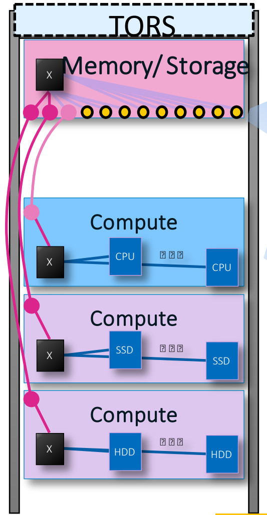
Dr. Debendra Das Sharma
Intel Fellow and Director of I/O Technology and Standards
Intel Corporation

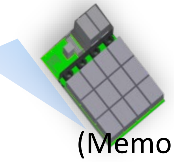
Agenda

- Interconnects in Memory, Storage, and Compute Landscape
- Load-Store I/O Evolution
- Memory, Storage, and Compute innovations with Load-Store I/O

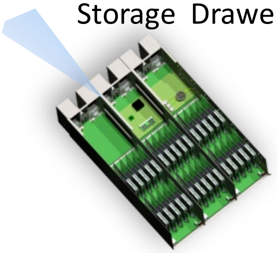
Compute Landscape today

(Rack of Servers)



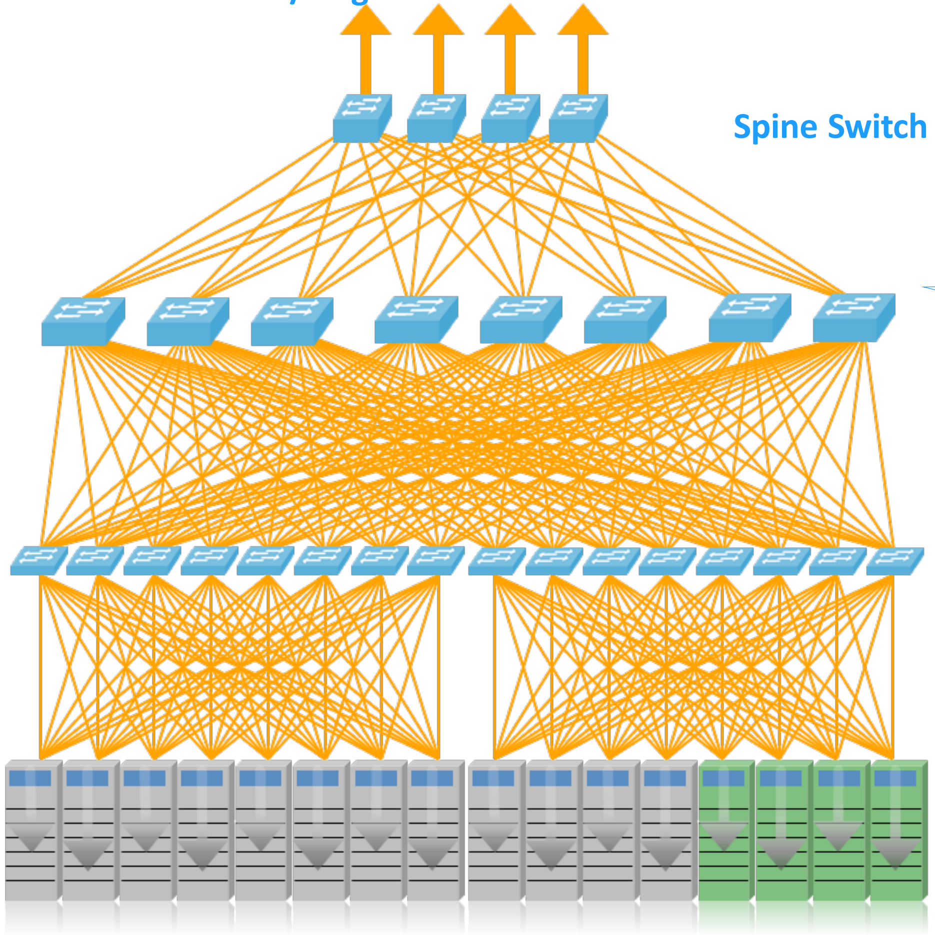


(Memory/ Storage Drawer)

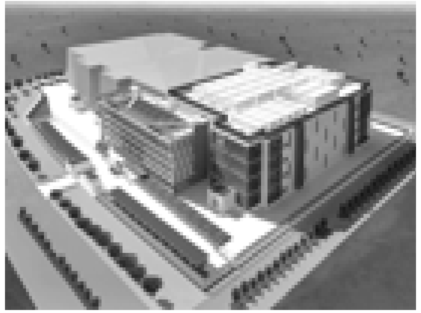


(Compute Drawer)

Ld/ St I/O inside drawer and Rack



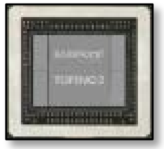
Hyperscale Data Center and Edge



Optical Modules

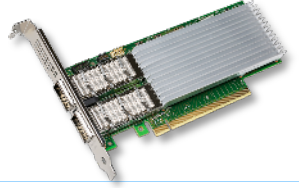
High-bandwidth connectivity at 100G and beyond

Leaf Switch



P4-programmable scale-out fabric with uncompromising performance

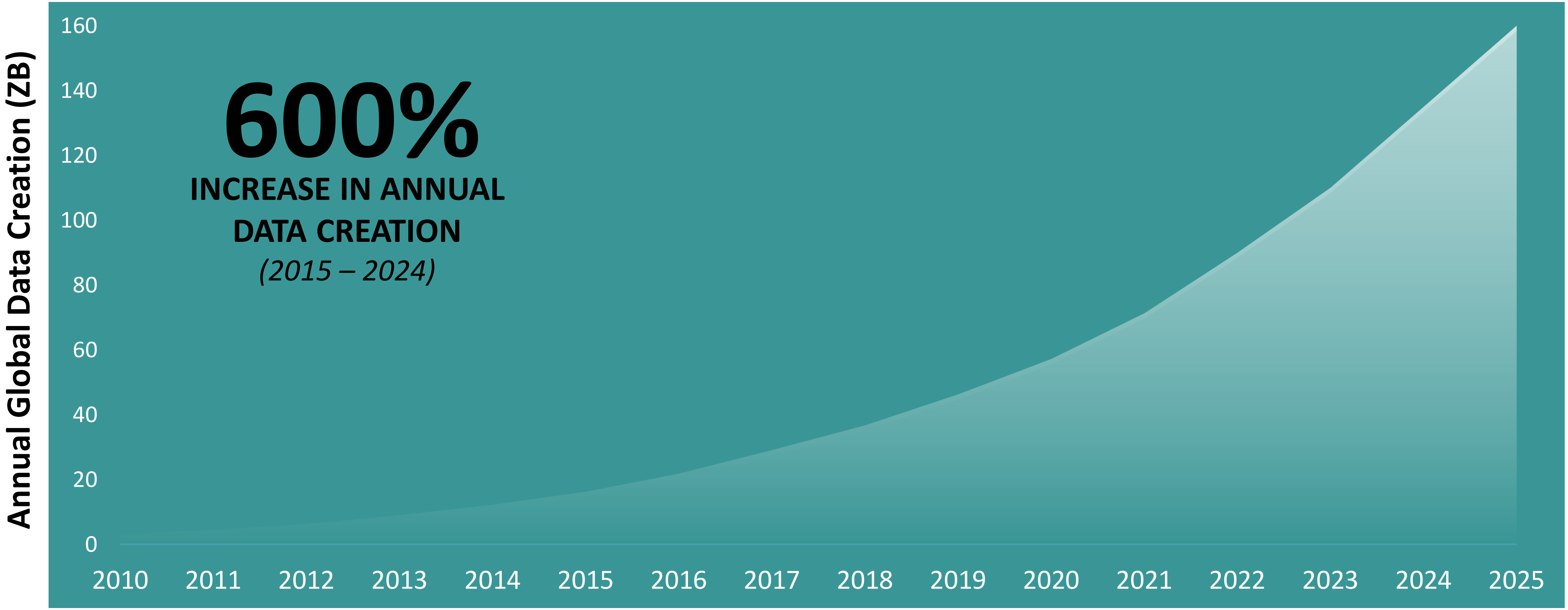
Rack of Servers



Programmable infrastructure acceleration for demanding data movement with Smart NIC

Industry Mega-trends: Proliferation of Cloud; Growth of AI and analytics; Cloudification of Network and Edge
Data Center as a Computer –Interconnects are key to driving warehouse scale efficiency!

Explosion of data enabling data-centric revolution



Source: IDC Data Age 2025

Drivers: Cloud, 5G, sensors, automotive, IoT, etc.. Large data sets with aggressive time to insight goals!
Scaling challenges: Latency, Bandwidth, Capacity all important!
Move faster, Store more, Process everything seamlessly, efficiently, and securely

Taxonomy, characteristics, and trends of interconnects

Category	Type and Scale	Data Rate/ Characteristics	PHY Latency (Tx + Rx)
Latency Tolerant (Narrow, very high speed)	Networking / Fabric Data Center Scale	56/ 112 GT/s-> 224 GT/s (PAM4) 4-8 Lanes, cables/ backplane	100+ ns w/ FEC (20ns+ w/o FEC)
Latency Sensitive (Wide, high speed)	Load-Store I/O Arch. Ordering (PCIe/ CXL / SMP cache coherency – PCIe PHY based) Node level (moving to sub-Rack level)	32 GT/s (NRZ) -> PCIe Gen6 64 GT/s (PAM4) Hundreds of Lanes Power, Cost, Si-Area, Backwards Compatible, Latency, On-board -> cables/ backplanes	<10ns (Tx+ Rx: PHY-PIPE) 0-1ns FEC overhead

Latency Sensitive I/O moving to PAM-4: innovations on track to meet latency, area, and cost challenges

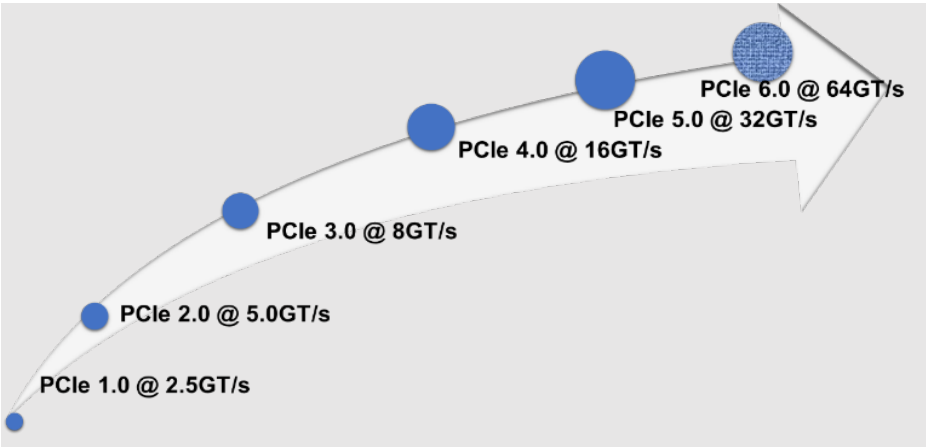
Agenda

- Interconnects in Memory, Storage, and Compute Landscape
- Load-Store I/O Evolution
- Memory, Storage, and Compute innovations with Load-Store I/O

Evolution of PCI-Express: Speeds and Feeds

- Double data rate every gen in ~3 years
- Full backward compatibility
- Ubiquitous I/O: PC, Hand-held, Workstation, Server, Cloud, Enterprise, HPC, Embedded, IoT, Automotive
- One stack / silicon, multiple form-factors
- Different widths (x1/ x2/ x4/ x8/ x16) and data rates fully inter-operable
  - a x16 Gen 5 interoperates with a x1 Gen 1!
- PCIe deployed in all computer systems since 2003 for all I/O needs
- Drivers: Networking, XPUs, Memory, Alternate Protocol – need to keep w/ compute cadence

Six generations of evolution spanning 2 decades! Supporting the Load-store interconnects seamlessly!

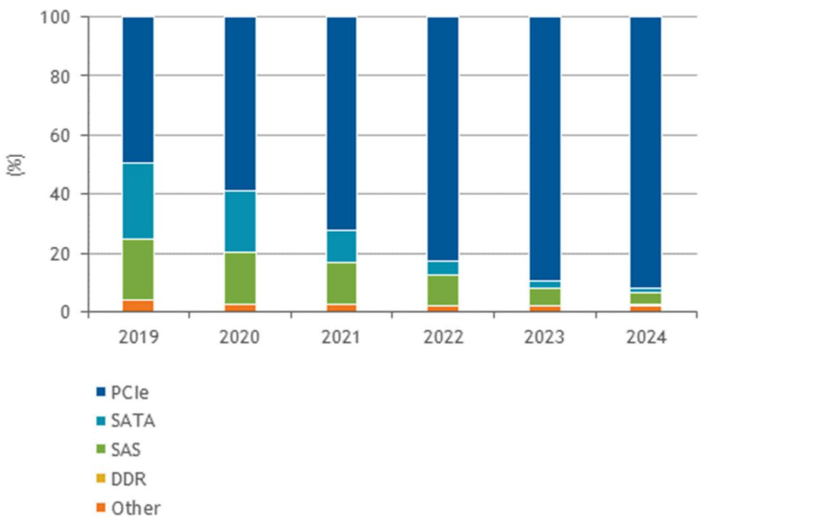


PCIe Specification	Data Rate(Gb/s) (Encoding)	x16 B/W per dirn**	Year
1.0	2.5 (8b/10b)	32 Gb/s	2003
2.0	5.0 (8b/10b)	64 Gb/s	2007
3.0	8.0 (128b/130b)	126 Gb/s	2010
4.0	16.0 (128b/130b)	252 Gb/s	2017
5.0	32.0 (128b/130b)	504 Gb/s	2019
6.0 (WIP)	64.0 (PAM-4, Flit)	1024 Gb/s (~1Tb/s)	2021*

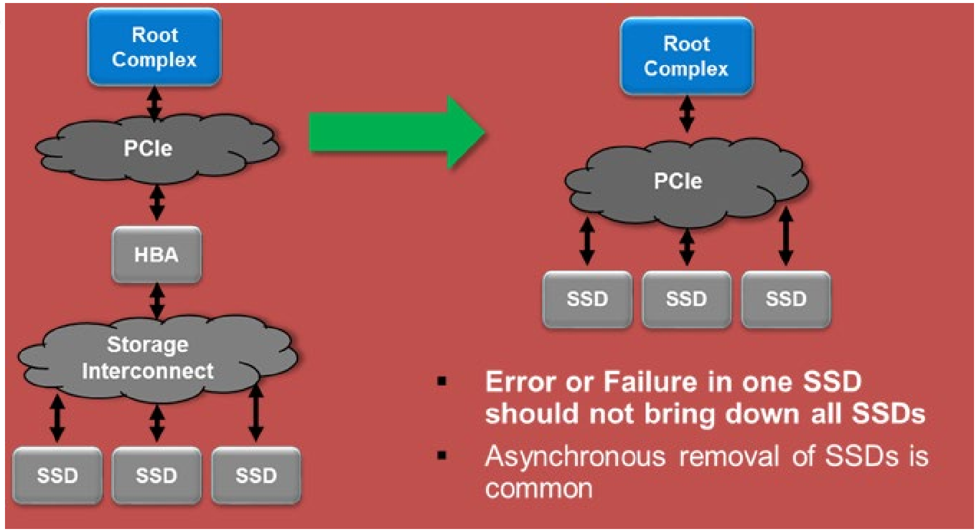
PCIe Features useful for Storage

- Predictable performance cadence
  - Low-latency, High Bandwidth, Scalability, backward compatibility – NVMe
- I/O Virtualization, RAS, and Hot-Plug Features
- Multitude of form factors including cabling support

Worldwide Enterprise SSD Capacity Shipment Share by Interface, 2019-2024

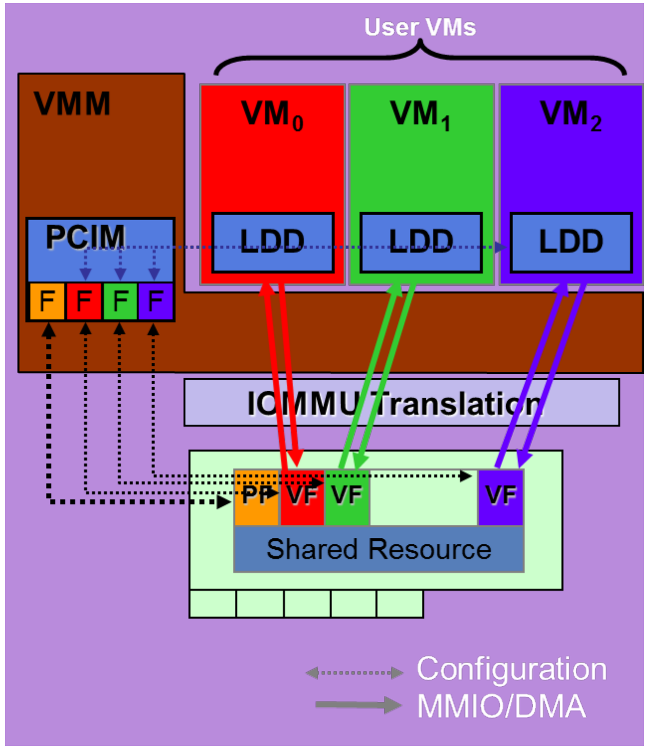


Source: IDC, 2020



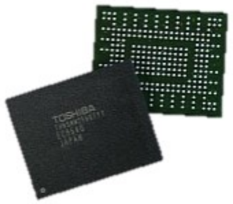
- Error or Failure in one SSD should not bring down all SSDs
- Asynchronous removal of SSDs is common

(RAS Enhancements: (e)DPC)



(IO Virtualization)

PCIe Form Factors



BGA

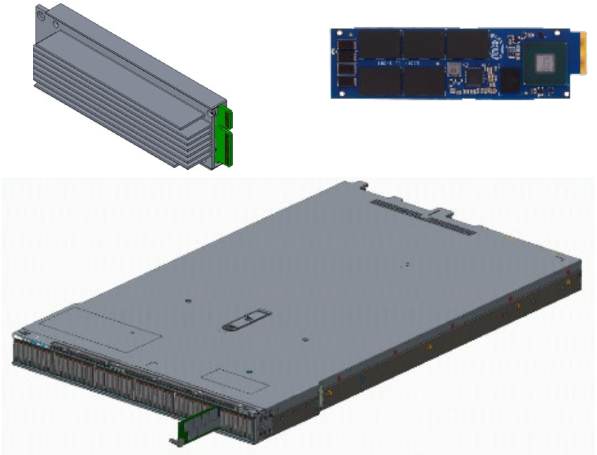
M.2

U.2 2.5in

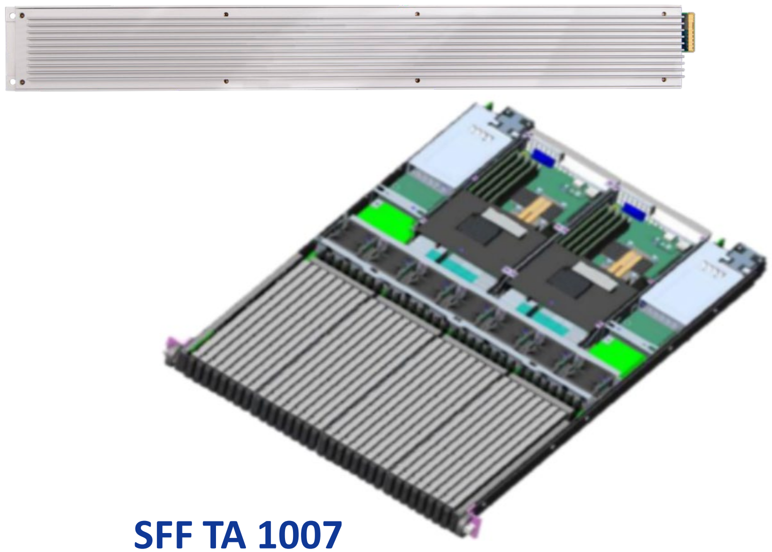
CEM Add-in-card

SD Express

SFF TA 1002



SFF TA 1006

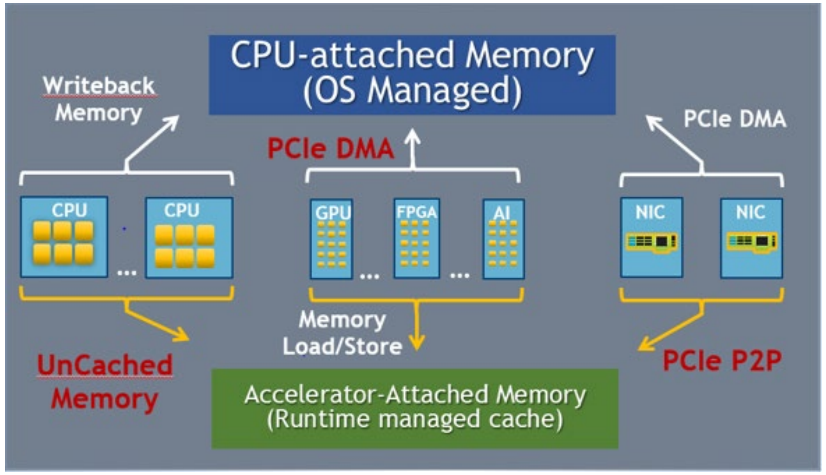


SFF TA 1007

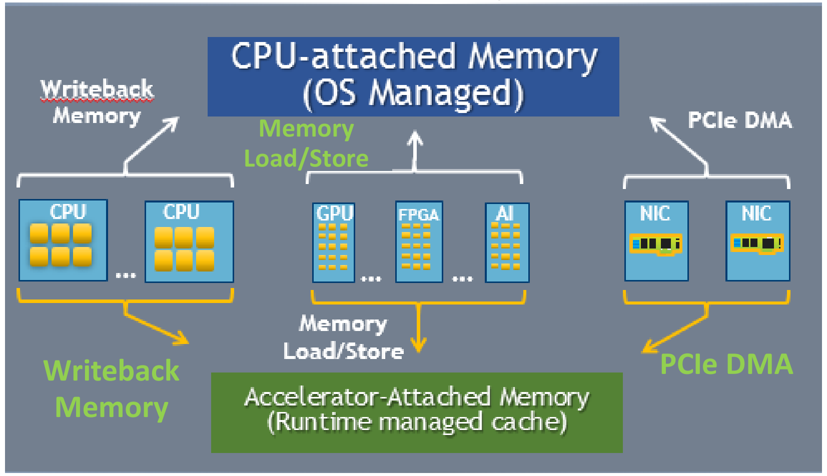
One PCIe specification, one PCIe stack, same silicon in multiple form-factors for different segments

CXL: A new class of open-standard interconnect

- Heterogenous computing and disaggregation
- Efficient resource sharing
- Shared memory – efficient access
- Enhanced movement of operands and results
- Memory bandwidth and capacity expansion
  - Memory tiering and different memory types
- CXL is an open industry standard interconnect with 150+ members
  - All CPU, GPU, Memory vendors in consortium
  - Tremendous momentum in the ecosystem
    - interop/ product announcements
  - CXL poised to be a game-changer in the industry!!



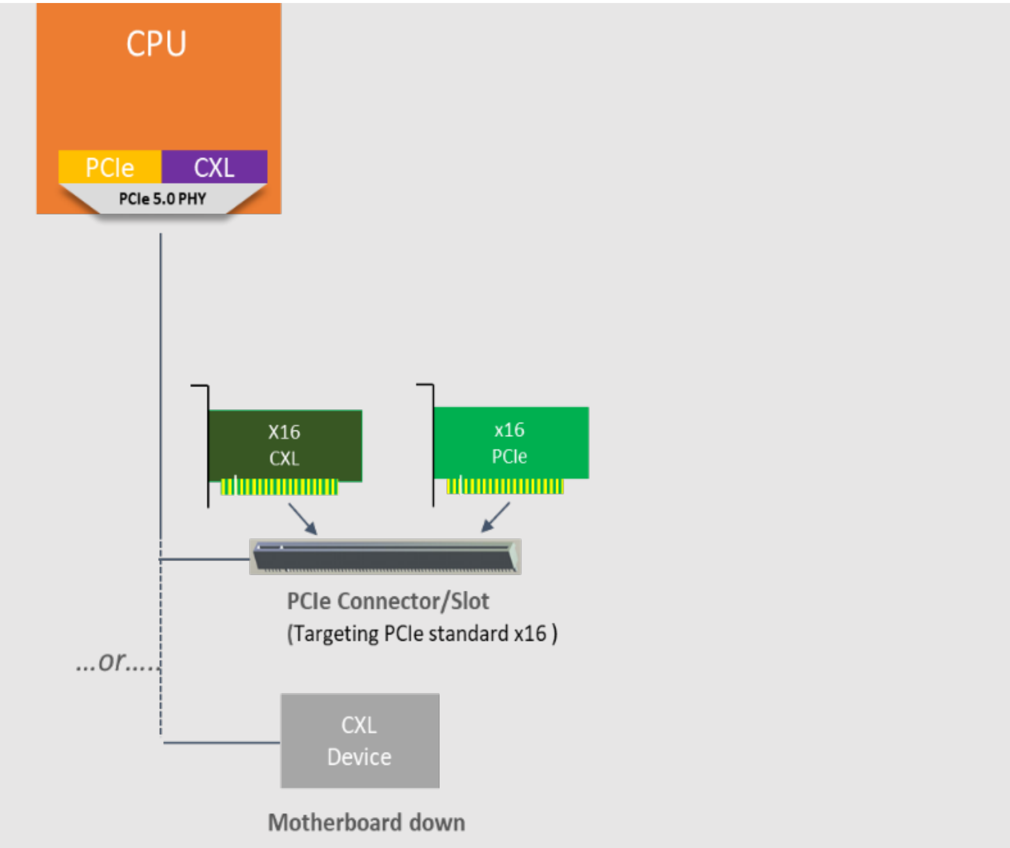
With PCIe-only



CXL Enabled Environment

CXL on PCIe® Infrastructure

- PCIe 5.0 PHY at 32 GT/s
  - Can down-grade to 8 / 16 GT/s
- Widths: x4, x8, x16
- Full Plug and play capable
  - Either a CXL card or a PCIe card
  - Protocol negotiated early in training
- Complete leverage of PCIe



Compute Express Link has the benefit of supporting both standard PCIe devices as well as CXL devices – all on the same Link

CXL approach

Coherent Interface

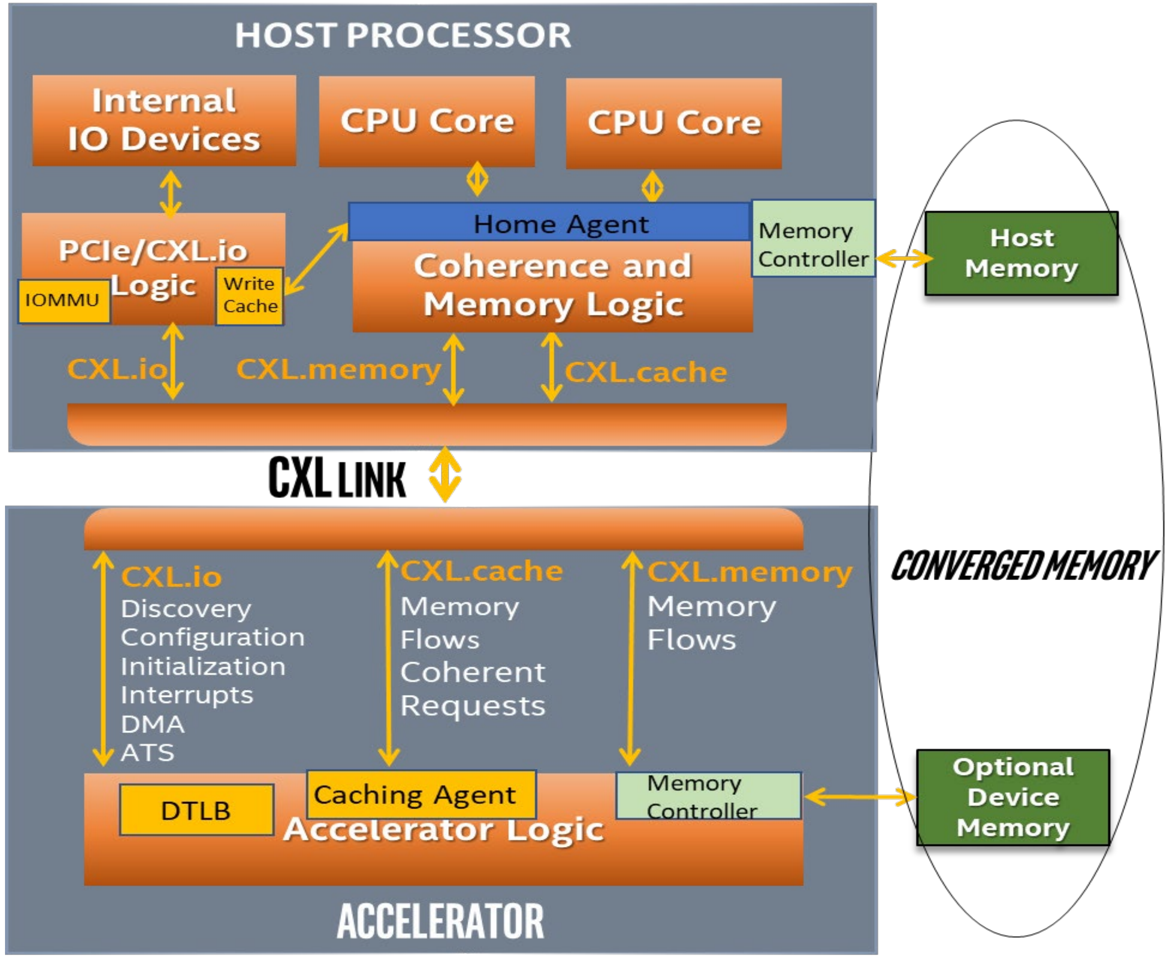
Leverages PCIe with 3 mix-and-match protocols
Built on top of PCIe infrastructure

Low Latency

.Cache and .Memory targeted at near CPU cache coherent latency (<200ns load to use)

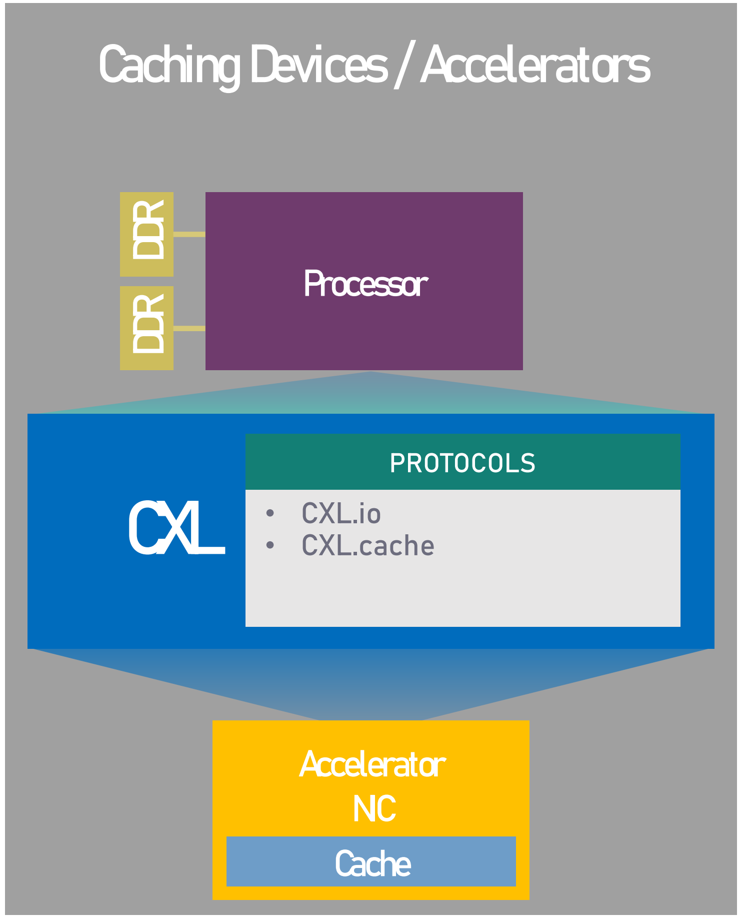
Asymmetric Complexity

Eases burdens of cache coherent interface designs



CXL 1.0 Usage Models

Caching Devices / Accelerators



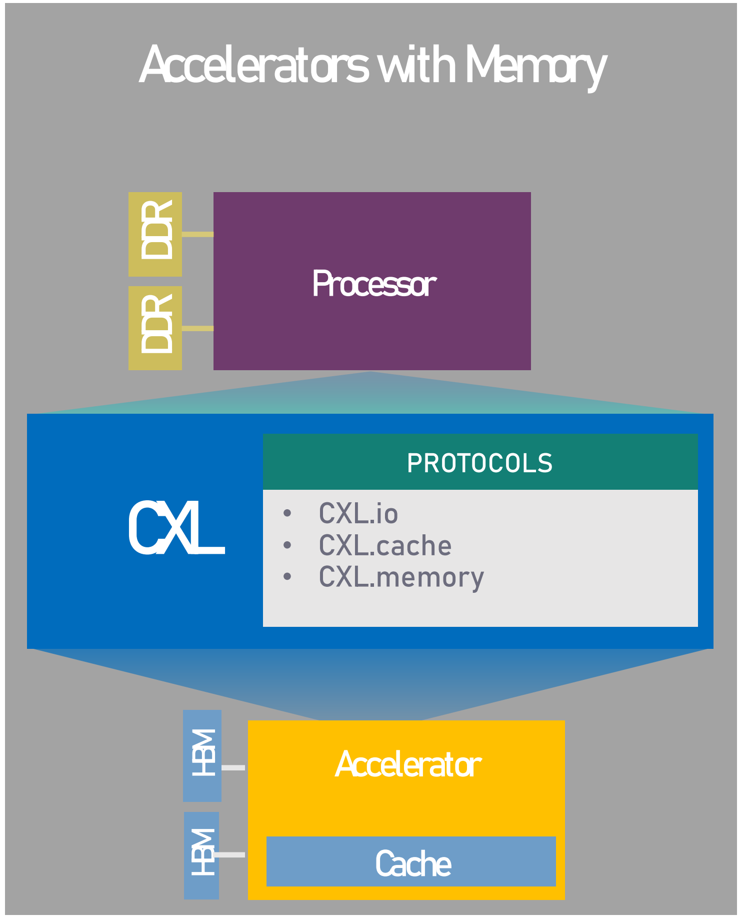
PROTOCOLS

- CXL.io
- CXL.cache

USAGES

- Smart NIC
- NIC atomics, PGAS, RDMA

Accelerators with Memory



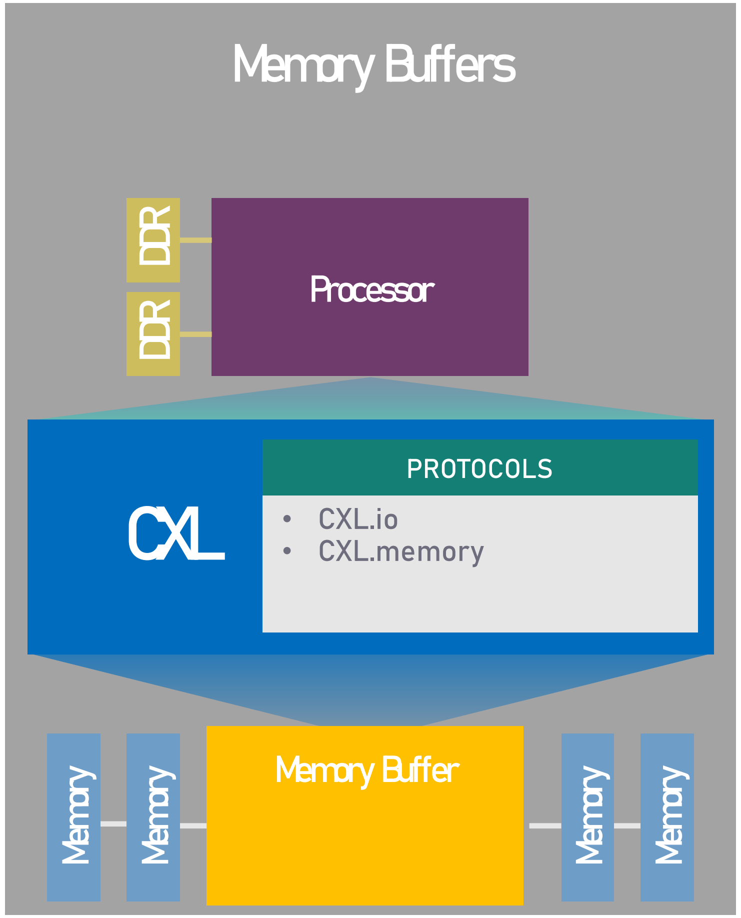
PROTOCOLS

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

USAGES

- GP GPU
- Dense computation

Memory Buffers



PROTOCOLS

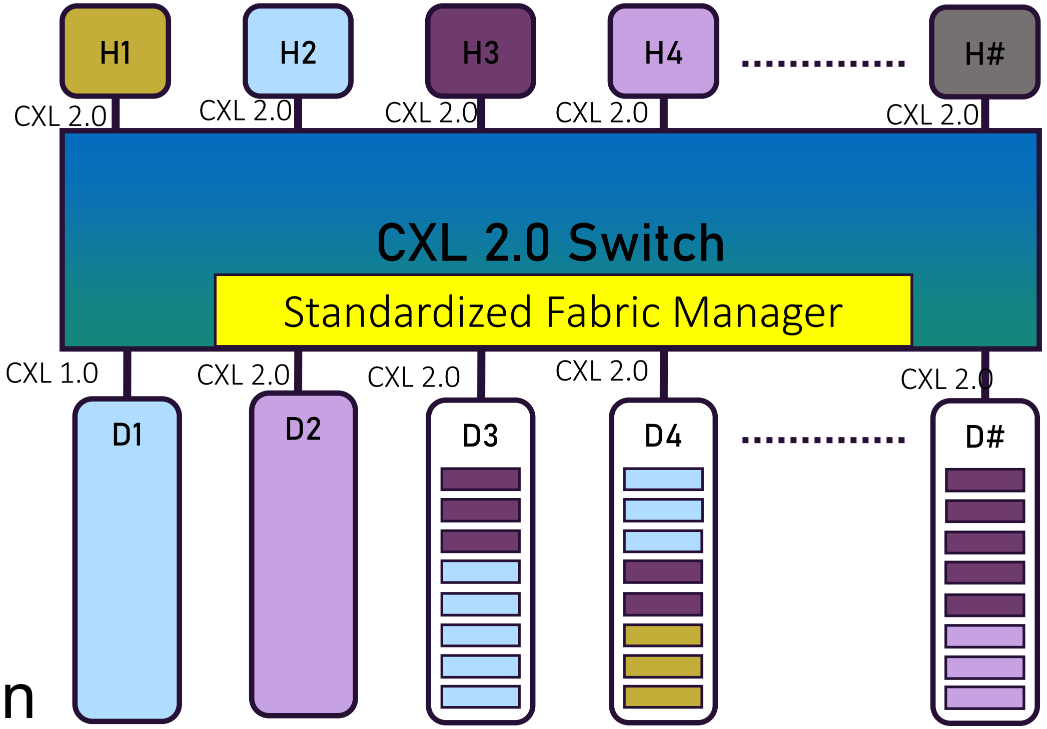
- CXL.io
- CXL.memory

USAGES

- Memory BW expansion
- Memory capacity expansion
- Storage class memory

CXL 2.0 enables resource pooling at rack level, Persistence Flows, and enhanced security

- Switching for fan-out and pooling
- Managed Hot-plug flows to move resources
- Persistence flows for persistent memory
- Type-1/ -2 device assigned to one host
- Type-3 device (memory) pooling across multiple hosts at Rack level
- Fabric Manager for managing resources
- Software API for devices
- Security enhancement: authentication, encryption
- Beyond node to Rack-level connectivity!!



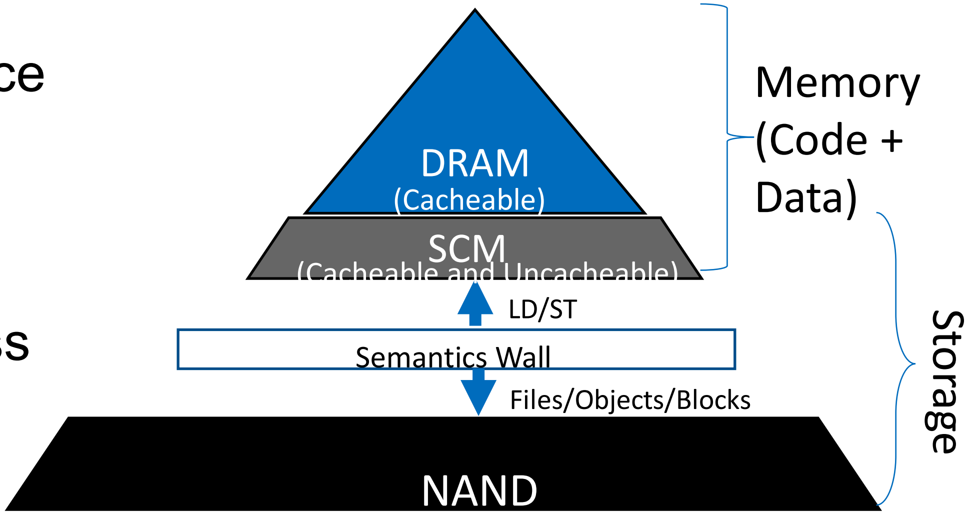
Dis-aggregated System with CXL optimizes resource utilization delivering lower TCO and power-efficiency

Agenda

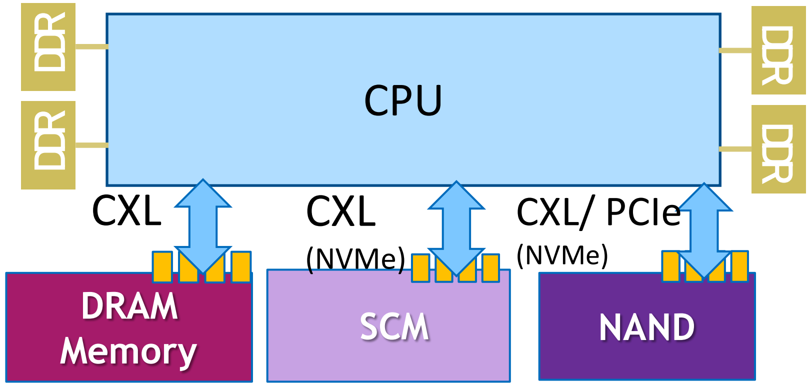
- Interconnects in Memory, Storage, and Compute Landscape
- Load-Store I/O Evolution
- Memory, Storage, and Compute innovations with Load-Store I/O

CXL implications on memory and storage

- CXL provides a media-independent, coherent memory interface
  - CXL.io preserves all PCIe functions / services (e.g., NVM Express)
  - Enables new compute and memory architectures
  - Spans DRAM, NRAM/ MRAM, and storage class memories
- Additive bandwidth and capacity over traditional DIMMs across multiple types of memory and hierarchy without interference
- PCIe form-factor enables higher power profiles (25+ W)
  - Lots of choices of form-factors and power profiles
  - Does not consume a DIMM slot
  - Unlike DIMM form-factor not constrained by 15-18 W
- Other benefits
  - Standard device discovery, configuration, and management
  - Software leverage: PCIe driver, ACPI – Heterogeneous Memory Attribute Table (HMAT) to describe properties of memory
  - DMA engine for data move – leverage PCIe
  - I/O Virtualization from PCIe

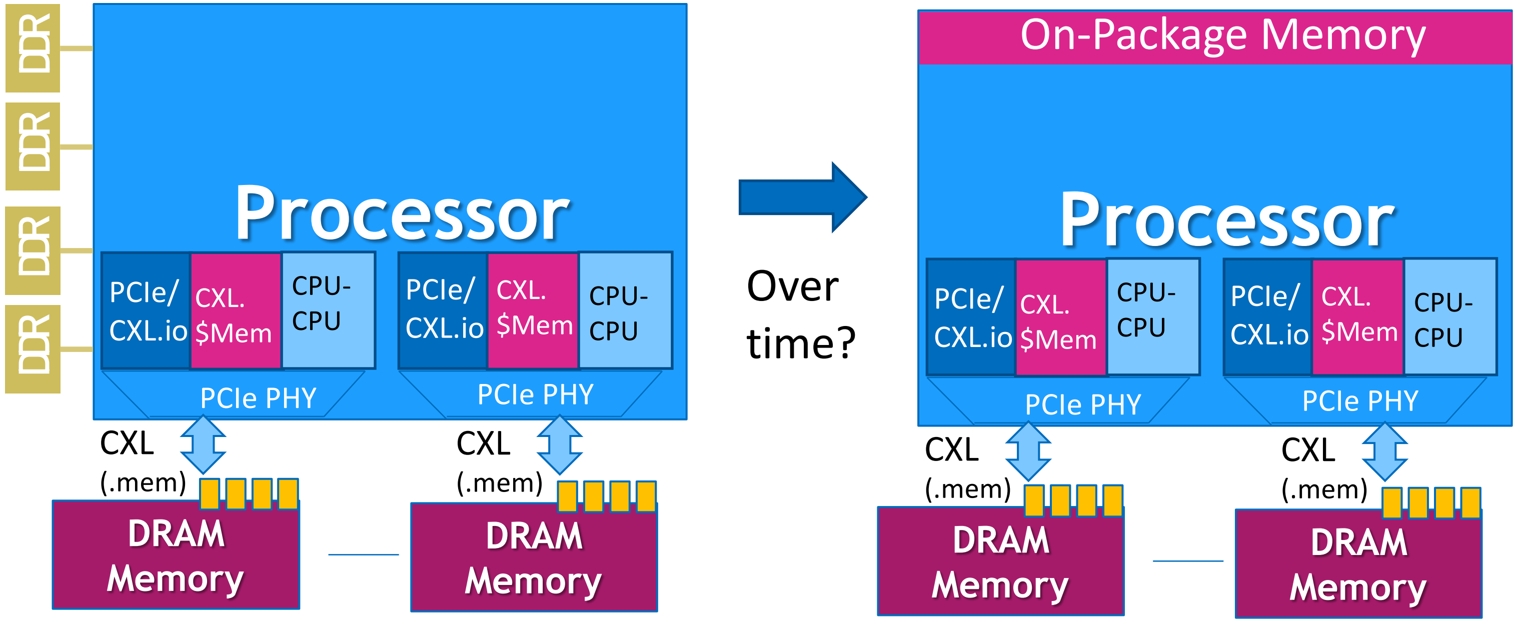


(CXL can span the entire memory hierarchy)



Capacity and Bandwidth Expansion with CXL-attached memory

- Common platform across wide usages
  - decoupling compute from traditional DIMM memory bandwidth/ capacity
- Scalable bandwidth (width and frequency), low latency, pin efficiency
  - X8 @ Gen 5, x4 @ Gen 6: 32 GB/s per direction
- Memory now serviceable with front-loaded form-factor
- Amount of memory in DIMM vs CXL?
  - NUMA domains are well established.
  - Would we see systems with only on-package memory and CXL memory?

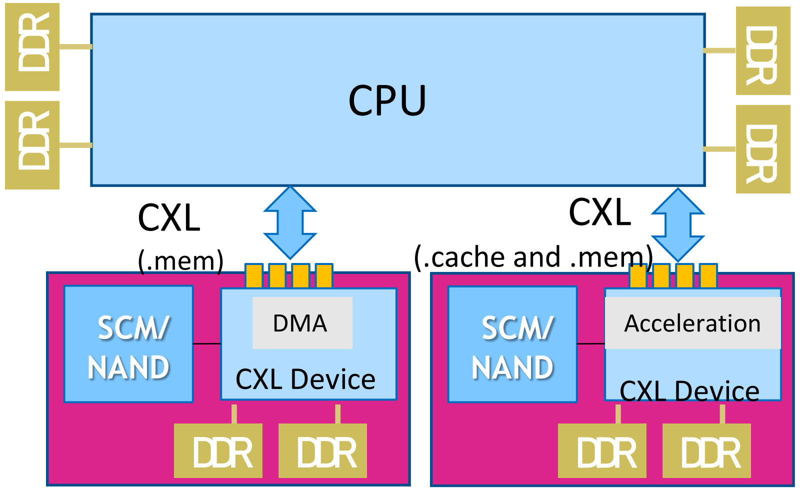


Memory Capacity and Bandwidth Expansion with CXL

CXL becomes the only external memory attach point

Persistent Memory innovations with CXL

- NVDIMM moves to CXL with DRAM backed up by SCM/ NAND
  - Pros: Serviceable, multi-headed, power profile, free up a DIMM slot
- Persistent Memory is now capable of being cacheable!
  - Multi-headed for fail-over
  - Serviceable - hot-plug
- Multi-level Memory hierarchy for larger capacity
  - DRAM as memory-side cache for lower latency
  - Mapping the entire SCM to cacheable memory – use the HMAT table and interleaving accordingly
  - DMA move engine for NVMe type usage
- Accelerator engines for near-memory processing

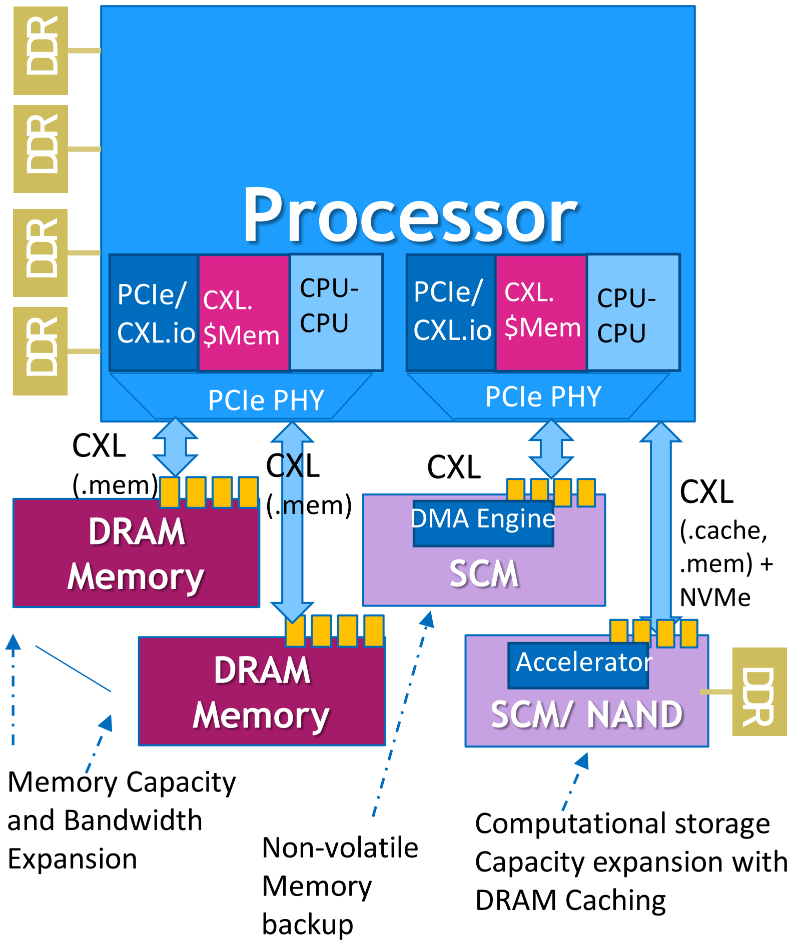


2LM for capacity
Memory backup

Computational storage
2LM for capacity

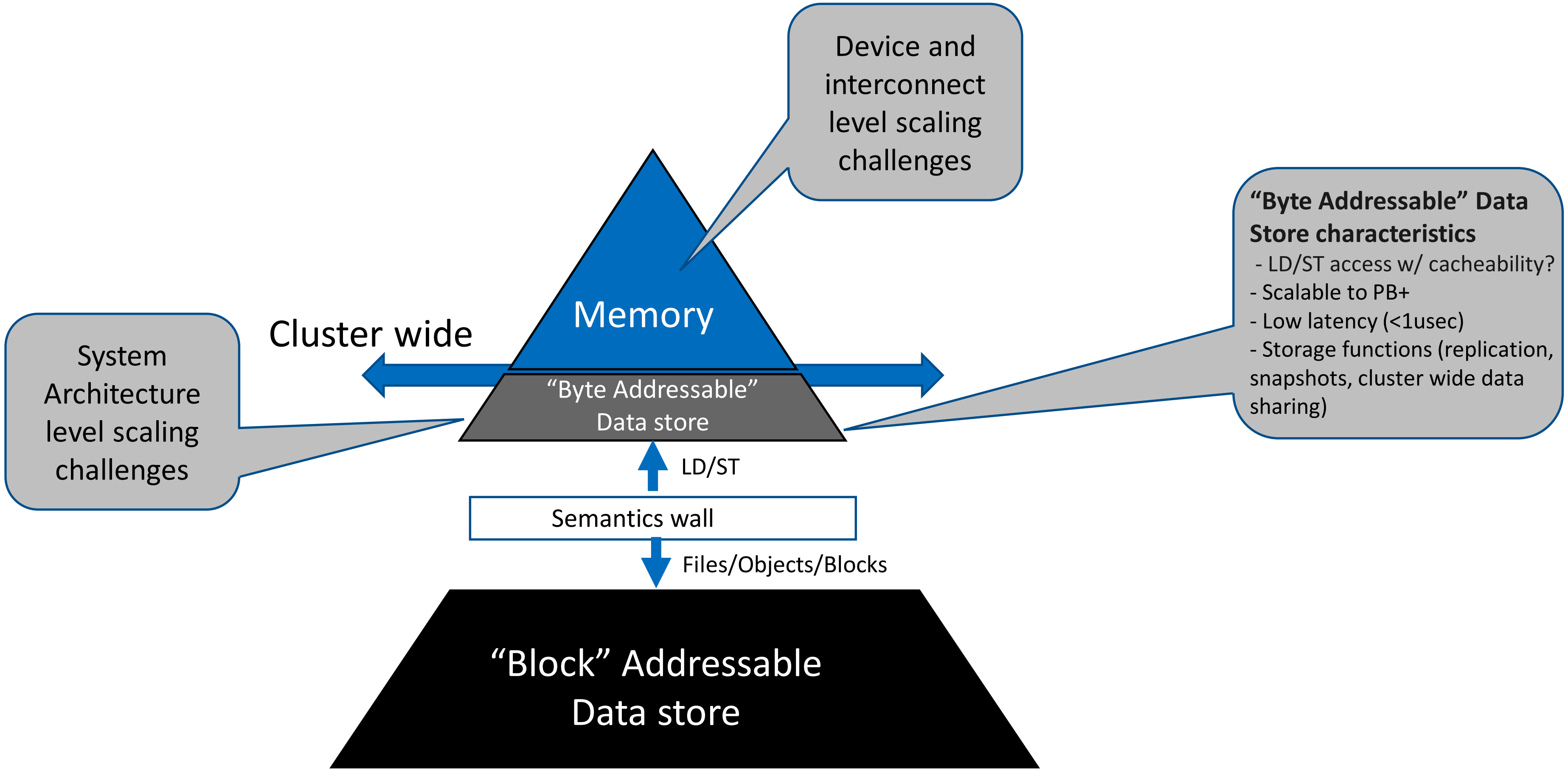
Computational Storage and Memory with Host Memory Sharing

- Accelerator in front with compute functions with caching semantics
  - compression, encryption, RAID, compaction for key-value store, search engine, or vector processing for AI/ML applications, etc.
- DMA engine for data move
- Leverage PCIe services, including NVM-Express
  - standard drivers and management framework that we have developed over the years in PCI Express.



CXL enables systems to scale with heterogeneous processing and memory with shared cacheable memory space accessible to all using same mechanisms

Cluster-wide storage and memory tier

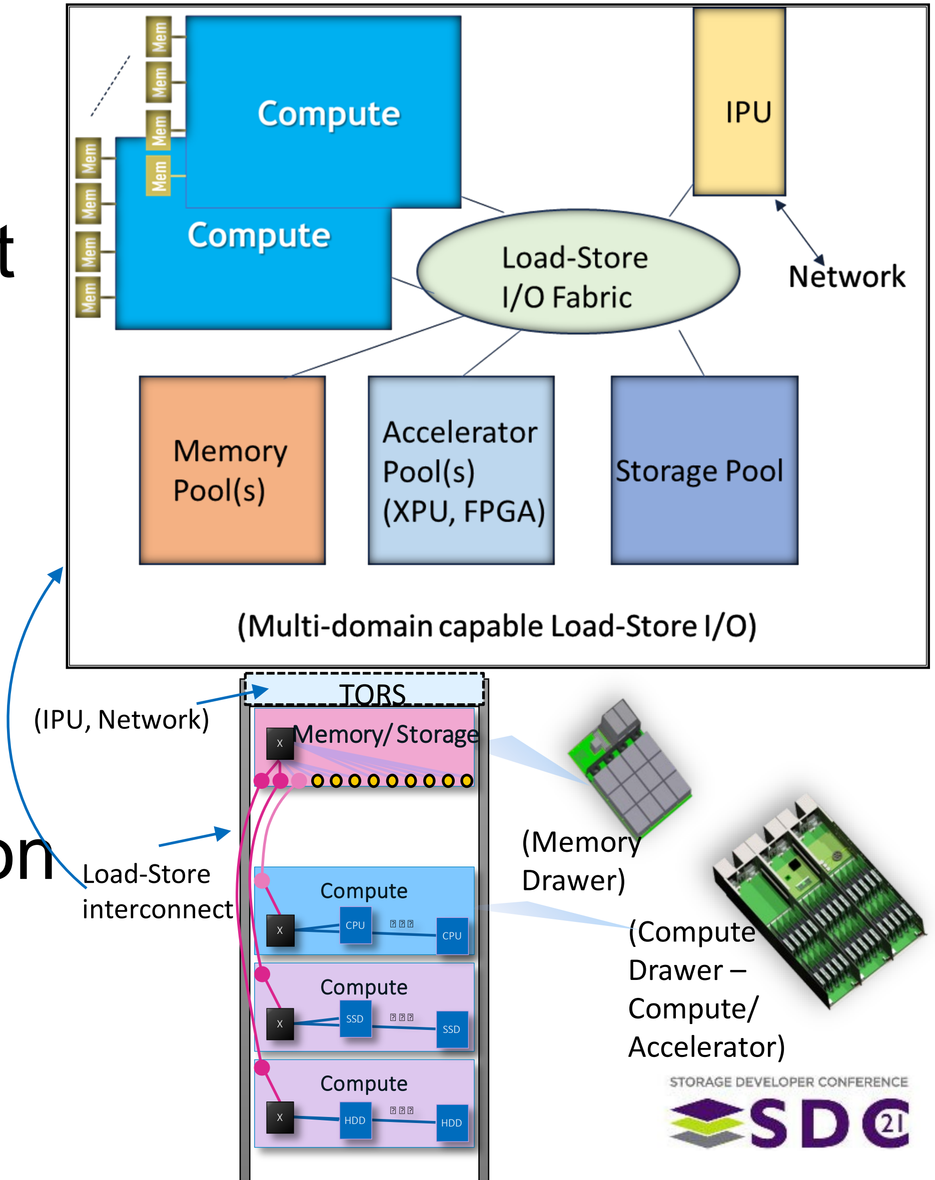

Device and interconnect level scaling challenges
Memory
Cluster wide
System Architecture level scaling challenges
"Byte Addressable" Data store
LD/ST
Semantics wall
Files/Objects/Blocks
"Block" Addressable Data store

"Byte Addressable" Data Store characteristics
- LD/ST access w/ cacheability?
- Scalable to PB+
- Low latency (<1usec)
- Storage functions (replication, snapshots, cluster wide data sharing)

Leveraged from SRC Round-Table 2021 presentation on Memory Scaling by Balint Fleischer, Micron

Rack-level disaggregation with CXL

- Heterogenous compute/ memory, storage, networking fabric resources
- High b/w, low-latency Load-Store Interconnect
- Iso power-performance as direct connect
- Multiple domains, shared memory, message passing, atomics, peer-to-peer accesses
- Memory protection through replication/ RAID
- Fabric Manager, Multi-head, multi-domain, Atomics, Persistence, Smart NIC, VM migration
- Address: Blast Radius, containment and QoS
- Software! Software! Software!



Please take a moment to rate this session.

Your feedback is important to us.