Exacomm 2024 Workshop

Compute Express Link (CXL)*: An open interconnect for HPC and AI applications

Dr. Debendra Das Sharma

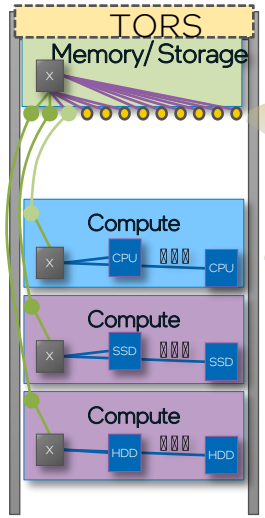
Intel Senior Fellow and Chief I/O Architect, Data Center and AI Group, Intel Corporation

Agenda

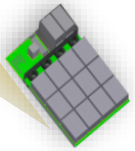
- Load-Store I/O in Compute Landscape
- Compute Express Link and its evolution
- Conclusions and Call to Action

Compute Landscape today

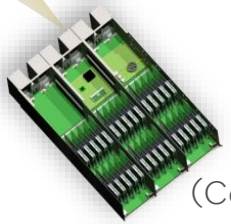
(Rack of Servers)



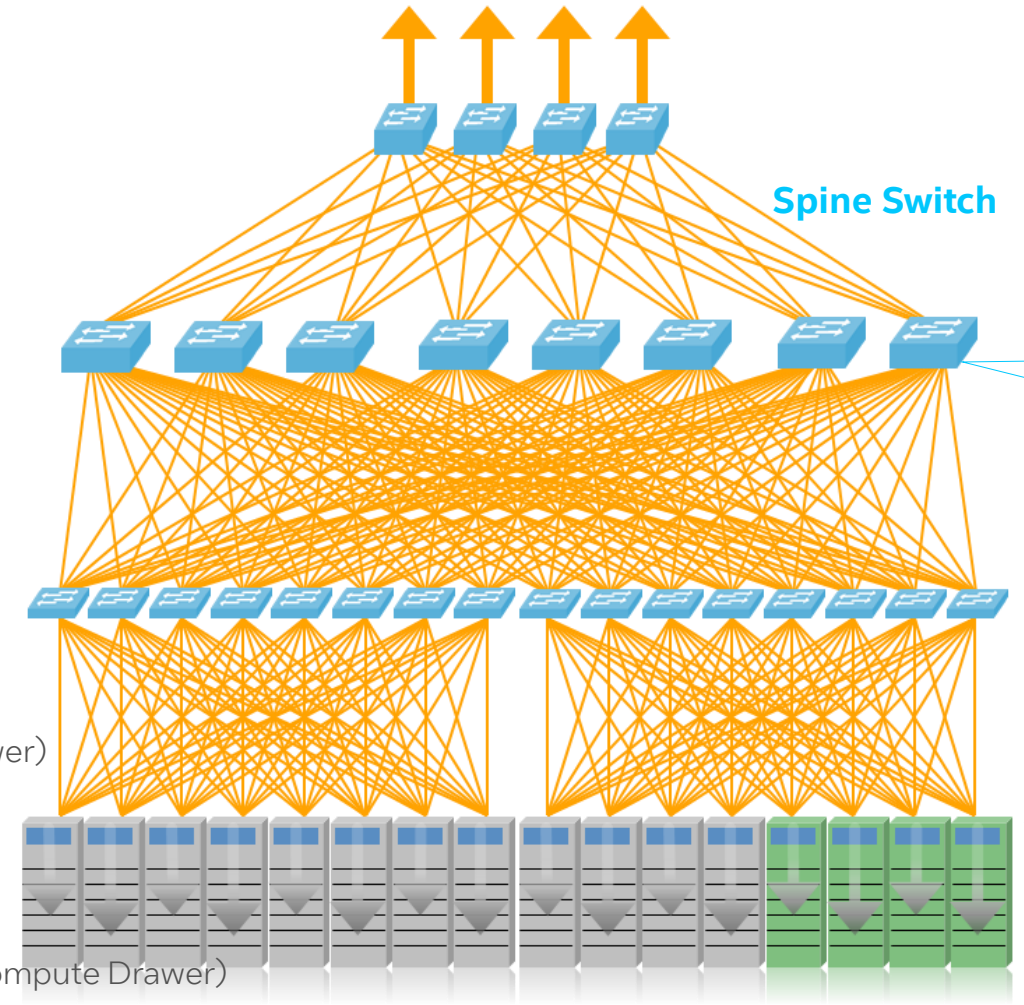
Ld/ St I/O inside drawer and Rack



(Memory/ Storage Drawer)



(Compute Drawer)



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

Data Center as a Computer –Interconnects are key to driving warehouse scale efficiency!

Evolution of PCI-Express: Speeds and Feeds

- Double data rate every gen in ~3 years since 2003
- 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!
- Drivers: Networking, XPUs, Memory, Alternate Protocol – need to keep w/ compute cadence

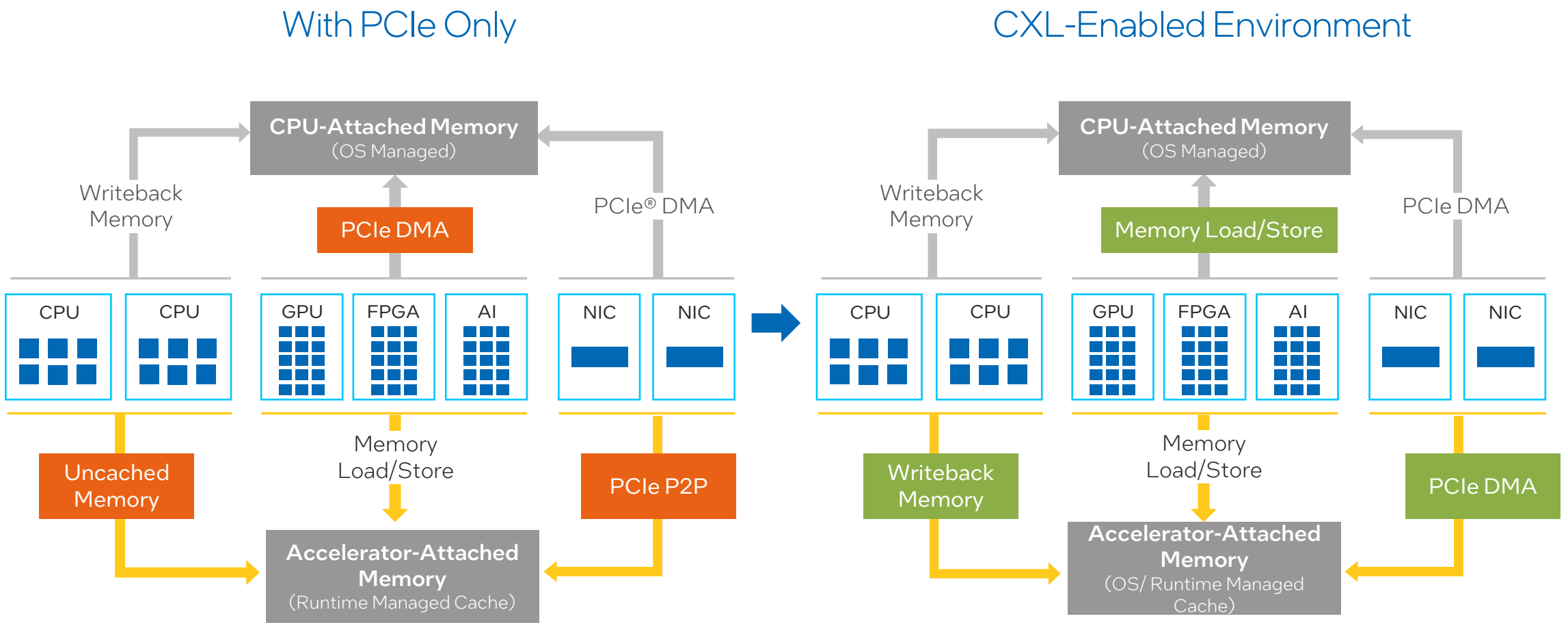
PCIe Specification	Data Rate(Gb/s) (Encoding)	x16 B/W per dirn**	Year
1.0	2.5 (8b/10b)	32 Gb/s	2003
2.0	5.0 (8b/10b)	64 Gb/s	2007
3.0	8.0 (128b/130b)	126 Gb/s	2010
4.0	16.0 (128b/130b)	252 Gb/s	2017
5.0	32.0 (128b/130b)	504 Gb/s	2019
6.0	64.0 (PAM-4, Flit)	1024 Gb/s	2022
7.0 (WIP)	128.0 (PAM-4, Flit)	2048 Gb/s	2025*

PCIe supporting the Load-store interconnects seamlessly! With cable and retimers, we can extend the reach from drawer to Rack

Agenda

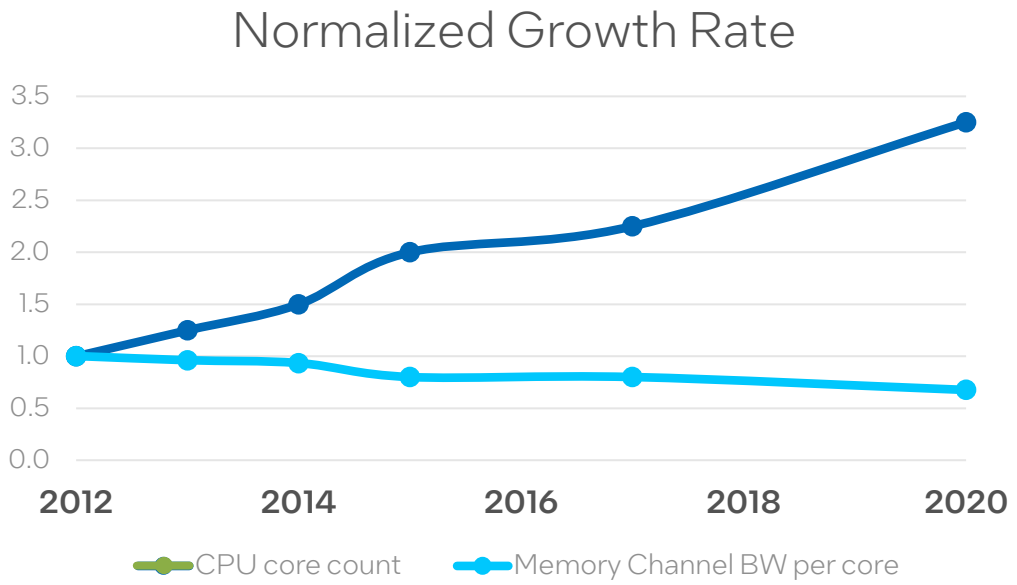
- Load-Store I/O in Compute Landscape
- Compute Express Link and its evolution
- Conclusions and Call to Action

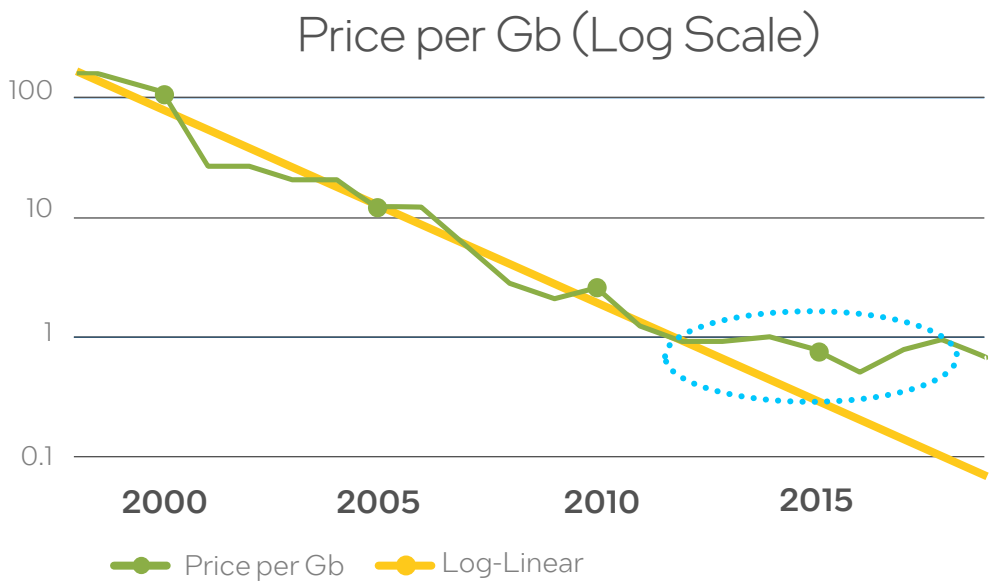
CXL™: A New Class of Open-Standard Interconnects



Challenge 1 addressed by CXL: Heterogeneous compute with coherent access to system and device memory

The System Memory Challenge





- Increasing core counts drives memory demand
- Increasing bandwidth and capacity
- Memory is not able to keep up –> more DDR channels (cost, power and feasibility challenges)

Challenge 2 addressed by CXL: Memory Scalability

- Memory is an increasing % of system power and cost
- Memory price (cost/bit) is flat due to scaling challenges
- Memory power scaling with speed

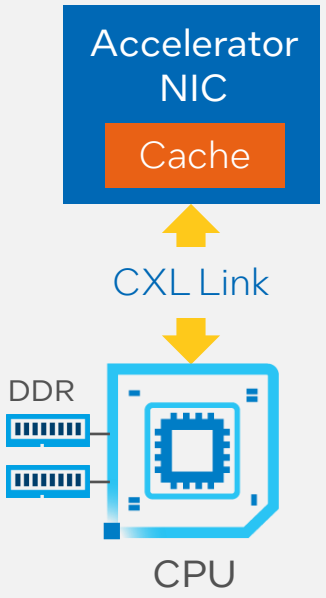
Source: De Dios & Associates

CXL™ 1.0/CXL 1.1 Usage Models

Type 1 Device
Caching Devices/Accelerators

Usages:
- PGAS NIC
- NIC atomics

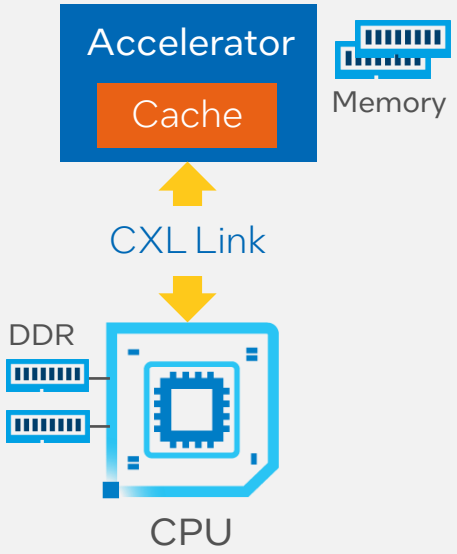
Protocols:
- CXL.io
- CXL.cache



Type 2 Device
Accelerators with Memory

Usages:
- GPU
- FPGA
- Dense
- Computation

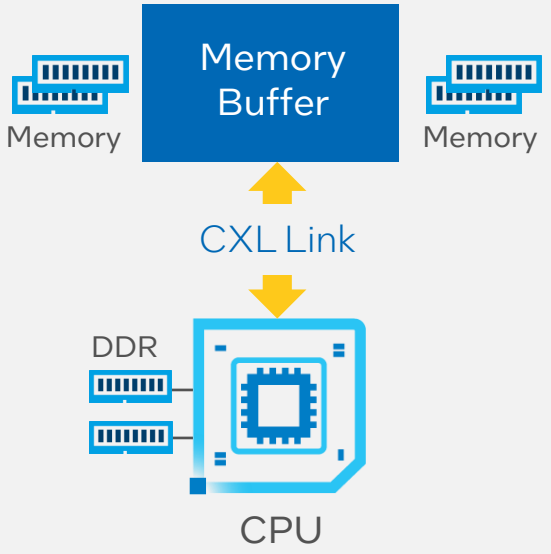
Protocols:
- CXL.io
- CXL.cache
- CXL.memory



Type 3 Device
Memory Buffers

Usages:
- Memory BW expansion
- Memory capacity expansion
- 2LM

Protocols:
- CXL.io
- CXL.mem

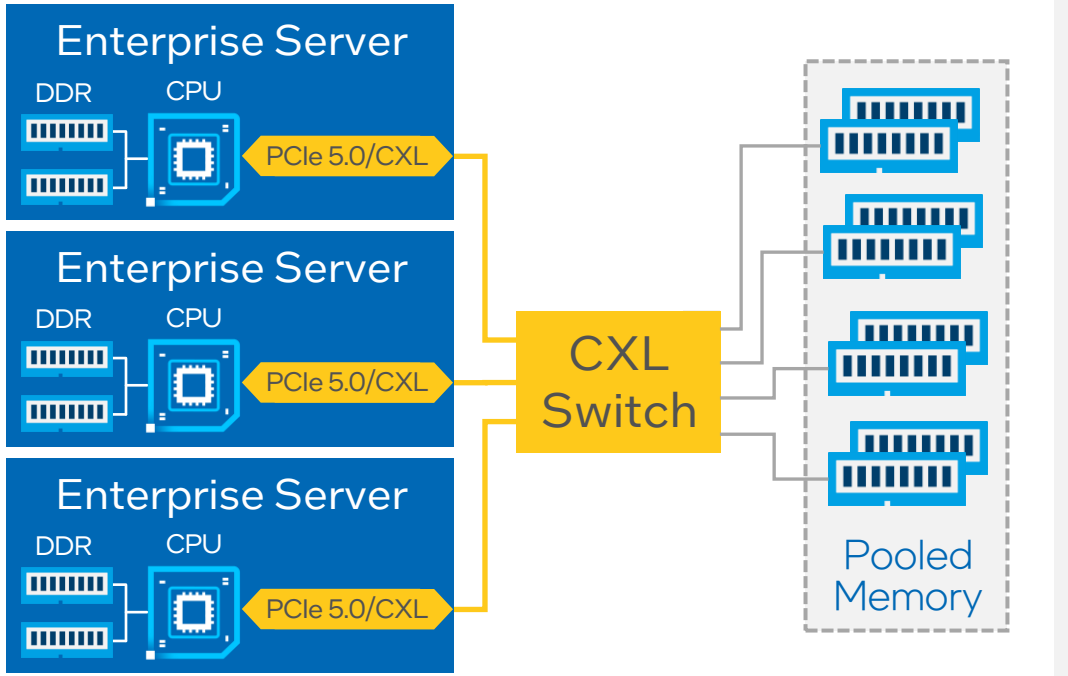


Challenge 1 : Heterogeneous compute addressed

Challenge 2 : Memory Scalability addressed

CXL™ 2.0: Resource Pooling at Rack Level, Persistent Memory Support and Enhanced Security

- Resource pooling/disaggregation
  - Managed hot-plug flows to move resources
  - Type-1/Type-2 device assigned to one host
  - Type-3 device (memory) pooling at rack level
  - Direct load-store, low-latency access – similar to memory attached in a neighboring CPU socket (vs. RDMA over network)
- Persistence flows for persistent memory
- Fabric Manager/API for managing resources
- Security: authentication, encryption
- Beyond node to rack-level connectivity!

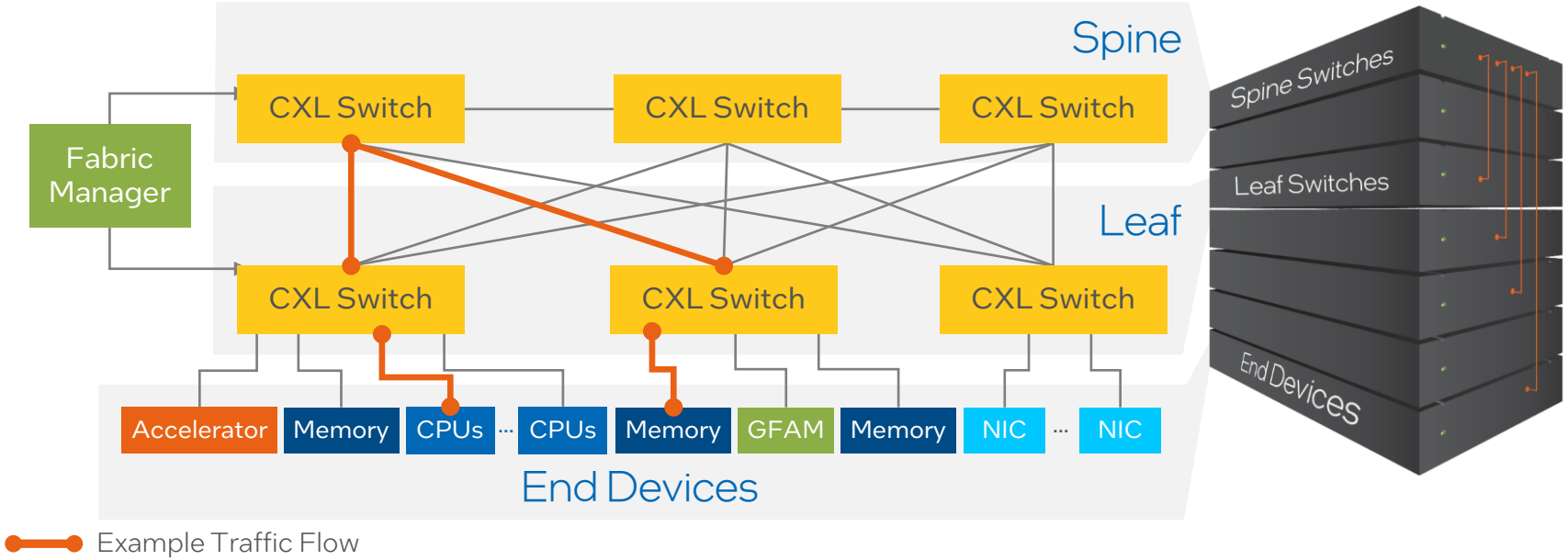


Challenge 3: Stranded resources in Data Center addressed by CXL 2.0

Disaggregated system with CXL optimizes resource utilization delivering lower TCO and power efficiency

CXL™ 3.0 Enhancements

- Bandwidth doubling with 64 GT/s at 0-latency add
- Protocol enhancements with direct peer-to-peer to HDM memory – Shared Memory
- Composable systems with spine/leaf architecture at rack/pod

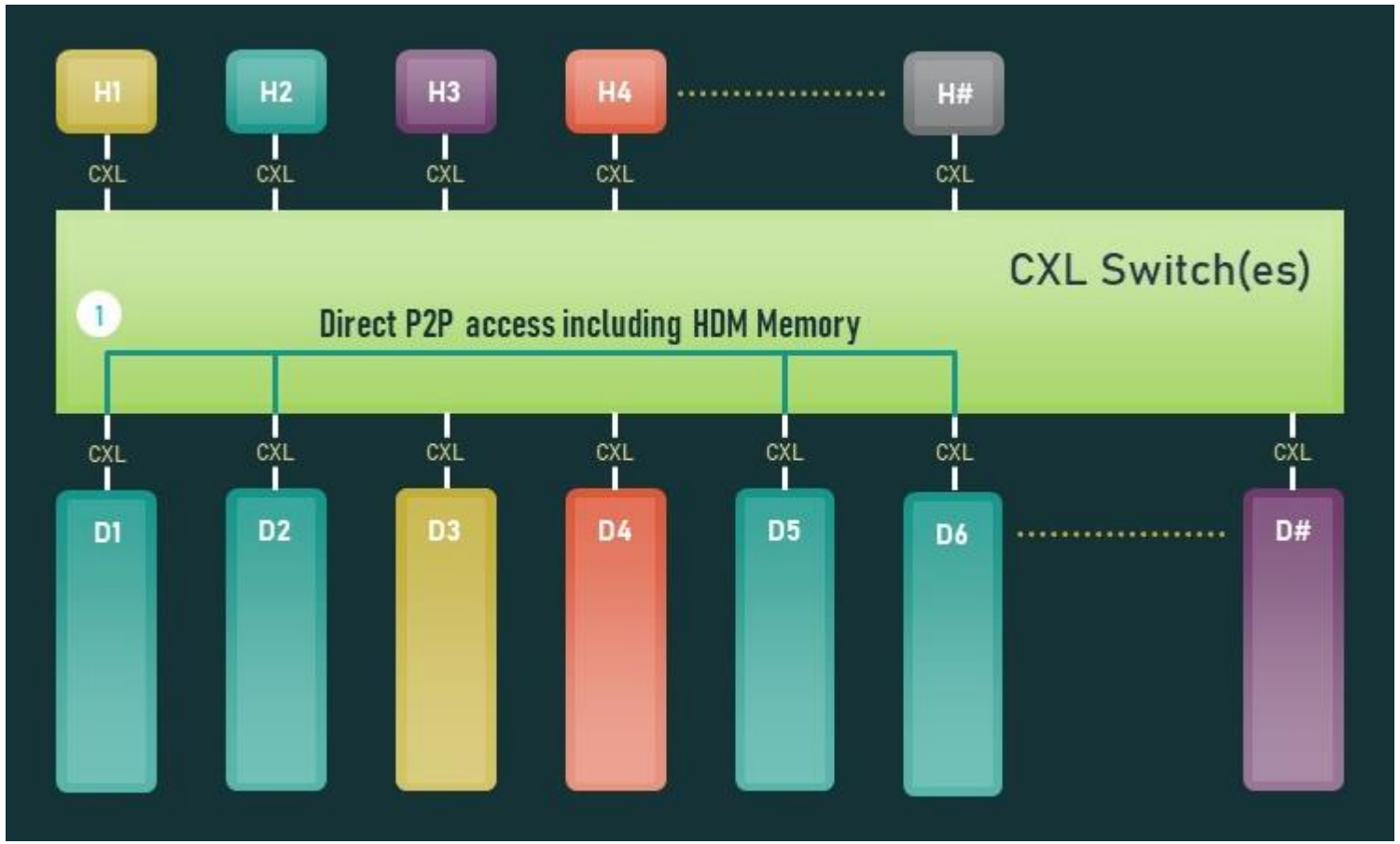


CXL 3.0 Fabric Architecture

- Interconnected spine switch system
- Leaf switch NIC enclosure
- Leaf switch CPU enclosure
- Leaf switch accelerator enclosure
- Leaf switch memory enclosure

Challenge 4: Fine-grained data sharing in large distributed system - addressed by CXL 3.0

CXL 3.0 Protocol Enhancements (UIO and BI) for Device to Device Connectivity



CXL 3.0 enables **non-tree topologies and peer-to-peer communication (P2P)** within a virtual hierarchy of devices

- Virtual hierarchies are associations of devices that maintains a coherency domain
- P2P to HDM-DB memory is I/O Coherent: a new Unordered I/O (UIO) Flow in CXL.io – the Type-2/3 device that hosts the memory will generate a new Back-Invalidation flow (CXL.Mem) to the host to ensure coherency if there is a coherency conflict

CXL: Health of the Ecosystem

Attribute	Status	Comments
Membership	255+	
Products	3 Compliance events since April 2023	4th: : May 4, 2024. 20 CXL1.1 devices in integrators list: 1 Type-1, 2 Type-2, 15 Type3, 2 Type1/2/3 Significant s/w dev. Linux Kernel 5.15 full support of T3 (Ubuntu 22.04.1 LTS/ Fedora Core 36 works) Multiple show-cases and demos in multiple conferences (SC, FMS, OCP, Memcon, etc.)
Heterogeneous Compute (Type1/2)	Deployed	UberNIC : low-latency (1/2) and high throughput (>2.5x) VM Migration
Memory (Type-3)	Deployed	Wide deployment. Both bandwidth and capacity expansion. Reduces loaded latency. Multiple media (DRAM and storage covered)
Pooling (CXL 2.0)	PoCs look promising	VM Elastic Memory demand: Pond showed 9% DRAM savings initially (still substantial; paper in ASPLOS 23) –likely to go up - direct attach. Estimates showing ~40K CXL switching costs are inaccurate both from cost as well as ignoring MHD. Data base elastic memory demand: SAP and Intel: works well for TPCC (negligible performance degradation even with switches). See paper.
Speeds	128G coming soon (2025)	128G PCIe PHY (demo'ed by several companies) and 112G Enet PHY on same process have almost identical b/w density. B/W density does improve on doubling rate but not 2X (e.g., 112G -> 224G is a 1.3x improvement). Need to consider platform/ channel reach also
Sharing/ Fabric	WIP (CXL3+)	CXL 3 silicon development in progress. S/W: Work actively continues on CXL 3.x (e.g., a patchset is to layer a filesystem on top of shared memory)

Agenda

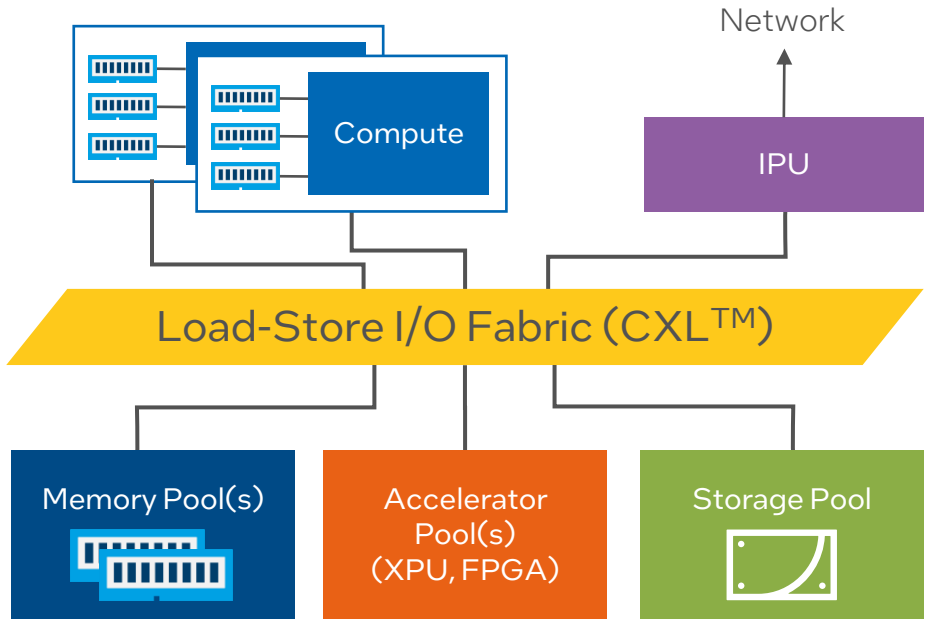
- Load-Store I/O in Compute Landscape
- Compute Express Link and its evolution
- Conclusions and Call to Action

Future Directions and Conclusions

- CXL™ enabling new usages beyond node
- Composable systems rack/pod level
  - Resource pooling/sharing
  - Multiple domains, shared memory, message passing, atomics, peer-to-peer
  - Fabric Manager
- High-bandwidth, low-latency CXL fabric
  - Low-latency switches
  - Cables with retimers/co-packaged optics
  - Iso power-performance as direct connect
- Challenges: blast radius, containment, QoS, memory reliability, fail-over, software

Multi-Domain Capable Load-Store I/O

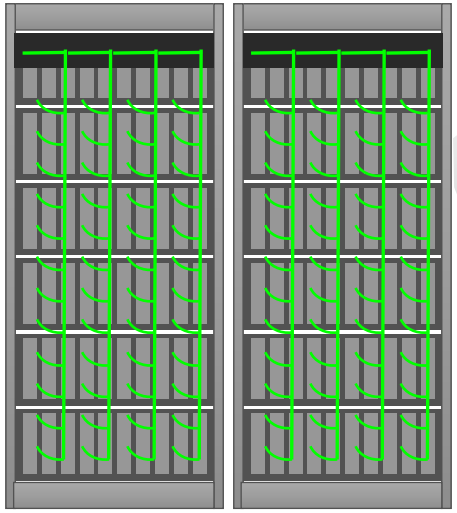
Vision: Load-Store I/O (CXL) as the fabric across the pod providing low-latency and high-bandwidth resource pooling/sharing as well as message passing



Future Directions and Conclusions: Rack/Pod-Level Resource Pooling/Sharing with CXL™ and UCIe

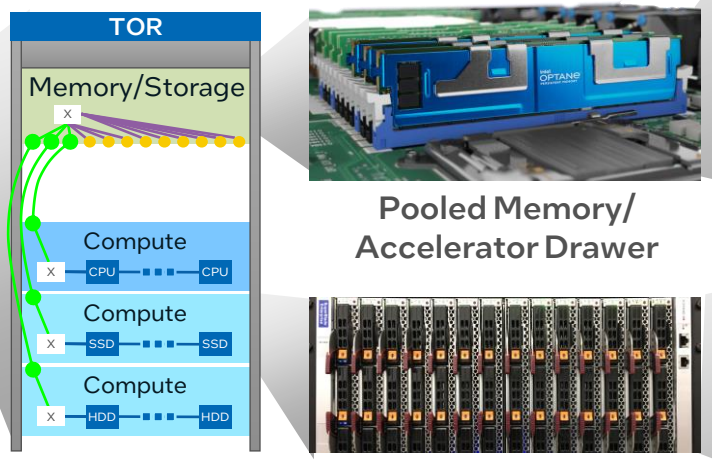
Pod of Racks

Physical connectivity using UCIe-Retimer-based co-packaged optics carrying CXL protocol



CXL Rack/Pod Level Connected Using Long-Reach Media

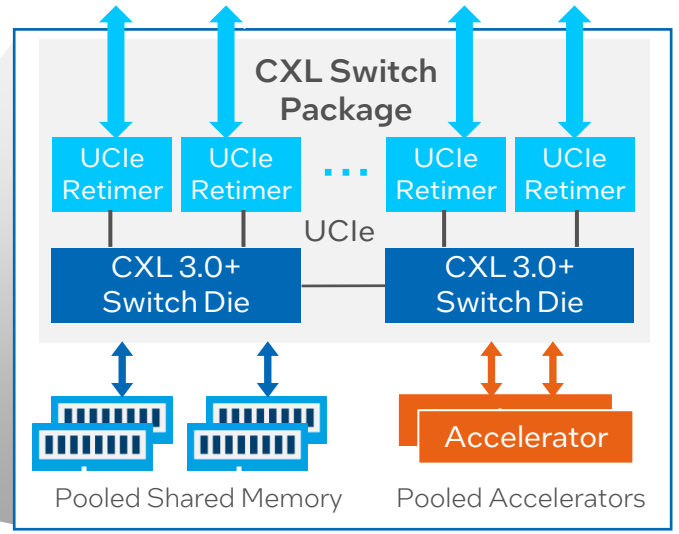
Electrical/optical/… through UCIe-Retimer (co-packaged optics)

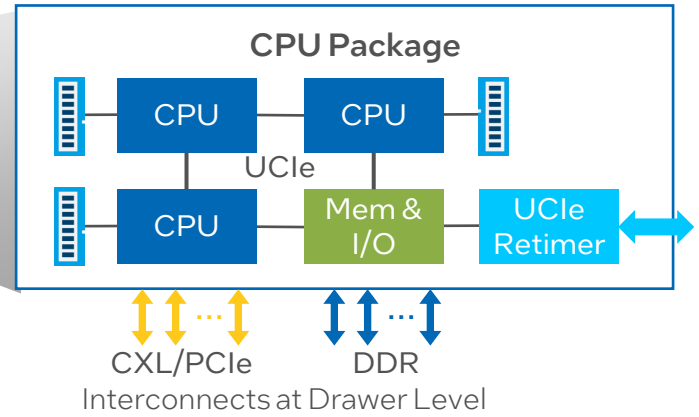


Pooled Memory/ Accelerator Drawer

Compute Drawer

Optical Connection for Intra-Rack and Pod





Thank You!