Opening Keynote

2021 OFA Virtual Workshop

EVOLUTION OF INTERCONNECTS AND FABRICS TO SUPPORT FUTURE COMPUTE INFRASTRUCTURE

Dr. Debendra Das Sharma

Intel Fellow and Director I/O Technologies and Standards

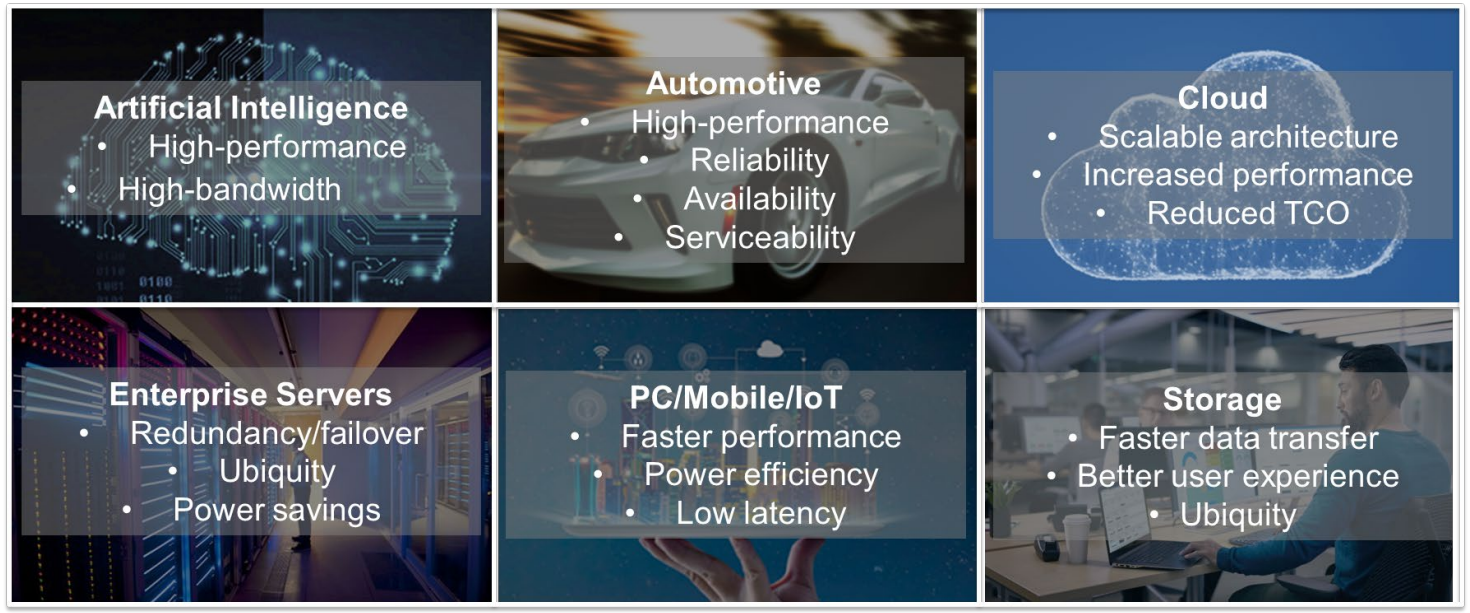
Intel Corporation

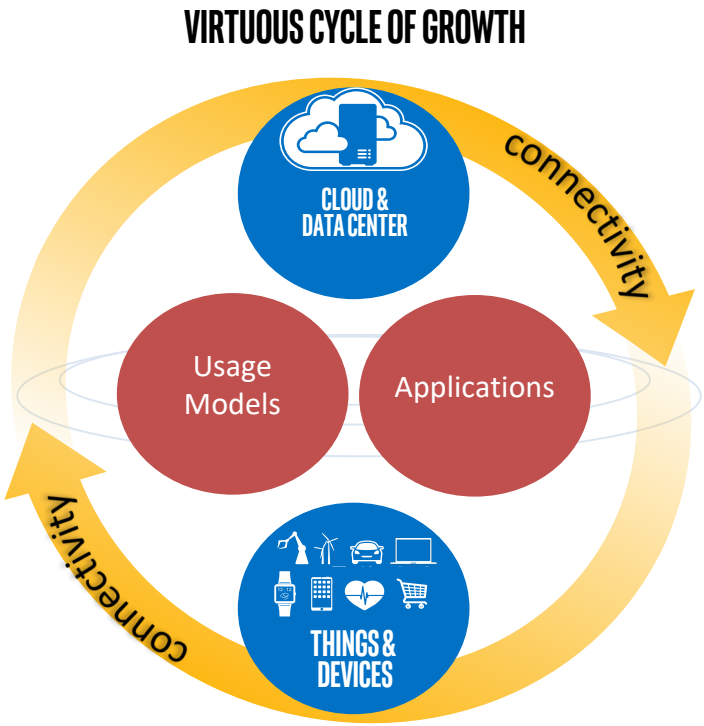
AGENDA

- Mega-Trends in compute landscape
- Interconnects and Fabrics - an important pillar of compute
- Evolution of Interconnects and Fabrics
- Future Directions

© OpenFabrics Alliance

MEGA-TRENDS IN THE COMPUTE LANDSCAPE

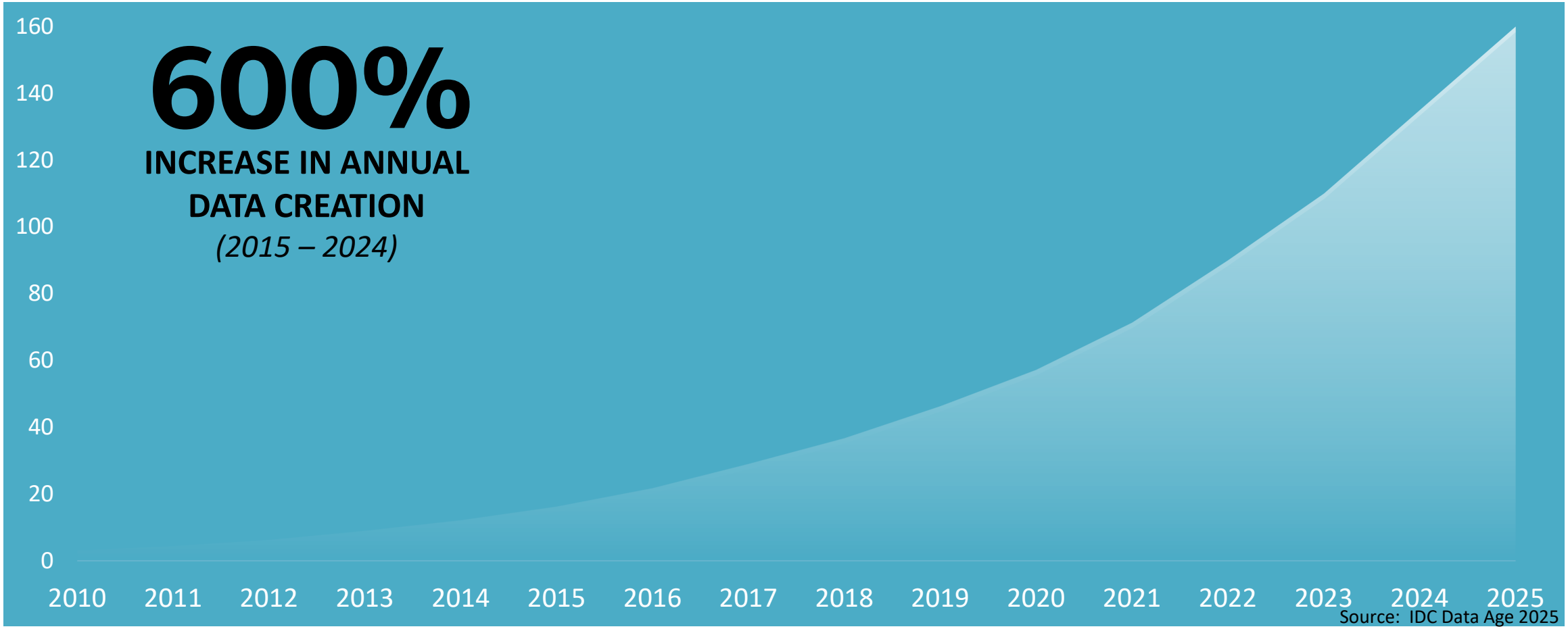




- **Insatiable demand for compute, storage, and data movement**
- **Innovative applications leading to more demand which in turn leading to more innovations**
- **Interconnect is an important pillar of compute**
  - Compute, storage/ memory, interconnect, software, process technology, security

© OpenFabrics Alliance

EXPLOSION OF DATA ENABLING DATA-CENTRIC REVOLUTION

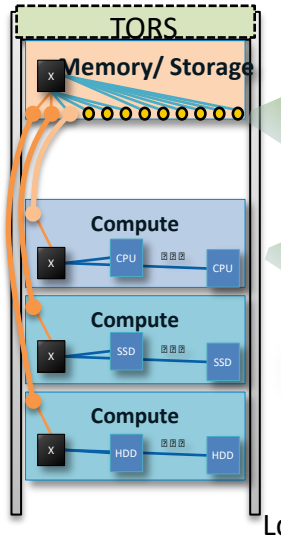


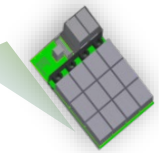
Explosion of data leading to rapid innovations. Move faster, Store more, Process everything seamlessly, efficiently, and securely

© OpenFabrics Alliance

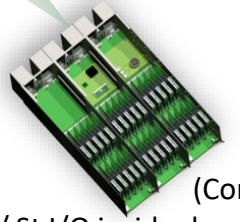
DELIVERING PERFORMANCE IN DATA CENTER

(Rack of Servers)



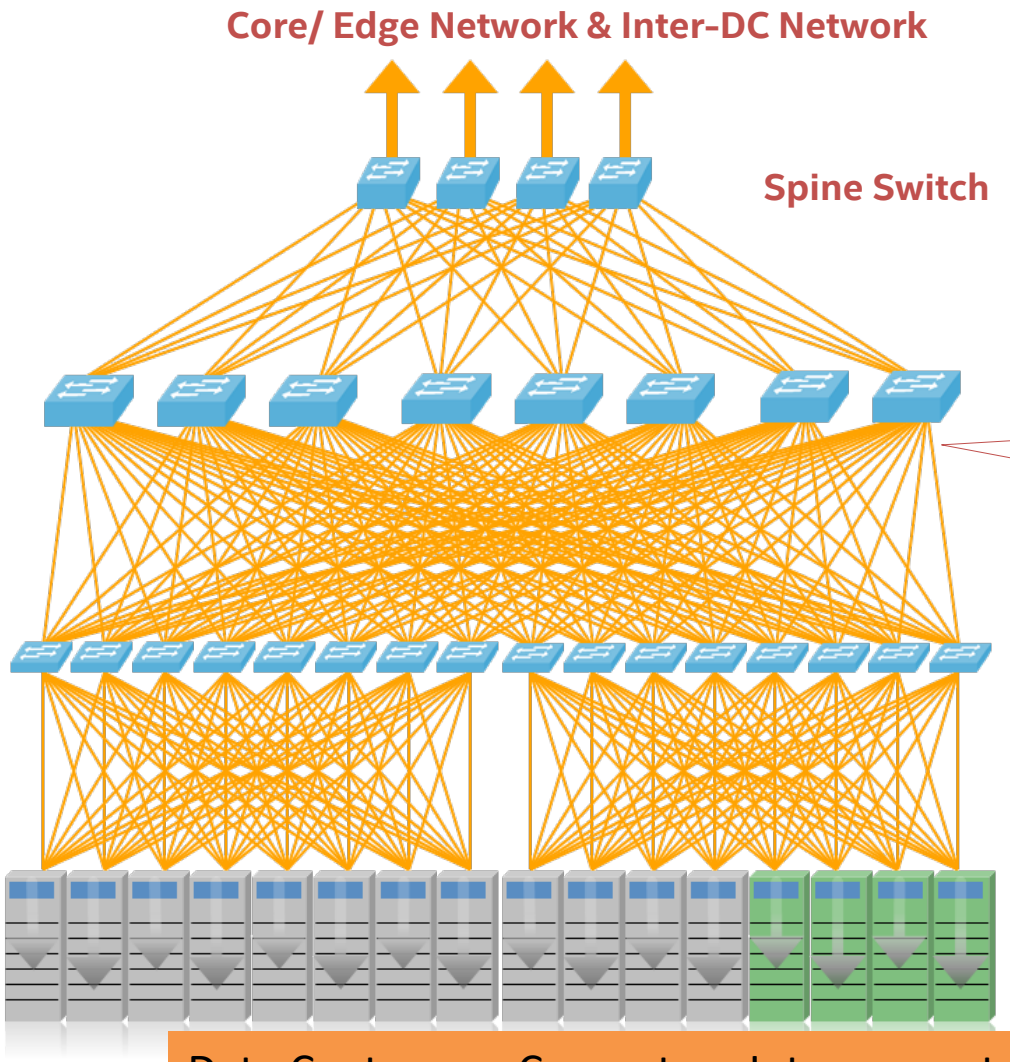


(Memory Drawer)



(Compute Drawer)

Ld/ St I/O inside drawer and Rack (Cu + optical)
Fabric at Rack level as well as at DC level

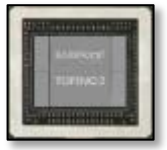


Hyperscale Data Center and Edge

Optical Modules

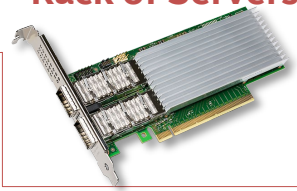
High-bandwidth connectivity at 100G and beyond

Leaf Switch



P4-programmable scale-out fabric with uncompromising performance

Rack of Servers

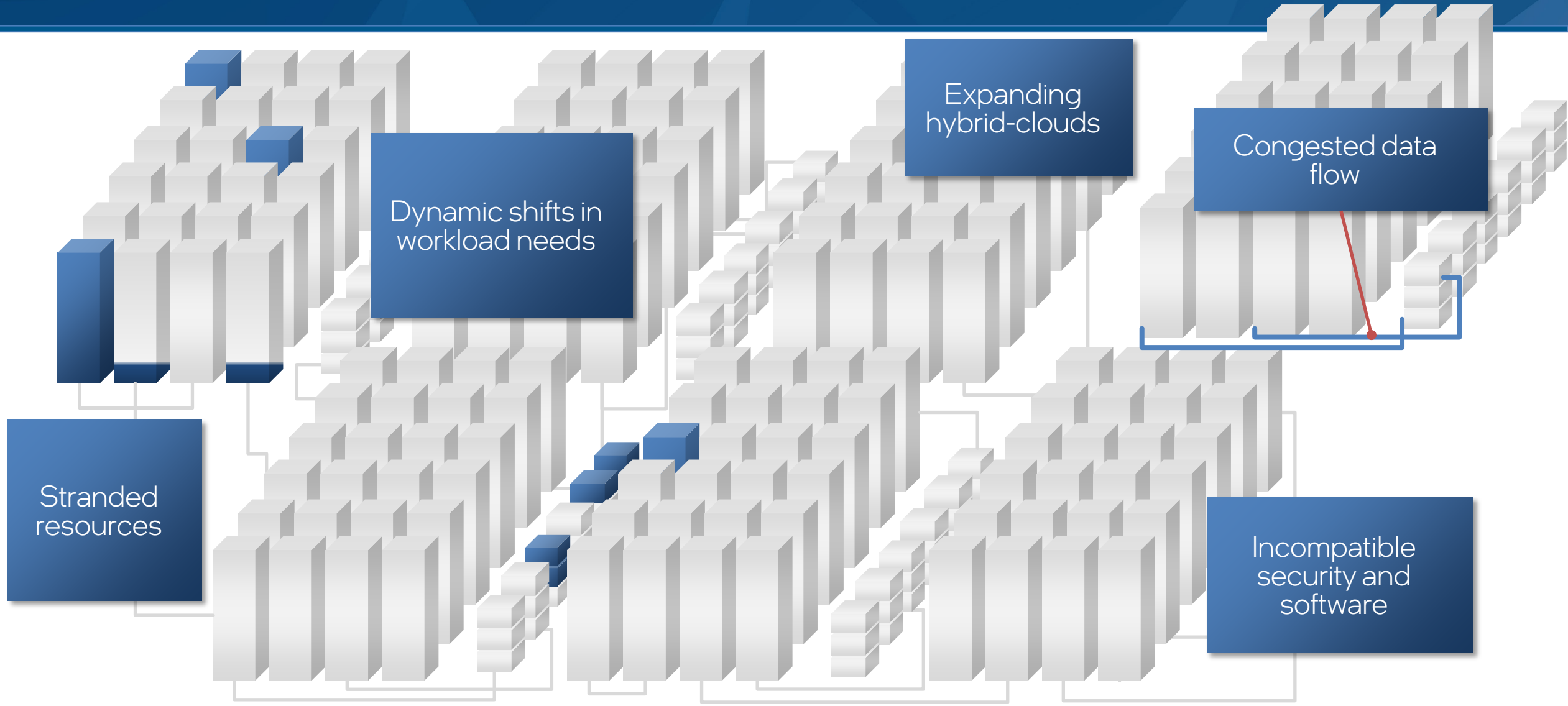


Programmable infrastructure acceleration for demanding data movement with Smart NIC

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

© OpenFabrics Alliance

CHALLENGES AT SCALE



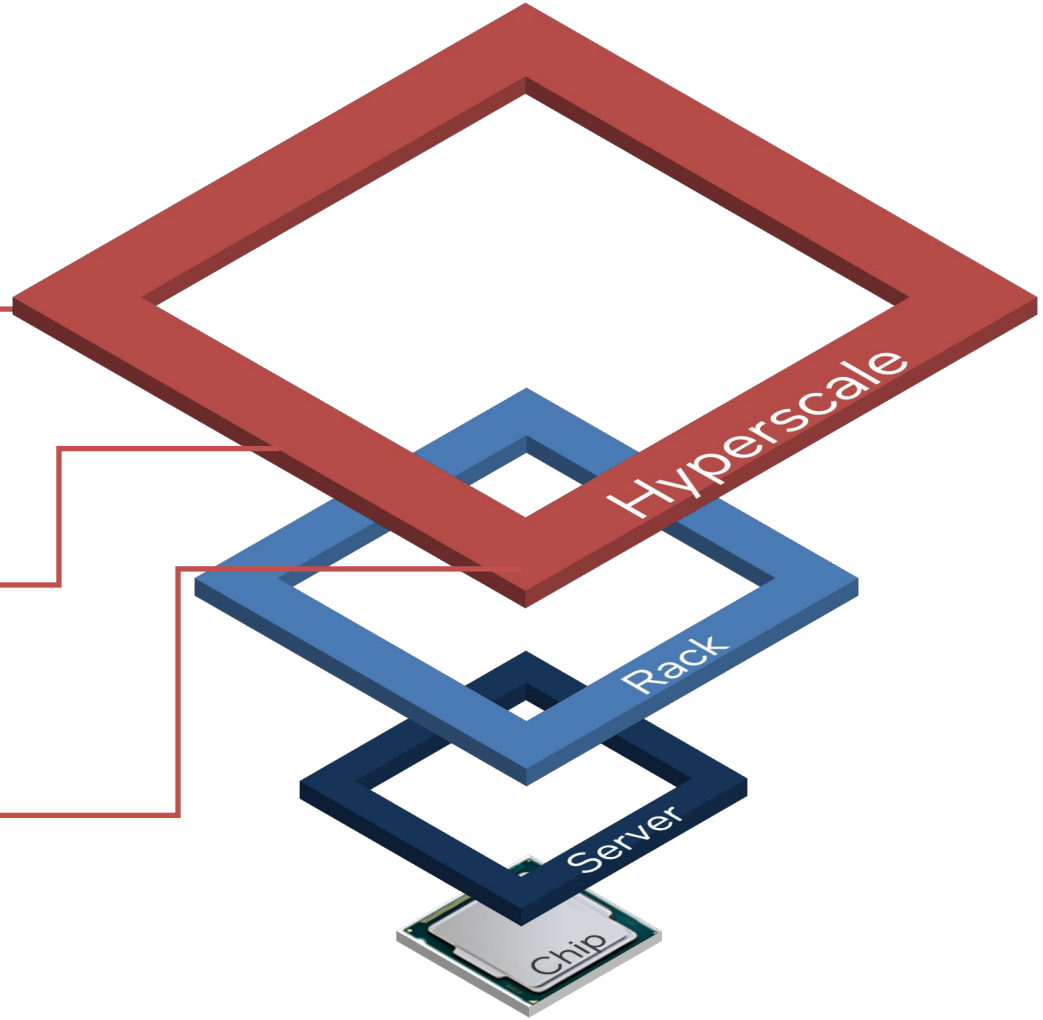
© OpenFabrics Alliance

THE VISION

Seamless
Edge to cloud experience

Predictable and secure
services anywhere

Optimized TCO
Hardware and Software



© OpenFabrics Alliance

TAXONOMY, CHARACTERISTICS, AND TRENDS OF INTERCONNECTS

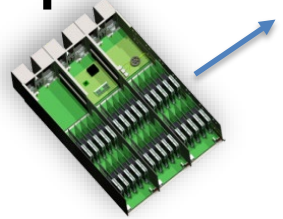
Category	Type and Scale	Current Data Rate/ Trend	PHY Latency (Tx + Rx)	Other Characteristics
Latency Tolerant	Networking Data Center Scale	56/ 112 GT/s-> 224 GT/s (PAM4)	100+ ns w/ FEC (20ns+ w/o FEC)	• Narrow Lane count (4 or 8) • Backplane usage w/ cables & retimers
Latency Sensitive	Load-Store I/O (PCIe/ CXL / SMP cache coherency) Node level (moving to sub-Rack level)	32 GT/s (NRZ) -> PCIe Gen6 64 GT/s (PAM4)	<10ns (Tx+ Rx: PHY-PIPE) 0-1ns FEC overhead	• 200-300 Lanes per CPU socket • Low-cost and HVM • Socket & mother board (12” reach) • Backwards compatibility (PCIe/ CXL) – single standard for all usages • Area/ power sensitive • Reliability (FIT << 1; Failure in Time – number of failures in 10^9 hours) • SMP coherency and memory access uber-latency sensitive

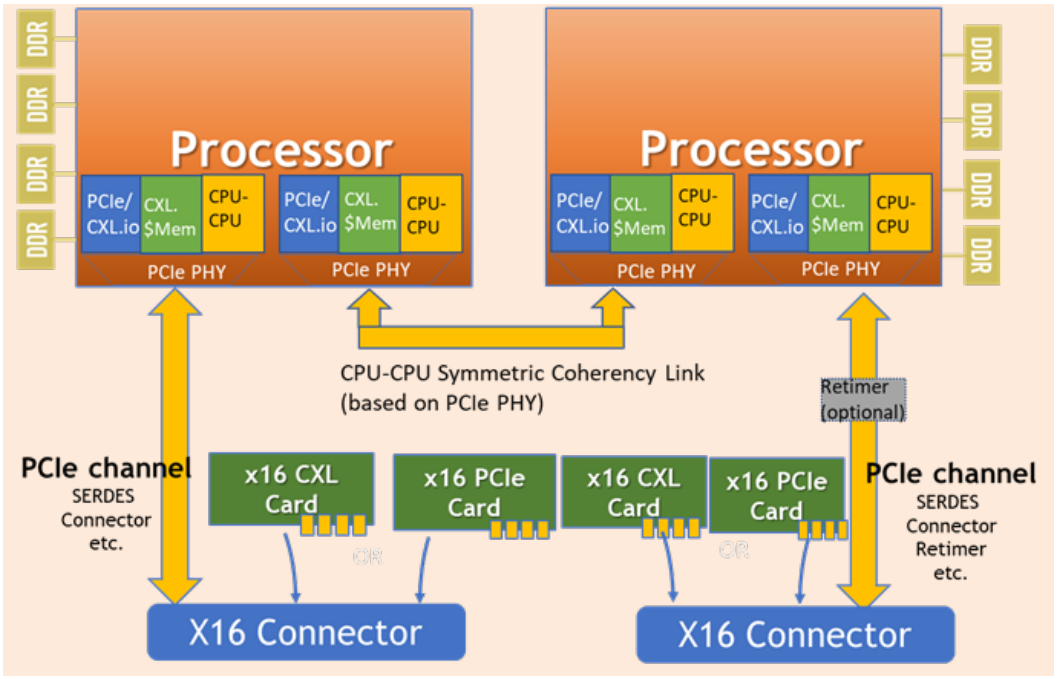
Latency Sensitive I/O moving to PAM-4: innovation needed to meet latency, area, and cost challenges for viability

© OpenFabrics Alliance

LOAD-STORE INTERCONNECT CHARACTERISTICS

- **Ability to directly access memory (CPU, I/O)**
- **Tightly coupled – small scale – Fabric through PCIe**
- **Memory mapped into system memory space**
  - Coherent or Non-coherent access
  - Accesses across PCIe non-coherent
  - Accesses across CXL can be either
- **Some form of ordering or cache coherency**
  - PCIe: Producer-Consumer Ordering Semantics
- **Transactions are <u>guaranteed</u> to be delivered and completed in a reasonable time**
  - No dropped packets, no software based retry
  - Hardware based link level replay on error
- **Timeout and Error reporting hierarchy**
  - Hardware based error containment guarantees





Device A	Device B
Write Data	Read Flag
Write Flag	Read Data

(Producer Consumer Ordering Model: Reading updated Flag guarantees reading updated Data)

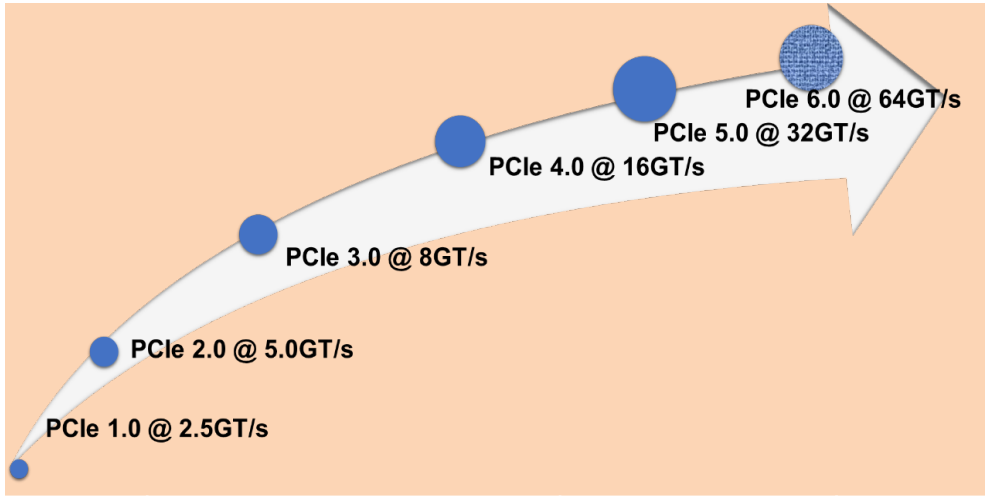
Load-store I/O evolving to meet memory innovations and expanding to Rack level for resource pooling

© OpenFabrics Alliance

EVOLUTION OF PCI-EXPRESS

- **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, AI**
- **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!
Need to keep KPIs in-tact!



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*

© OpenFabrics Alliance

KEY METRICS FOR LD/ST 64.0 GT/S W/ PAM-4: REQUIREMENTS

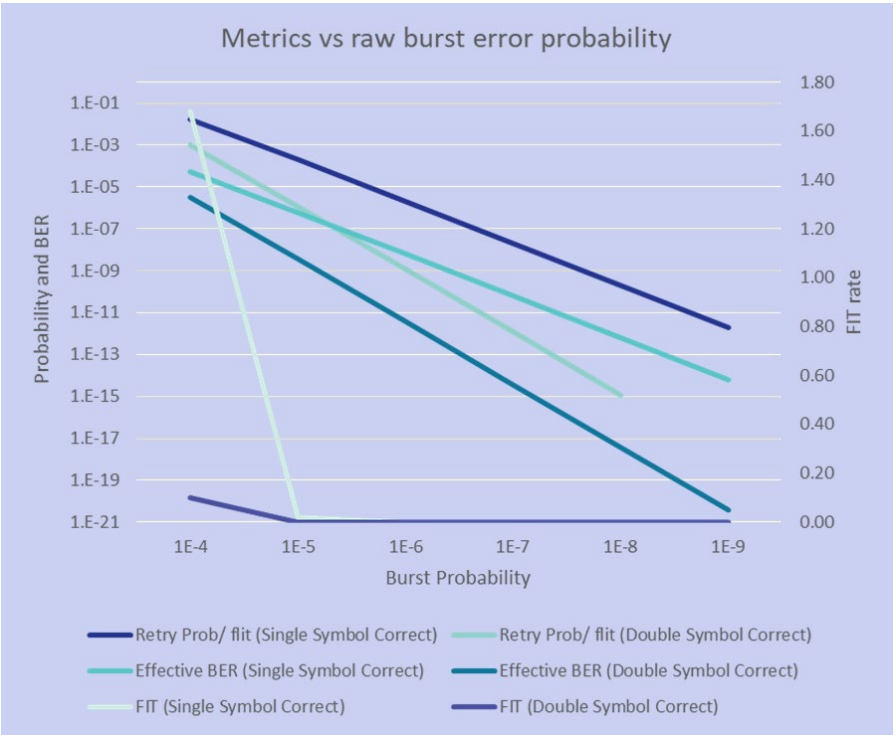
Metrics	Requirements
Data Rate	64 GT/s, PAM4 (double the bandwidth per pin every generation)
Latency	<10ns adder for Transmitter + Receiver (including Forward Error Correct, FEC) for PCIe CXL / Memory/ SMP coherency interconnects need less than 1ns adder (Ld/St can not afford the 100ns FEC latency as networking does – okay at DC-level scale)
Bandwidth Inefficiency	<2 % adder over PCIe 5.0 across all payload sizes
Reliability	0 < FIT << 1 for a x16 (FIT – Failure in Time, number of failures in 10^9 hours)
Channel Reach	Similar to PCIe 5.0 under similar set up for Retimer(s) (maximum 2)
Power Efficiency	Better than PCIe 5.0
Low Power	Similar entry/ exit latency for L1 low-power state Addition of a new power state (L0p) to support scalable power consumption with bandwidth usage without interrupting traffic
Plug and Play	Fully backwards compatible with PCIe 1.x through PCIe 5.0
Others	HVM-ready, cost-effective, scalable to hundreds of Lanes in a platform

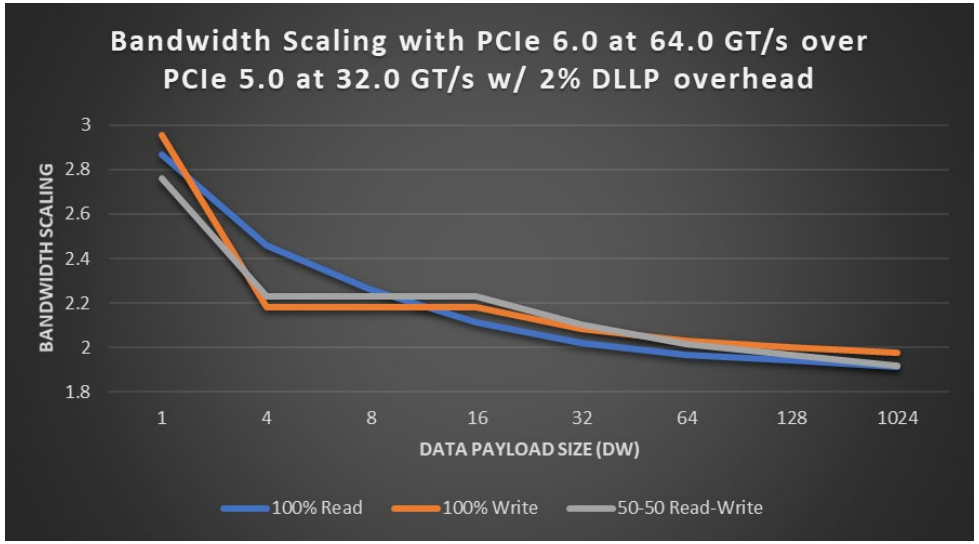
64.0 GT/s PAM-4 is a major inflection point for Load-store I/O - PCIe 6.0 is on track to meet each of these metrics!

© OpenFabrics Alliance

PCIE 6.0 - A LOW-LATENCY APPROACH

- Light-weight FEC + Low-latency Link level replay
- A combination of 10^{-6} FBER (First Burst Error Rate) with a 3-way interleaved single symbol correct FEC keeps retry rate low
- Spec defined mechanisms for low-latency replay
- Strong CRC (Cyclic Redundancy Check) for low FIT
- Flit (Flow-Control Unit) mode results in better link efficiency than before!





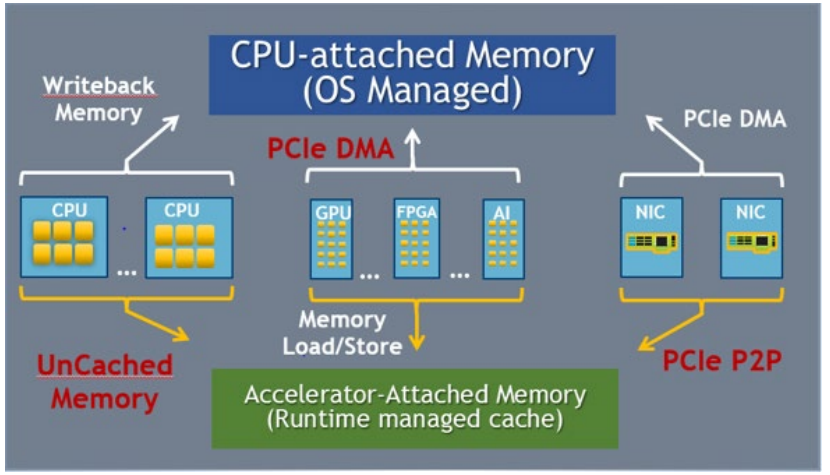
FBER/ Retry Time	10^{-6}/ 100ns	10^{-6}/ 200ns	10^{-6}/ 300ns	10^{-5}/200ns
Retry probability per flit	$5x10^{-6}$	$5x10^{-6}$	$5x10^{-6}$	0.048
B/W loss with go-back-n (%)	0.025	0.05	0.075	4.8
FIT	4×10^{-7}	4×10^{-7}	4×10^{-7}	4×10^{-4}

Low-latency, low-power, backward-compatible doubling of bandwidth journey continues!

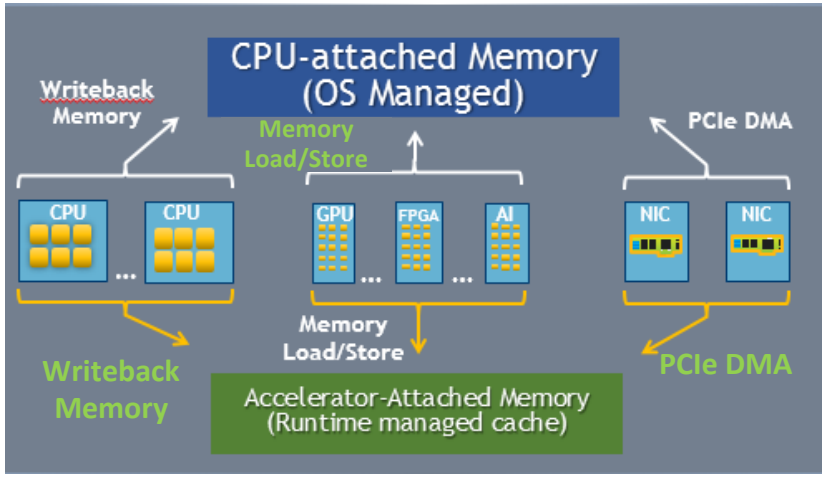
© OpenFabrics Alliance

CXL: A NEW CLASS OF 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
  - In-memory processing



With PCIe-only



CXL Enabled Environment

© OpenFabrics Alliance

CXL APPROACH

Coherent Interface

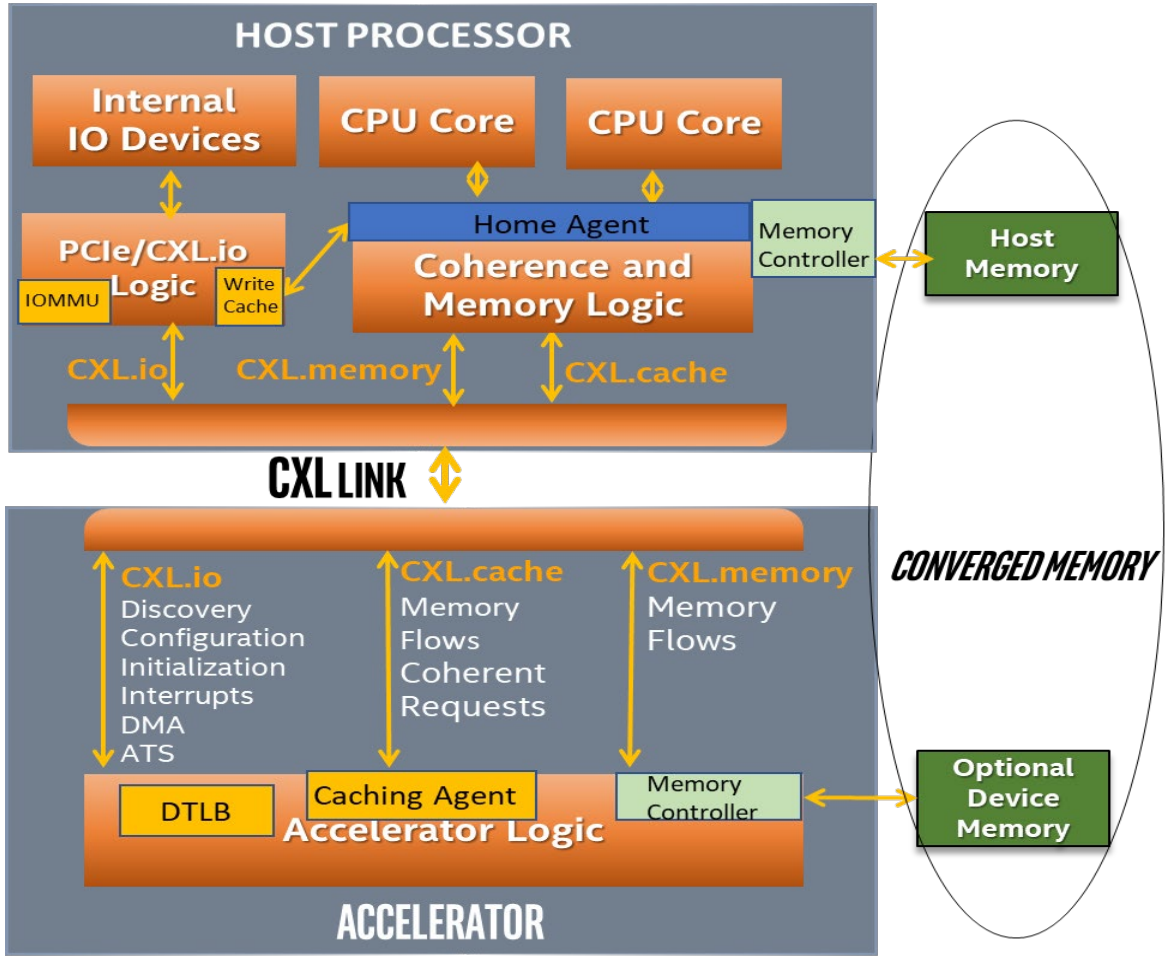
Leverages PCIe with 3 mix-and-match protocols

Low Latency

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

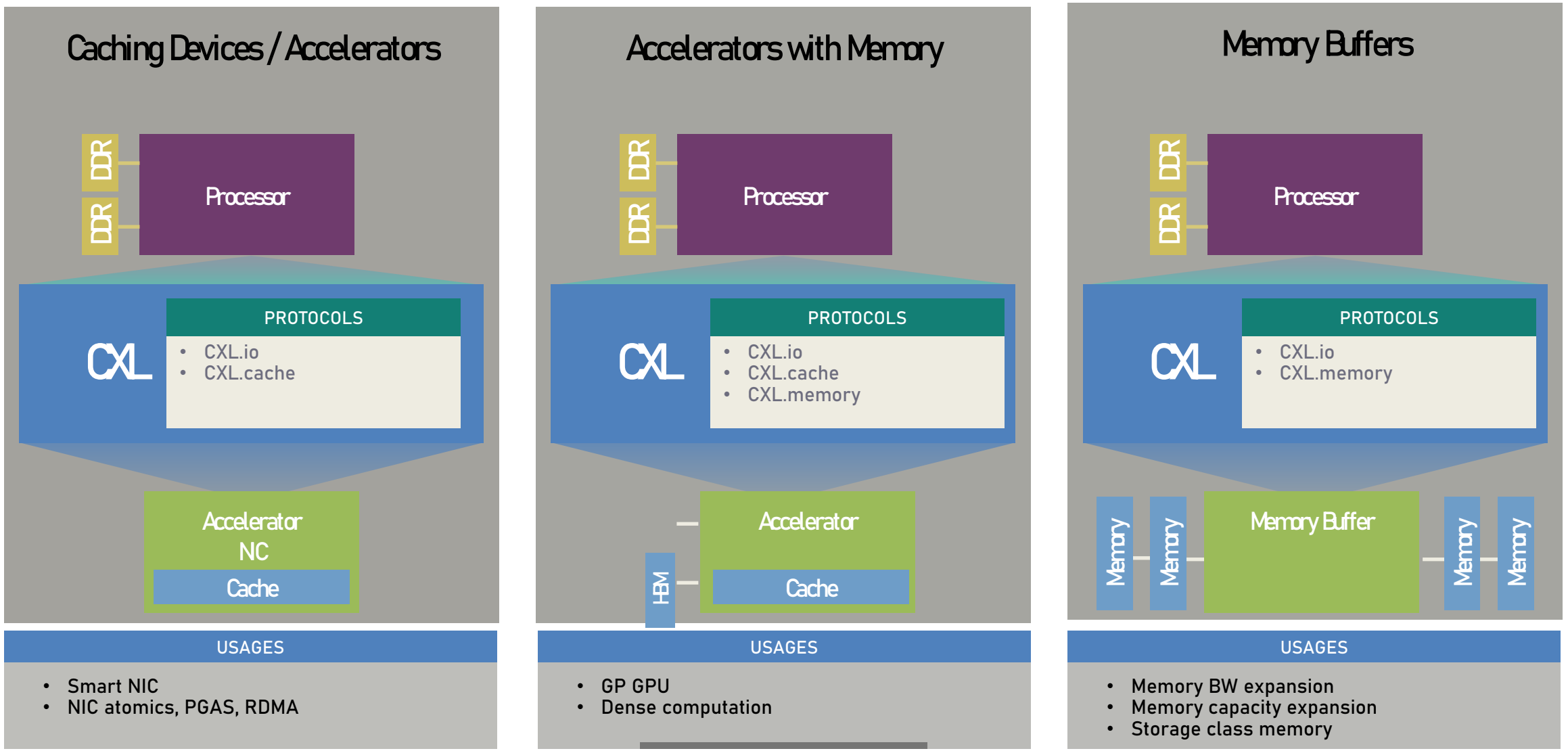
Asymmetric Complexity

Eases burdens of cache coherent interface designs



© OpenFabrics Alliance

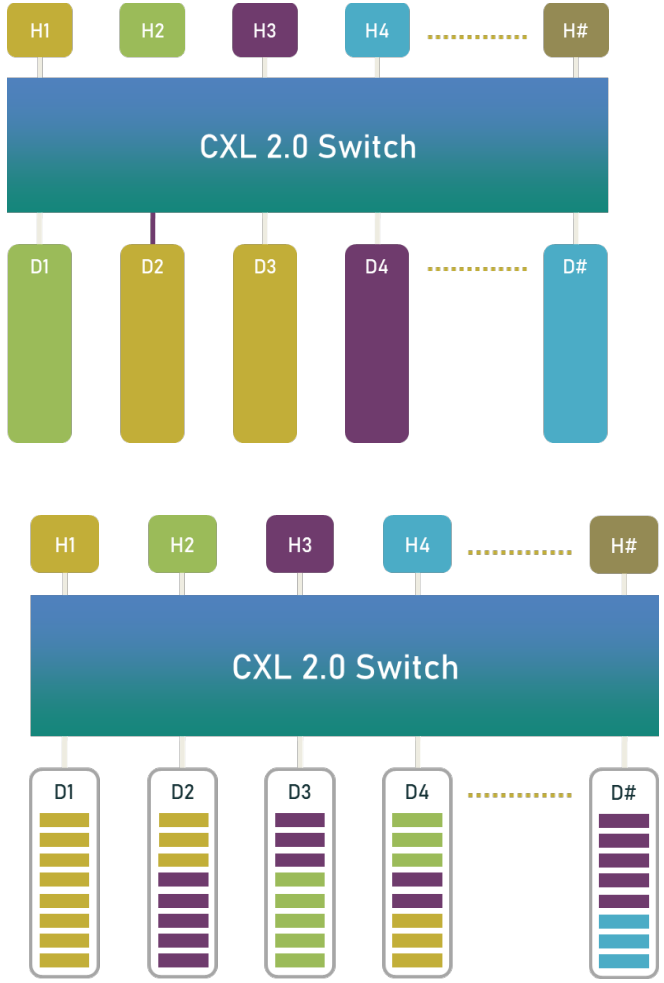
CXL 1.1 USAGE MODELS



© OpenFabrics Alliance

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 supporting persistent memory**
  - Covers larger latency memory in addition to DRAM
- **Type-1 and Type-2 device (accelerator) assigned to one host**
- **Type-3 device (memory) can be pooled across multiple hosts**
- **Fabric Manager for managing resources**
- **Software API for devices**
- **Enhanced security: DTLB, device authentication, link encryption**
  - Working with DMTF, PCI-SIG for synergies
  - Spans devices and switch

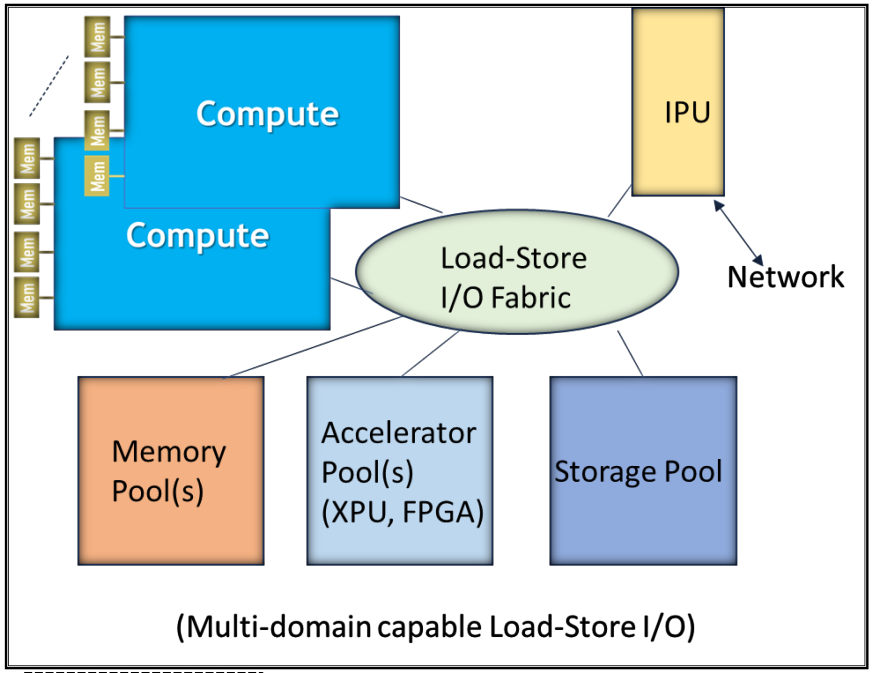


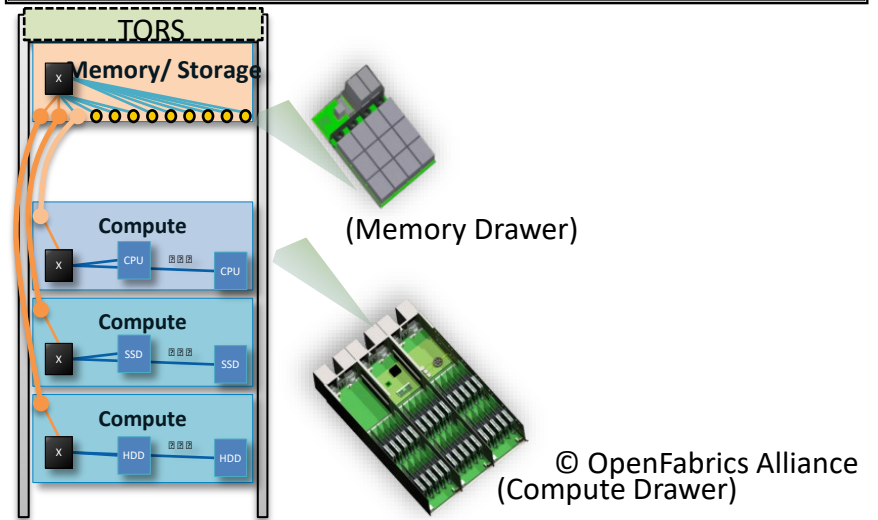
Dis-aggregated System with CXL optimizes resource utilization delivering lower TCO and power-efficiency

© OpenFabrics Alliance

FUTURE DIRECTIONS

- **Composable Disaggregated Infrastructure at Rack level**
  - Heterogenous compute/ memory, storage, networking fabric resources
  - connected through high bandwidth, low-latency Load-Store Interconnect
  - delivering almost-identical performance per watt as independent servers
  - w/ multiple domains w/ shared memory, message passing, atomics
- **Synergy between Networking and Load-store**
  - Expect boundaries to be fungible
  - Fabric Manager, Multi-head, multi-domain, Atomics support, Persistence flows, Smart NIC with optimized flows to access system memory without involving host, VM migration
- **Challenges:**
  - Latency: NUMA optimization, low-latency switch
  - Bandwidth demand: higher rate helps
  - Power Efficiency
  - Blast Radius – containment and QoS
  - Scaling: Moore's law and Dennard-scaling
  - Copper-Optical transition point
  - Software!





© OpenFabrics Alliance

2021 OFA Virtual Workshop

THANK YOU

Dr. Debendra Das Sharma

Intel Fellow and Director I/O Technologies and Standards

Intel Corporation