

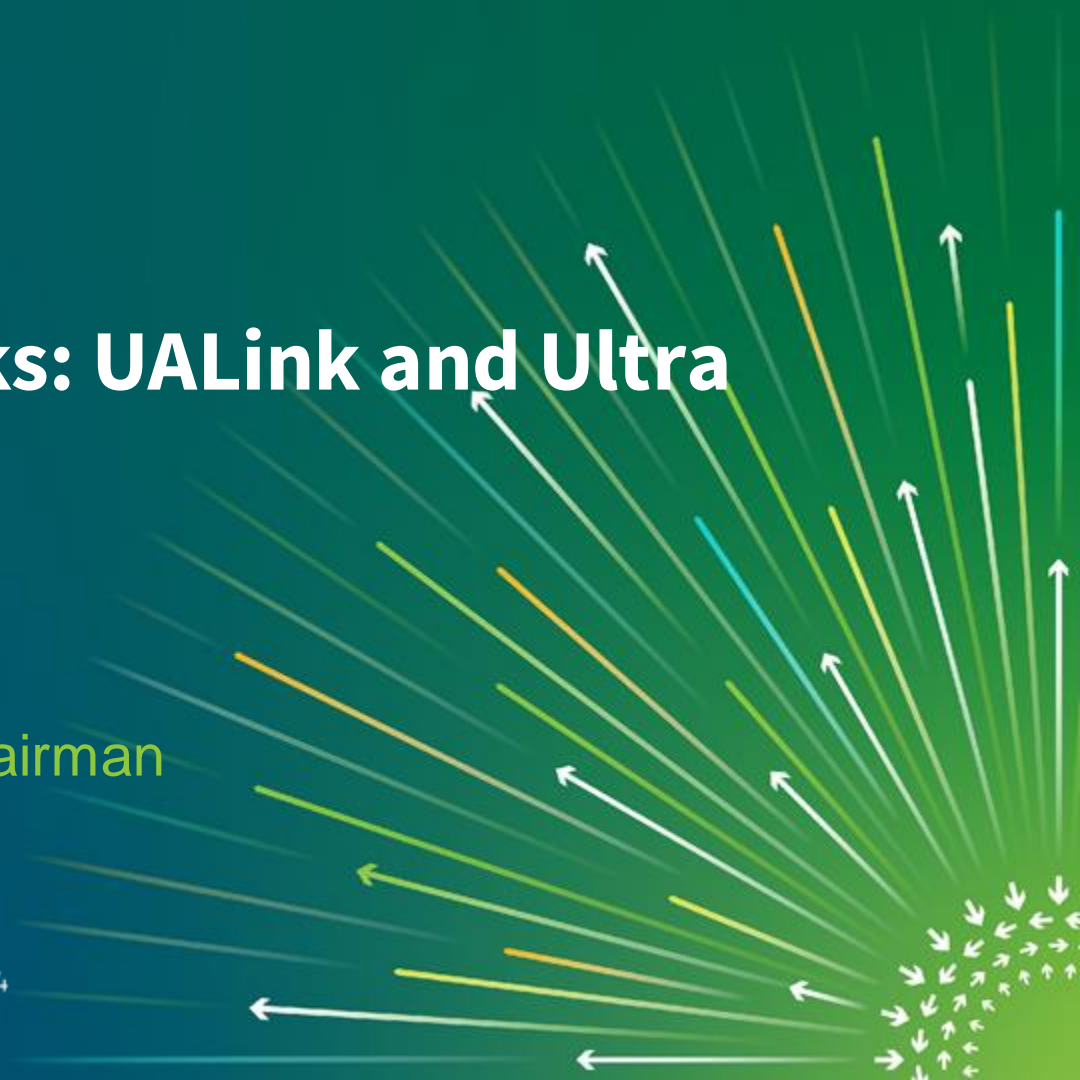
Future of AI Networks: UALink and Ultra Ethernet

J Metz – UEC Chairman

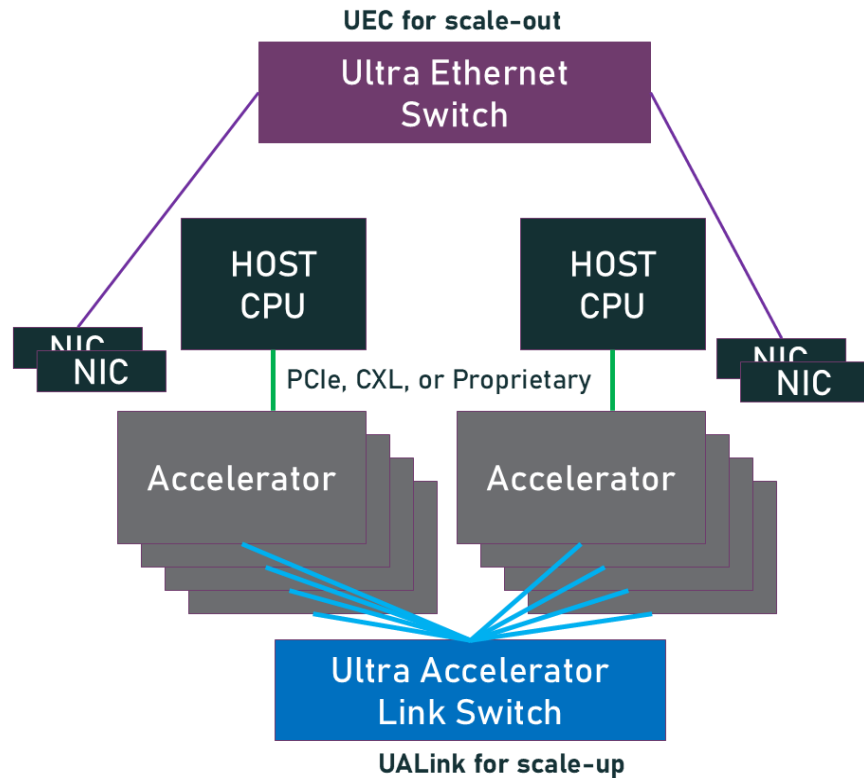
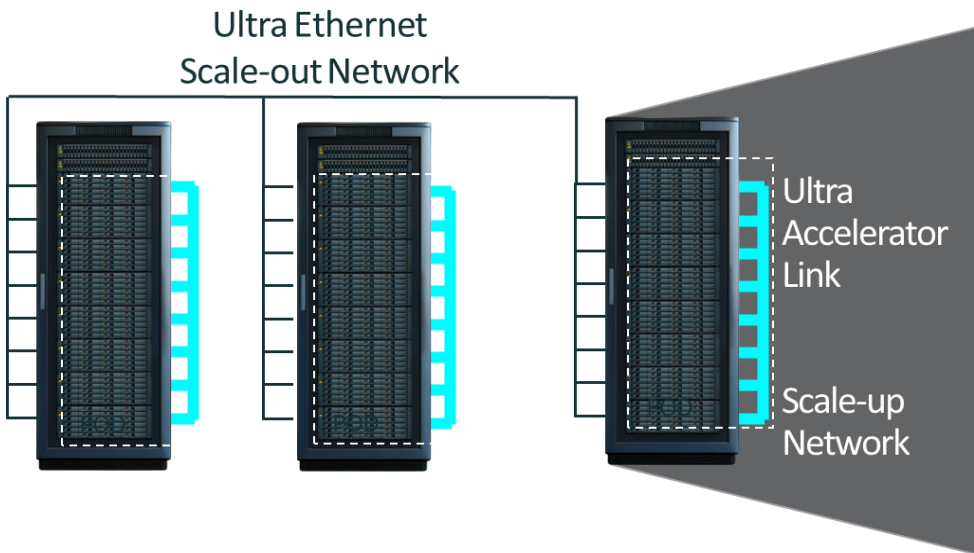
Kurtis Bowman – UALink Chairman



OCT 15-17, 2024
SAN JOSE, CA



The Future is ULTRA



AI is driving unique challenges for Networking

Large scale AI requires scaling up to hundreds/thousands of GPUs

Large scale AI requires scaling out to tens/hundreds of thousands of GPUs

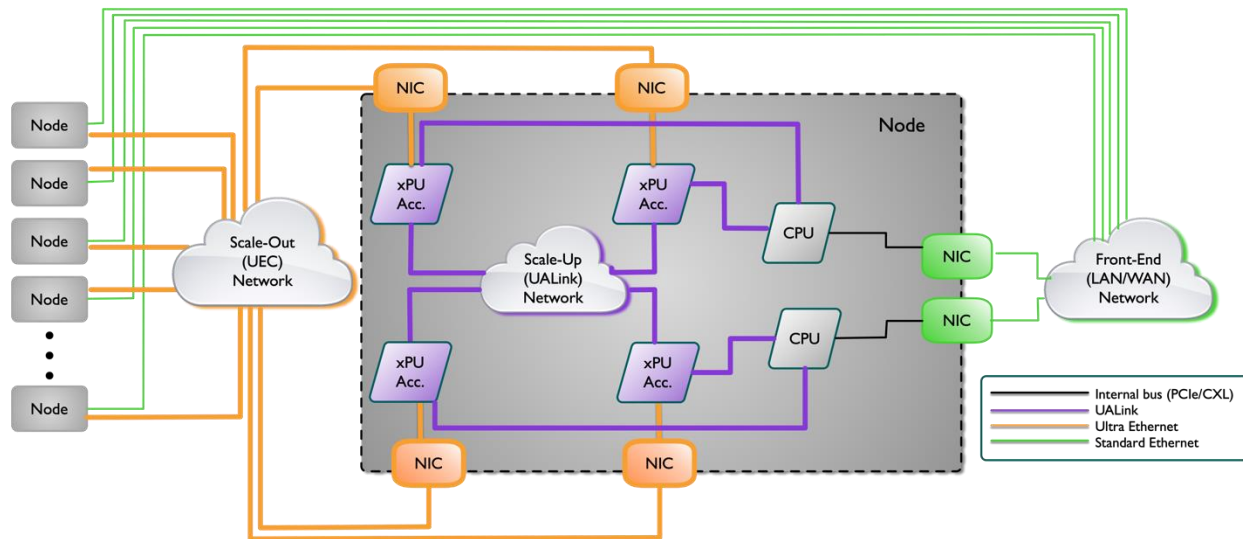


Ultra Ethernet

Pieces of the AI Puzzle

- AI environments have multiple networks
 - As systems scale, dedicated-purpose fabrics are required
 - Main, general-purpose network
 - Dedicated, **Scale-Up** network
 - Dedicated, **Scale-Out** Network
- How are they different?

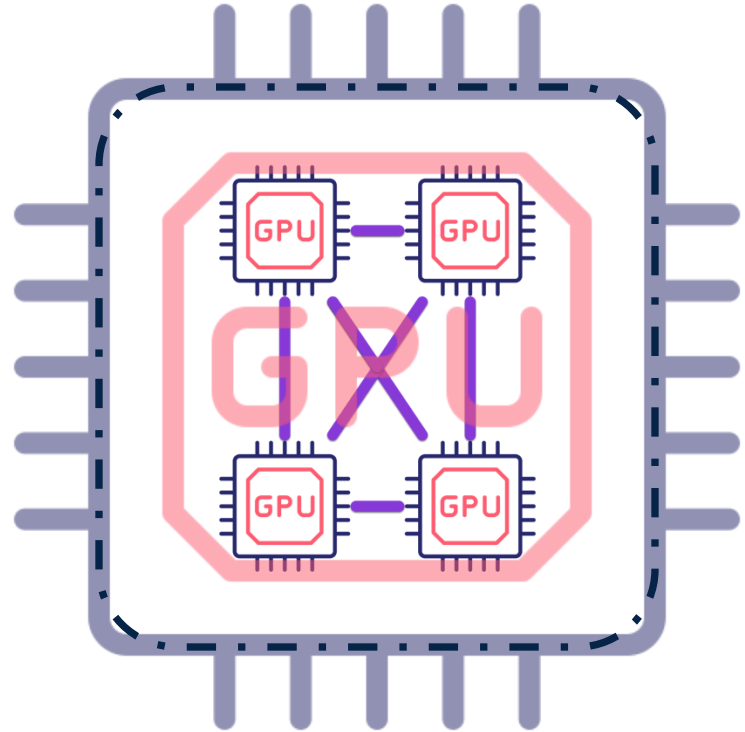
General Purpose vs. Scale-Up (UALink) versus Scale-Out (UEC) Networks



Scale-Up Network

Or, how to make one giant GPU

- As model and parameter sizes increase, it is more difficult to fit inside of a single GPU memory
- Memory needs to be shared across GPUs, but it needs to act as a single GPU
 - Load/Store operations
- **Scale-Up** refers to the ability to make several GPUs act like a one giant GPU to complete the task





Ultra Accelerator Link

Partner group of innovators for scale up AI infrastructure



Google



intel.

∞ Meta



High Performance

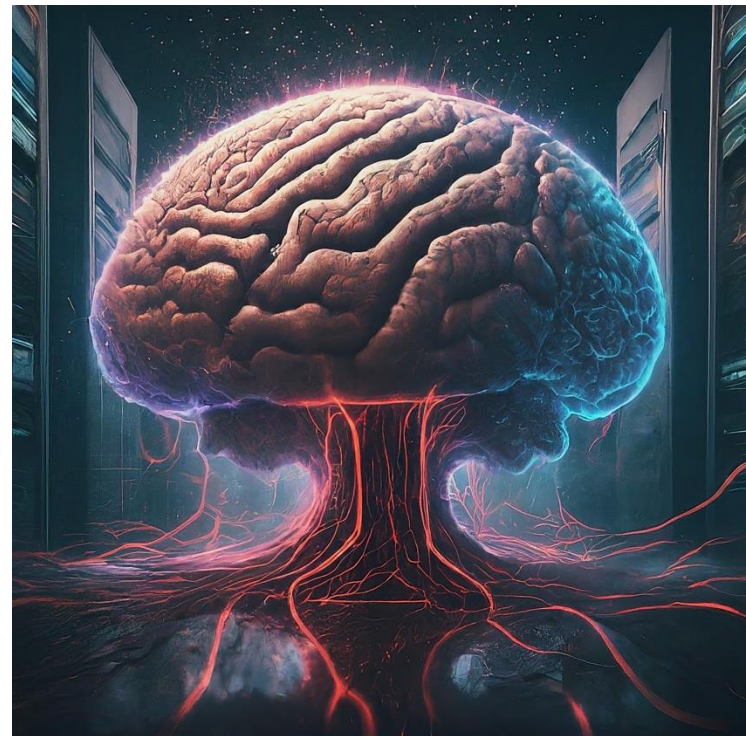
Open

Scalable

Ultra Accelerator Link (UALink)



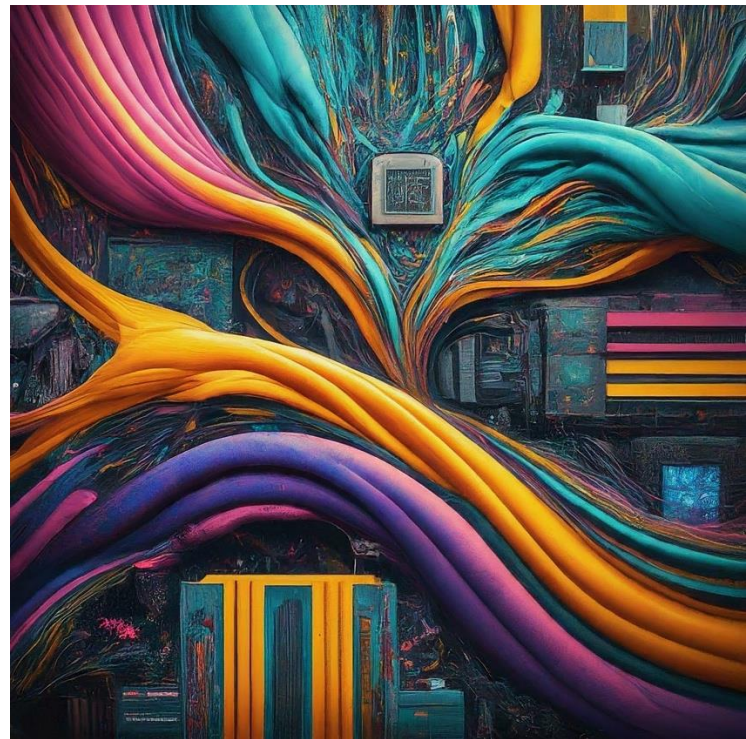
- UALink creates an open ecosystem for scale-up connections of many AI accelerators
 - Effectively communicate between accelerators using an open industry standard protocol
 - Easily expand the number of accelerators in a pod up to 1K
 - Optimize the performance needed for compute intensive workloads now and in the future
- An open scale up memory semantic fabric has significant advantages
 - Bandwidth, Latency, Power, and Efficiency
- The Consortium plans to open for members soon
 - The focus of the organization is to release the 1.0 specification by the end of the year



Ultra Accelerator Link Overview



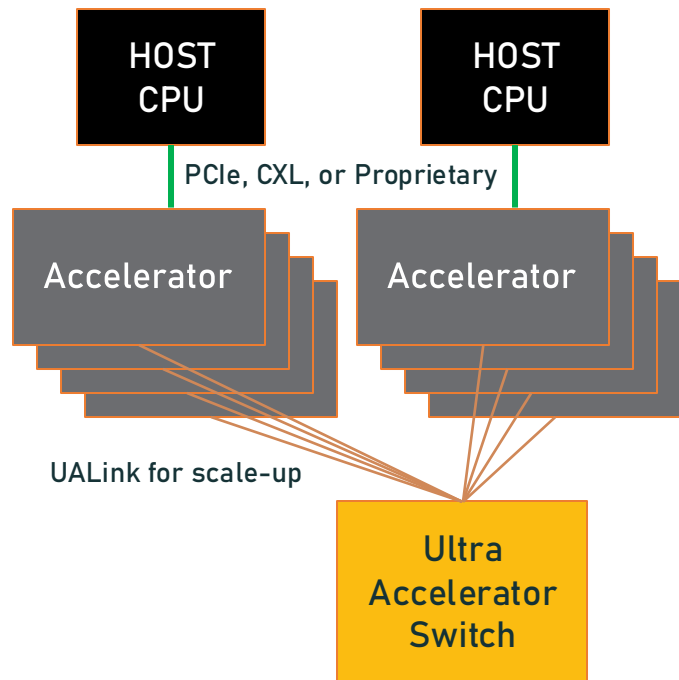
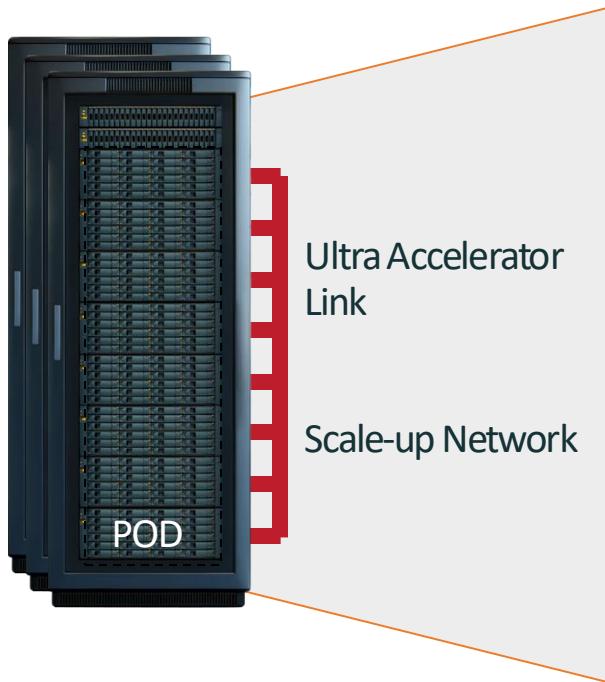
- The UALink interconnect is for scale-up Accelerator-to-Accelerator communication
 - The initial focus will be sharing DDR & HBM memory among accelerators
- Direct load, store, and atomic operations between accelerators (i.e. GPUs)
 - Low latency, high bandwidth fabric for 100's of accelerators in a pod
 - Simple load/store semantics with software coherency
- Supports data rates up to state-of-the-art 200Gbps per lane
- The UALink spec taps into the experience of the Promoters developing and deploying a broad range of accelerators and leverages the proven Infinity Fabric™ protocol
- Complementary with scale-out approaches such as Ultra Ethernet Consortium (UEC)



UALink Creates the Scale-up Pod



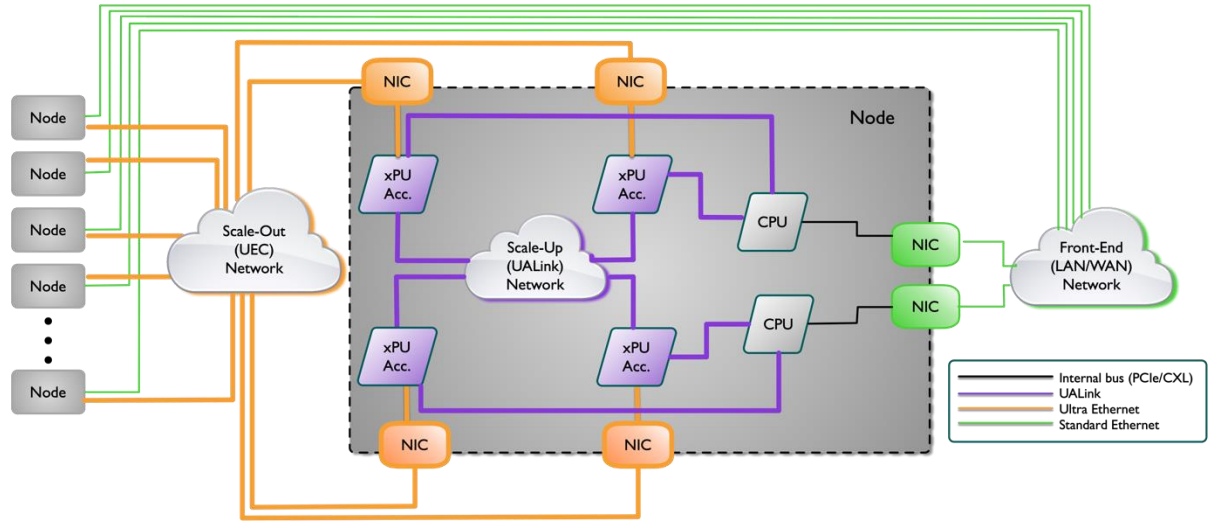
- **Low latency stack**
 - Protocol, Transaction, Link, & Physical
- **Lower power**
 - The simplified UALink stack leads to lower power than Ethernet switching for the same bandwidth
- **Lower latency switch**
 - Latency <100ns pin-to-pin
- **Lower die area**
 - Optimizing the Data Layer and Transaction Layer saves significant die area and reduces cost and TCO



Revisiting the AI Puzzle

- What about the Scale-Out network?

General Purpose vs. Scale-Up (UALink) versus Scale-Out (UEC) Networks

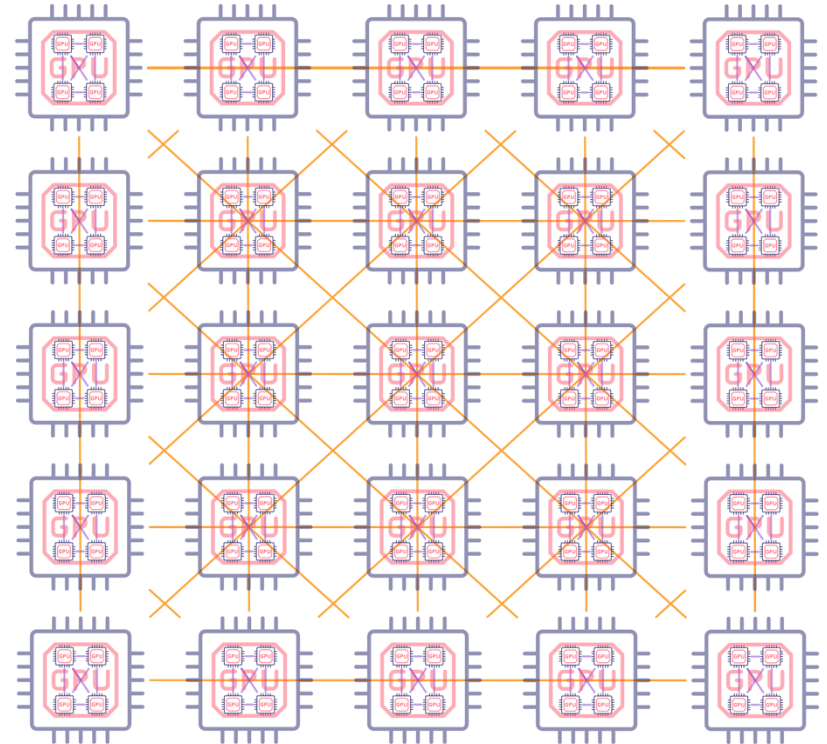


Scale-Out Network

Or, how to connect several giant GPUs

- GPU-GPU communication is critical and requires special consideration at large scale
- Typically DMA and packetized I/O
- Immediate focus of Ultra Ethernet

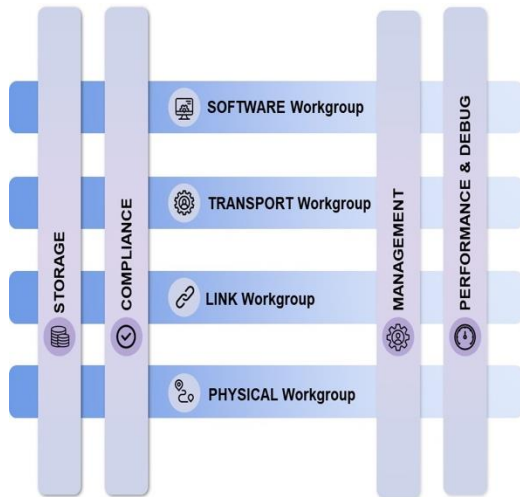
Ultra *Ethernet*



*topology not to scale

2024 Organization and Member

- Full Standards Development Organization
- (One of the?) Fastest growing projects in Linux Foundation
- 100+ Companies
- 1300+ individual active contributor volunteers
- 8 Workgroups
 - Physical
 - Link Layer
 - Transport
 - Software
 - Storage
 - Management
 - Compliance & Test
 - Performance & Debug



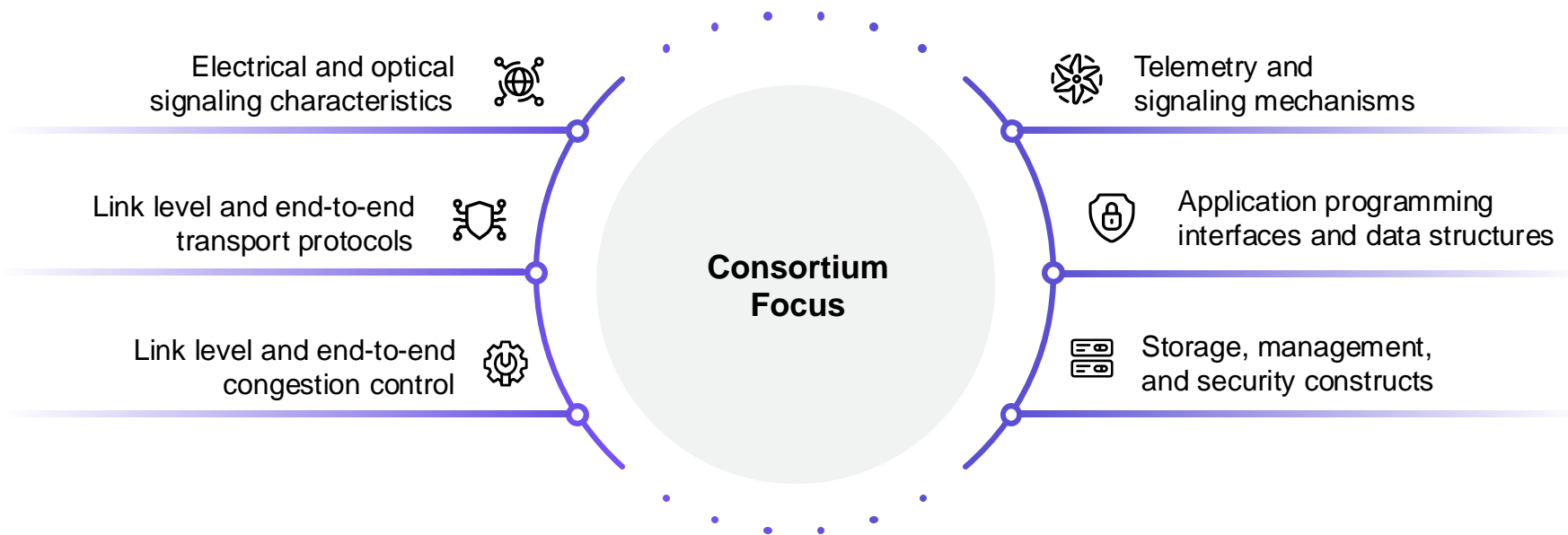
Ultra Ethernet Consortium 2024









*Please note that not all members are displayed on this page.

UEC Technical Goals

Open specifications, APIs, source code for optimal performance of AI and HPC workloads at scale.



UEC Addresses AI Network Needs

	Traditional RDMA-Based Networking	<i>UltraEthernet</i> Consortium
	Required In-Order Delivery, Go-Back- <i>N</i> recovery	Out-of-Order packet delivery with In-Order Message Completion
	Security external to specification	Built-in high-scale, modern security
	Flow-level multi-pathing	Packet Spraying (packet-level multipathing)
	DC-QCN, Timely, DCTCP, Swift	Sender- and Receiver-based Congestion Control
	Rigid networking architecture for network tuning	Semantic-level configuration of workload tuning
	Scale to low tens of thousands of simultaneous endpoints	Targeting scale of 1M simultaneous endpoints



MEMORY FABRIC
FORUM



Brought to you by
MemVerge

