REGIONAL SDC 25

BY Developers FOR Developers

Ultra Ethernet Consortium

Regional SDC Denver
April 30, 2025

UEC Overview

Craig W. Carlson

SNIA

Introducing: The Promise of Ultra Ethernet

https://*ultra*ethernet.org/

THE NEW ERA NEEDS A NEW NETWORK

Ultra Ethernet

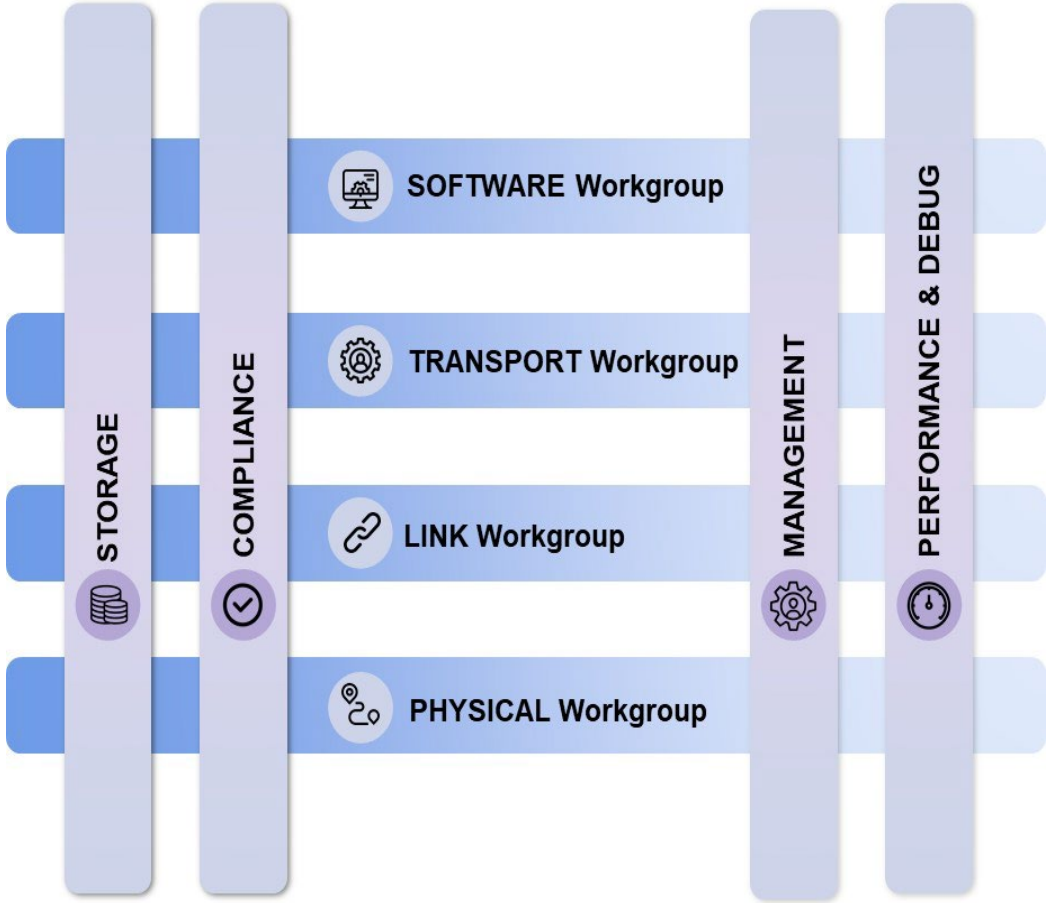
As **performant** as a supercomputing interconnect

As **ubiquitous** and **cost-effective** as Ethernet

As **scalable** as a cloud data center

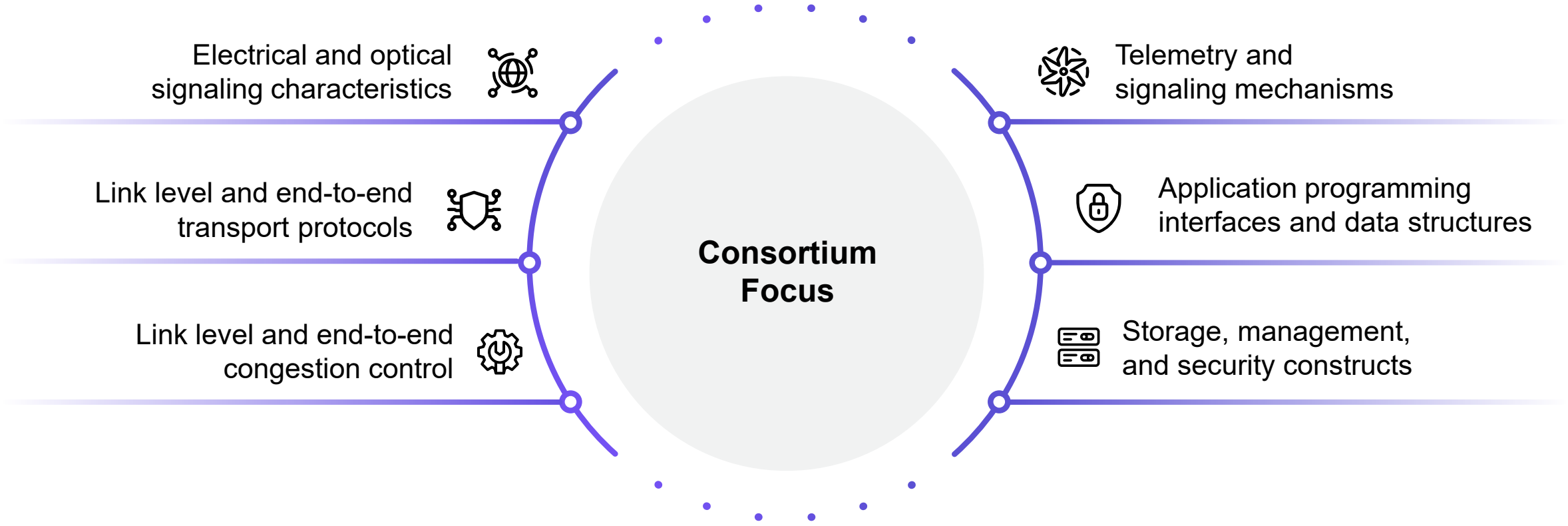
2025 Organization

- Full Standards Development Organization
- (One of the?) Fastest growing projects in Linux Foundation
- 120+ Companies
- 1500+ individual active contributor volunteers
- 8 Workgroups
  - Physical
  - Link Layer
  - Transport
  - Software
  - Storage
  - Management
  - Compliance & Test
  - Performance & Debug



UEC Technical Goals

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



UEC background

Ultra Ethernet Consortium

Who why?

*not all members listed
snapshot as of 2024-10

Mission:
Advance an Ethernet-Based Open, Interoperable, High-Performance Full-Stack architecture to meet the Growing Demands of AI and HPC at Scale

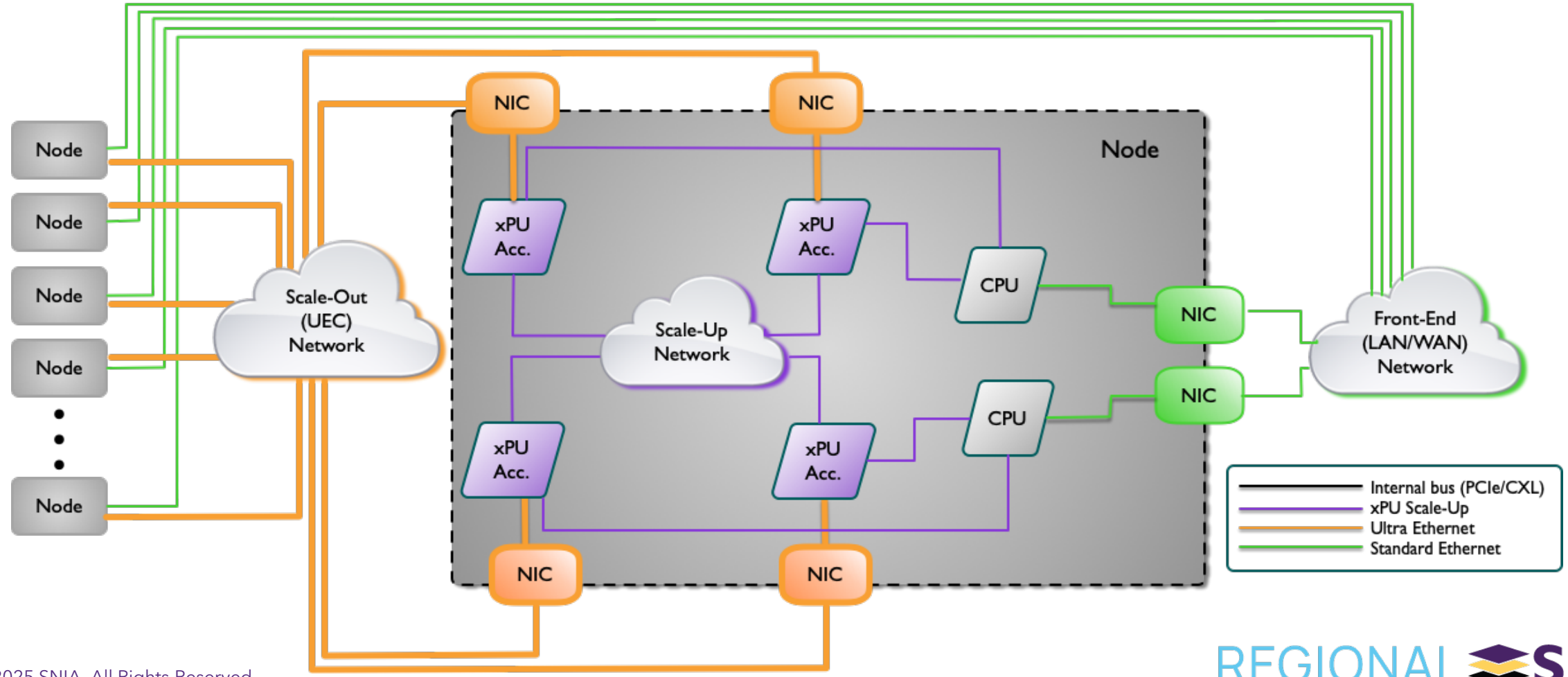
>100 member companies
>1300 active participants

REGIONAL SDC

Source:ultraethernet.org

Which Network?

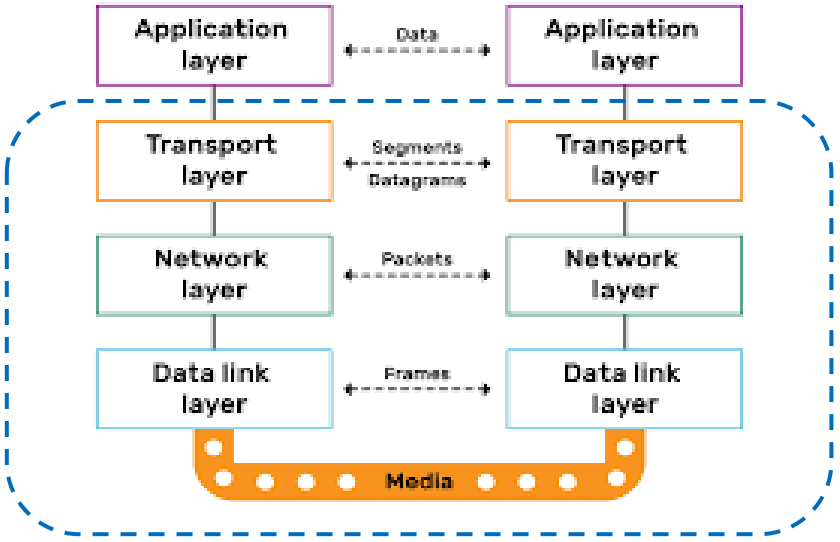
General Purpose vs. Scale-Up versus Scale-Out (UEC) Networks

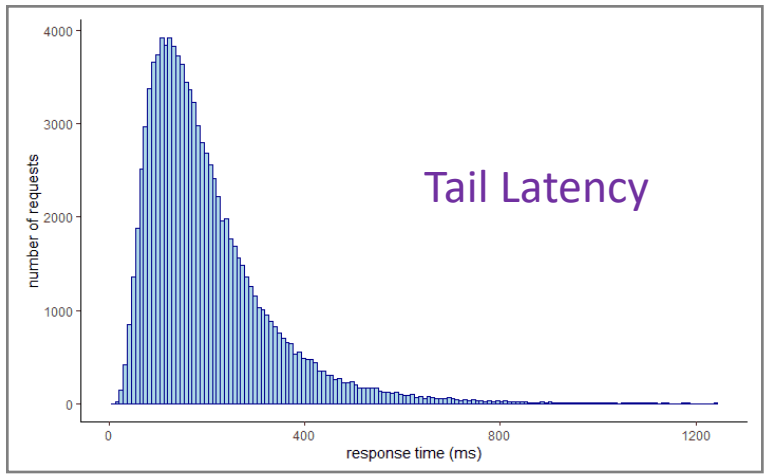


Modern Transport and RDMA Services for AI and HPC

Requirement	UEC Transport	Legacy RDMA	UEC Advantage
Multi-Pathing	Packet spraying	Flow-level multi-pathing	Higher network utilization
Flexible Ordering	Out-of-order packet delivery with in-order message delivery	N/A	Matches application requirements, lower tail latency
AI and HPC Congestion Control	Workload-optimized, configuration free, lower latency, programmable	DCQCN: configuration required, brittle, signaling requires additional round trip	Incast reduction, faster response, future-proofing
In Network Collective	Built-In	NONE	Faster Collective operation, lower latency
Simplified RDMA	Streamlined API, native workload interaction, minimal endpoint state	Based on IBTA Verbs	App-level performance, lower cost implementation
Security	Scalable, 1st class citizen	Not addressed, external to spec	High scale, modern security
Large Scale with Stability and Reliability	Targeting 1M endpoints	Typically, a few thousand simultaneous end points	Current and future-proof scale

AI/HPC Common Requirements





- Transport primitives for
  - Large Scale
  - Multi pathing
  - Relaxed ordering
  - Modernized Congestion Control
  - Optimized RDMA
  - Performance – bandwidth, latency, tail latency, Packets/S
  - High network utilization
  - Stability and Reliability

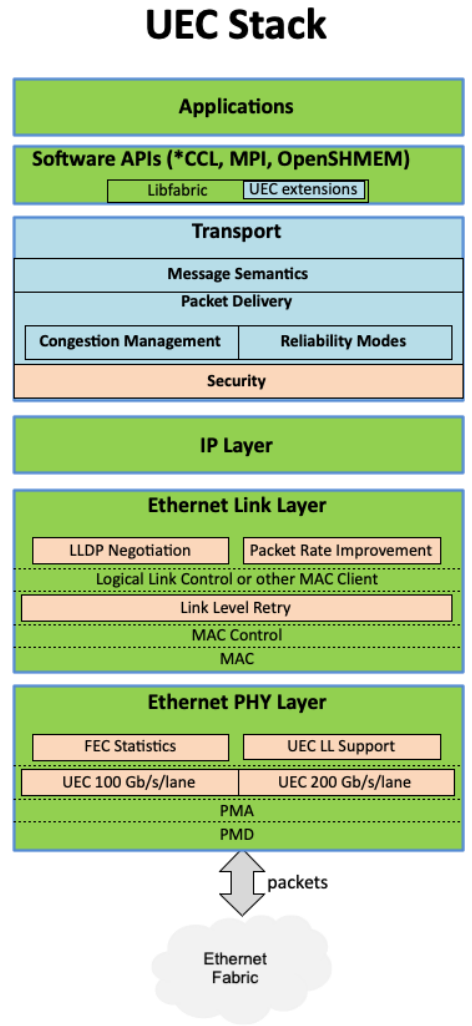
Key goals: high utilization & low tail latency!

UEC Stack Overview (partial feature list)

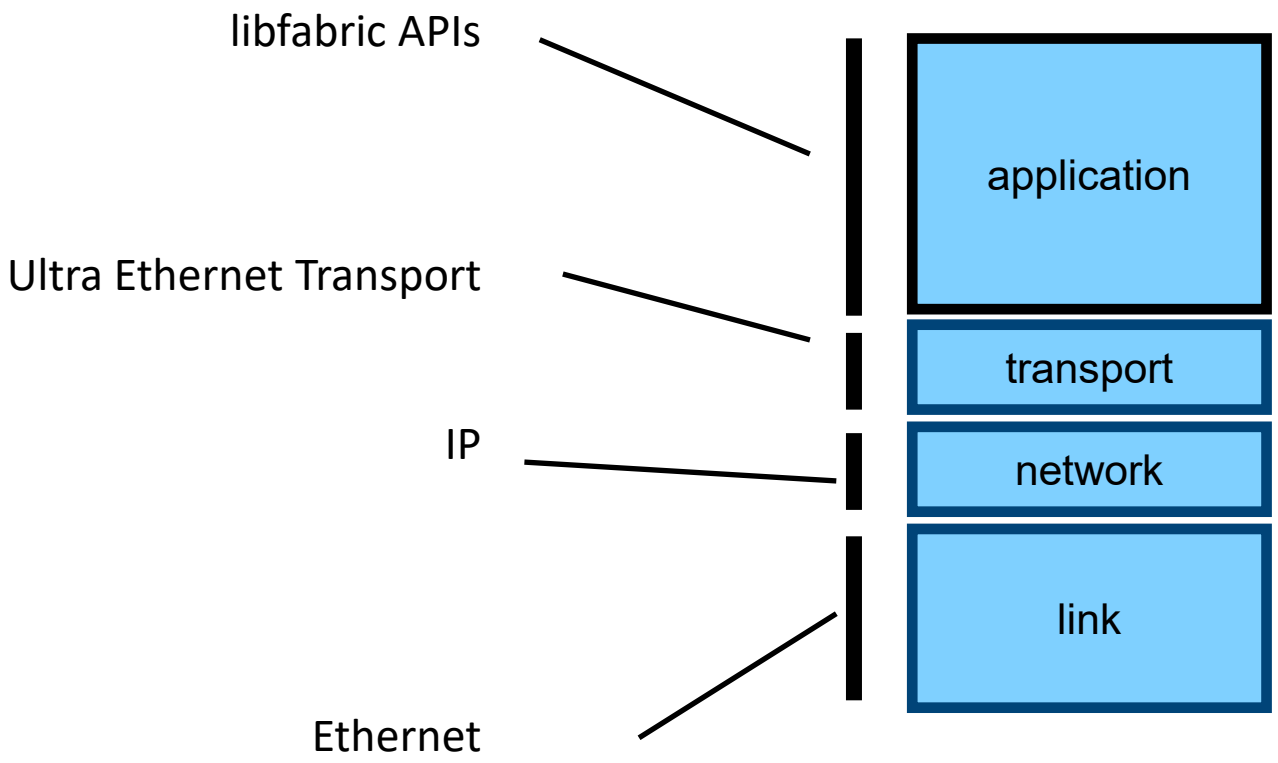
- Software API
  - Libfabrics 2.0 with extensions
- New Transport Layer
  - Multi-pathing
  - Packet spraying
  - Ordered (ROD) and un-ordered (RUD)
  - Lossy (no PFC) or Lossless
  - Congestion Control: Enhanced Tx and new Rx
  - Trimming
  - In Network Collective
- Network Layer
  - IP v4/v6
  - ECN

- Data Link Layer
  - Negotiation – LLDP
  - Link Level Retry - LLR
  - Header Efficiency Improvements
- Physical Layer
  - IEEE Compliant 100G Signaling

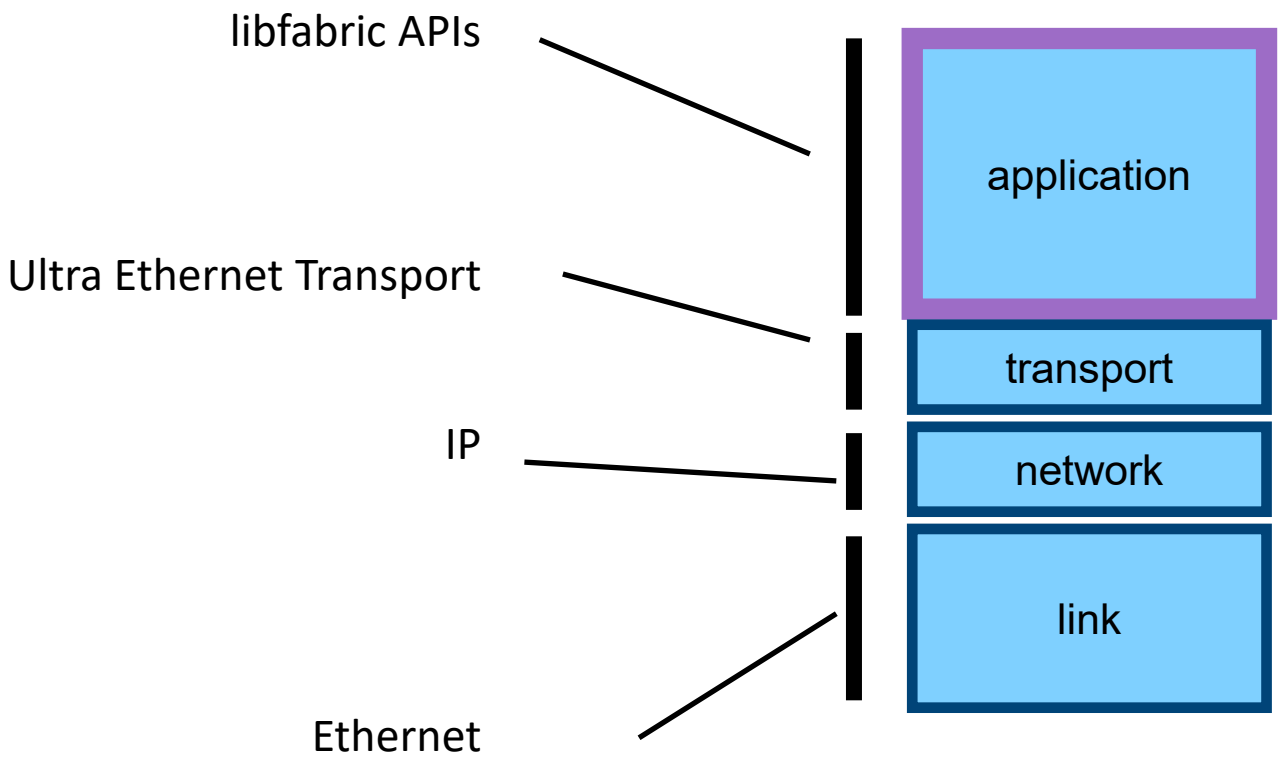
- AI and HPC Profiles



Ultra Ethernet up and down the stack

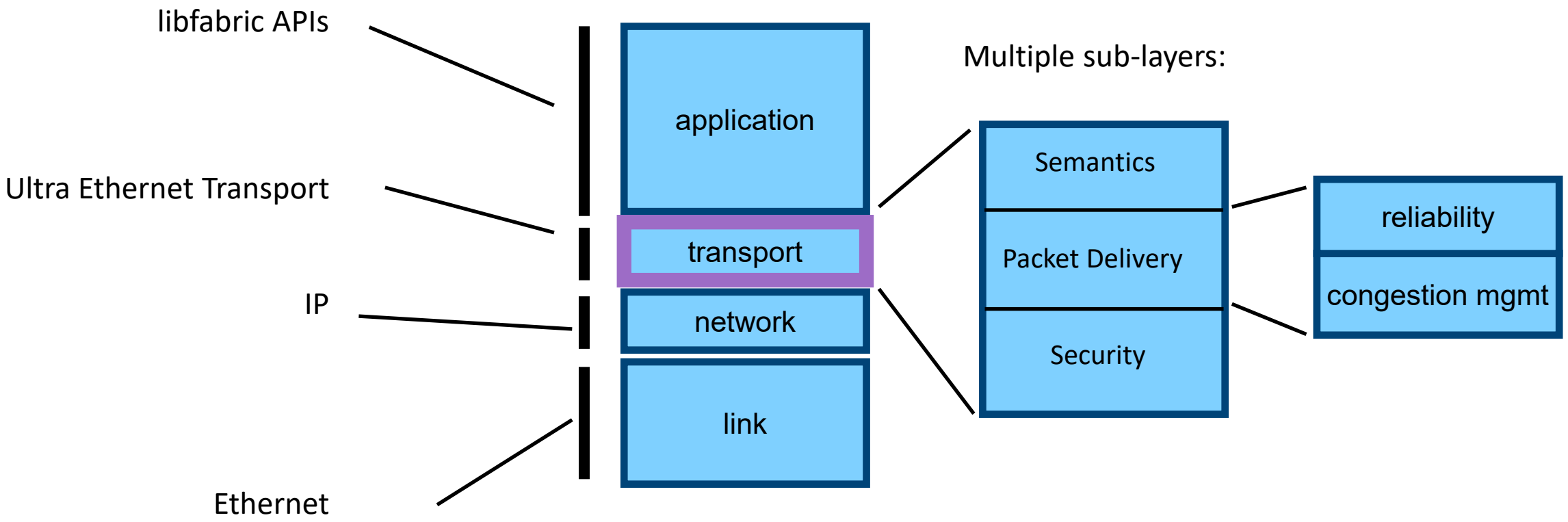


Ultra Ethernet up and down the stack

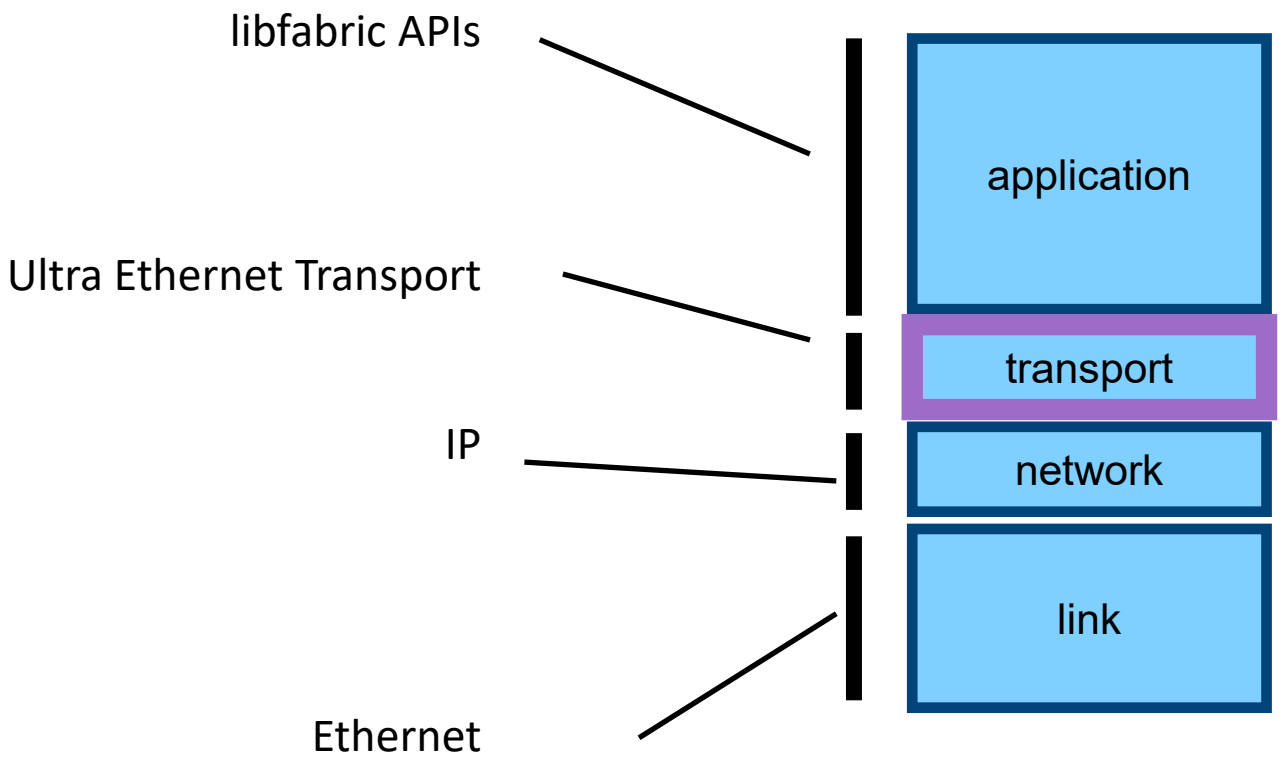


- Modern RMA API for HPC, AI
- Richer than sockets

Ultra Ethernet up and down the stack

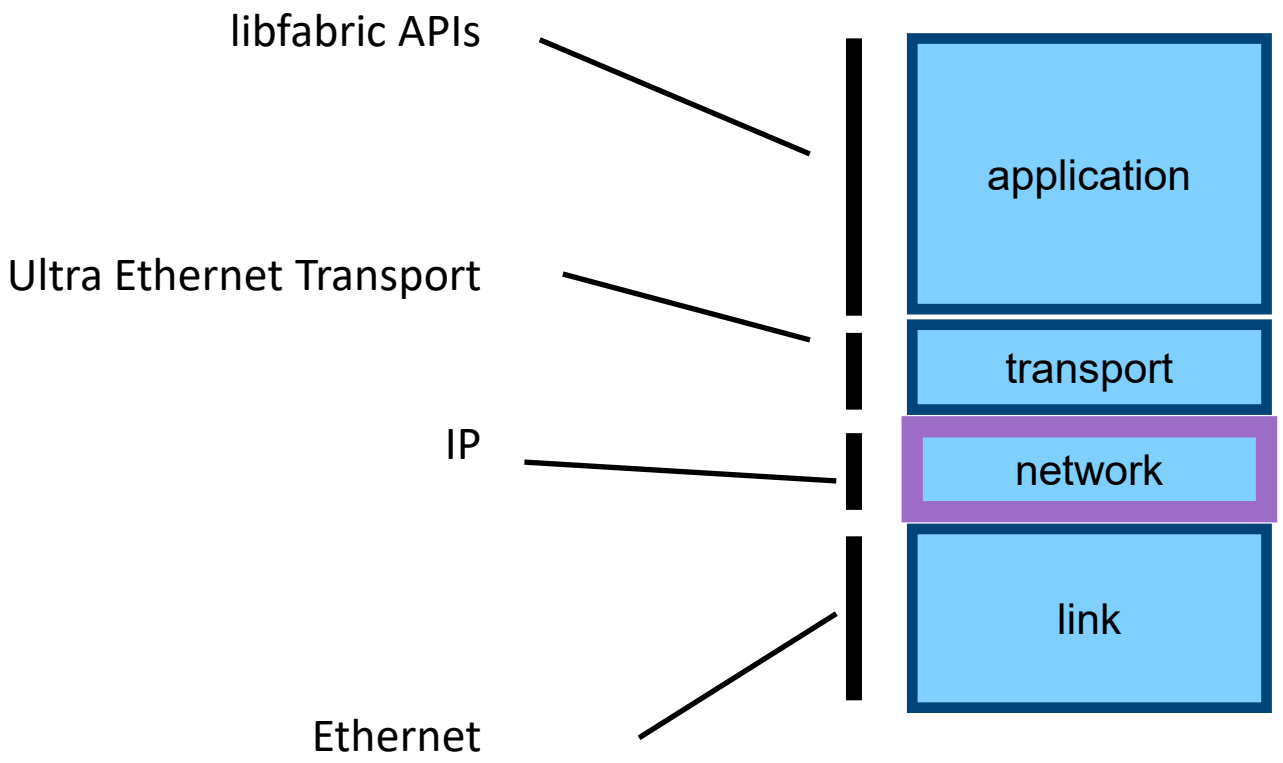


Ultra Ethernet up and down the stack



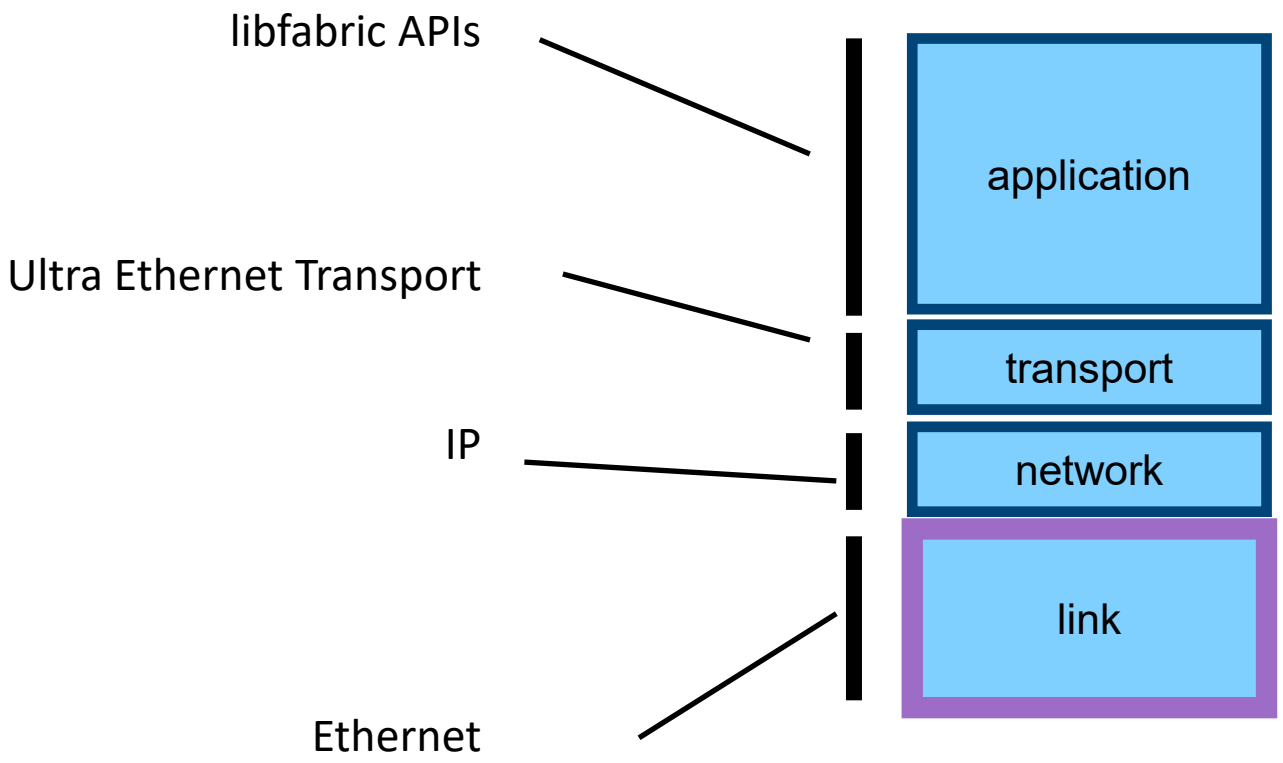
- RMA
- Multipathing
- Out-of-order delivery
- Security

Ultra Ethernet up and down the stack



- **Standard IP** - v4 or v6
  - DSCP, QoS
- Packet Trimming (optional)

Ultra Ethernet up and down the stack



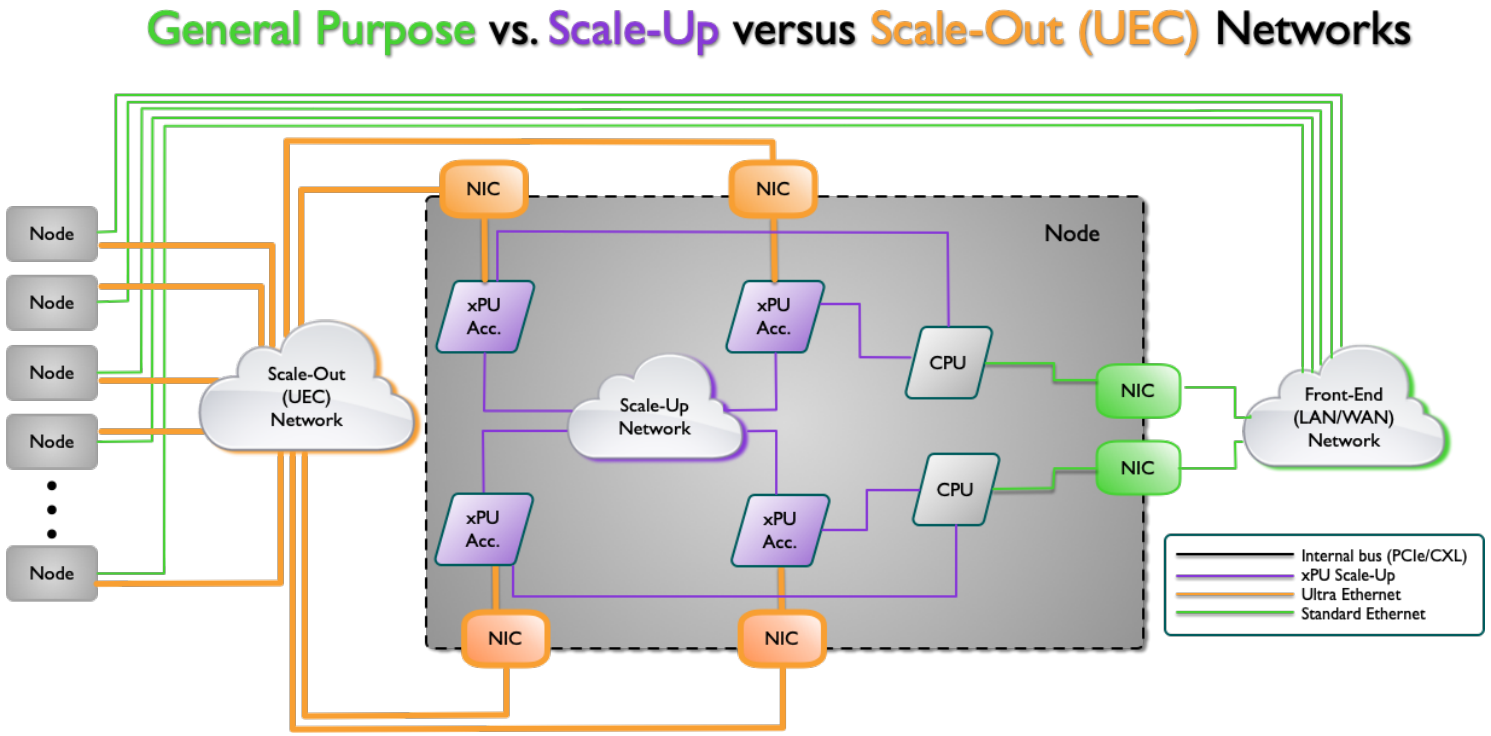
- **Standard Ethernet**
- Optional Link-layer Retransmit

UEC and Storage

- Storage Workgroup formed summer ‘24
- 3 tiers of storage being considered
  - Object
  - File
  - Block
- First project has been started for block storage
  - NVMe over Ultra Ethernet

Enter NVMe over UE

- Why NVMe over UE?
- For AI datacenters, Ultra Ethernet is designed to serve as the Scale-Out network connecting PODs of GPUs
- Part of the AI training process requires regular checkpointing of the data to storage nodes on the Scale-Out network
- NVMe-oF could fill this use case very well (there are other storage options as well, but NVMe-oF is being looked at as one of the big ones)
- NVMe over UE could also have a use in HPC networks (for which Ultra Ethernet is to be used as well)

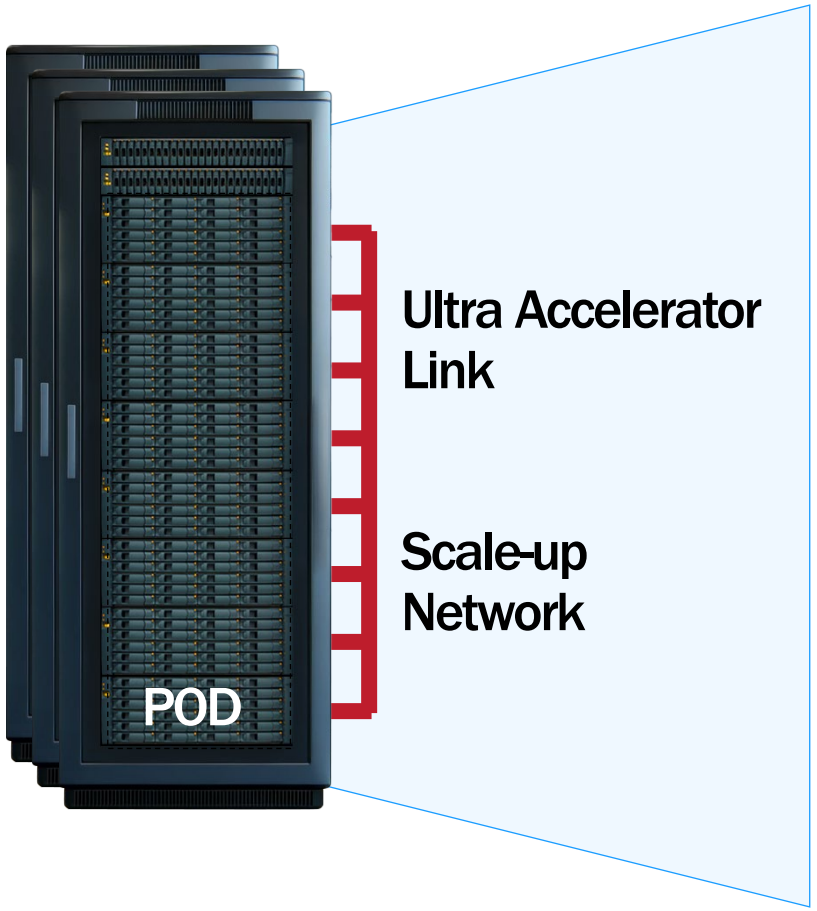


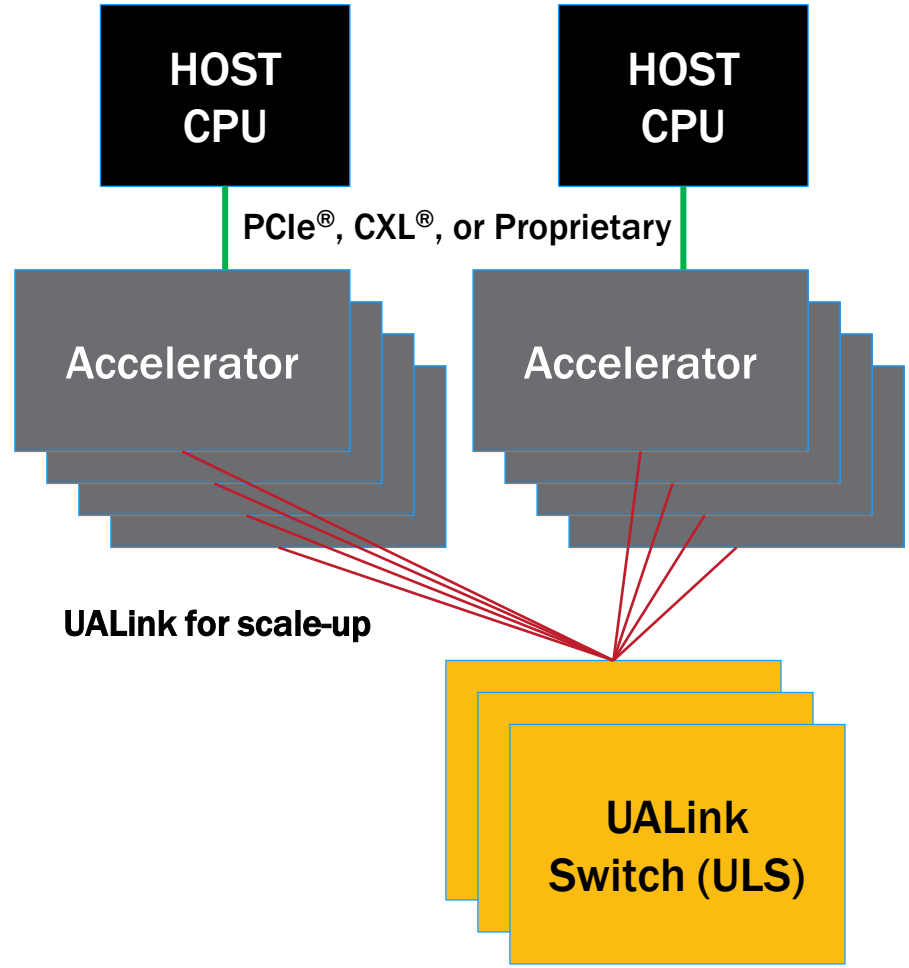
NVMe over Ultra Ethernet (NVMe/UE)

- First storage project started within UEC Storage WG
  - MOU has been completed between UEC and NVMe
- Should be similar to NVMe/RDMA
  - Semantics should be similar even if some of the details (such as connection management may be UE specific)
  - Goal is to work in concert with NVMe on specifcation

UALink Creates the Scale-up Pod

- High performance
  - 800Gbps per accelerator, up to 1,024 accelerators
- Low latency
  - Optimized protocol, transaction, link & physical
- Low power
  - The simplified UALink stack leads to lower power solutions
- Low die area
  - Optimized data layer and transaction layer saves significant die area





UALink 200G 1.0 Specification Overview

- The UALink interconnect is for Accelerator-to-Accelerator communication
  - The initial focus is sharing 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 (up to 1K)
  - Simple load/store/atomics semantics with software coherency
- The initial UALink specification taps into the experience of the Promoters developing and deploying a broad range of accelerators and seeded with the proven Infinity Fabric protocol