Ultra Ethernet Consortium

Powering AI's Future with Ethernet

Or, what is this "Ultra Ethernet" I keep hearing about?

AI for Networking, or Networking for AI?

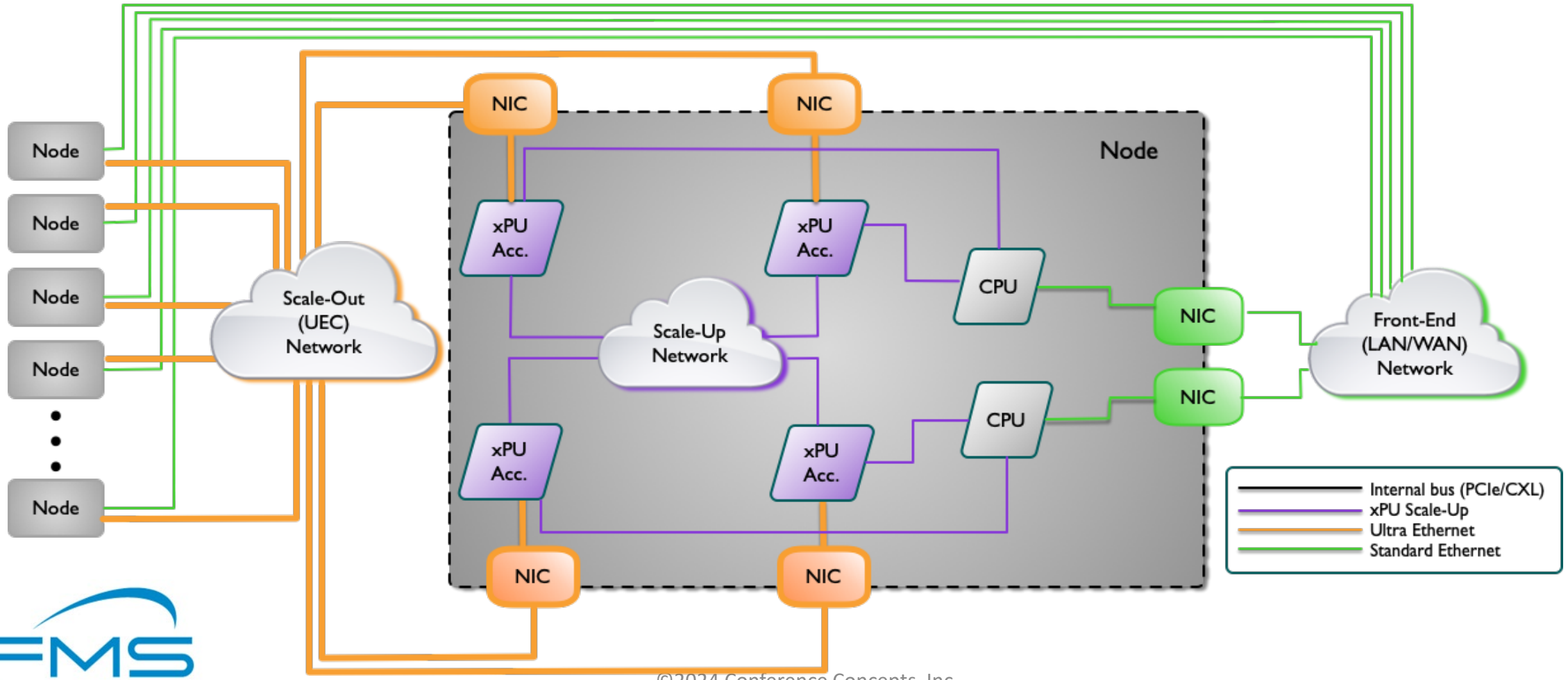
- Many articles/blogs have talked about how AI can change the networking infrastructure
  - … but what network infrastructure do you need to have enough AI to change the networking infrastructure?
  - Is it more than just superfast speeds and feeds?
  - Massive data sets, parallel processing requirements
  - Where does the data need to be, and when?

The AI Monster

- AI workloads need
  - Ever-increasing Memory Bandwidth
  - Ever-increasing Memory Capacity
  - (Near) Instantaneous Data Access (Exabytes)
- Intermittent data surges
- ”Straggler” data (tail latency) significantly impacts completion time
- Extended operation duration (hours, days)

Which Network?

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



Bandwidth and Latency

- Training is highly *latency*-bound, where tail latency negatively impacts the frequent computation and communications phases
  - Generation stage is maximum contribution to latency; 60-80% of total
  - Latency increases with # of output tokens
- Large models (e.g., from 175B parameters in GPT-3 to 1T+ in GPT-4) drive larger messages on the network
- Underperforming networks therefore underutilize expensive resources

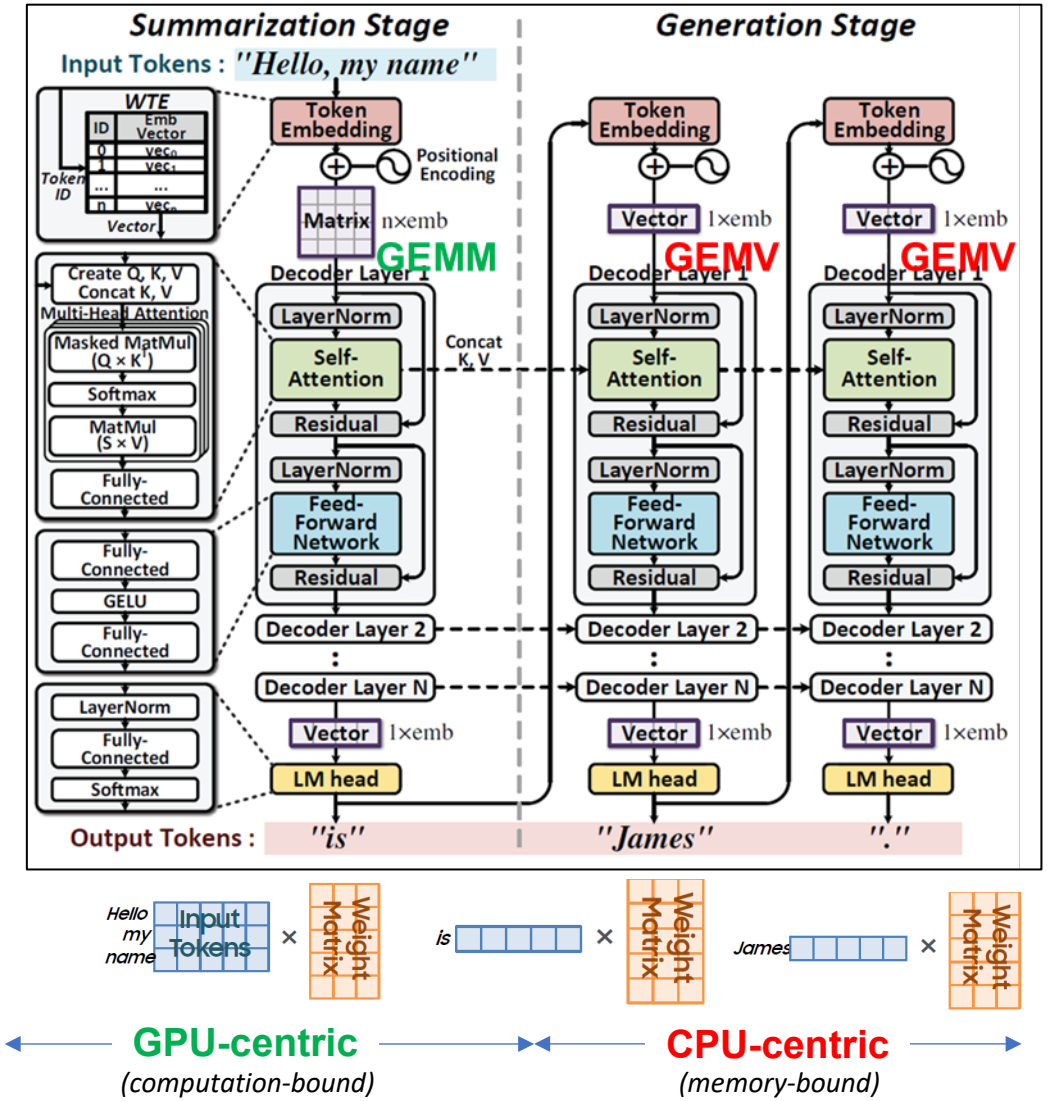


Image credit: Hong, Seongmin, et al. "DFX: A Low-latency Multi-FPGA Appliance for Accelerating Transformer-based Text Generation." *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2022.

Introducing: Ultra Ethernet Consortium (UEC)

Ultra Ethernet Consortium

FMS the Future of Memory and Storage

Introducing: The Promise of Ultra Ethernet

https://ultraethernet.org/

THE NEW ERA NEEDS A NEW NETWORK

Ultra Ethernet

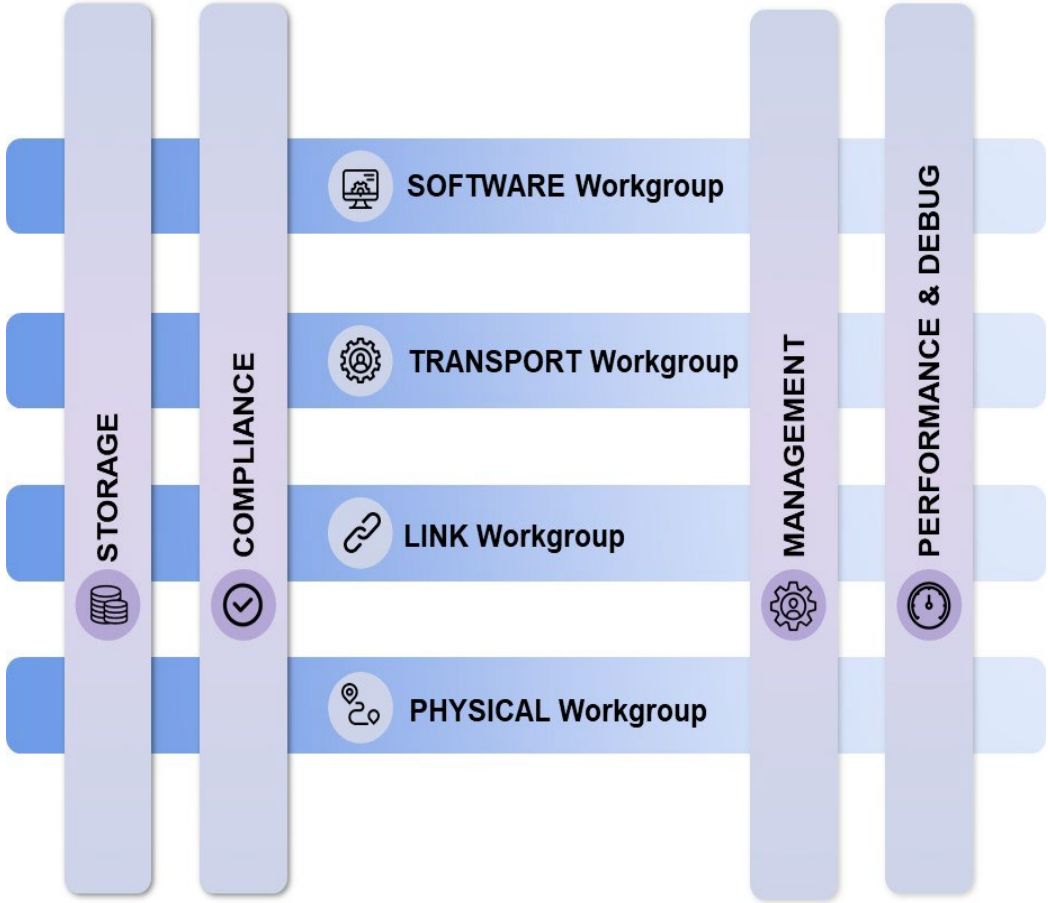
As **performant** as a supercomputing interconnect

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

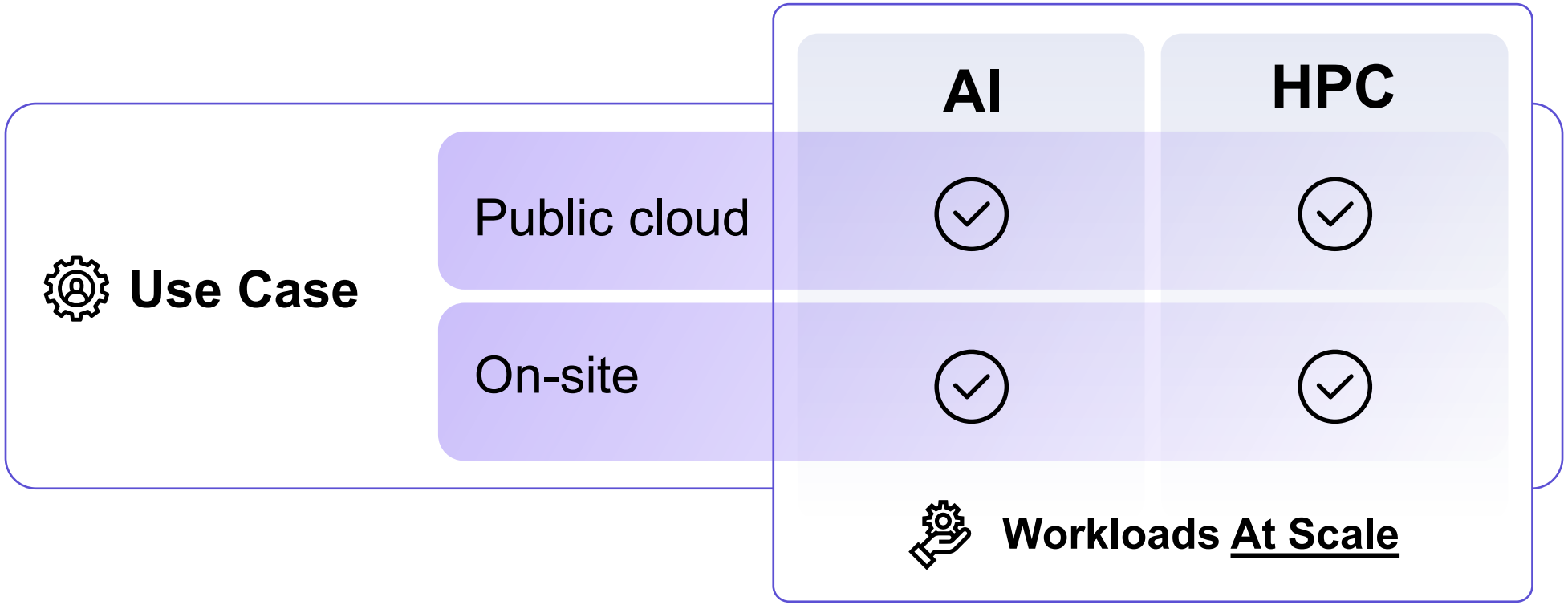
As **scalable** as a cloud data center

2024 Organization

- Full Standards Development Organization
- (One of the?) Fastest growing projects in Linux Foundation
- 90+ Companies
- 1200+ individual active contributor volunteers
- 8 Workgroups
  - Physical
  - Link Layer
  - Transport
  - Software
  - Storage
  - Management
  - Compliance & Test
  - Performance & Debug



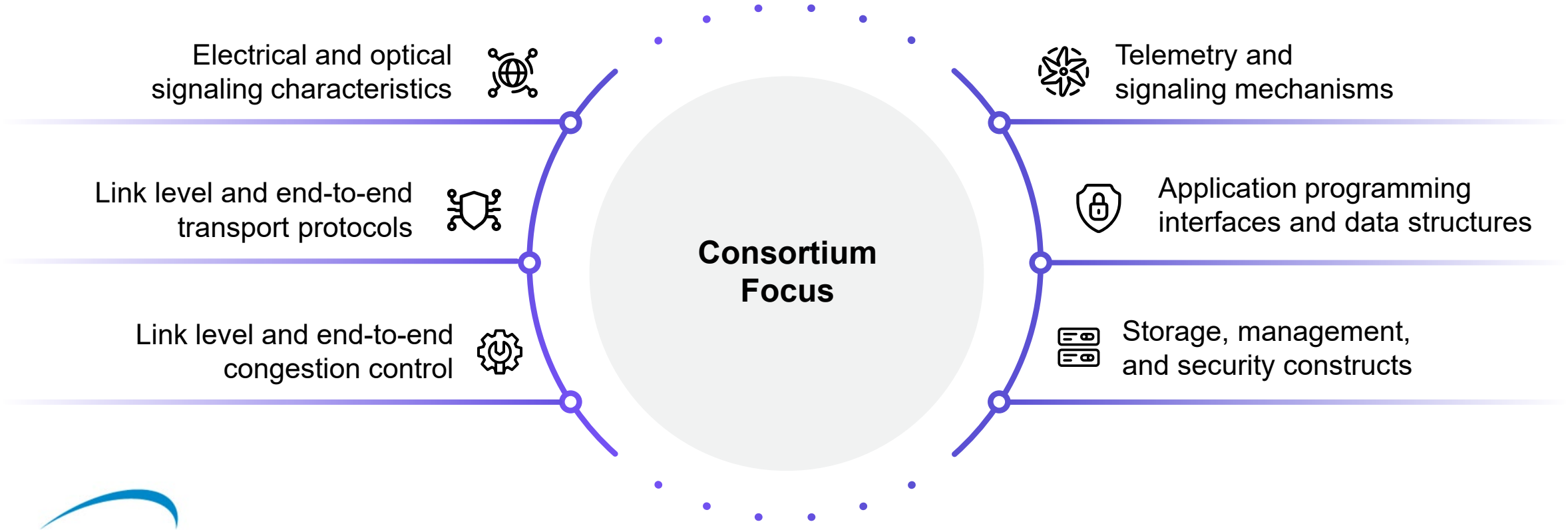
Target Deployment Models/Use Cases



Profiles defined for AI and HPC use cases

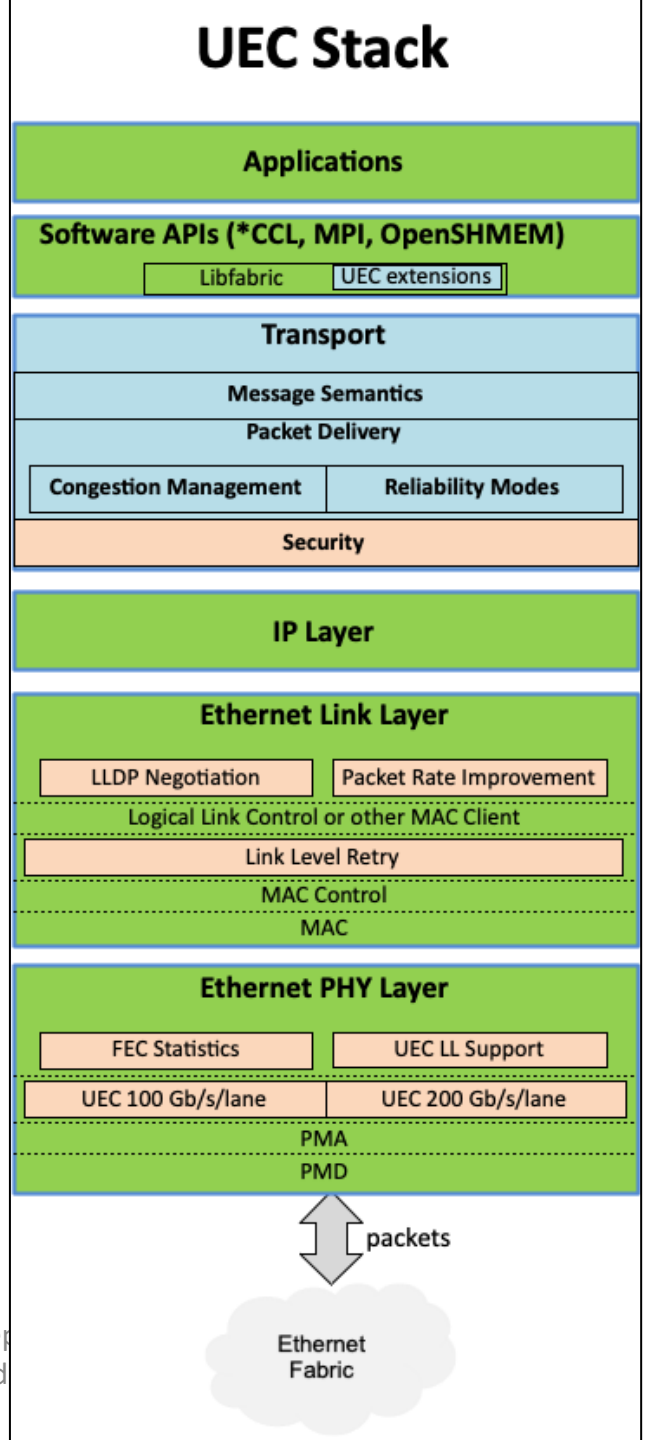
UEC Technical Goals

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



Understanding the UEC Stack

- Backwards-compatible
  - Uses libfabric as its north-bound API
  - Designed to integrate into existing frameworks where libfabric is commonly utilized
- Key driving force is in the Ultra Ethernet Transport (UET)
  - Supplemented by optional functions and features, depending upon the profile



Legend:

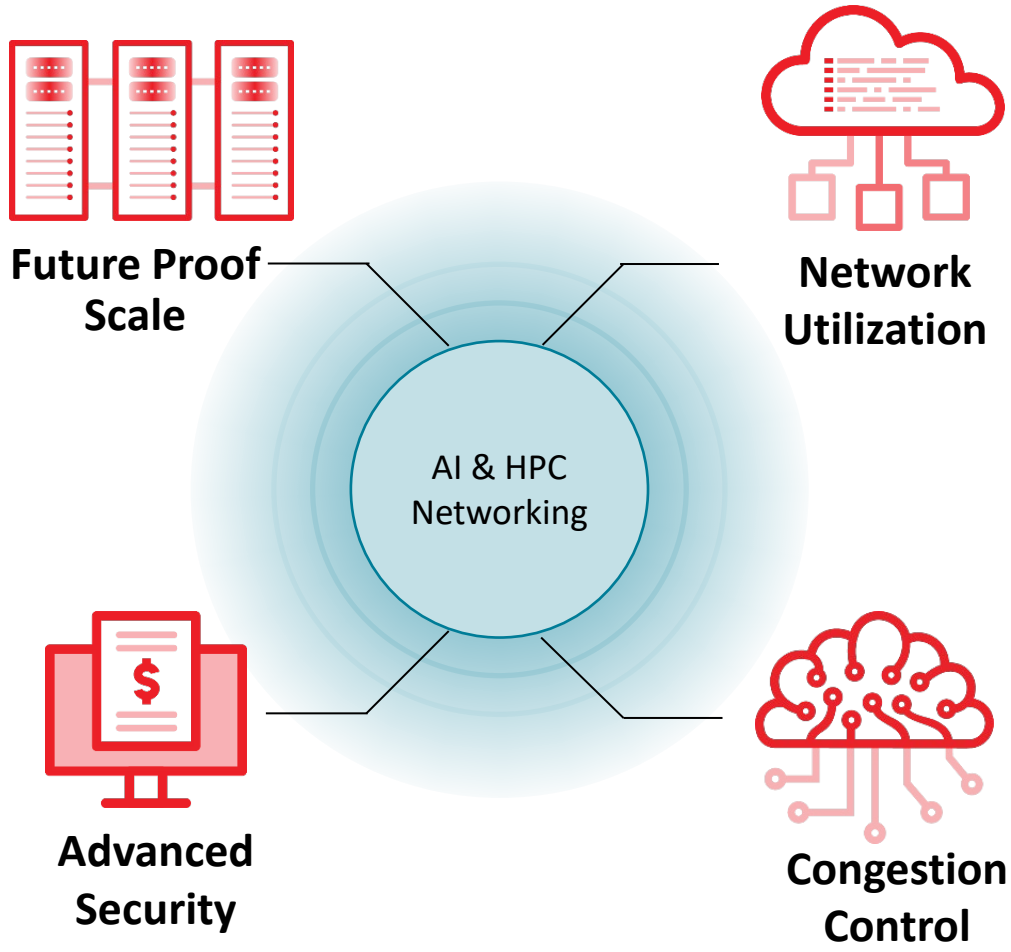
Mandatory

Optional

Untouched

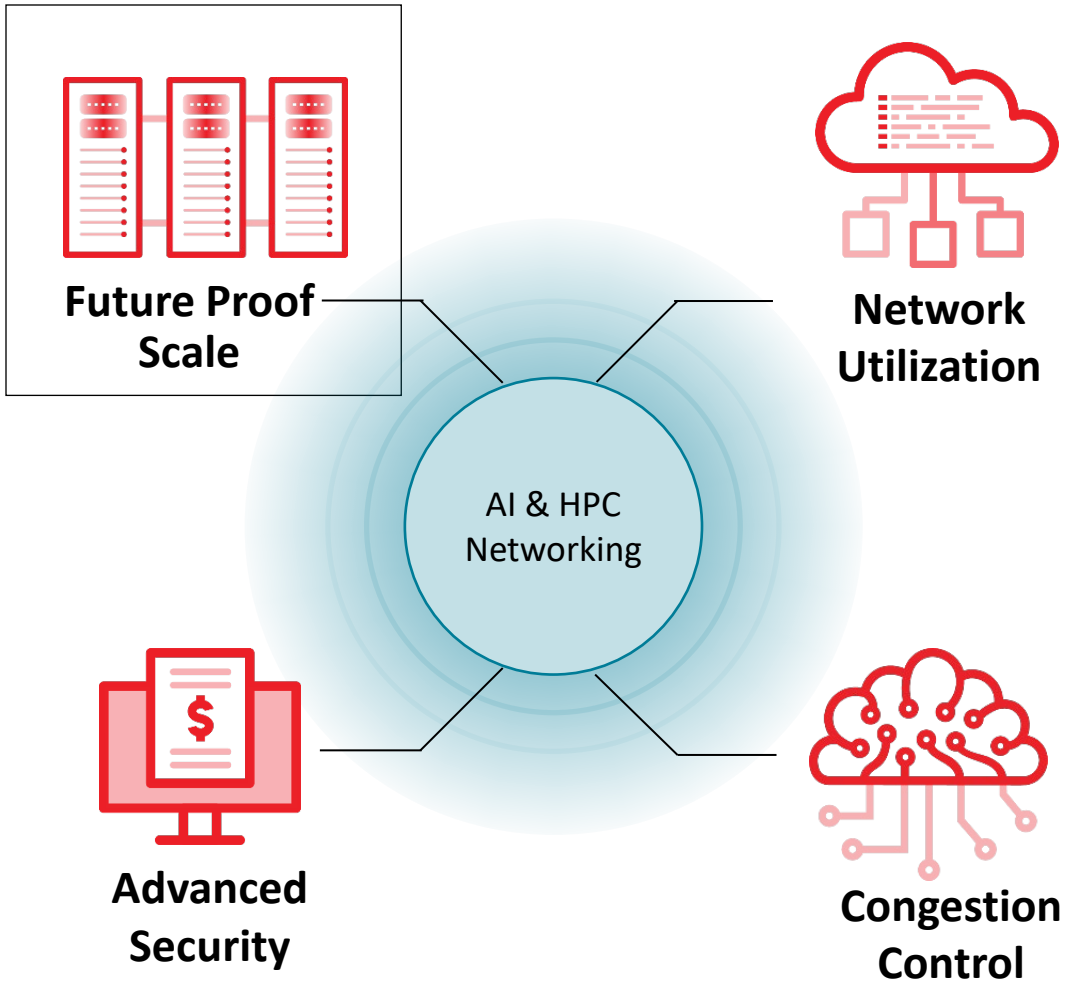
UEC Transport Addresses Grand Challenges

- Future proof system scale with 1M endpoints
- Improved network utilization with multi-path routing
- Lower tail latency with flexible packet ordering
- Security built-in from the beginning
- AI and HPC congestion control require faster response times
- End-To-End telemetry provides improved network visibility



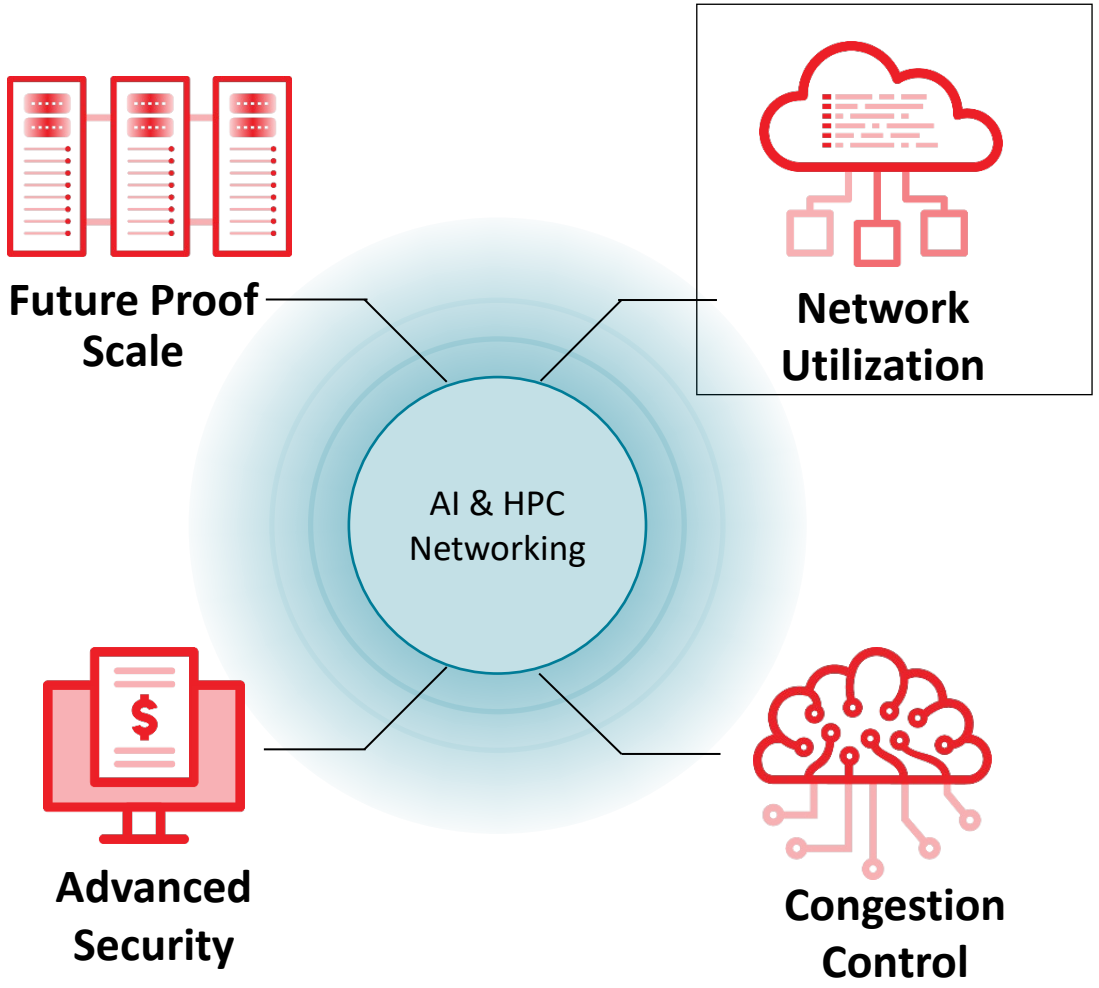
Future-Proof Scale

- Determinism and predictability become more difficult as systems grow
  - New methods needed to achieve holistic stability & visibility
- Simultaneous packet-based multipathing/“packet spraying”
  - Every flow simultaneously access all paths
  - Achieves more balanced use of entire network



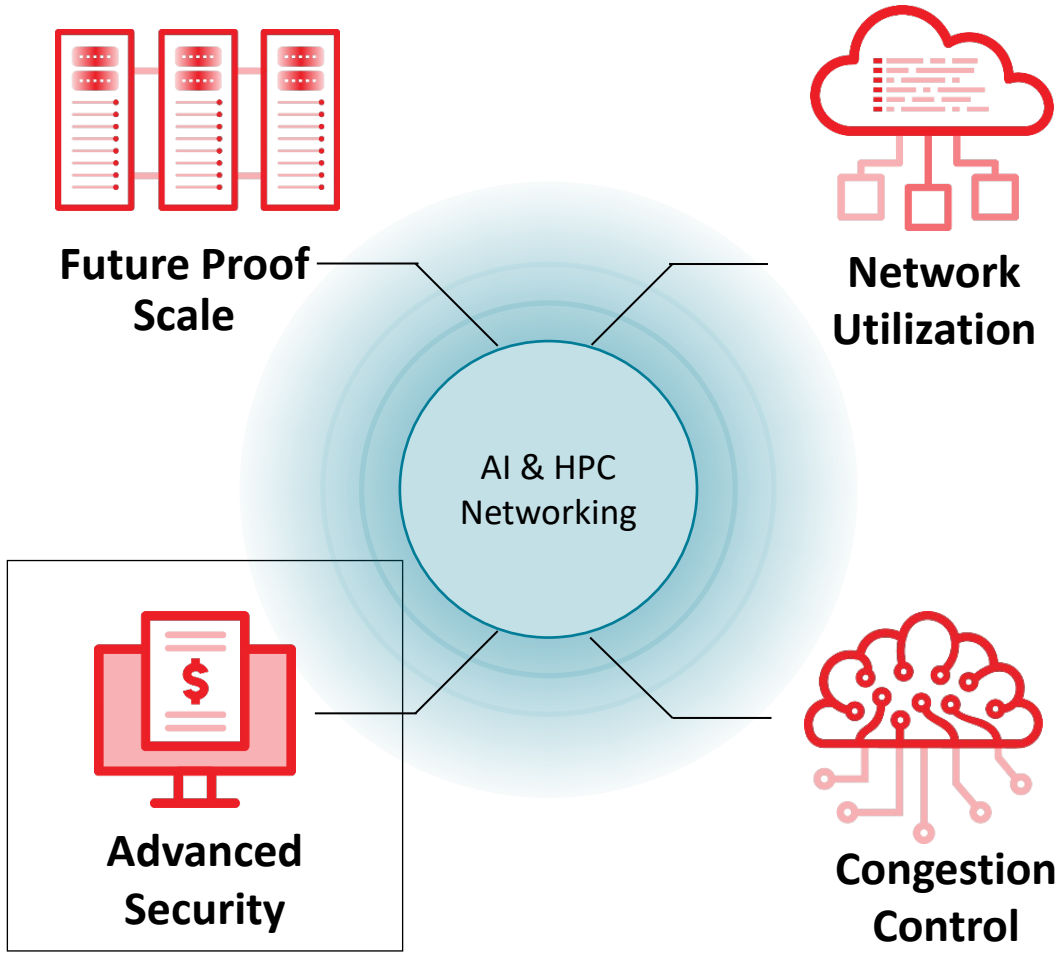
Network Utilization

- From Rigid to Flexible Ordering
  - Rigid ordering enables "go-back-n" recovery and in-order delivery, but restricts network utilization and increases tail latencies
  - Flexible ordering enables packet-spraying in bandwidth-intensive collective operations; eliminates to reorder packets
  - Supports modern APIs that relax the packet-by-packet ordering requirements for applications where it's critical to curtail tail latencies



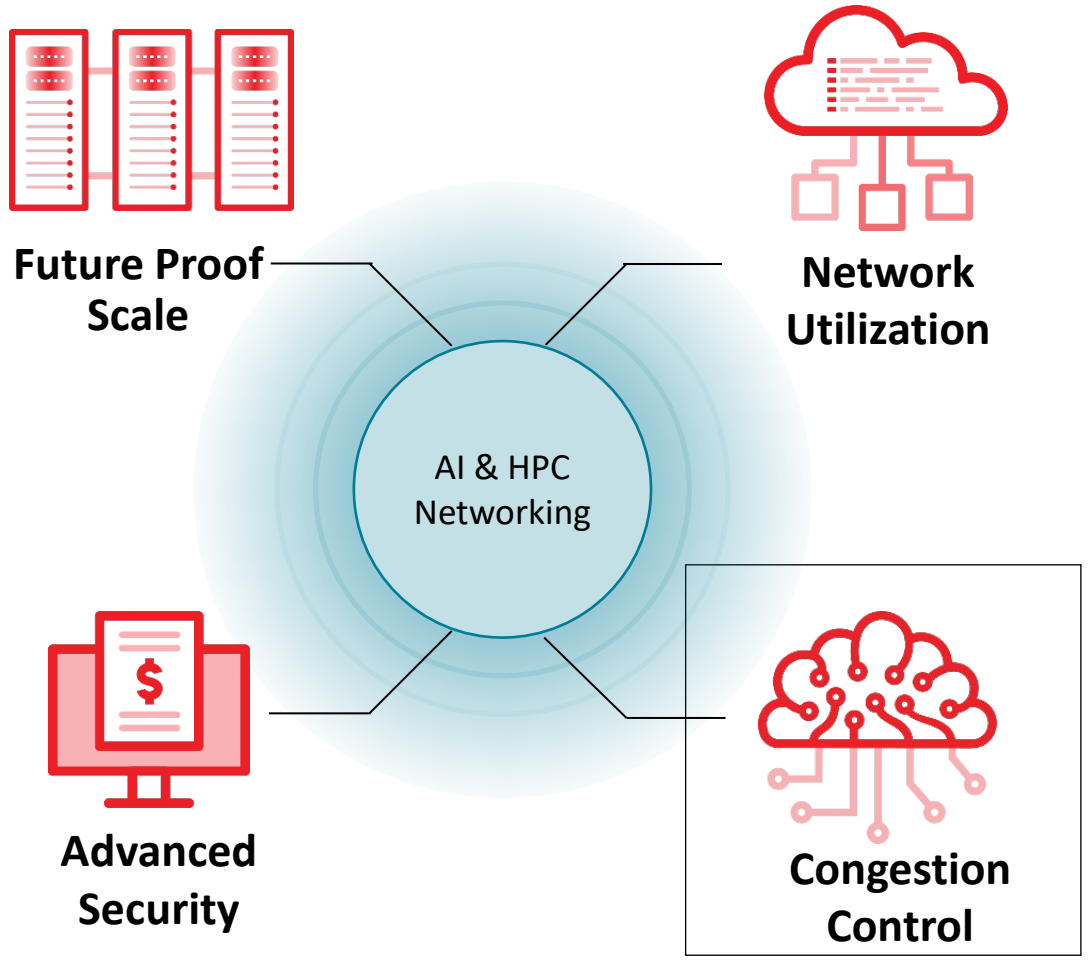
Advanced Security

- Advanced Security
  - Encryption support that doesn't balloon the session state in hosts and network interfaces
  - Similar conditions in AI and HPC



Congestion Control (and Telemetry)

- Congestion
  - Must work with packet spraying
  - Must coordinate with scheduling algorithms on sending host
- Telemetry
  - Congestion information originating from the network can advise the participants of the location and cause of the congestion
  - Robust end-to-end telemetry enables optimized congestion control algorithms
  - Shortening the congestion signaling path and providing more information to the endpoints allows more responsive congestion control



UEC Addresses AI Network Needs

	Traditional RDMA-Based Networking	Ultra Ethernet Consortium
	Required In-Order Delivery, Go-Back-*N* 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

Summary

- Ambitious, full-stack solution for Ethernet-based AI and HPC networking
- Massive scale and performance-sensitive
- Anticipated 1.0 integrated spec by end of CY 2024

Ultra Ethernet Consortium

LEARN MORE AT

www.ultraethernet.org


Ultra Ethernet
Consortium