



Ethernet Evolved: Powering AI's Future with the Ultra Ethernet Consortium

J Metz, Ph.D, *Ultra Ethernet Consortium*

SNIA COMPUTE, MEMORY, AND STORAGE SUMMIT

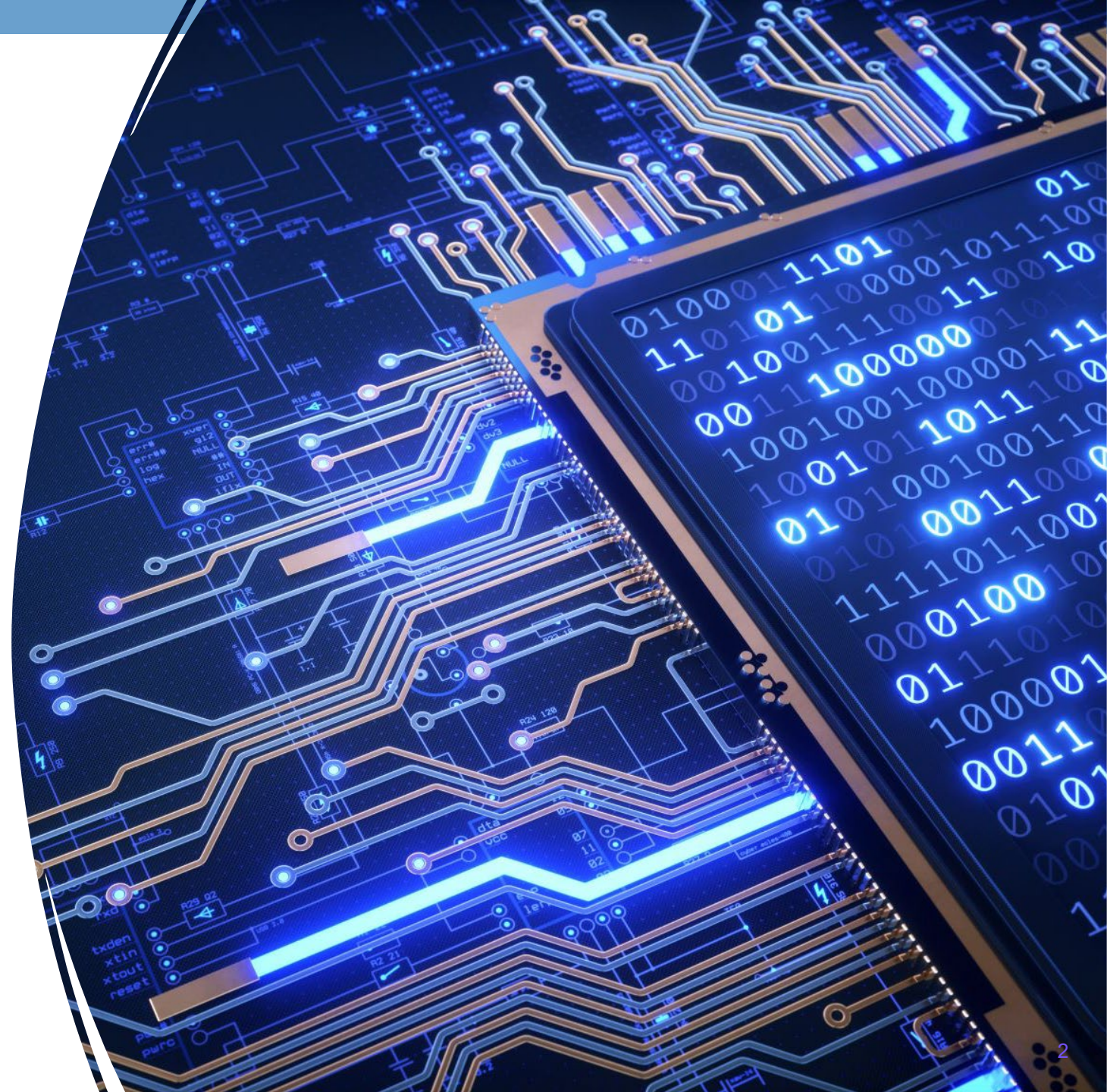
Solutions, Architectures, and Community
VIRTUAL EVENT, MAY 21-22, 2024

www.snia.org/cms-summit



“Ultra Ethernet?” - Agenda

- Why “Ultra Ethernet?”
- Who is Ultra Ethernet Consortium (UEC)?
- Philosophy of UEC
- Overview of projects
- Where to get more information



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?

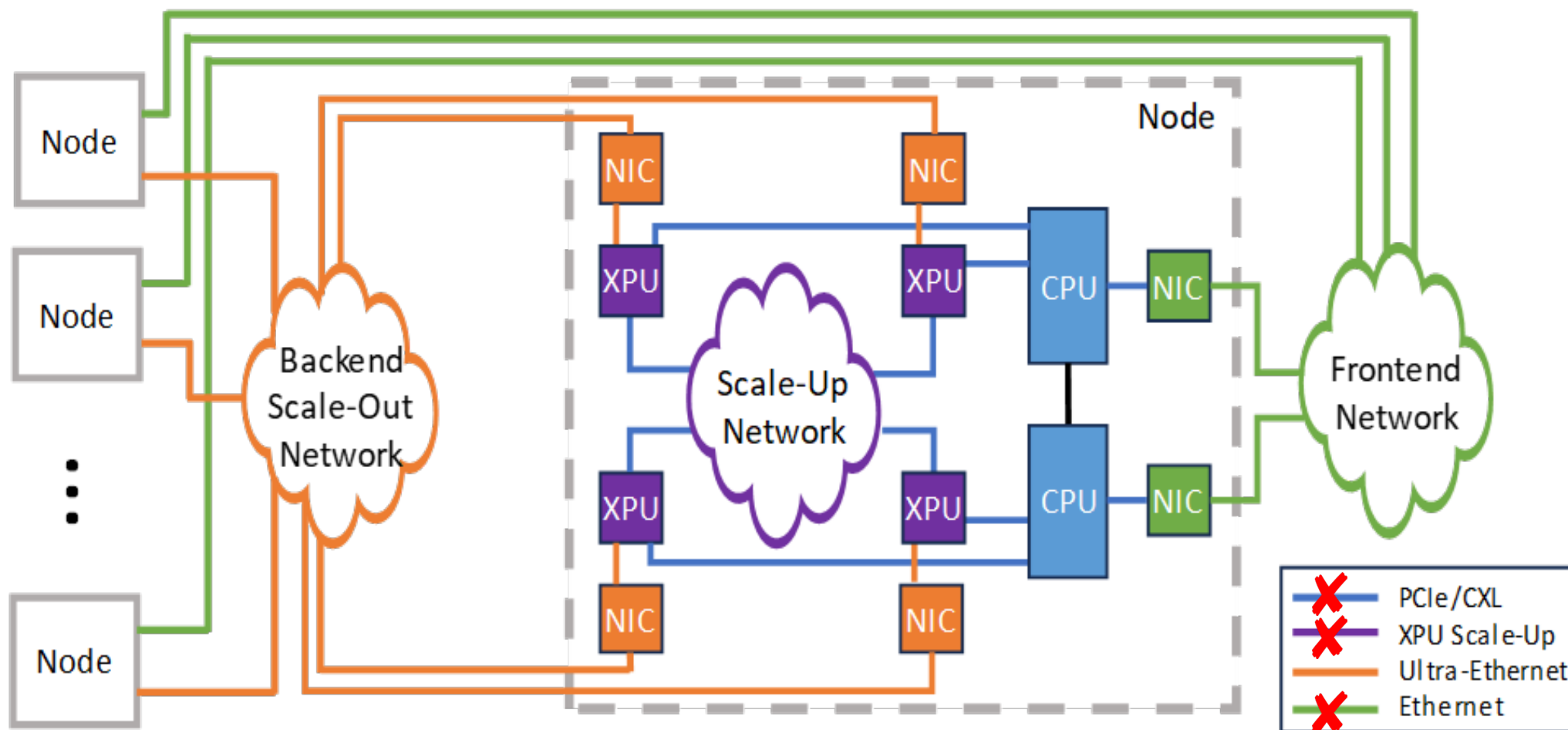
The AI Monster

- 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?

- UEC = Scale-Out



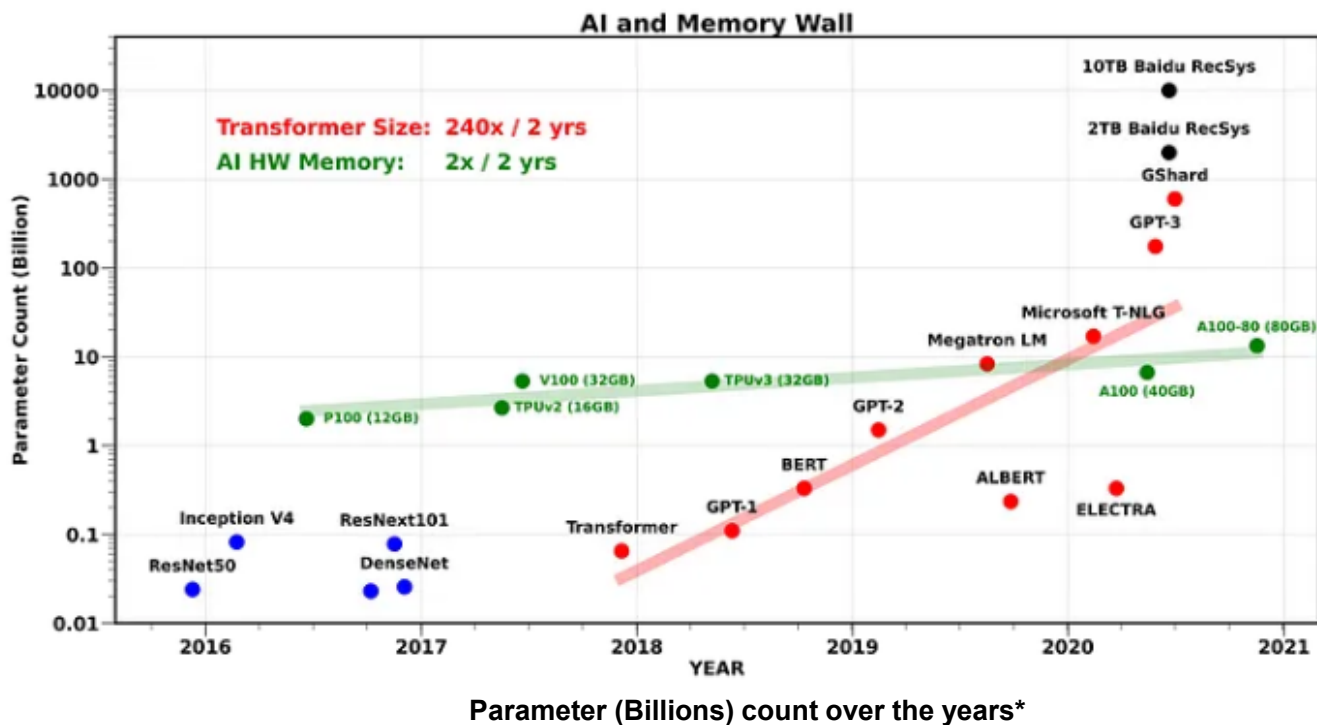
AI Problems To Solve



- Memory Bandwidth vs. Capacity vs. Latency
- Computation-Bound Workloads
 - E.g., Summarization: processes large input context simultaneously (parallel)
 - Higher weights reusability
- Memory-Bound Workloads
 - E.g., Generation: produces single word at a time (iteration)
 - Lower weight reusability
- Recommendation Workloads spend almost 60% of time in Network I/O*

*Meta, OCP 2022 Global Summit

The Goodput Dilemma



- Logarithmic expansion of the relationships between data stresses networks
- Compute, Memory, and Bandwidth constraints
 - What's the impact on data movement?
 - What happens when you hit 1 Million endpoints?
- Things that break (or, at least, hurt):
 - Congestion signaling, notification, spreading and mitigation (e.g., reaction time)
 - Data ordering and sequencing
 - Timely telemetry
 - Multipath flow-hashing and load-balancing
 - Best practices that require manual tuning
 - Recovery methods
 - Management techniques
 - I/O Amplification
 - Security

Remote Access to Memory



■ Issues

- Verbs API limits efficiency by preventing OOO packet data from being delivered straight through the network to the application buffer (final destination)
- Go-Back- N recovery methods retransmit N packets for any single packet loss

■ Impact

- Ties up network bandwidth for recovery
- Causes under-utilization of available links
- Increases tail latencies

■ Ideal Solution

- All links are used; order is only enforced when the AI workload requires it

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

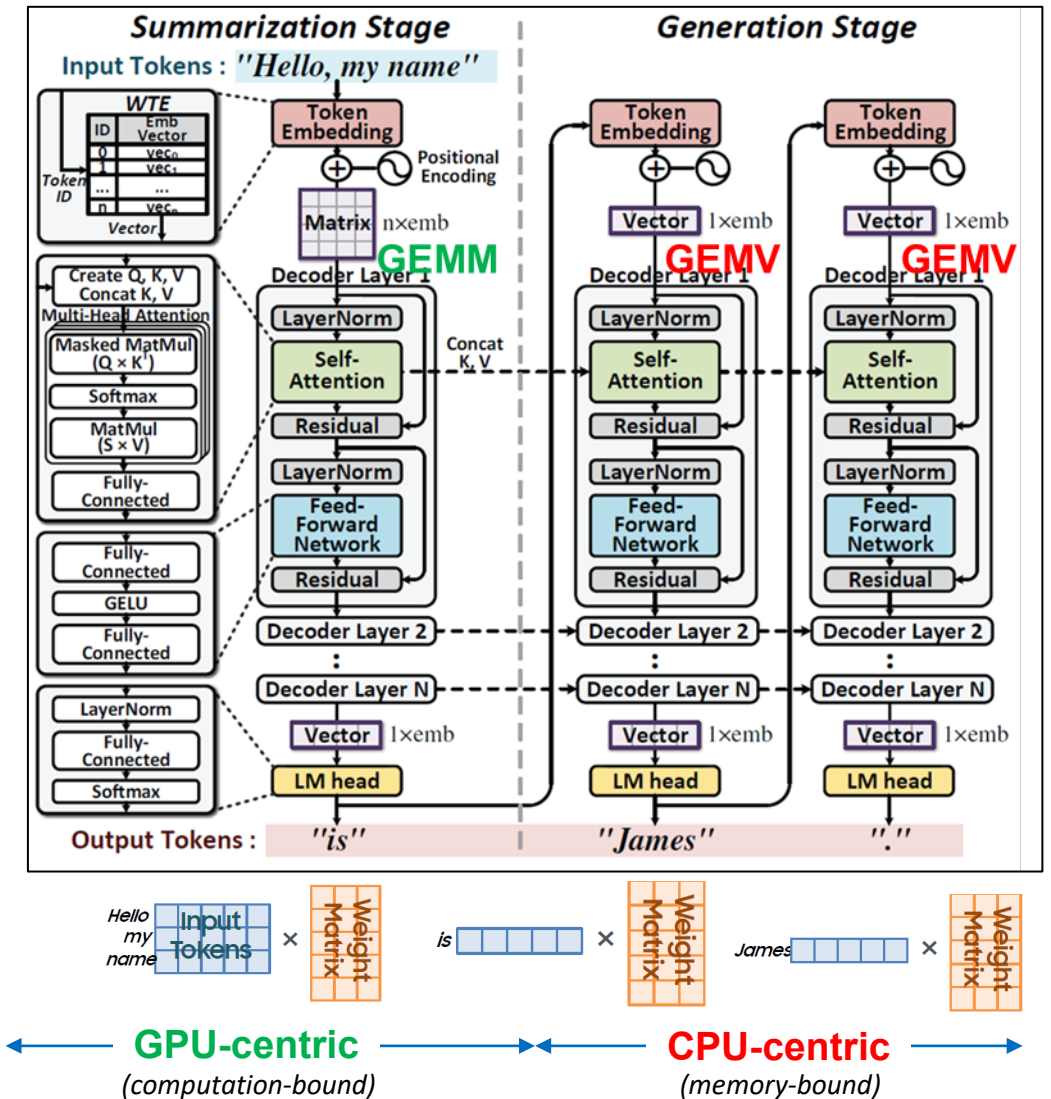


Image credit: Hong, Seongmin, et al. "DFX: A Low-latency Multi-FPGA Appliance for Accelerating Transformer-based Text Generation." 2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO). IEEE, 2022.

Introducing: Ultra Ethernet Consortium (UEC)

Ultra Ethernet
Consortium

SNIA COMPUTE, MEMORY,
AND STORAGE SUMMIT

Solutions, Architectures, and Community
VIRTUAL EVENT, MAY 21-22, 2024

INTRODUCING: THE PROMISE OF ULTRA ETHERNET

<https://ultraethernet.org/>

**THE NEW ERA
NEEDS A
NEW NETWORK**

*Ultra***Ethernet**

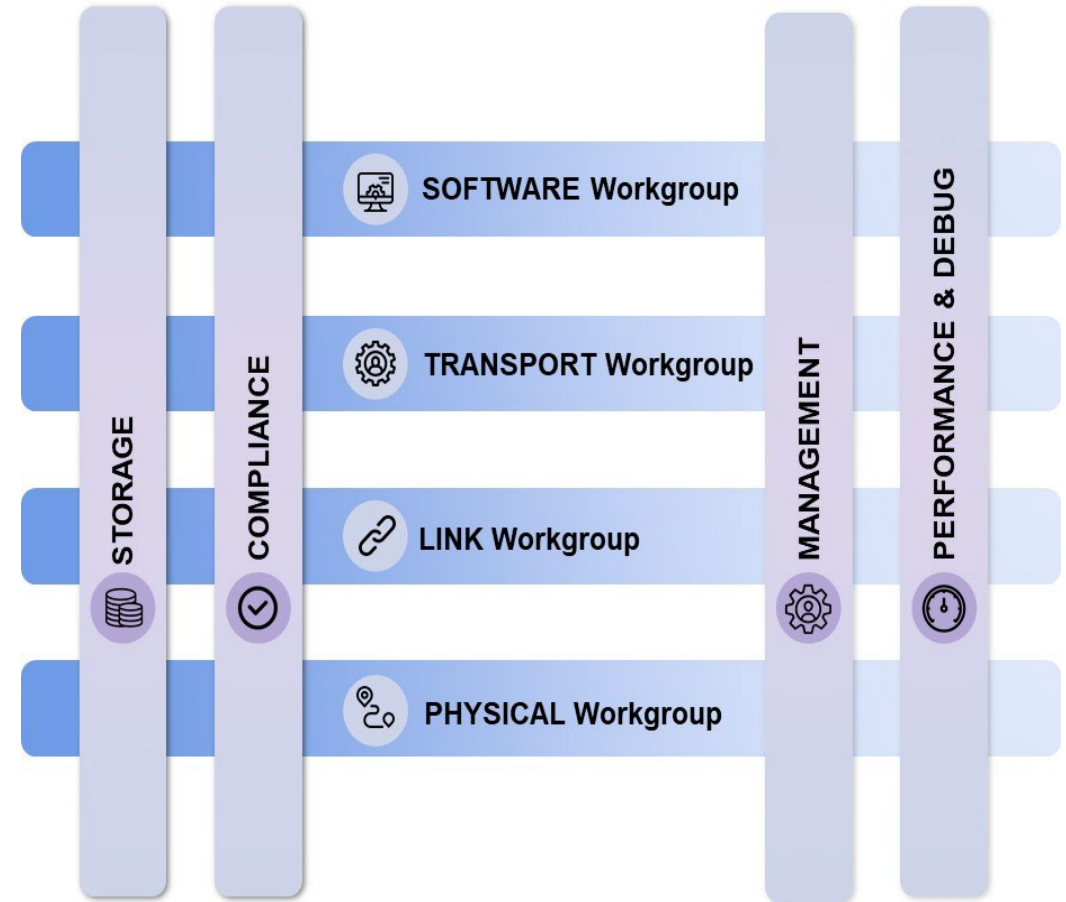
As **performant** as a
supercomputing interconnect

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

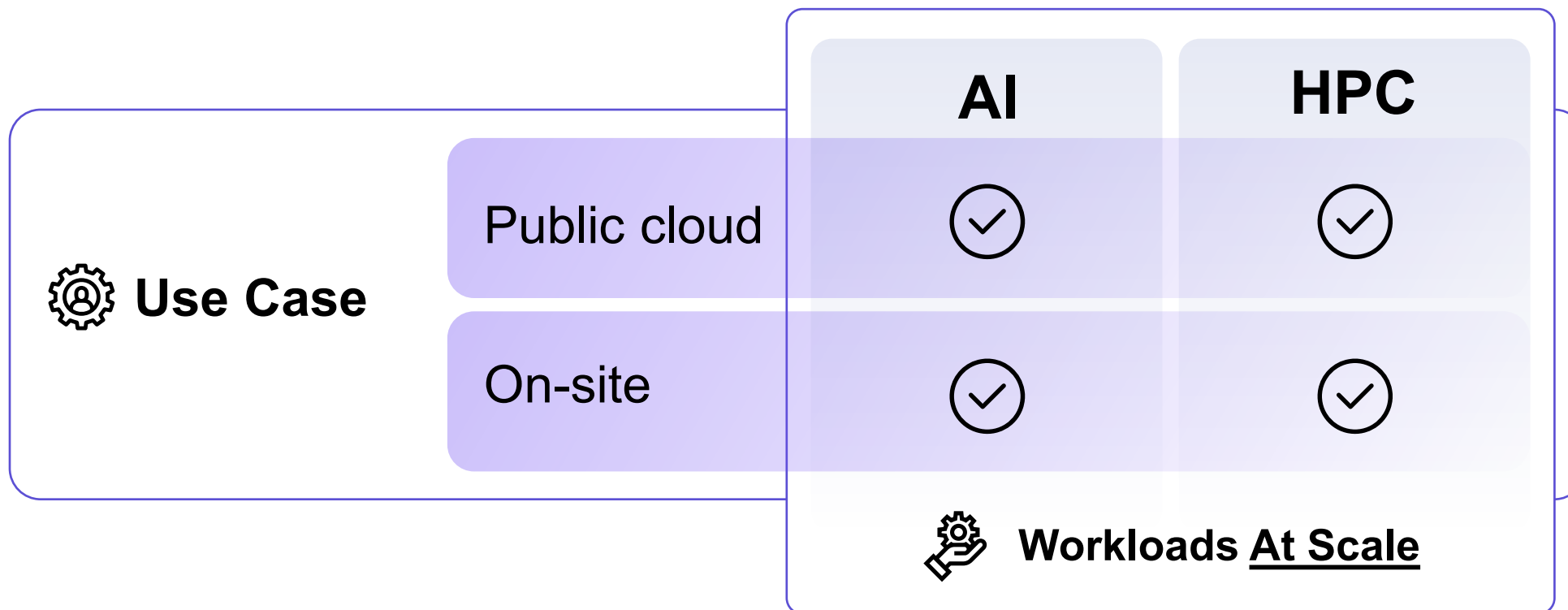
As **scalable** as a cloud data
center

2023 Organization

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



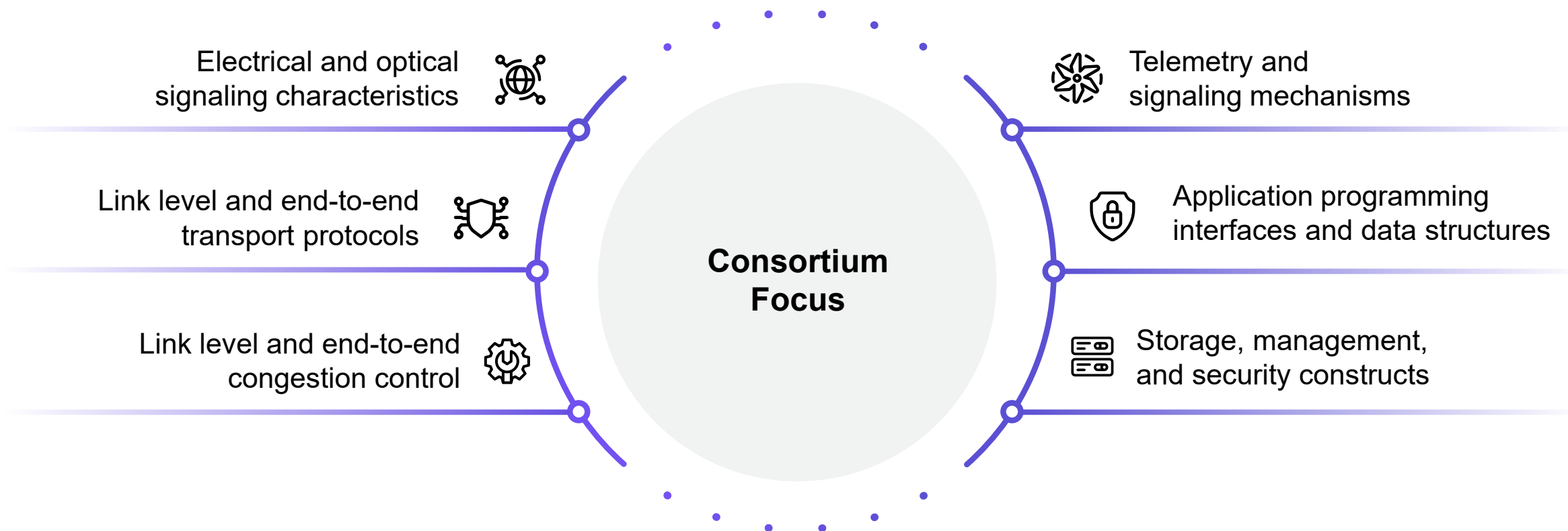
TARGET DEPLOYMENT MODELS / USE CASES



Profiles defined for AI and HPC use cases

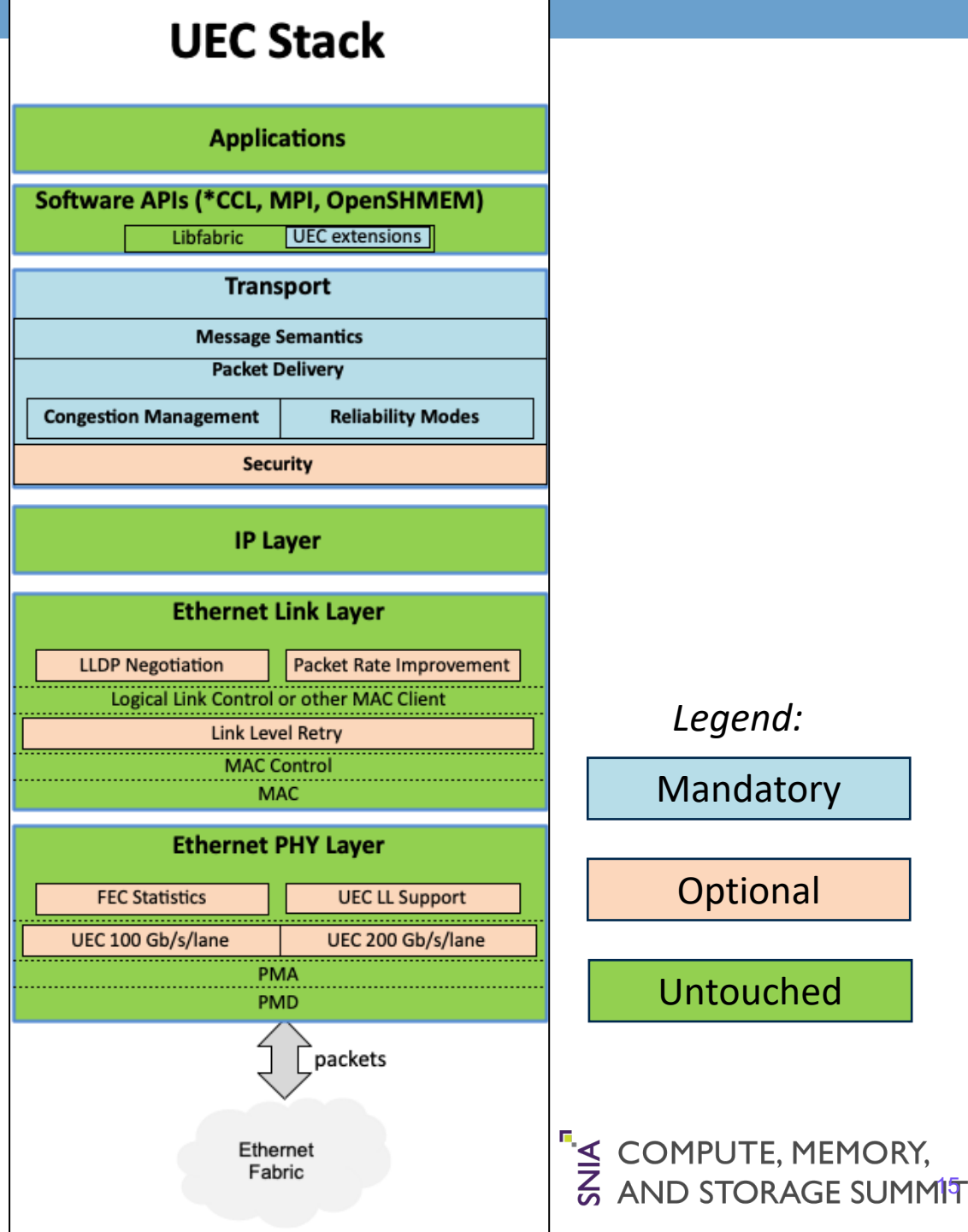
TECHNICAL GOALS

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



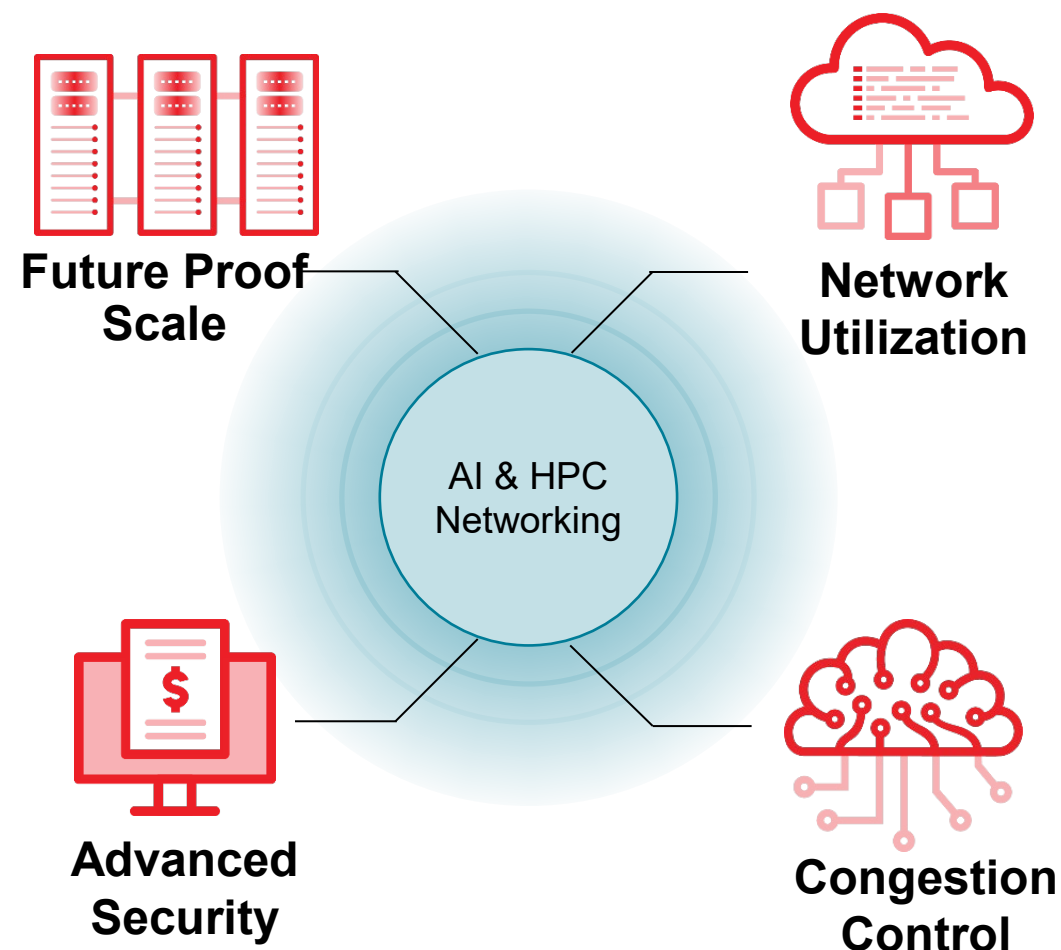
Understanding the UEC Stack

- 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



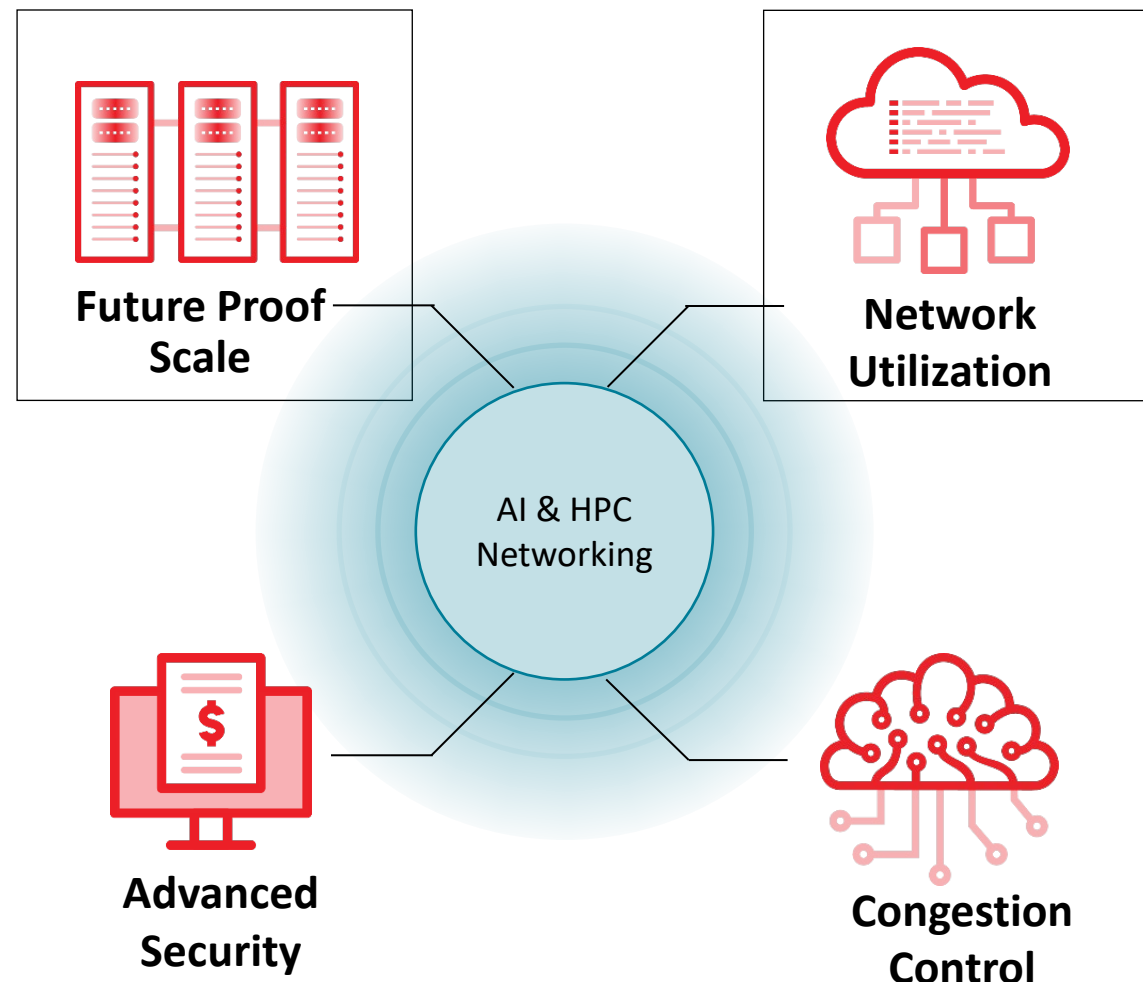
UEC TRANSPORT ADDRESSES GRAND CHALLENGES

- 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



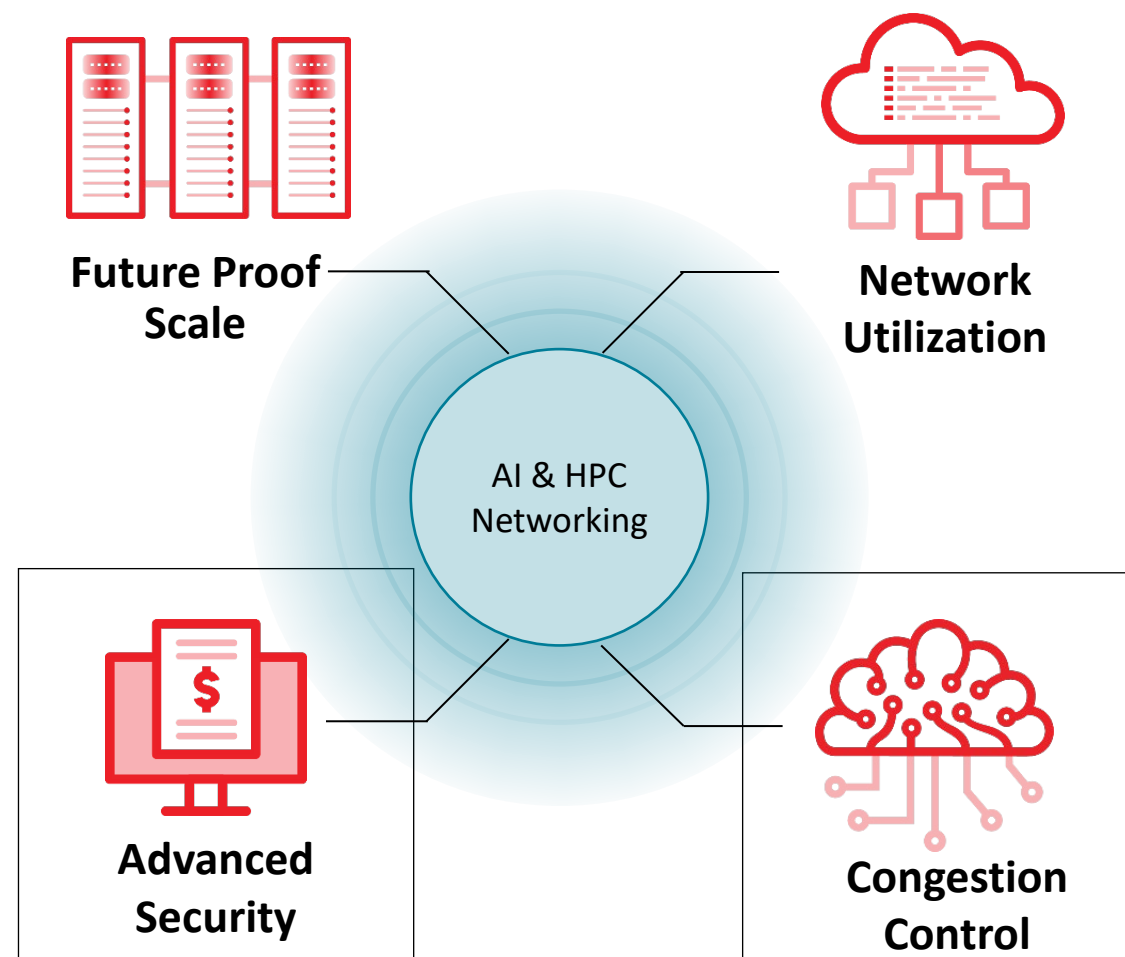
FUTURE PROOF SYSTEM SCALE & NETWORK UTILIZATION

- 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
- 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, CONGESTION CONTROL & TELEMETRY

- **Advanced Security**
 - Encryption support that doesn't balloon the session state in hosts and network interfaces
 - Similar conditions in AI and HPC
- **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	<i>Ultra Ethernet</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



LEARN MORE AT 

www.ultraethernet.org

Please take a moment
to rate this session.

Your feedback is important to us.



SNIA COMPUTE, MEMORY,
AND STORAGE SUMMIT

Solutions, Architectures, and Community
VIRTUAL EVENT, MAY 21-22, 2024