THE LINUX FOUNDATION
OPEN SOURCE SUMMIT
INDIA

53 Years of Ethernet: Evolving With Open Standards for AI Infrastructure

Kapil Mehta
Technical Leader, Cisco Systems

The Network Is More Powerful Than the Node

Chuck Robbins, Cisco Systems

15+ Years of experience in Plan, Design, Implementation & operations of Service Provider & Enterprise networks

Masters in Networks Technology & Management

Speaker @ Cisco Live, Cisco TV , Cx Lighthouse on AI Data Centre Technologies

Kapil Mehta

Technical Leader @ Cisco

Agenda

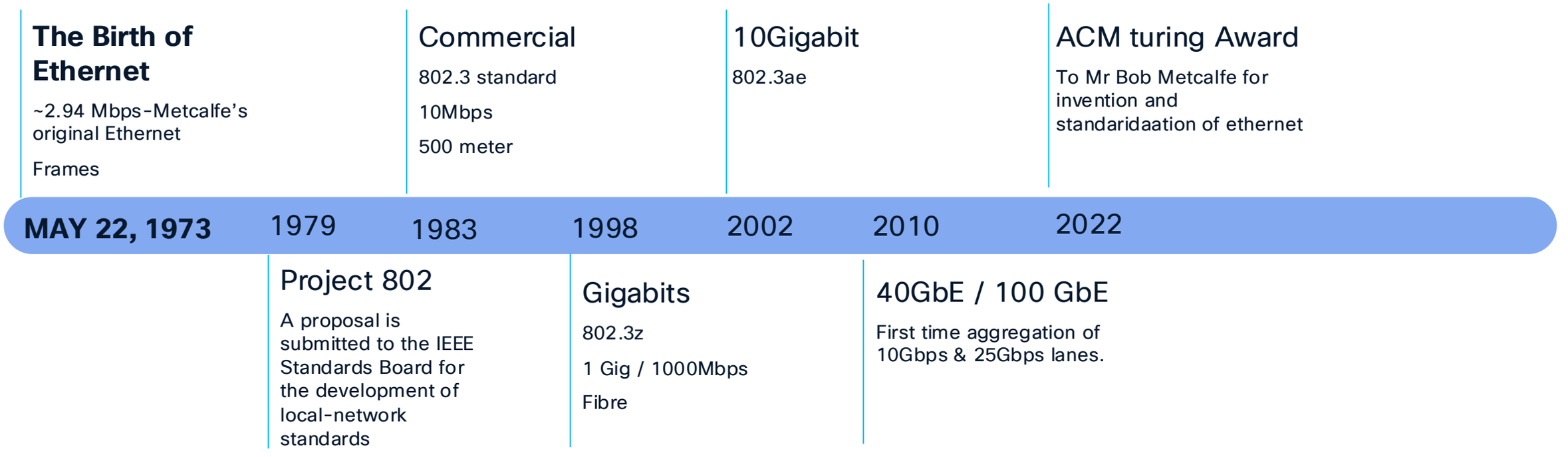
01 Ethernet History

02 AI Infrastructure Challenges

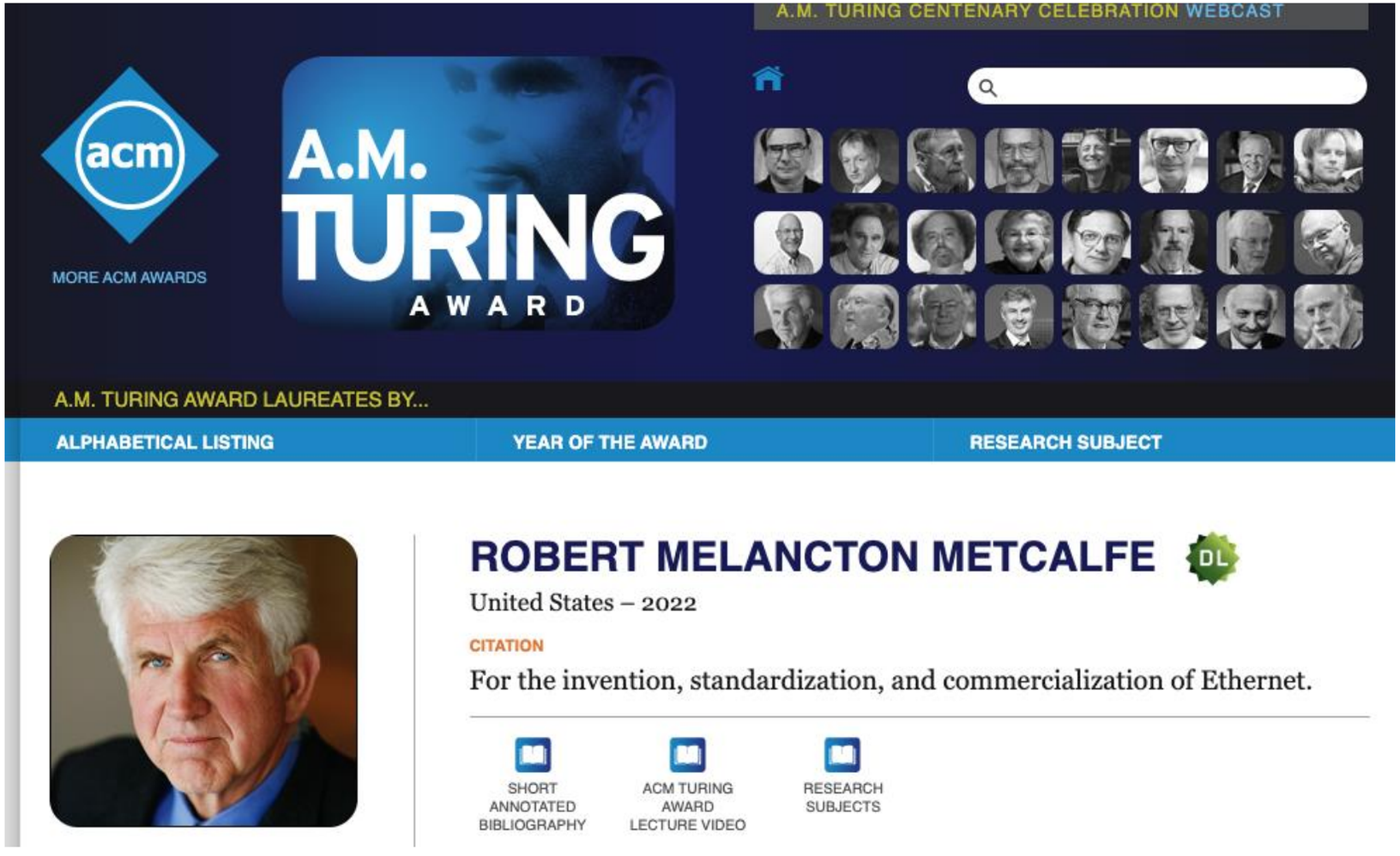
03 What’s Next in Open Standards

Ethernet History

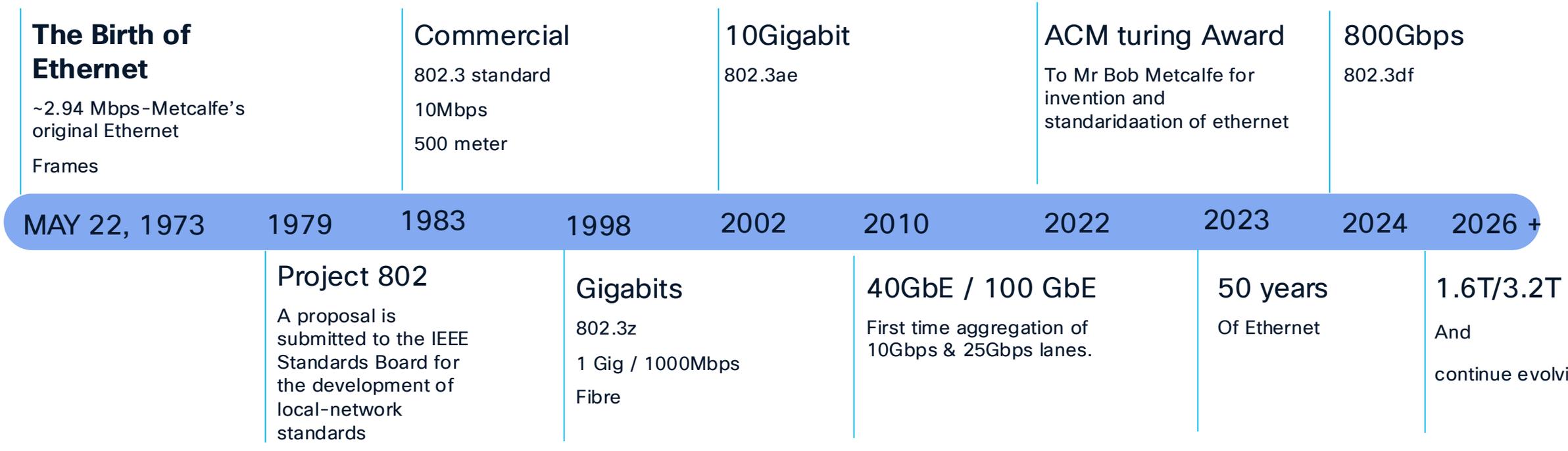
Ethernet Speed Standards Evolution



Source:-ethernet-50th-anniversary



Ethernet Speed Standards Evolution



Source:-ethernet-50th-anniversary

The Power of Ethernet

Mutli Vendor Ecosystem

Management Tools

A lot of tools for testing , measuring , deploying & operating

Competitive ecosystem and economics, helps to reduce opex/capex cost

IEEE Ethernet Standards

Strategic Value of Open Standards

Healthy multi-vendor ecosystem

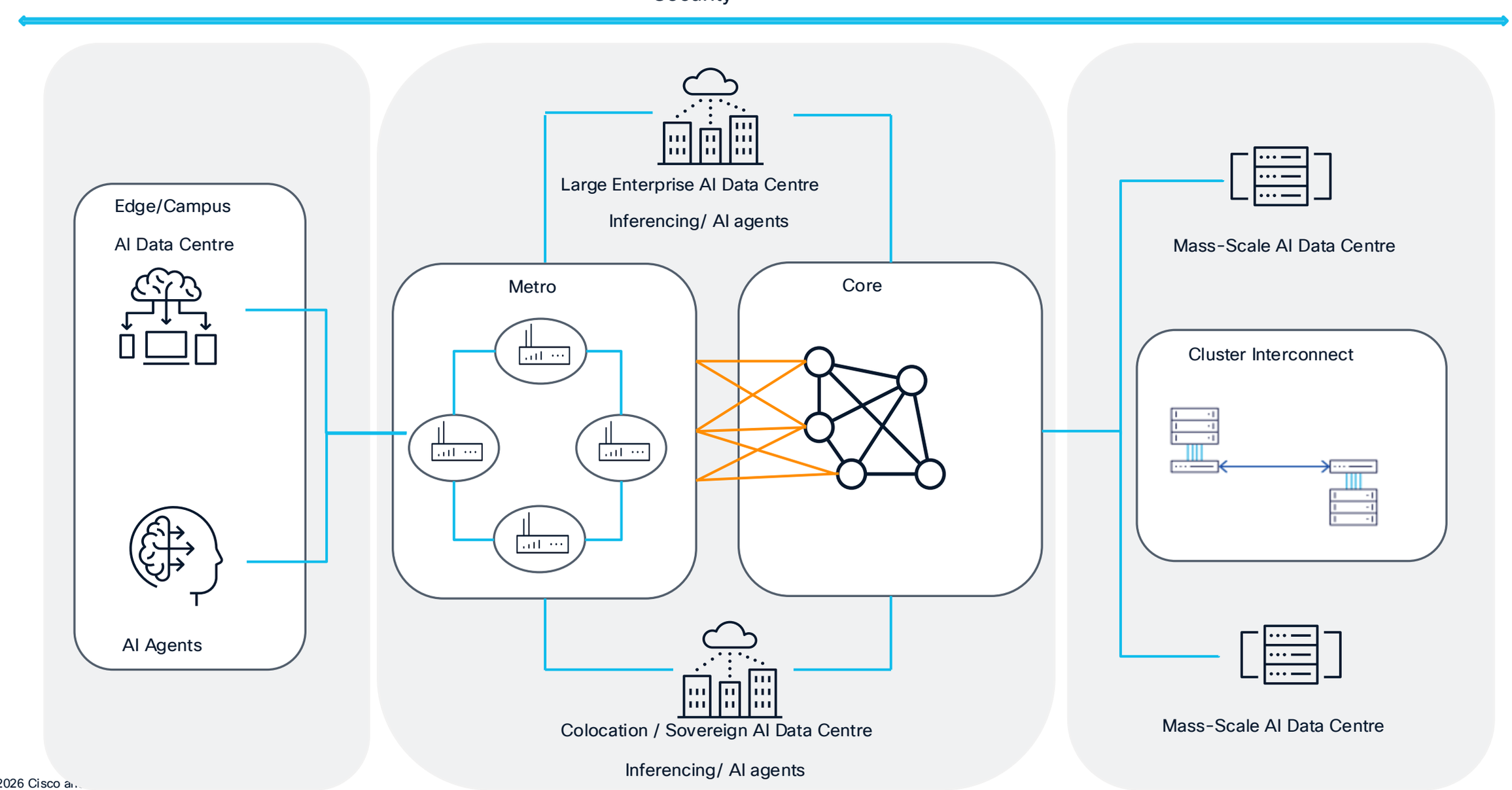
Industry-wide interoperability

Cost efficiency and supply chain benefits

Future-proofing complex AI networks

AI at the Helm:
Driving the Next-Gen Ethernet Network Evolution

Networks for the AI Agentic Age

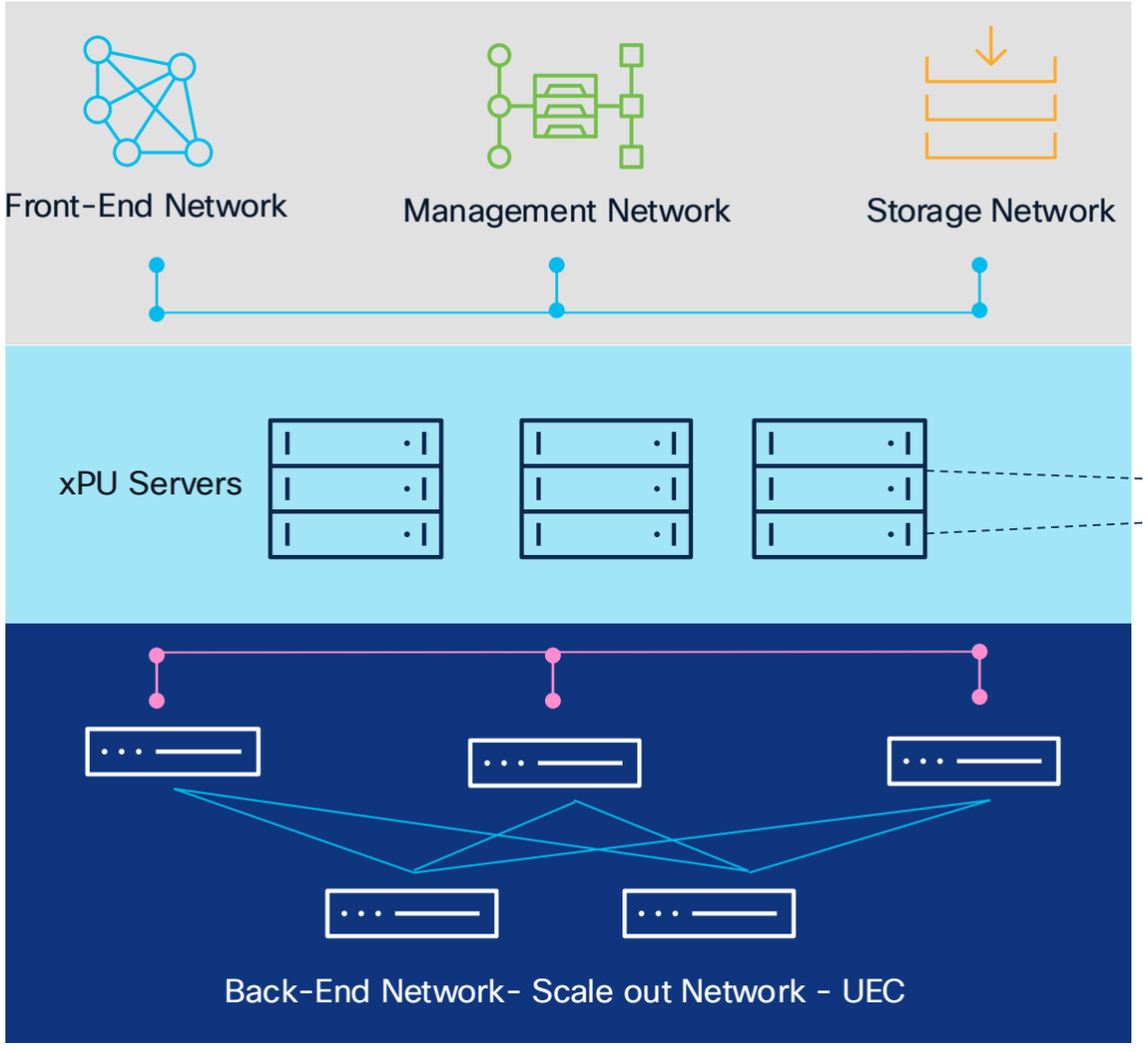


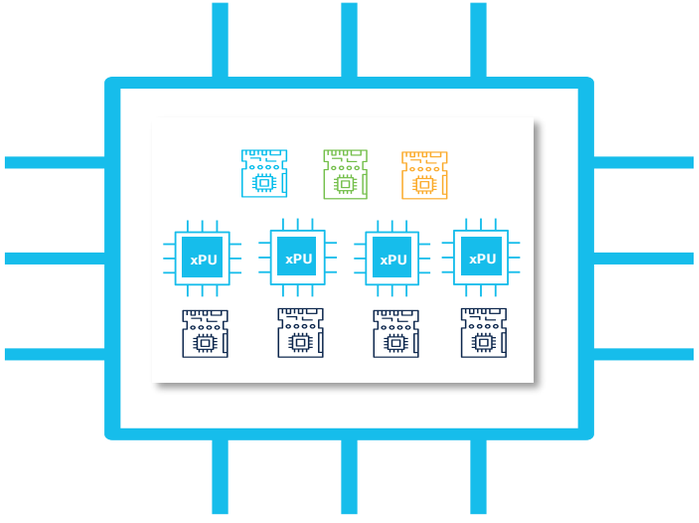
Edge / Campus Connectivity

Inter Data Centre Connectivity

Inter-Cluster Connectivity

Scale-Up & Scale Out Network



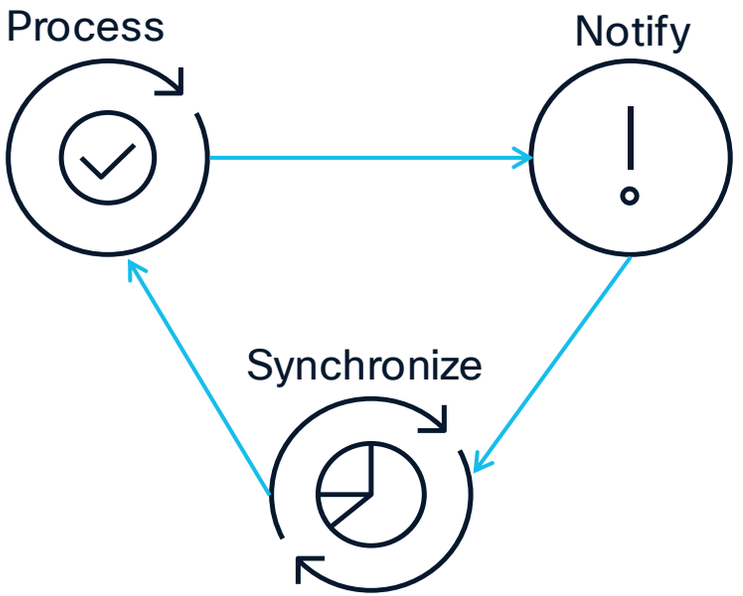


Scale Up- to act as a single xPU , share memory - UA link/ NV link/ESUN

The AI/ML Workload Cycle

xPU Execute Instructions

High Bandwidth capable xPUs
can saturate network links



Send results of computation

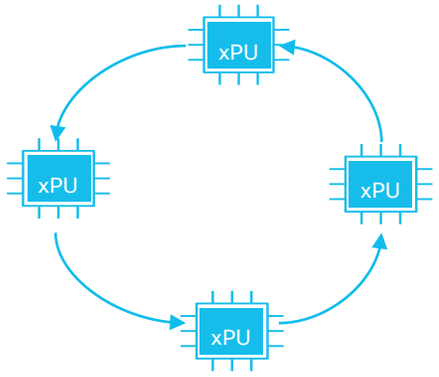
Different collective communication patterns
everyone's data
and send to everyone

Wait for all xPUs to complete

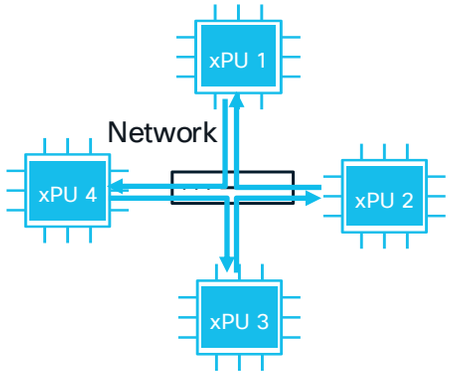
Synchronizes all xPUs
Compute stalls, waiting for the slowest path

Collective Communication via RDMA Operations

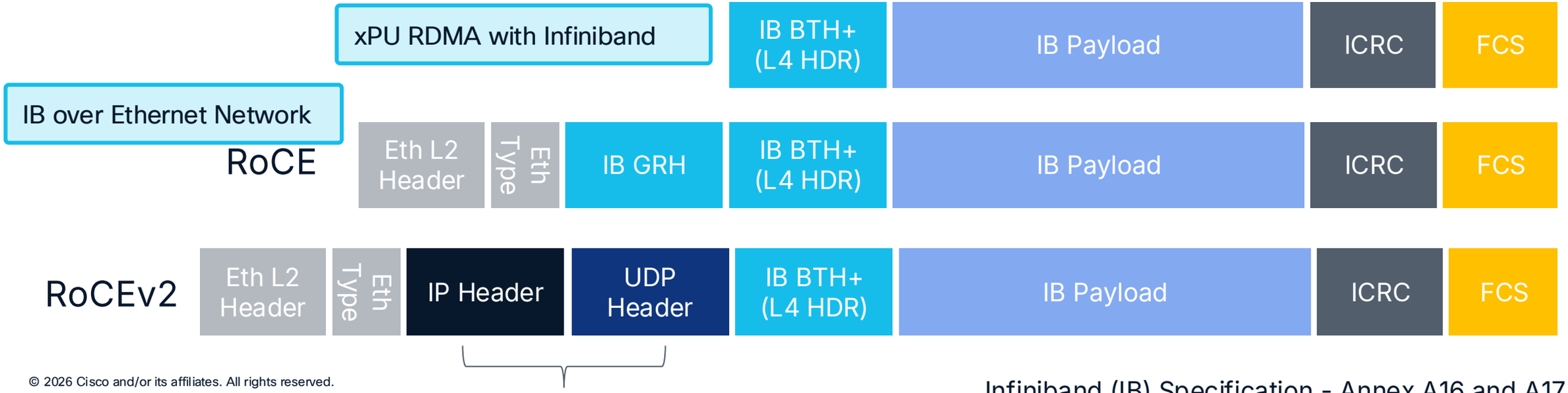
Concept



Real Traffic Flow



Infiniband

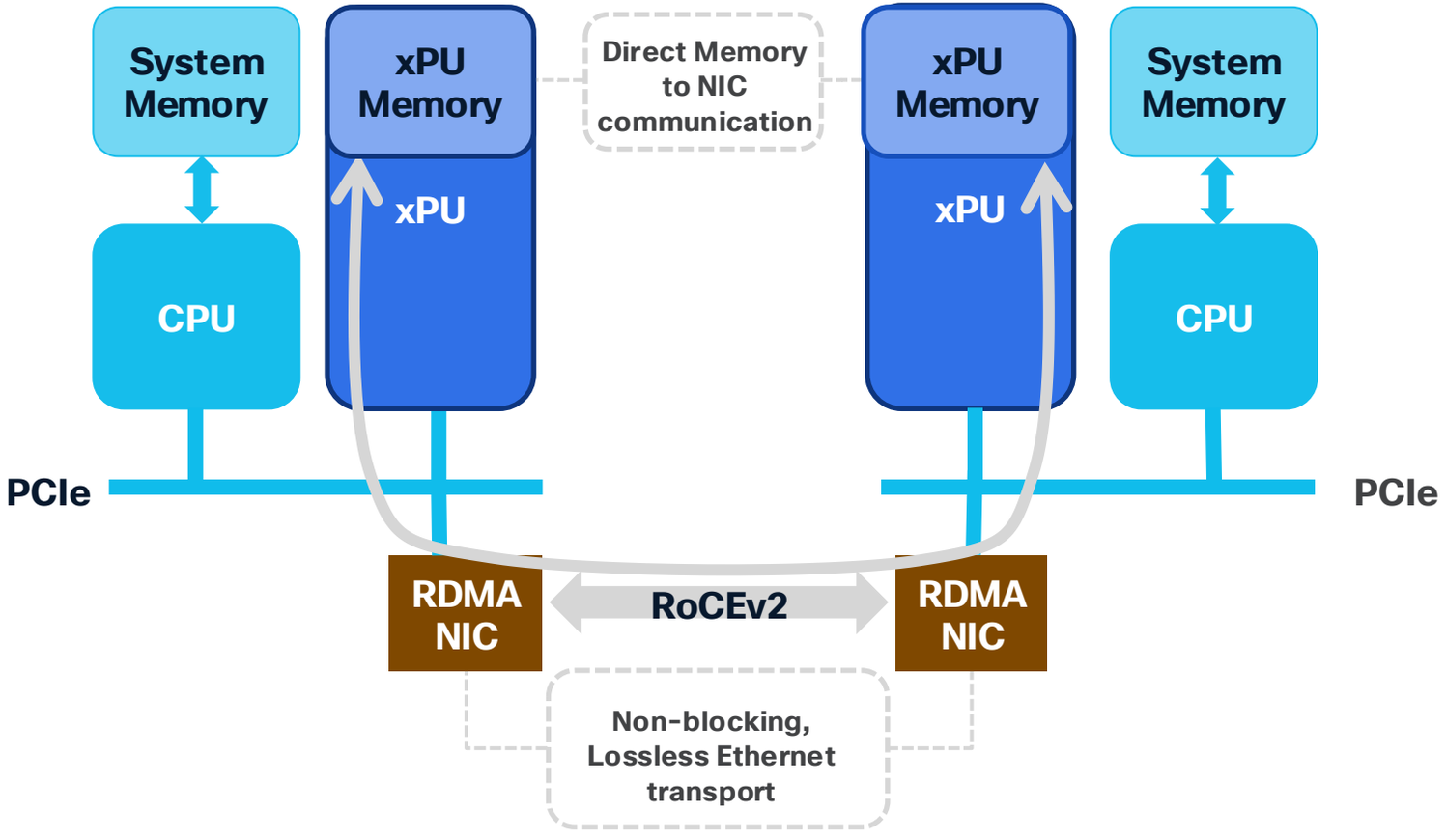


Infiniband (IB) Specification - Annex A16 and A17
https://www.infinibandta.org/ibta-specification/5

Remote Direct Memory Access (RDMA)

over Converged Ethernet

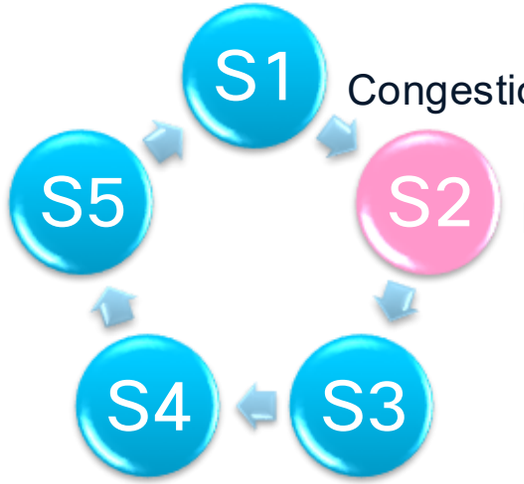
- High throughput
- Low latency transmission
- NIC Driven
- Zero copy
- kernel bypass



Challenges in the AI World

Tail Latency & Goodput Matters

- Slowest link/Node matters for a job completion due to ring or tree collective operations.
- expensive xPU's are sitting idle due to slow node/link.



Congestion / loss on a link between S1/S2

Fail or retransmission

Waiting for S2 to complete

worst-case tail latency can impact Training and inferencing

What Causes Packet Loss

3 C’s

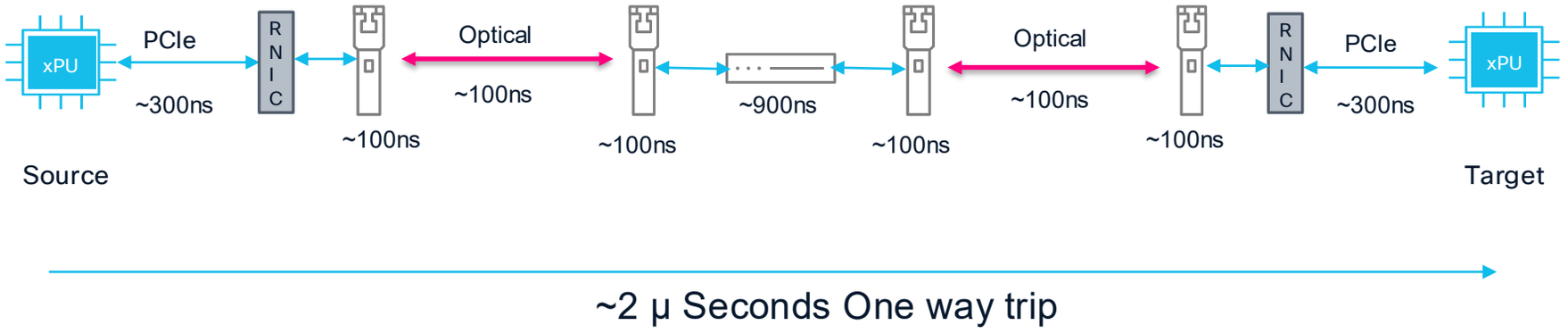
Congestion Drops

Corruption Drops or
Bit errors (CRC)

Configuration Drops

Latency from one xPU-xPU

Every network component introduces processing latency as it handles packets



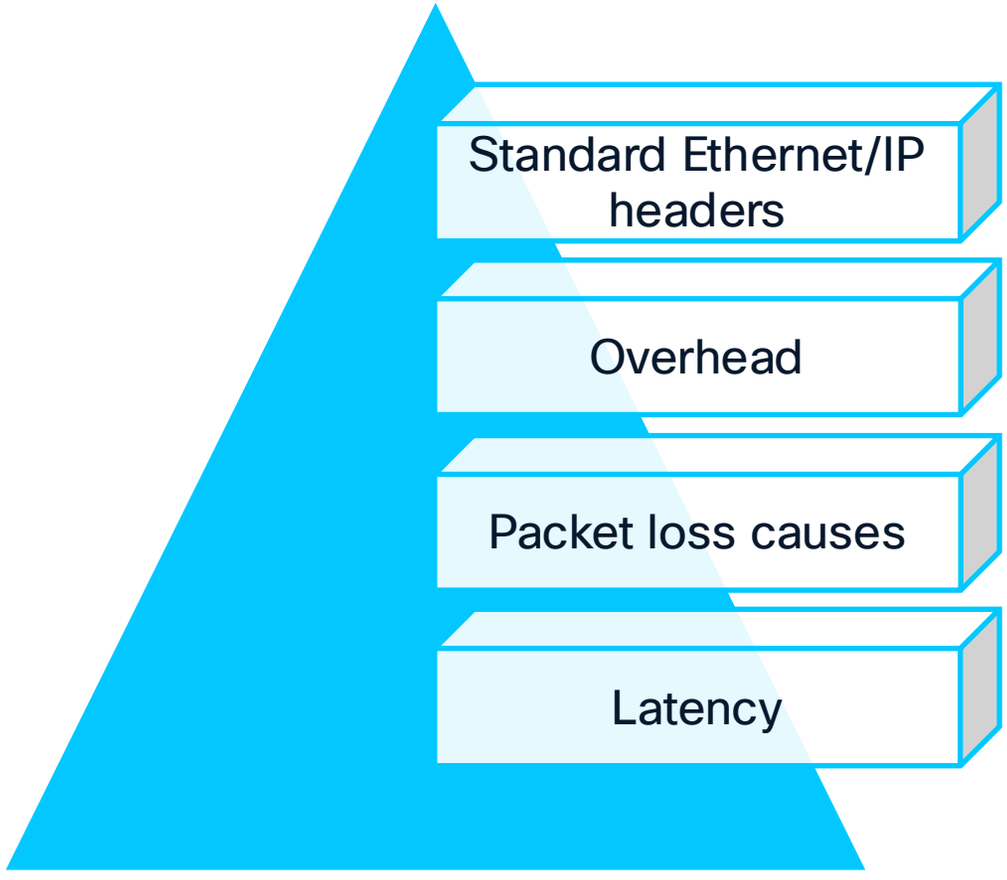
Network Latency Considerations

~5 ns/m one-way propagation delay for fiber optics (varies based on cable and SFP type)

Latency can increases due to:

- Longer cable lengths
- More complex network topologies / increased number of intermediate devices

Limitations of Standard Ethernet



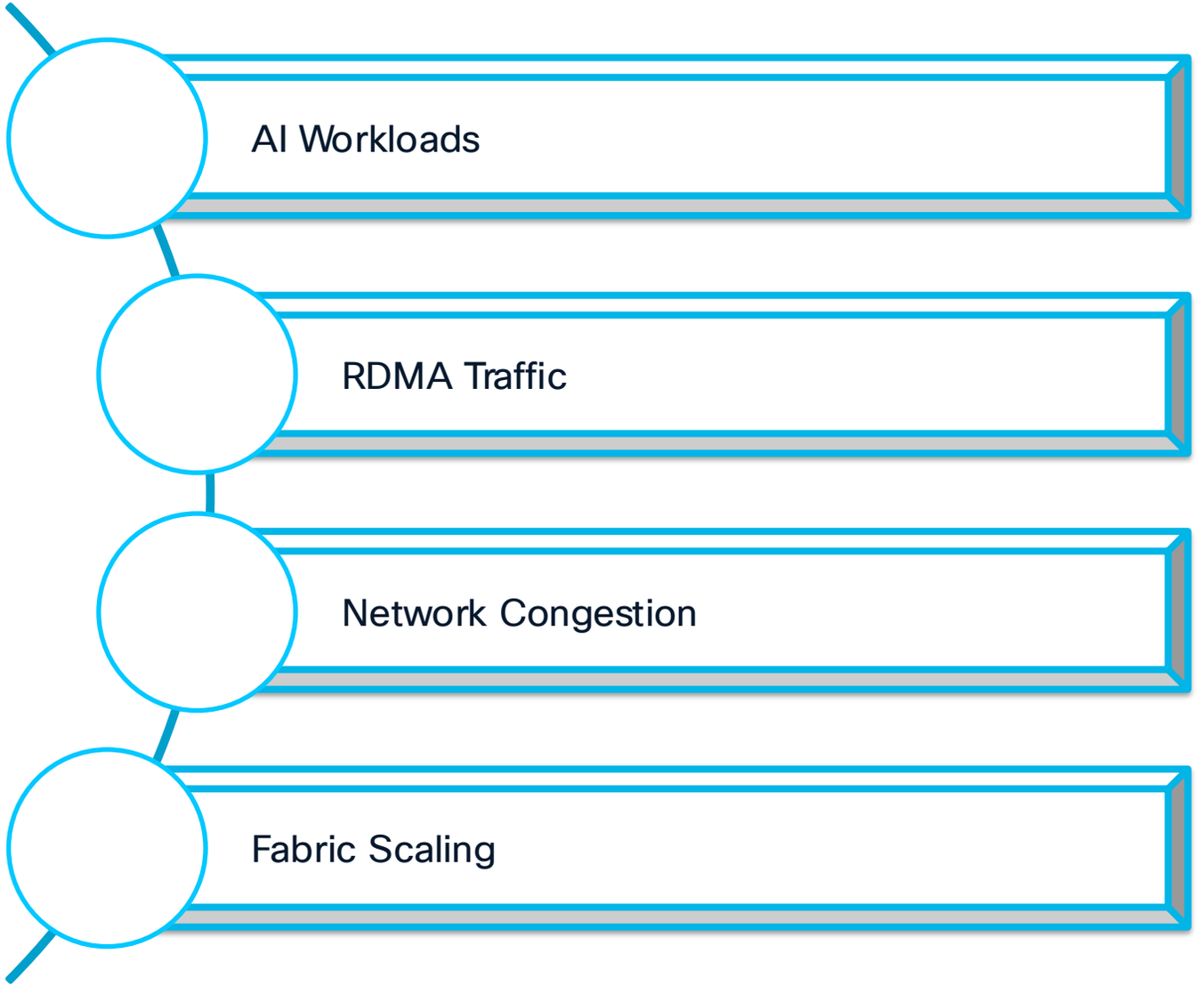
Standard Ethernet/IP headers designed for general networking, Best Effort Delivery, inefficient for AI fabrics

Overhead from IP headers and protocols like TCP/UDP, Ipsec

Packet loss causes: congestion, corruption, configuration errors

Latency introduced by network components and protocols

AI Era Networking Challenges



AI workloads require lossless, ultra-low latency communication

RDMA-driven traffic patterns stressing networks

Network congestion and tail latency impact job completion time

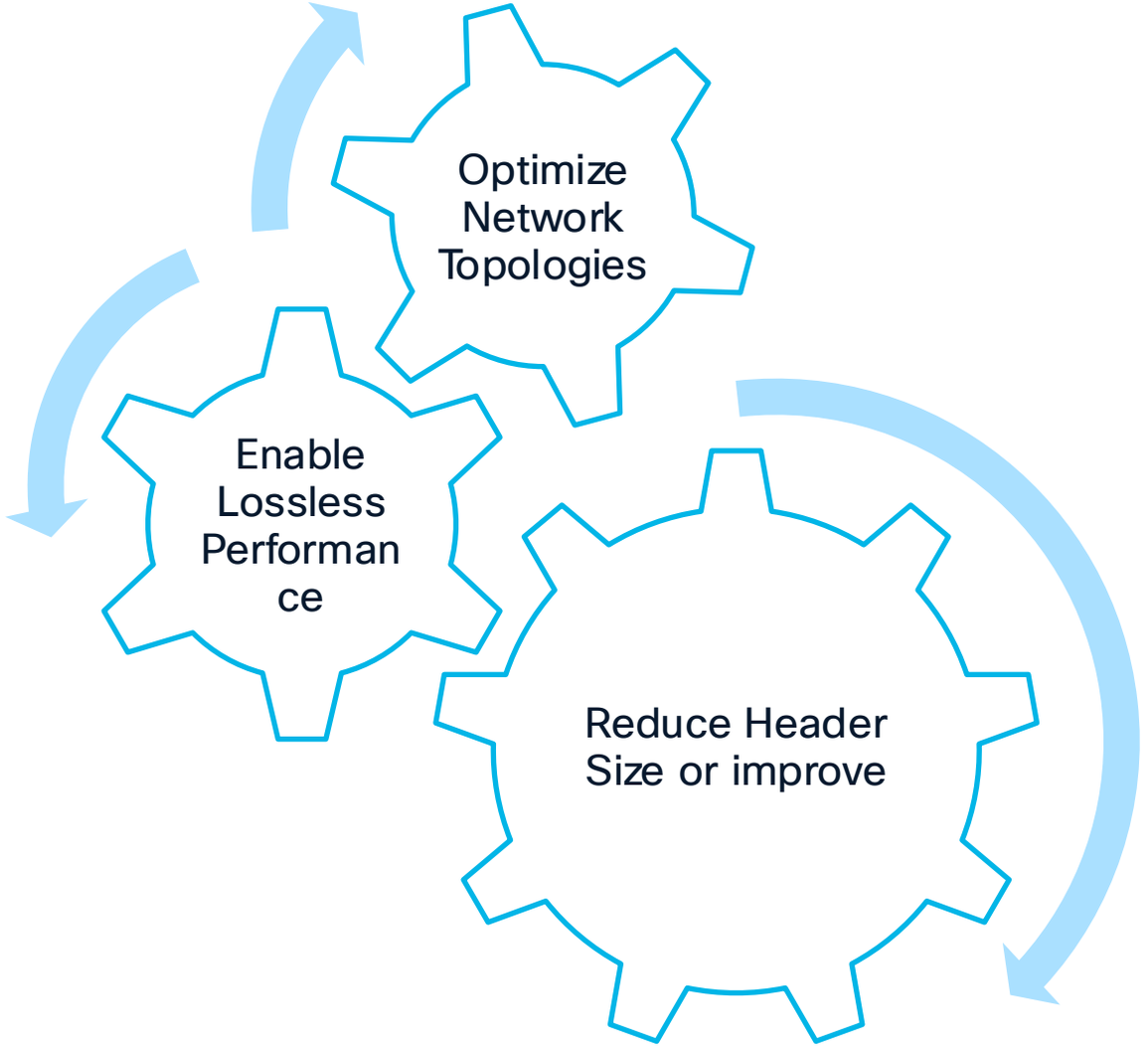
Scale-up and scale-out fabrics with tightly coupled xPUs/xPUs

Ethernet and InfiniBand Trade-offs

Category	Ethernet	Infiniband
Speed	Fast, best effort packet delivery	Faster than ethernet , native lossless
Vendors	Multiple vendors, competitive ecosystem and economies of scale.	Single vendor.
Flexibility	High interoperability (switches, NICs, cables, optics, software).	No interoperability with other vendors.
Adoption	Proven, wide adoption across diverse data centre networks.	Niche, developed primarily for HPC network use.
Cost & Supply Chain	Lower costs, multiple vendor offerings.	Higher costs and longer lead times due to single vendor.
Standardization	IEEE standards, evolves regularly across physical/optical layers.	Infiniband Trade Association, limited membership.

What’s Next

The Challenge: Standard Ethernet in AI



The Problem

Standard Ethernet/IP headers were designed for general-purpose networking and are often inefficient for the specific needs of massive AI fabrics.

The Solution

Implementation of custom header overlays to:

- **Reduce header size or improve**
- **Enable lossless fabric performance**
- **Optimize network topologies for AI workloads**

Ethernet for AI Networks: Evolution

(partial list)

IEEE 802.3

IEEE P802.3dj is writing the 200G/lane standard, target September 2026

Ultra Ethernet Consortium(UEC)

Scale-Out Ethernet networks – UE 1.0.1 specification published June 2025

Ultra Accelerator Link™ (UAL)

Scale-up XPU to XPU – UALink 1.0 (released April 2025) 200G/lane for 1K XPUs in a pod

ESUN Open Compute Project(OCP)

Ethernet Scale-up Networking (ESUN) workstream (First Spec Released Feb 2026)

Many Consortia are Relying on Ethernet

To support the breadth of AI applications, multiple organizations exist:

Ultra Ethernet Consortium

Deliver an Ethernet based open, interoperable, high performance, full-communications stack architecture

ULTRA ACCELERATOR LINK™

An open, high-bandwidth and low latency interconnect for connecting AI accelerators (such as GPUs) and switches

OPEN Compute Project®

SUE and ESUN build on Ethernet by introducing evolutionary enhancements that improve throughput, latency, and reliability to meet the demands of Scale-Up GPU connectivity.

All are looking for IEEE 802.3 to define the Ethernet Physical Layer Specifications

Ultra Ethernet

The New Era Needs a New Network

As performant as a supercomputing interconnect

As ubiquitous and cost-effective as Ethernet

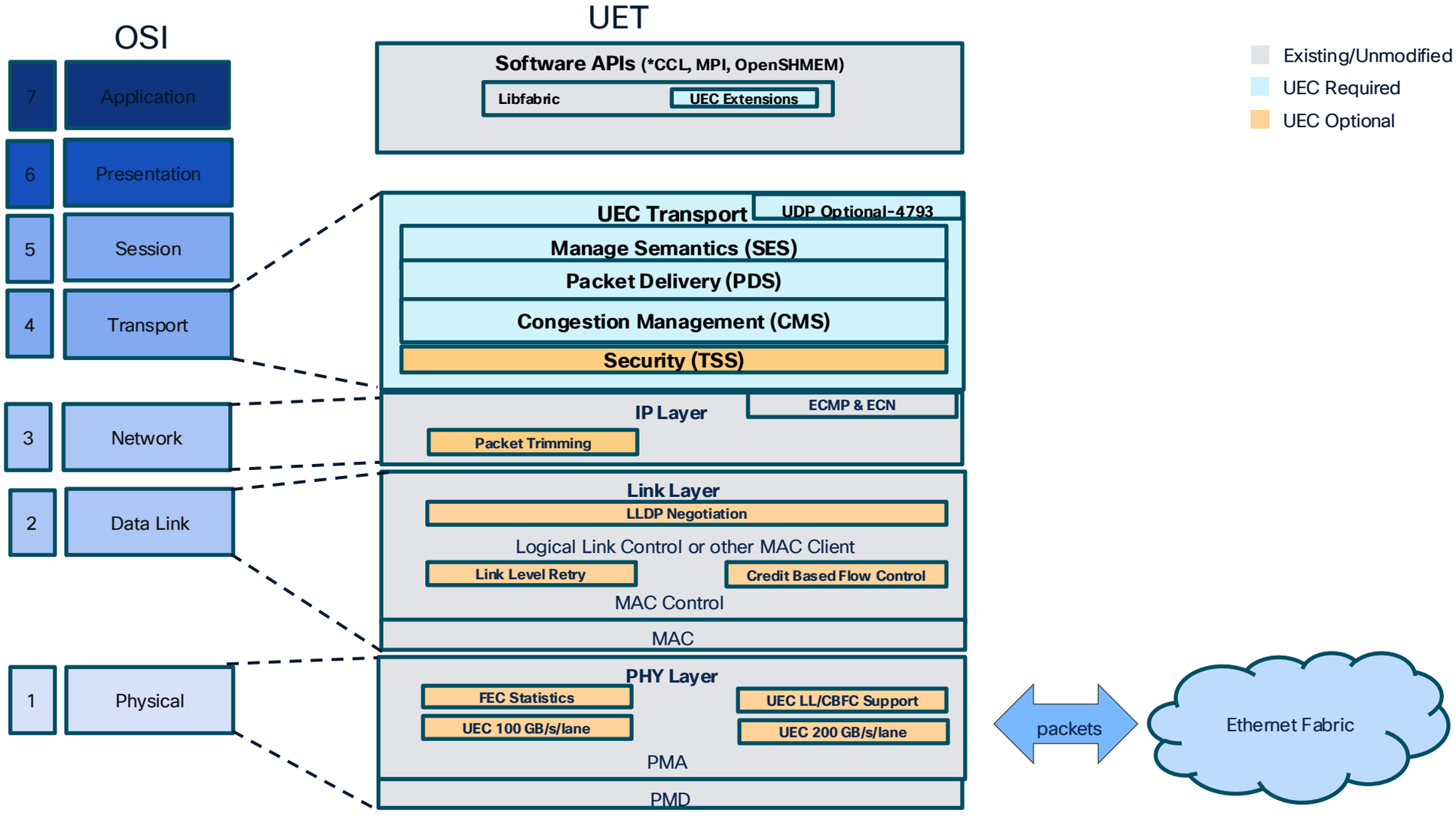
As scalable as a cloud data center

Our Mission

Deliver an Ethernet based open, interoperable, high performance, full-communications stack architecture to meet the growing network demands of AI & HPC at scale

LATEST NEWS | DOWNLOAD 1.0.2 SPECIFICATION | READ 1.0 WHITE PAPER

Ultra Ethernet – Specification 1.0 Overview for Scale Out Networks



Ultra Ethernet Advantage over Legacy RDMA- Summary

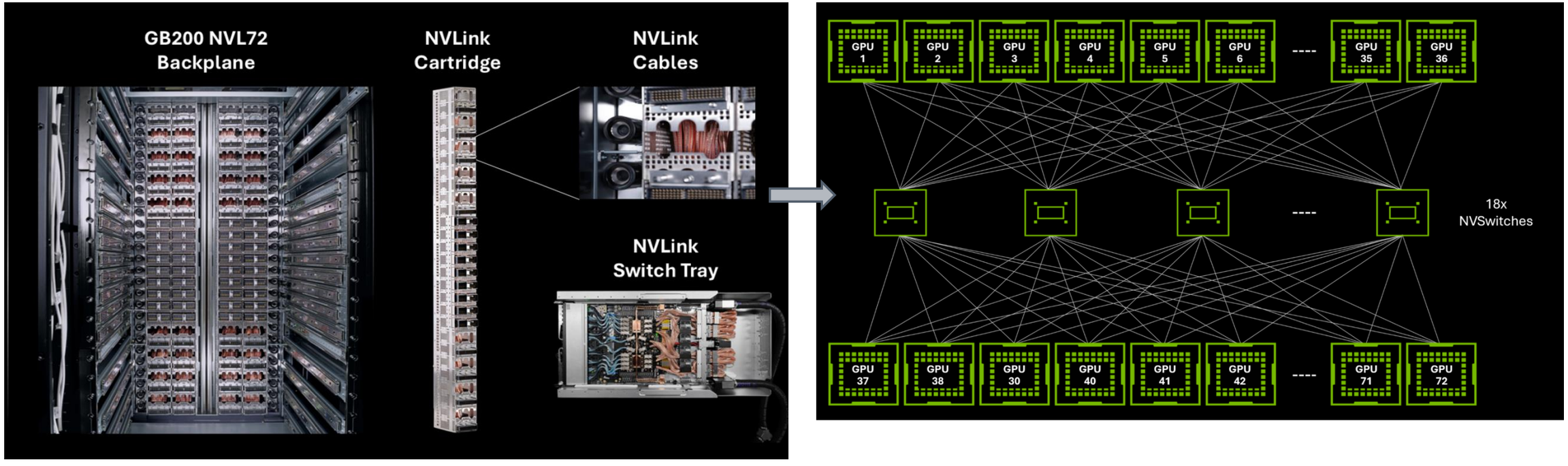
Requirement	UEC Transport	Legacy RDMA	UEC Advantage
Multi-pathing	Packet-spraying	Flow-level multipathing	Higher network utilization
Flexible ordering	Out-of-order packet delivery with in-order message delivery	N/A	Matches application requirements, lower tail latency
AI and HPC Congestion Control	Workload-optimized, configuration-free, lower latency, programmable	DCQCN; configuration required, brittle, signaling requires additional round trip	Incast reduction, faster response, future-proofing
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 endpoints	Current and future-proof scale

UALink: Ultra Accelerator Link

High-Performance Interconnect for AI-Scale Up Systems

Scale Up Example: NVIDIA NVL72 Rack Designs

Grace Blackwell (GB), Vera Rubin (VR)



Source: https://developer.nvidia.com/blog/nvidia-contributes-nvidia-gb200-nvl72-designs-to-open-compute-project/

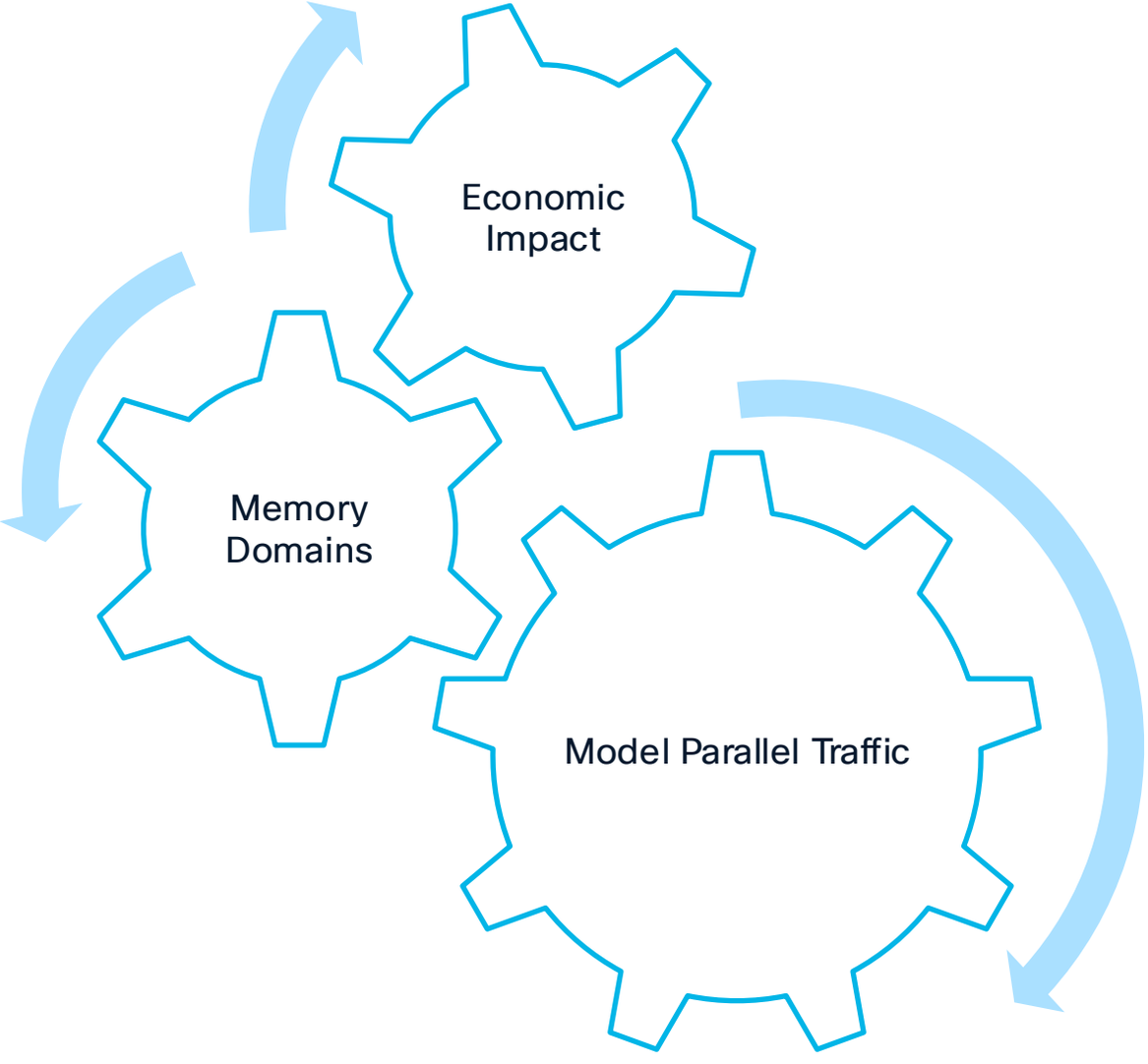
Trends Driving Scale-Up Networks

Model Parallel Traffic
High bandwidth/low latency needed for smaller message sizes

Memory Domains
Treat all xPUs in a cluster as a single, large memory domain

Economic Impact
Scaling issues impact training uptime

Performance Loss
20% bandwidth drop in 128-accelerator domain causes 5% performance loss



Introduction to UALink

Unified Access Link

UALink serves as the foundational connectivity layer for next-generation scale-up architecture. It streamlines data flow between edge devices and core infrastructure, ensuring low-latency performance and robust security protocols.

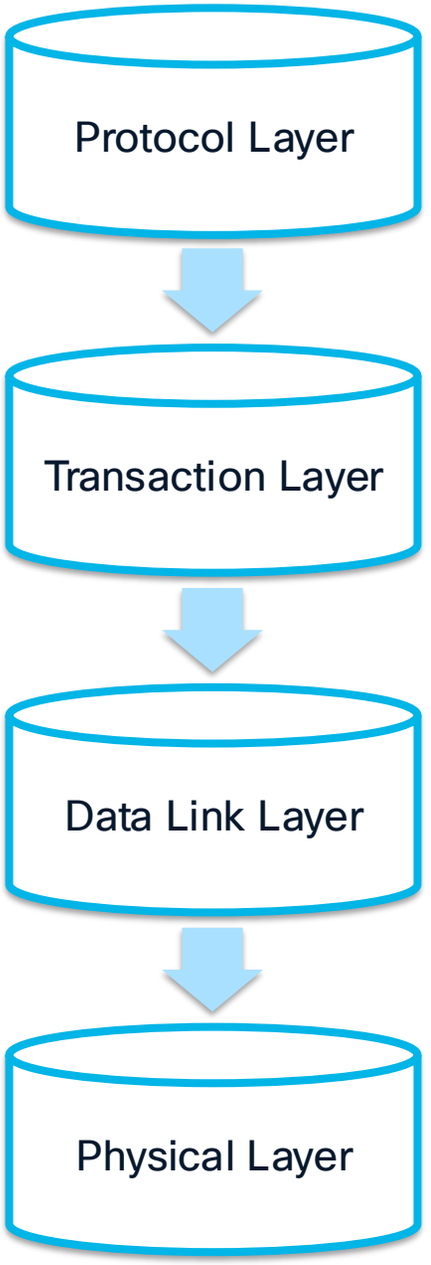
UALink Protocol Layers

Protocol Layer (UPLI)
logical signaling interface

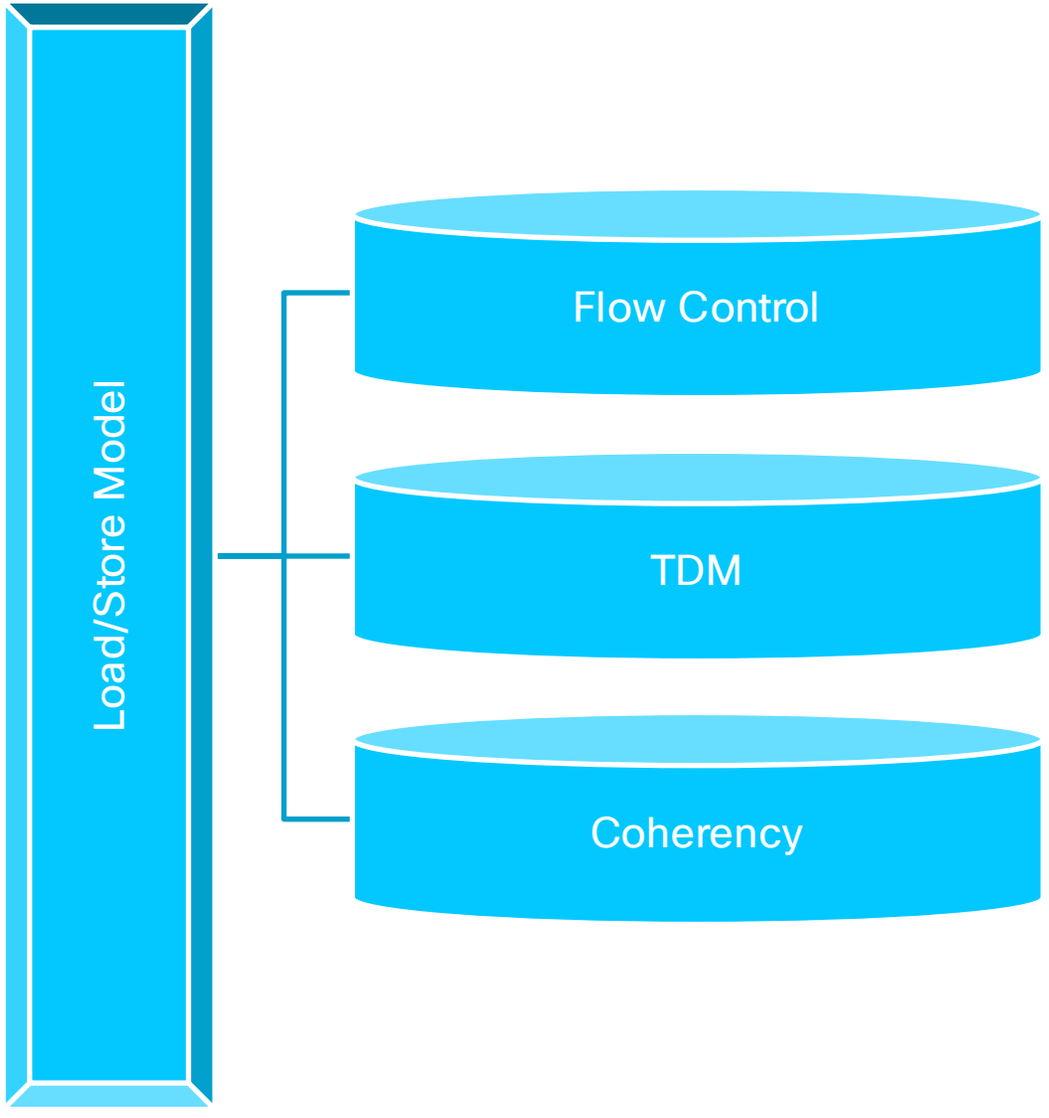
Transaction Layer
flit packaging/unpacking

Data Link Layer
framing, CRC, UART services

Physical Layer
IEEE 802.3dj-based high-speed serial interface



UPLI Interface & Operation



Load/Store Model
Supports direct read, write, and atomic transactions.

Flow Control
Credit-based mechanism to ensure lossless delivery.

TDM
Time Division Multiplexing used to manage multiple ports within a station.

Coherency
Hardware coherency is handled within the System Node; UALink focuses on software-managed coherence.

Introduction to ESUN

What is OCP ESUN?

Ethernet for Scale-Up Networks (ESUN) group represents the next generation of high-performance interconnects for scale up AI infrastructure announced in Oct 2025 at Open Compute Summit , OCP ESUN 1.0 Spec released Feb 2026.

Definition

An evolution of standard Ethernet designed to provide the low-latency, reliable high-bandwidth requirements of tightly coupled AI training and inference workloads for Scale-Up GPU connectivity (1K+ GPUs)

Objective

To improve goodput, latency, and reliability for tightly coupled GPU domains by optimizing the transport layer for scale-up topologies.

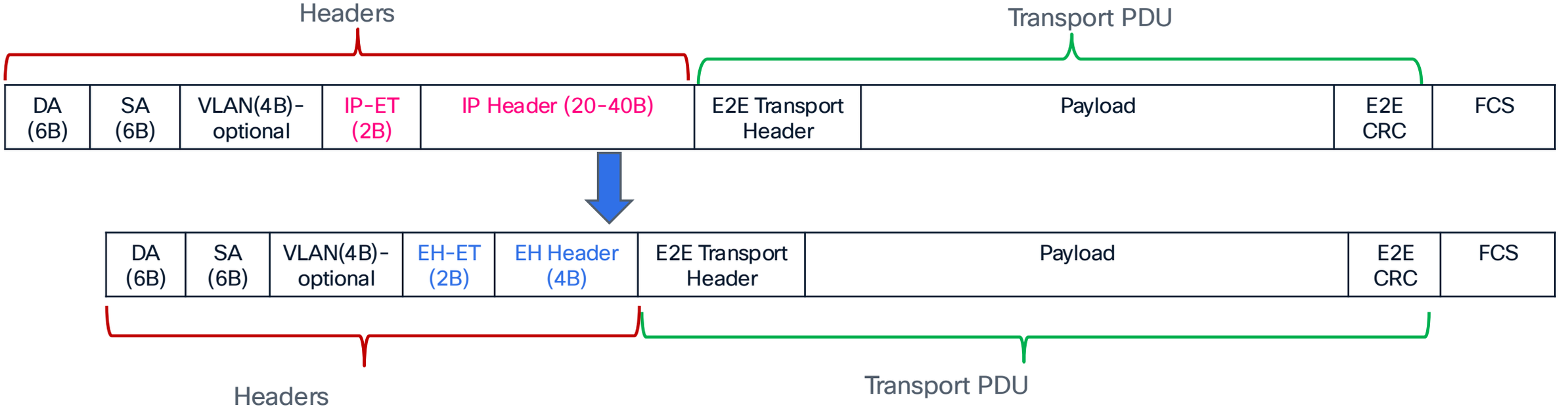
Core Strategy

Introduces the ESUN Header (EH) to optimize headers for smaller, high-performance topologies, reducing overhead and increasing effective throughput.

What's needed for Scale-Up (cont.) ?

Improve small packet performance

Replacing IP header with new 4B ESUN Header(EH)



EH Fields

Rev – Revision Number
F - Flow Label Valid
EH-CoS – Class of Service
EH-ECN – Explicit Congestion Notification
Flow label – load balancing entropy
TTL - Time To Live
UD - User Defined – end to end signaling

Rev (2b)	F	EH-CoS (3b)	EH-ECN (2b)	Flow label (16b)	TTL (4b)	UD (2b)	RSVD (2b)

Compare between Scale up solutions

Feature Category	UALink Specification	OCP ESUN Specification
Core Architecture	Memory-semantic (Load/Store/Atomics)	Message-passing (Ethernet based)
Data Structure	Fixed-size flits (64B TL, 640B DL)	Variable-size Ethernet frames
Network Overhead	Header compression via Tx/Rx address caching	4-byte ESUN Header replacing IP/UDP stack
Hardware Compatibility	Requires dedicated UALink Switches (ULS)	Fully compatible with commercial L2/L3 silicon
Flow Control	Credit-based backoff mechanism	Credit-Based Flow Control (CBFC) & PFC
Error Recovery	Link-Level Replay (LLR) via Data Link Layer	Link Layer Retry (LLR) at MAC layer
Collectives Offload	Native In-Network Compute (INC) in UALink 2.0	Managed via endpoint software
Target Scale	up to ~1k accelerators	up to ~1k accelerators

Protocol Comparison: UALink, ESUN, UET, Ethernet

Feature Category	UALink	ESUN	UET	Legacy Ethernet
Transport Model	Memory-semantic	Message-passing	RDMA optimized	General purpose
Header Efficiency	Dedicated header	4B ESUN header	Overlay headers	IP + Ethernet
Flow Control	Credit-based	PFC + CBFC	Programmable congestion	PFC
Reliability	Link-level retry	Link-level retry	Advanced reliability Link Level retry optional	Basic
Scale	Up to 1K XPUs	Up to 1K XPUs	Millions of endpoints	General

What’s next after Ethernet is Ethernet

Bob Metcalfe, Co-inventor of Ethernet

References

IEEE 802.3 working group :- https://www.ieee802.org/3/

OCP ESUN:- https://www.opencompute.org/wiki/Networking/ESUN

Ultra Ethernet :- https://ultraethernet.org/

UA link:- https://ualinkconsortium.org/

Videos :- https://www.youtube.com/@OpencomputeOrg

THE LINUX FOUNDATION

OPEN SOURCE SUMMIT INDIA