UALink: Why Now?

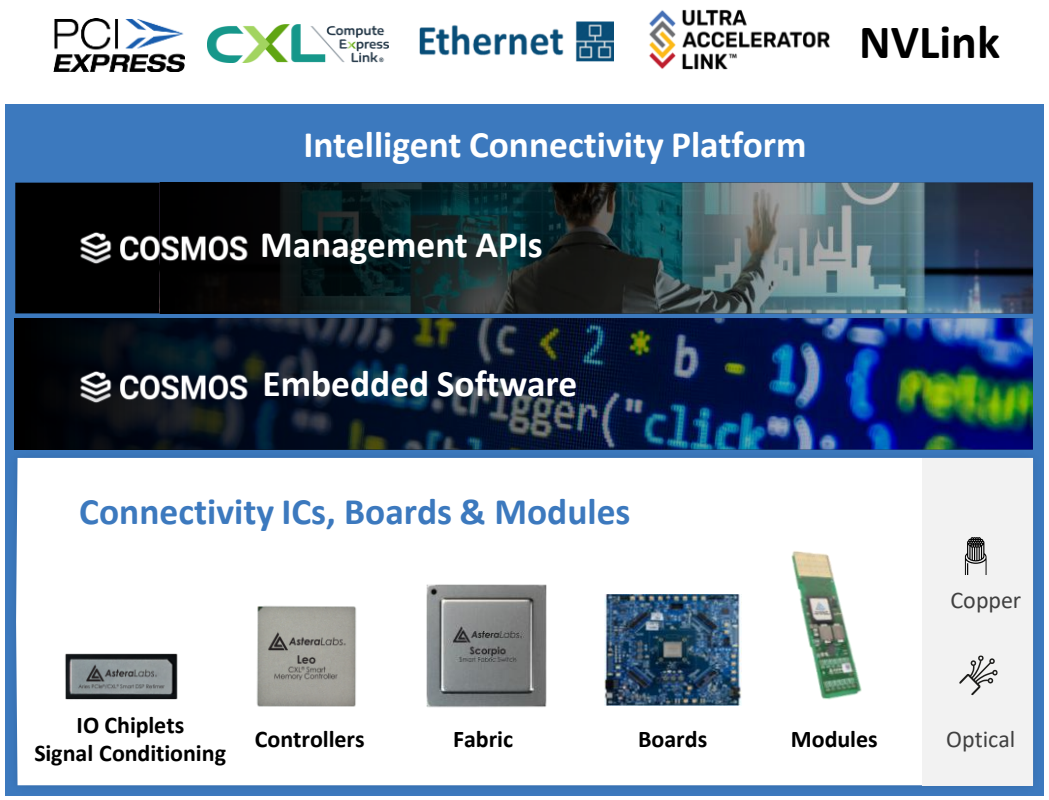
UALink Emerges as a Leading Standard for Scale Up Architectures

Mike Hendricks
AVP Solutions Marketing & Ecosystem Partnerships

FMS – August 7, 2025

the Future of Memory and Storage

Astera Labs & Speaker Introduction



- Speaker: Mike Hendricks
- Role: AVP Solutions & Ecosystem
- Previous companies:
  - Amazon
  - Intel / Altera (FPGA)
  - Texas Instruments / National Semi
- Previous Roles:
  - GM / Business Unit Lead
  - Business Development
  - Product Marketing / Management

Why Now for an Open Scale-up Connectivity Solution?

General Compute Servers

Source: HPE

- 2 to 8 CPUs scaled-up
- In-the-box
- Proprietary, CPU-to-CPU
- Simple memory semantics

AI Servers

Source: AMD

- 8 GPUs scaled-up
- In-the-box
- Proprietary, GPU-to-GPU
- Simple memory semantics

AI Server → Rack (aka Pod)

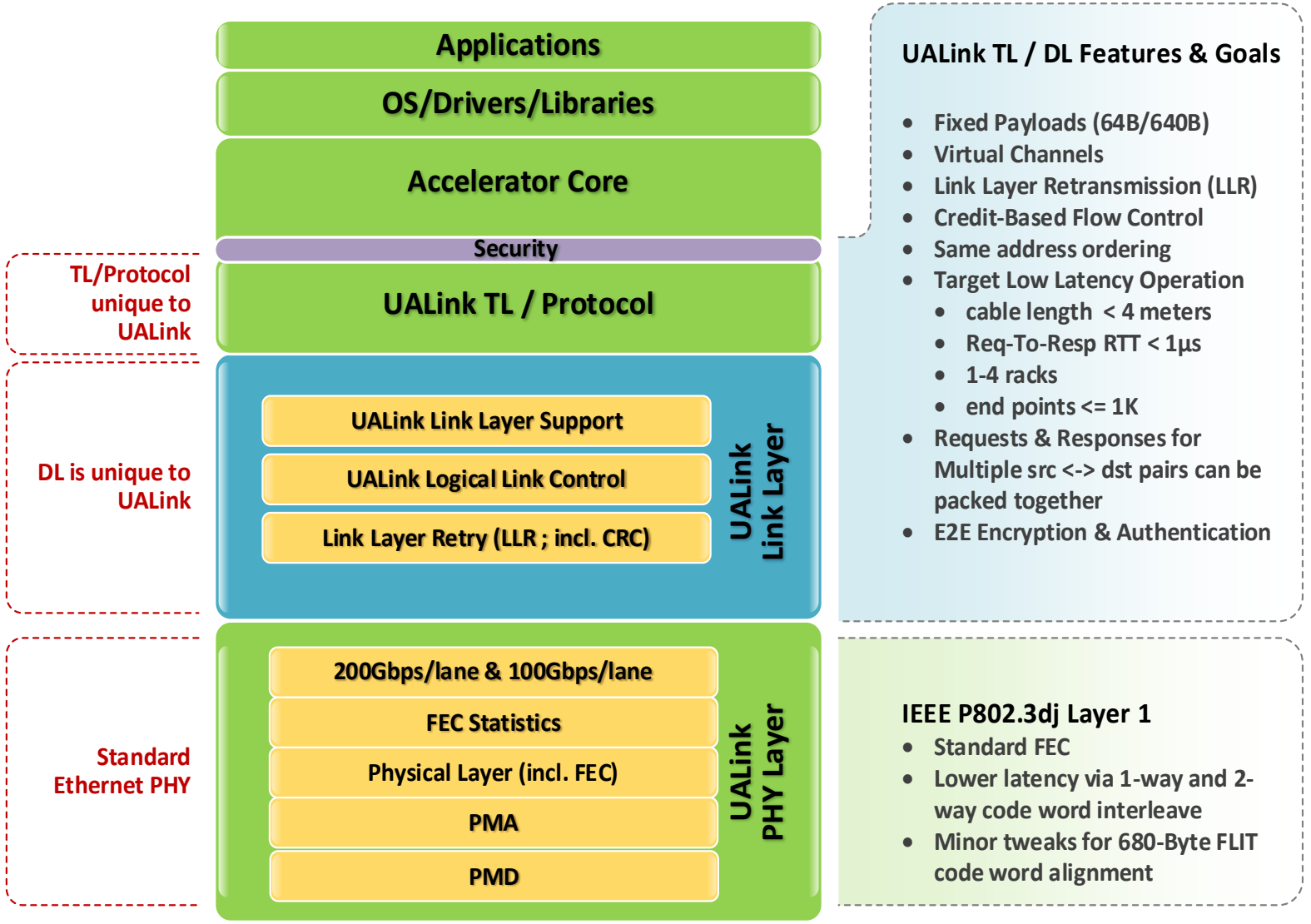
Source: Supermicro

- **36/72 → 576 → 1Ku+(!) GPU/XPUs**
- **In-the-box → rack & rack-to-rack**
- Proprietary or Ethernet-based **w/ switch(!)**
- Simple memory semantics OR more complex network semantics?

Modern AI infrastructure requires an open, multi-processing rack-level scale-up solution

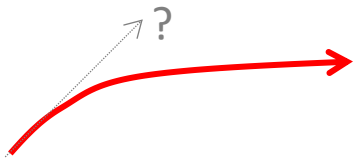
UALink Stack & Features

- Purpose built for scale-up
- Low latency
- High bandwidth
- Memory semantic
- Direct load, store, atomic operations
- Up to 1K accelerators in a pod



Why Now: AI Scale-Up Connectivity Challenges and Solutions

Challenge



Large cluster scaling

Wasted resources burden TCO

High integration cost & slow TTM

Solution

Today's Discussion

ULTRA ACCELERATOR LINK™

Native Memory Semantics
Simple Switching
Open Innovation

ULTRA ACCELERATOR LINK™

Higher Link Utilization
Efficient Silicon Design
Optimized Power Consumption

ULTRA ACCELERATOR LINK™

Open Specification
Interoperable
Supply Chain Resilience

UALink addresses the critical scale-up connectivity challenges for next generation AI infrastructure deployment

UALink: Purpose Built for Scale-up Application

Memory Semantics vs. Network Semantics

Semantics Comparison

Feature	Memory Semantics	Network Semantics
SW development Model	**Shared memory** (load/store)	**Message passing** (send/receive)
Messaging Type	**Implicit** – data only, no protocol	**Explicit** – requests and defines information
Memory Access	**Simple** - avoid DMA engine and network stack	**Complex –** must program DMA engine and network stack
BW Utilization	**High utilization** - no addressing/routing	**Lower utilization** due to addressing/routing overhead
Access Latency	**Low latency** - direct access to remote memory	**Higher latency** - serialization, pack/unpack, messaging
Common Use Cases	**Scale up** - Connectivity within a node **High-speed, real-time** applications **Shared context** in parallel compute	**Scale out** - connectivity between nodes **Non-time critical**; large amounts of data **Exchange context** between compute units
Value	**Memory-vs-time trade-offs** **In network compute** (INC)	**Add / remove individual nodes**

UALink is the only OPEN standard with memory semantics

Data Transfer Flows in SW/HW Stack

Memory Semantics Stack

Application
(Memcpy, malloc, implicit mem accesses, …)

Compiler Optimization
(GCC, CUDA, …)

Memory Management
(Local Mem, Remote Mem, Devices, …)

Connectivity Devices
(UALink, NVLink, PCIe, …)

- Simple and Low Latency Stack
- Small/Flexible/**Implicit** Transfers
- Optimized for parallel processing transfers

Network Semantics Stack

Application
(Pytorch, TensorFlow, …)

Collective Libraries
(*CCL, MPI, …)

Network Libraries
(Libibverbs, …)

Memory Management
(Page pinning, page faults, memory regi, …)

Network Device Driver

Network Devices
(SmartNIC, RoCE/RDMA offload, …)

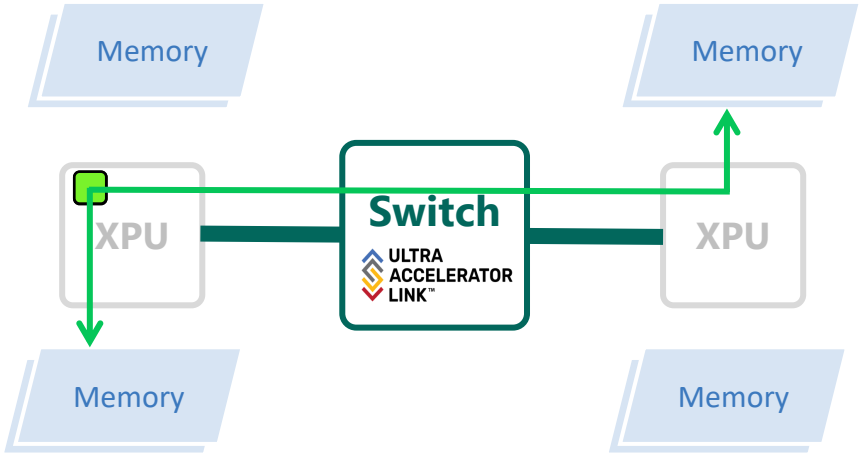
- Complex and Higher Latency Stack
- **Explicit** Transfers
- Optimized for large DMA transfers

Store Example

Key

- Data
- Network Message Handlers (overhead)

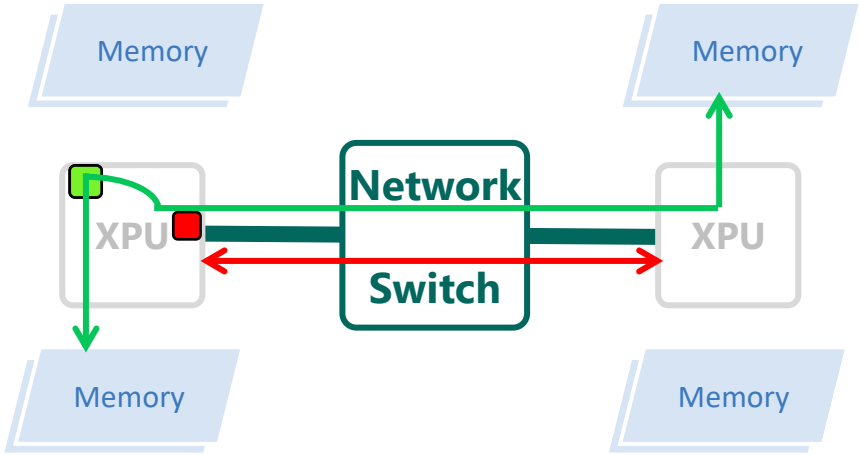
Native Memory Semantics



One Step:

1. **Local & Remote memory:** XPU Store (same for both)

Network Semantics



Multiple Steps:

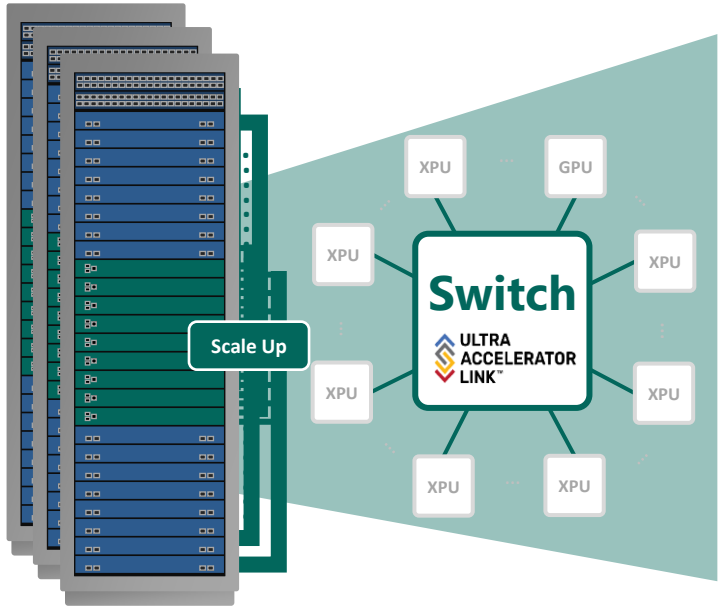
1. **Local memory**: XPU Store
2. **Remote memory**:
   a) Initiate send
   b) Networking message protocol
   c) Data transfer

Ultra Accelerator Link™ (UALink™) Consortium

Board of Directors

Alibaba, AMD, Apple, AsteraLabs, aws, cisco, Google, Hewlett Packard Enterprise, intel, Meta, Microsoft, SYNOPSYS

100+ members



- Open, standard, accelerator-to-accelerator communication
- Simple memory semantic based protocol up to 1Ku nodes
- Single-tier switch and cluster management
- Data rates of Ethernet, low latency of PCIe, low power
- Lower TCO & fast time-to-market

Oct 2024	Apr 2025	Target Q3 25	Target Q4 25	...
Consortium Incorporates	UALink 200G 1.0 Spec.	UCIe IO Chiplet Spec.	In-Network Collectives Spec.	

Switch

Call to Action: Download UALink 1.0 specification and join the UALink Consortium!

Thank You