OIF 448G AI WORKSHOP

ULTRA ACCELERATOR LINK™

UALink™ Consortium

Advancing AI Across Data Centers

ULTRA ACCELERATOR LINK™

AI models continue to grow requiring more compute and memory to efficiently execute training and inference on large models

The industry needs an **open** solution that enables efficient distribution of models across many accelerators within a pod

Large inference models will require scale-up of 10's – 100's of accelerators in pods

Large training models will require scale-up and scale-out from 100's – 10,000's of accelerators by connecting multiple pods

Board of Directors

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

Contributor Members

ALPHAWAVE SEMI, anapass, arm, auradine, AyarLabs, BizLink, ByteDance 字节跳动, cadence, centec, celestial AI, Clustered Systems Faster Networks, CORNELIS NETWORKS, CREDO, enfabrica, etopus we accelerate, FUJITSU, GIGAIO, H3C, KEYSIGHT TECHNOLOGIES, kiwi moore, Lenovo, LIGHTELLIGENCE, LIGHTMATTER, 灵可达 Linkdata, MARVELL, MEDIATEK, MICROCHIP, MIPS, MONTAGE Technology, NEUREALITY, panmnesia, Parade, Qualcomm, Rivos, TE connectivity, TELEDYNE LECROY Everywhereyoulook, Tencent Cloud, tenstorrent, Unifabrix, 合见工软 UNIVISTA, XCONN TECH, xFUSION, ZTE中兴

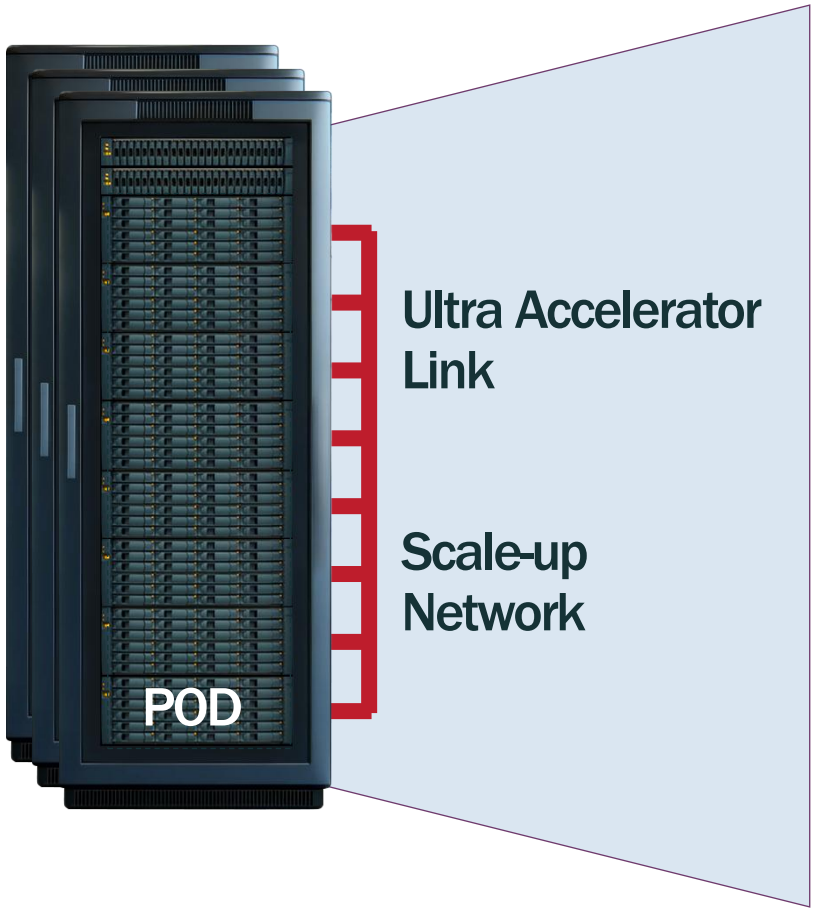
85+ Members

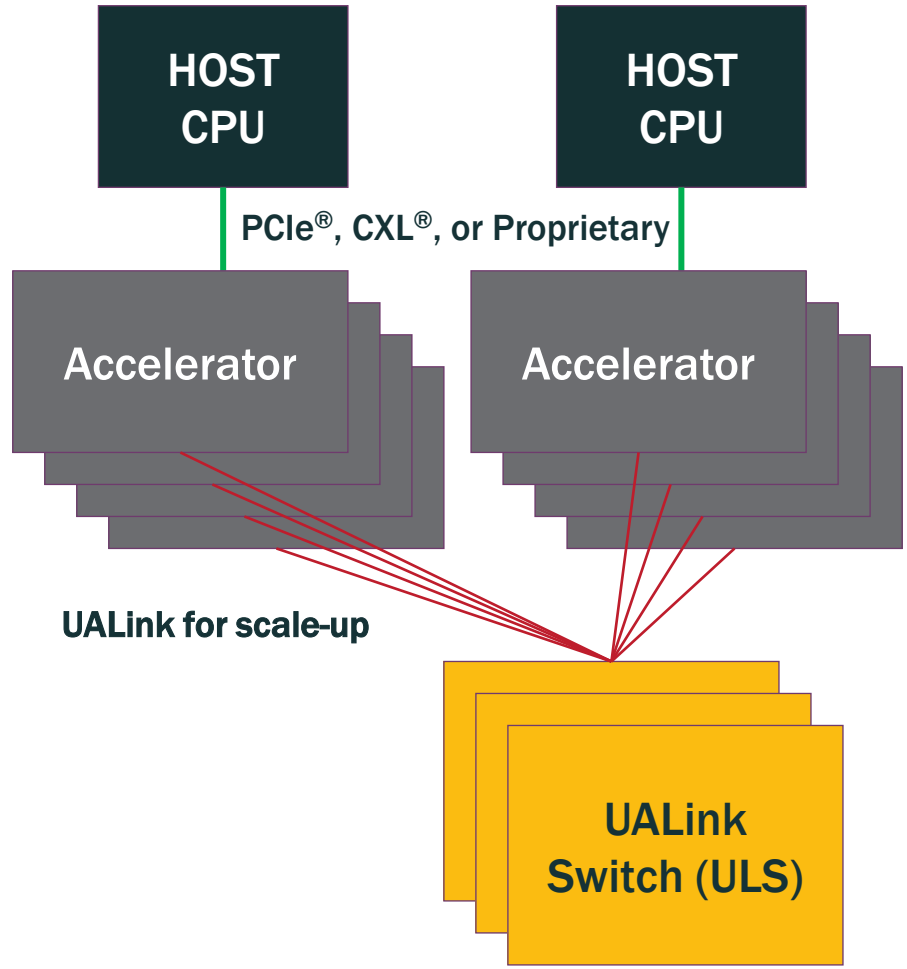
Ultra Accelerator Link Timeline

ULTRA ACCELERATOR LINK™

UALink Creates the Scale-up Pod

- High performance
  - Up to 1.6Tbps bi-directional bandwidth per accelerator
  - Support for 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 Boosts The Performance of AI Systems

ULTRA ACCELERATOR LINK™

- UALink's open standard significantly enhances the efficiency of accelerator-to-accelerator communication
- AI performance is highly sensitive to memory bandwidth
  - UALink provides best-in-class scale-up bandwidth for massive AI data transfers
- AI performance is also sensitive to memory latency
  - UALink provides ultra-low latency for scale-up environments allowing accelerators to collaborate more efficiently
- Direct accelerator-to-accelerator communication is beneficial for large-scale AI models
- UALink enables efficient scaling of AI systems to large numbers of accelerators
  - Allowing for the creation of pods that can address the most demanding AI workloads
- Open and Standardized
  - UALink harnesses the innovation of member companies to drive leading-edge features into the specification and interoperable products to the market

Higher Bandwidth Is Required

ULTRA ACCELERATOR LINK™

- 448Gbps transfers support faster data movement required by AI workloads
  - Reducing training time
  - Allowing for real-time inferencing
  - Making accelerator-to-accelerator communication more efficient
- Increasing the lane bandwidth reduces IO needs for the same external bandwidth or allows more ports per device
- Trade-offs will be required
  - Fast vs wide
  - Power
  - Cost
  - Latency
- Faster interconnect speeds must be a community effort that drive interoperability

Summary

- UALink addresses industry demand for a scale-up communication empowering efficient, scalable AI applications
  - Facilitates direct load/store for AI accelerators
  - Advances large AI model training
- Increasing interconnect speeds is important to UALink and the workloads it supports
- Visit our website for more info UALink www.ualinkconsortium.org
- Follow us on social media for updates
  - LinkedIn: Ultra Accelerator Link Consortium
  - X: @UltraAccelLink

THANK YOU

ULTRA ACCELERATOR LINK™