ULTRA ACCELERATOR LINK™

UALink 200G 1.0 Specification Overview

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

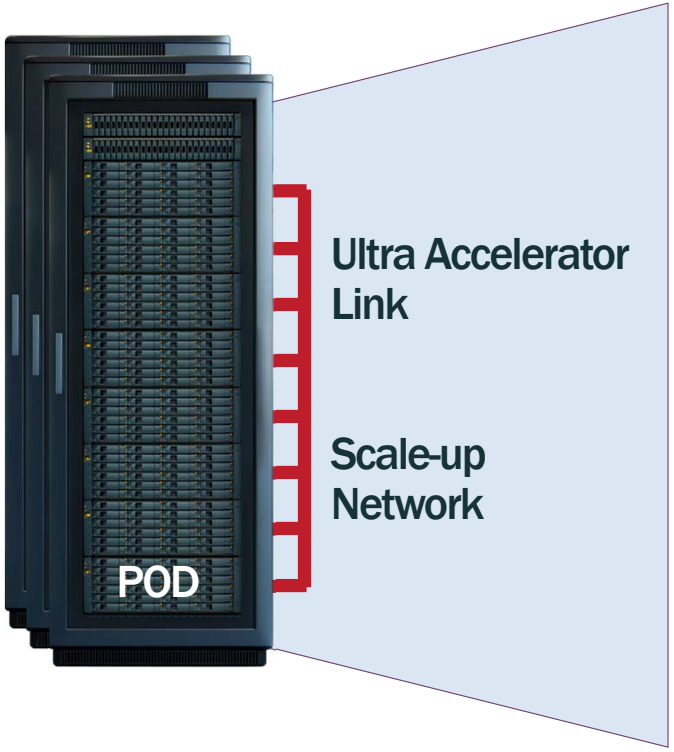
ALPHAWAVE SEMI, anapass, arm, auradine, AyarLabs, BizLink, ByteDance 字节跳动, cadence, centec, celestial AI, Clustered Systems Faster Networks, CORNELIS NETWORKS, CREDO, enfabrica, etopus we accelerate, FUJITSU, GIGAIO, H3C, KEYSIGHT TECHNOLOGIES, kiwi moore, Lenovo, LIGHTELLIGENCE, LIGHTMATTER, 灵可达 Linkdata, 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

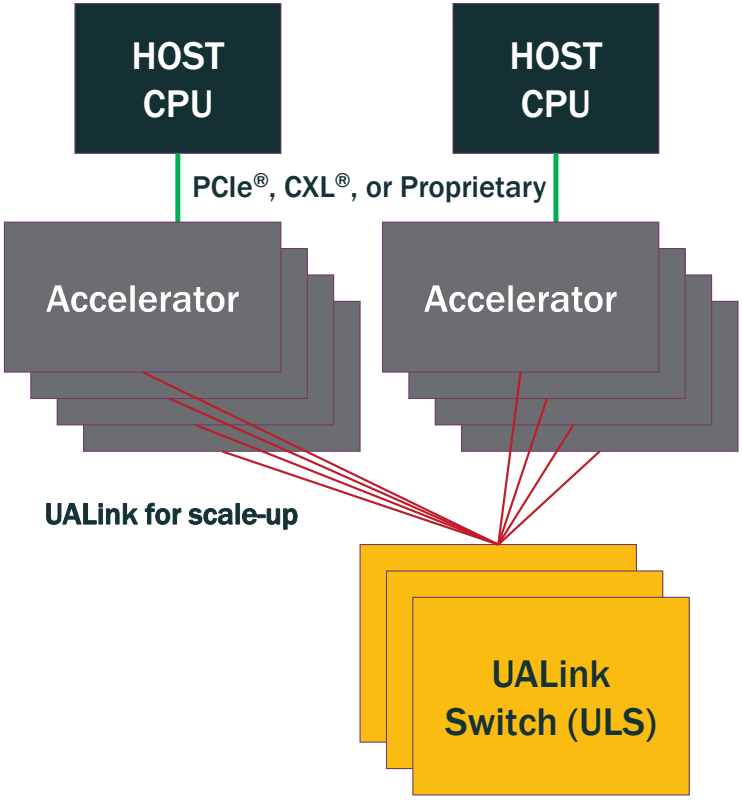
85+ Members

Ultra Accelerator Link Timeline

UALink Creates the Scale-up Pod

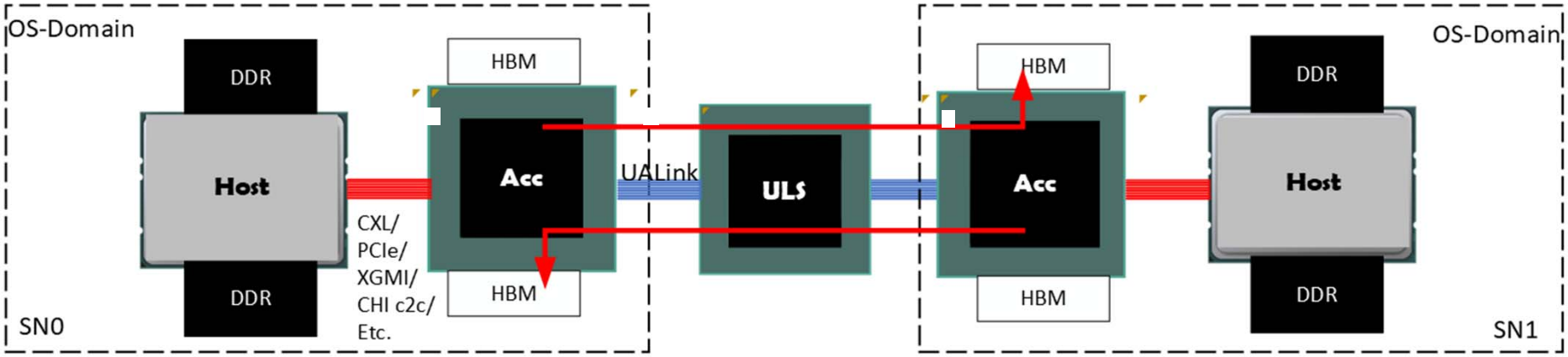
- High performance
  - 800Gbps per link, various links/accelerator, up to 1,024 accelerators
- Low latency
  - Optimized protocol, transaction, link & physical
- Low power
  - The simplified UALink stack leads to lower power solutions
- Low die area
  - Optimized data layer and transaction layer saves significant die area





UALink 200G 1.0 Specification Overview

- The UALink interconnect is for Accelerator-to-Accelerator communication
  - The initial focus is sharing memory among accelerators
- Direct load, store, and atomic operations between accelerators (i.e. GPUs)
  - Low latency, high bandwidth fabric for 100's of accelerators in a pod (up to 1K)
  - Simple load/store/atomics semantics with software coherency
- The initial UALink specification taps into the experience of the Promoters developing and deploying a broad range of accelerators and seeded with the proven Infinity Fabric protocol



UALink 200G 1.0 Specification Benefits

- High Performance
  - Low-latency, high-bandwidth interconnect for hundreds of accelerators in a pod
  - Features the same raw speed as Ethernet with the latency of PCIe® switches
  - Designed for deterministic performance achieving 93% effective peak bandwidth
- Low Power
  - Enables a highly efficient switch design that reduces power and complexity with small packets, fixed FLIT sizes, ID based routing, and overall simplicity
- Cost Efficient
  - Uses significantly smaller die area for link stack, lowering power and acquisition costs
  - Increased bandwidth efficiency further enables lower TCO
- Open and Standardized
  - UALink harnesses the innovation of member companies to drive leading-edge features into the specification and interoperable products to the market

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
- Creating efficient, low-latency and high bandwidth interconnect across hundreds of accelerators within a few racks
  - Open industry standard enables advanced models across multiple AI accelerators
- The UALink 200G 1.0 Specification is available for download at: www.ualinkconsortium.org

THANK YOU