ULTRA ACCELERATOR LINK

Ultra Accelerator Link (UALink)

Kurtis Bowman – AMD – Co-Lead, UALink Promoter Group

FMS the Future of Memory and Storage

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 these large models
- The industry needs an open solution that allows for distributing the models across multiple accelerators
- Large inference models will require scale-up of 10’s – 100’s of accelerators in pods
- Large training models will require scale-out of 100’s – 10,000’s of accelerators by connecting multiple pods

ULTRA ACCELERATOR LINK

Ultra Accelerator Link

Partner group of innovators for scale up AI infrastructure

AMD | BROADCOM | cisco | Google | Hewlett Packard Enterprise | intel | Meta | Microsoft

---

High Performance | **Open** | **Scalable**

Introducing Ultra Accelerator Link (UALink)

ULTRA ACCELERATOR LINK

- AMD, Broadcom, Cisco, Google, HPE, Intel, Meta, and Microsoft have collaborated in the formation of a Promoter's Group to form a new industry standard, UALink, to create the scale-up ecosystem

- UALink creates an open ecosystem for scale-up connections of many AI accelerators
  - Effectively communicate between accelerators using an industry standard protocol
  - Easily expand the number of accelerators in a pod
  - Optimizes the performance needed for compute intensive workloads now and in the future

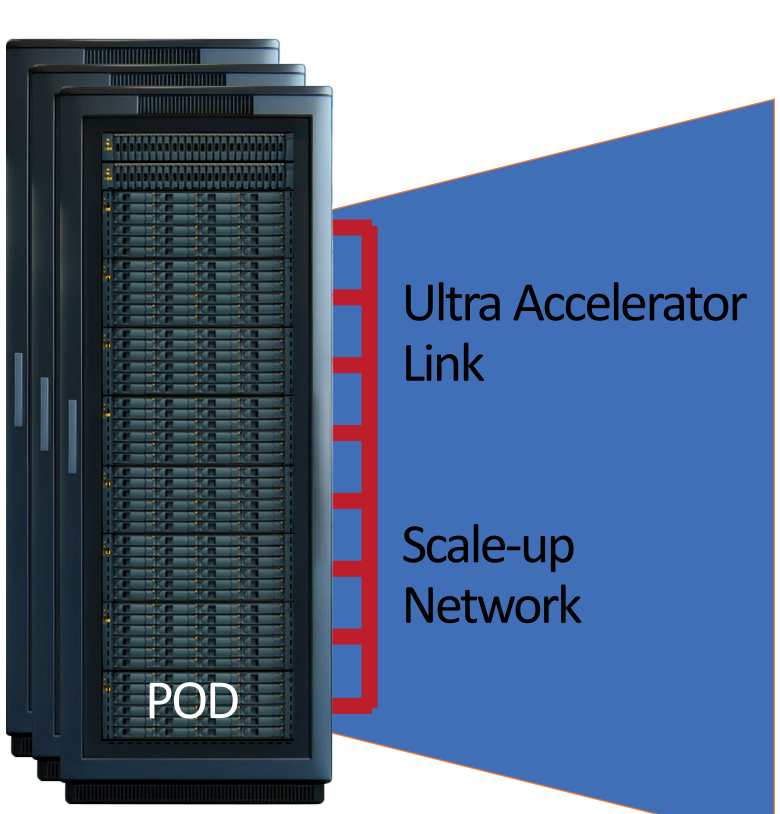
- An open scale up memory semantic fabric has significant advantages

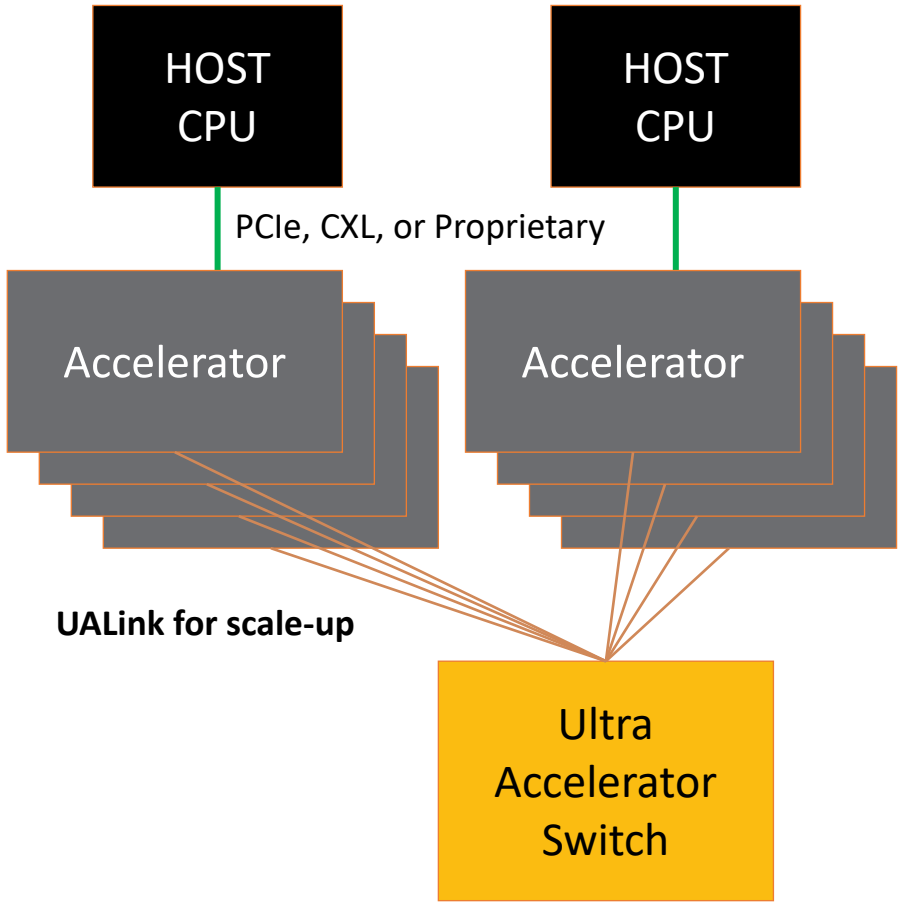
- Complimentary with scale out approaches such as Ultra Ethernet Consortium (UEC)

Ultra Accelerator Link Overview

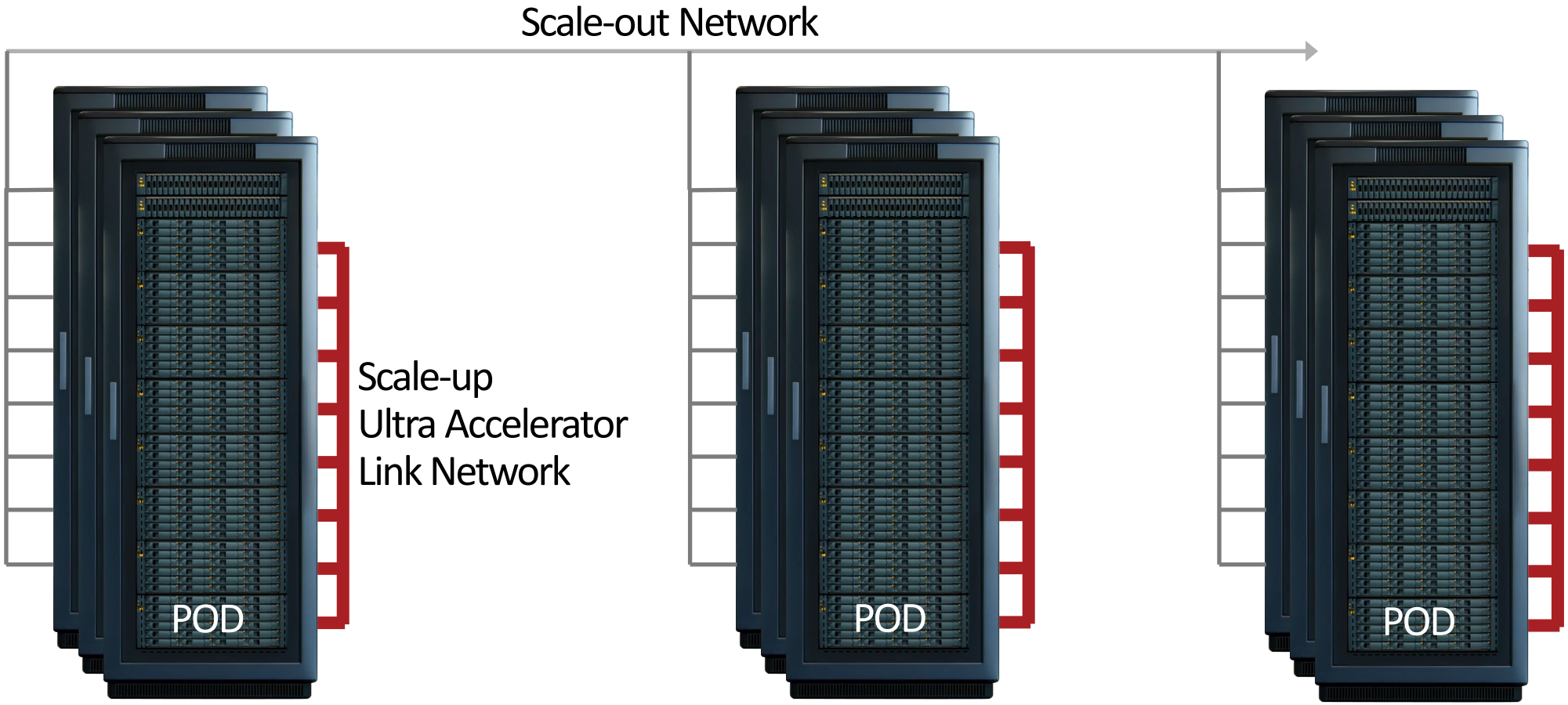
- The group plans to launch the organization in 3Q24
  - Focus of the organization is to release the initial specification by end of year
- The interconnect is for Accelerator-to-Accelerator communication
- Direct load, store, and atomic operations between AI Accelerators (i.e. GPUs)
  - Low latency, high bandwidth fabric for 100's of accelerators in a pod
  - Simple load/store semantics with software coherency
- The initial UALink spec taps into the experience of the Promoters developing and deploying a broad range of accelerators and leverages the proven Infinity Fabric protocol

UALink Creates the Scale-up Pod





Multiple UALink Pods Can Be Connected Via a Scale-Out Network



Summary

- UALink is an open solution allowing AI models to be deployed across multiple accelerators
- UALink creates an open ecosystem for scale-up connections of many AI accelerators
- The UALink Consortium plans to launch the organization in 3Q24
  - Initial specification will be released by end of year

Thank You