



UALink 200G 1.0 Specification Overview

Presenters



- Peter Onufryk, UALink Consortium President, Intel
- Nathan Kalyanasundharam, UALink Consortium Technical Task Force Co-Chair, AMD
- Chris Petersen, UALink Consortium Director, Astera Labs



Introduction

Peter Onufryk, Intel

UALink Consortium President

Advancing AI Across Data Centers



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



Contributor Members



90+ Members

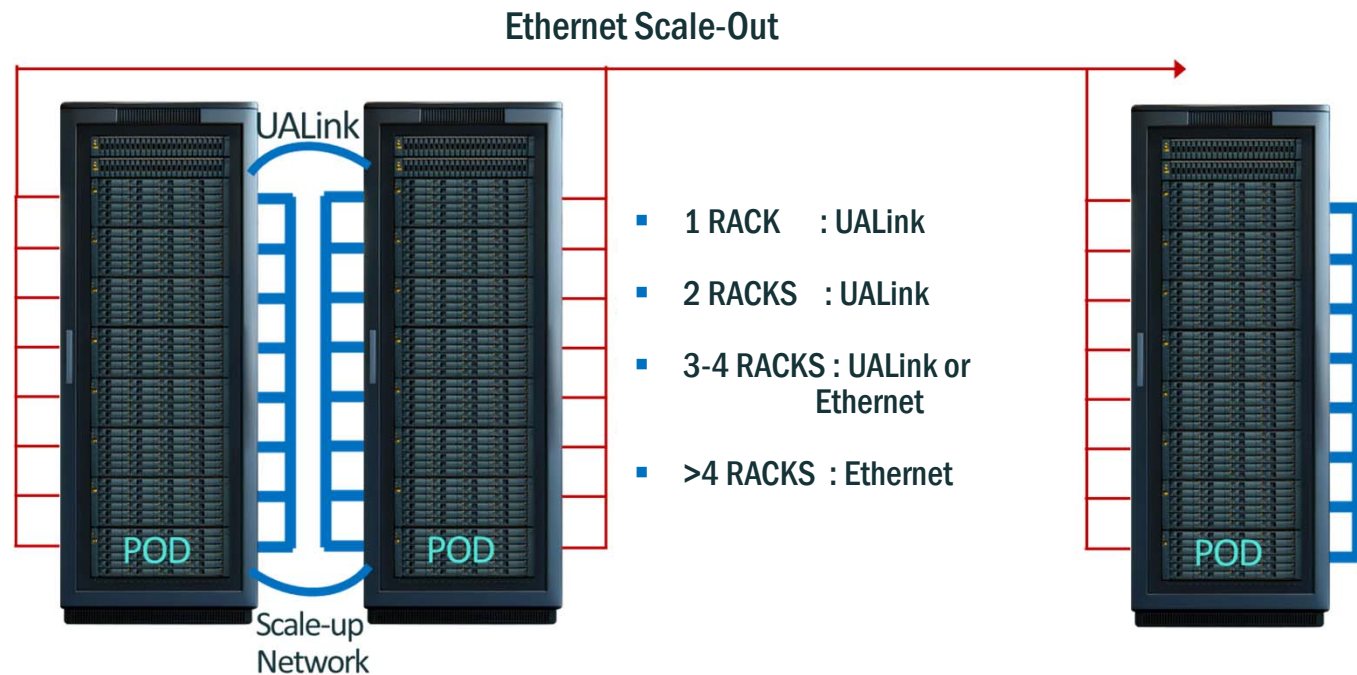
Ultra Accelerator Link Timeline



UALink Creates the Scale-up Pod



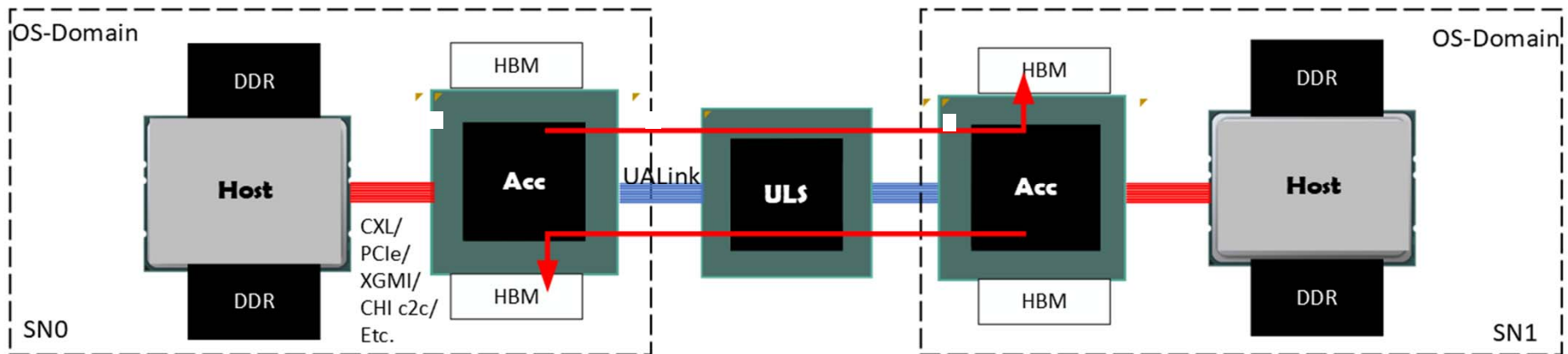
- High performance
 - Up to 800Gbps per Port, scalable ports per accelerator, Up to 1,024 accelerators
- Low latency
 - Optimized protocol, transaction, link & physical
- Low power
 - The simplified UALink stack leads to lower power solutions
- Low die area
 - Optimized data layer and transaction layer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



UALink1.0 focus is to deliver optimized scale-up solutions with single tier switching

UALink 200G 1.0 Specification

- The UALink interconnect enables 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 Benefits



- **Performance, Power & Efficiency**

- 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
- Enables a highly efficient switch design that reduces power and complexity with small packets, fixed FLIT sizes, ID based routing, and overall simplicity
- 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

- **Leverages ubiquitous Ethernet infrastructure**

- Cables, Connectors, Retimers, Management Software, and more.



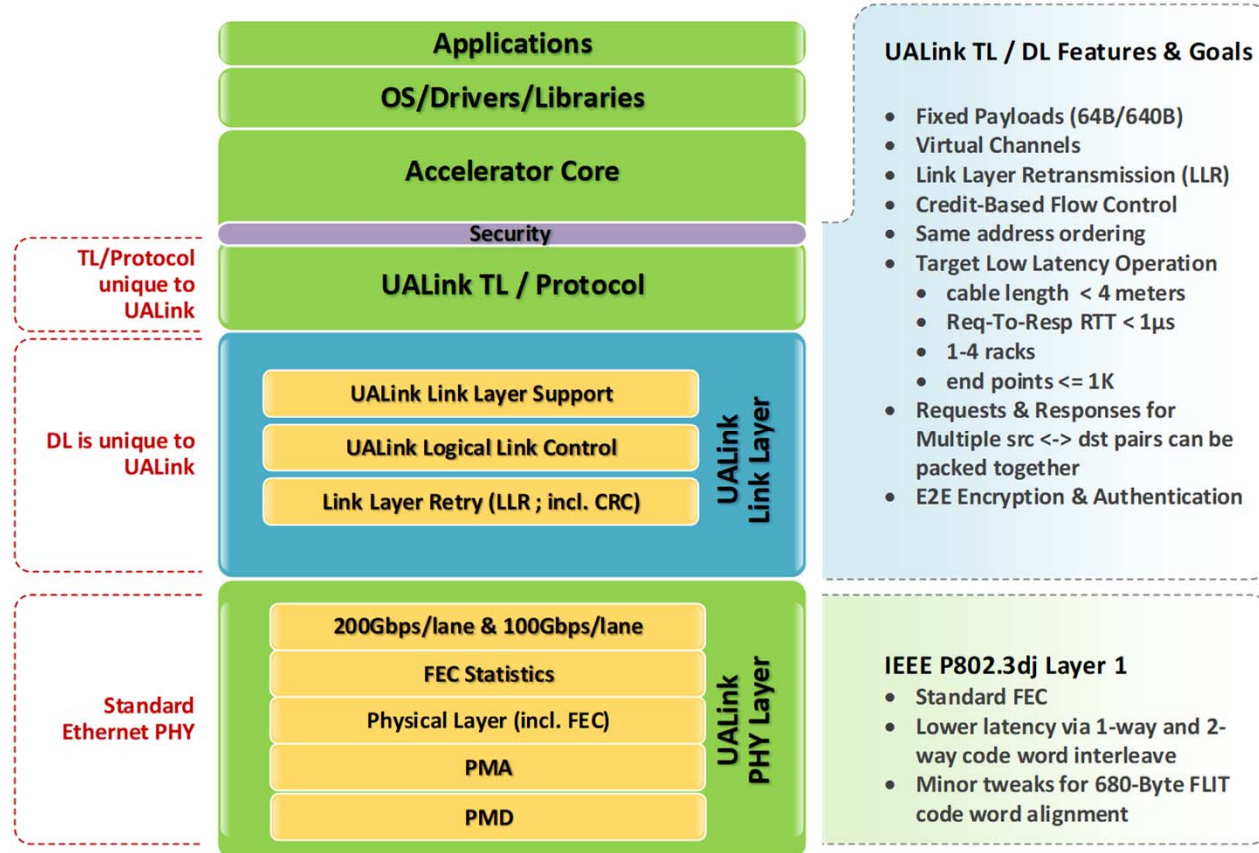
Technical Overview

Nathan Kalyanasundharam, AMD
UALink Technical Task Force Co-Chair

UALink Stack Features & Goals



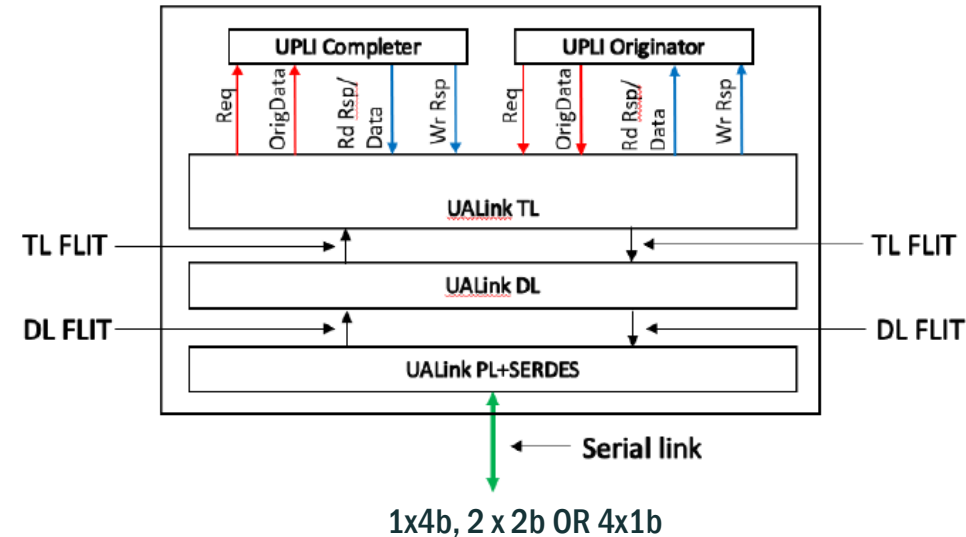
- Standard Ethernet Physical
- UALink DL
- UALink TL
- UALink Protocol



UALink Protocol Interface (UPLI)



- Simple symmetric interface protocol
 - Request
 - Request Data
 - Read Response + Data
 - Write Response
- Originator interface sends requests to other accelerators and receives responses.
- Completer interface receives requests from other accelerators and returns responses
- Src/Dst Identifier(ID) based routing
- Provisioned to enable multiple address spaces
- Same address ordering for Requests; Completions unordered



Transaction Layer (TL)



- TL Flit organized as sixteen 4-byte Sectors
- TL Flit is also divided into Upper and Lower 32-byte Half Flits
- Control half-flit is used for
 - Requests, read responses, write responses, flow control and NOP indication
- Data uses half & full Flits
 - Read response data, Write data and byte mask, Atomic operand data and byte mask
- Requests & responses may be compressed
 - Uncompressed Requests = 16B
 - Compressed Requests = 8B
 - Uncompressed Responses = 8B
 - Compressed Responses = 4B

Eff
92.3%

	64-byte TL Flit															
	Upper TL Half-Flit								Lower TL Half-Flit							
	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
Flit																
0	Req0.Data.0								Req1 (WriteFull)				Req0 (WriteFull)			
1	Req0.Data.2								Req0.Data.1							
2	Req0.Data.4								Req0.Data.3							
3	Req0.Data.6								Req0.Data.5							
4	Req1.Data.0								Req0.Data.7							
5	Req1.Data.2								Req1.Data.1							
6	Req1.Data.4								Req1.Data.3							
7	Req1.Data.6								Req1.Data.5							
8	Req1.Data.7								Req2 (Write Full)				CWR C	CWR B	CWR A	F
9	Req2.Data.1								Req2.Data.0							
10	Req2.Data.3								Req2.Data.2							
11	Req2.Data.5								Req2.Data.4							
12	Req2.Data.7								Req2.Data.6							

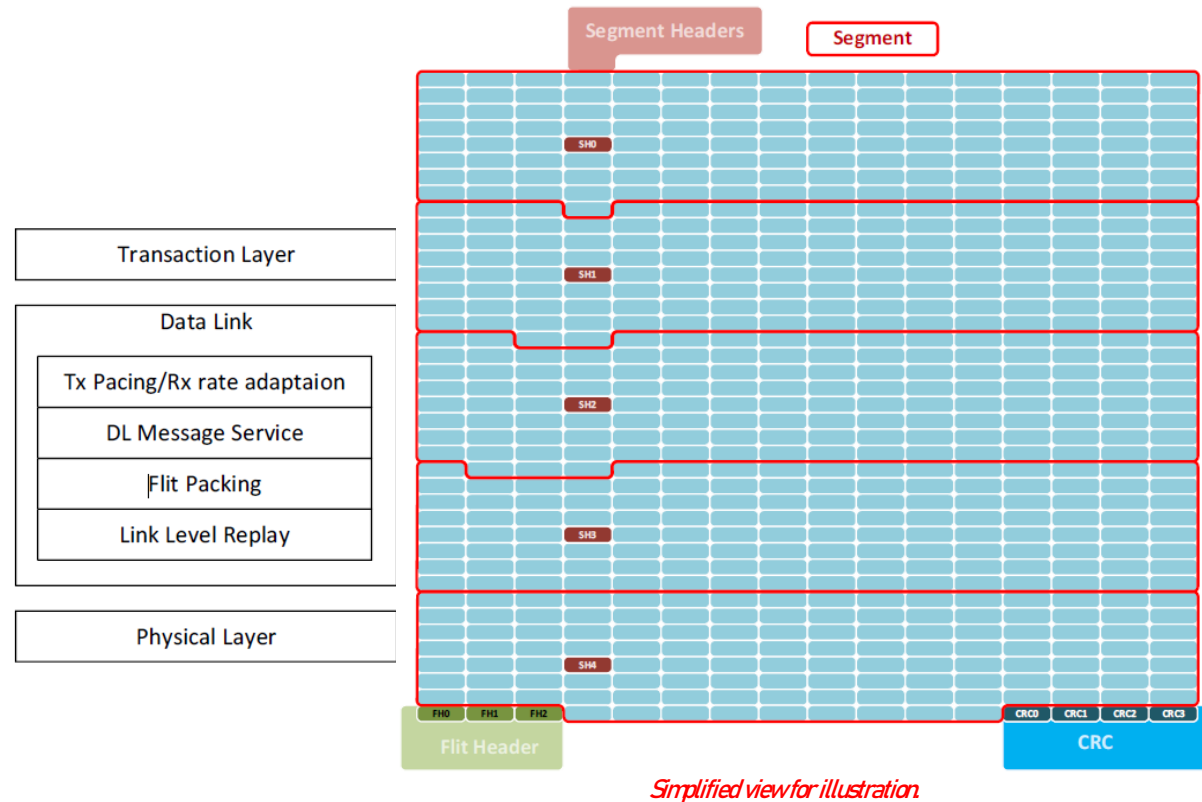
Eff.
95.2%

64-byte TL Flit															
Upper TL Half-Flit								Lower TL Half-Flit							
15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
Flit															
0	Req0.Data.0							CWReq2	CWReq1	CWR _B	CWR _A	CWReq0			
1	Req0.Data.2							Req0.Data.1							
2	Req0.Data.4							Req0.Data.3							
3	Req0.Data.6							Req0.Data.5							
4	Req1.Data.0							Req0.Data.7							
5	Req1.Data.2							Req1.Data.1							
6	Req1.Data.4							Req1.Data.3							
7	Req1.Data.6							Req1.Data.5							
8	Req2.Data.0							Req1.Data.7							
9	Req2.Data.2							Req2.Data.1							
10	Req2.Data.4							Req2.Data.3							
11	Req2.Data.6							Req2.Data.5							
12	Req2.Data.7							CWReq4	CWReq3	CWR _E	CWR _D	CWR _C	FC		
13	Req3.Data.1							Req3.Data.0							
14	Req3.Data.3							Req3.Data.2							
15	Req3.Data.5							Req3.Data.4							
16	Req3.Data.7							Req3.Data.6							
17	Req4.Data.1							Req4.Data.0							
18	Req4.Data.3							Req4.Data.2							
19	Req4.Data.5							Req4.Data.4							
20	Req4.Data.7							Req4.Data.6							

Note: For illustration

Data Link Layer (DL) – 640B

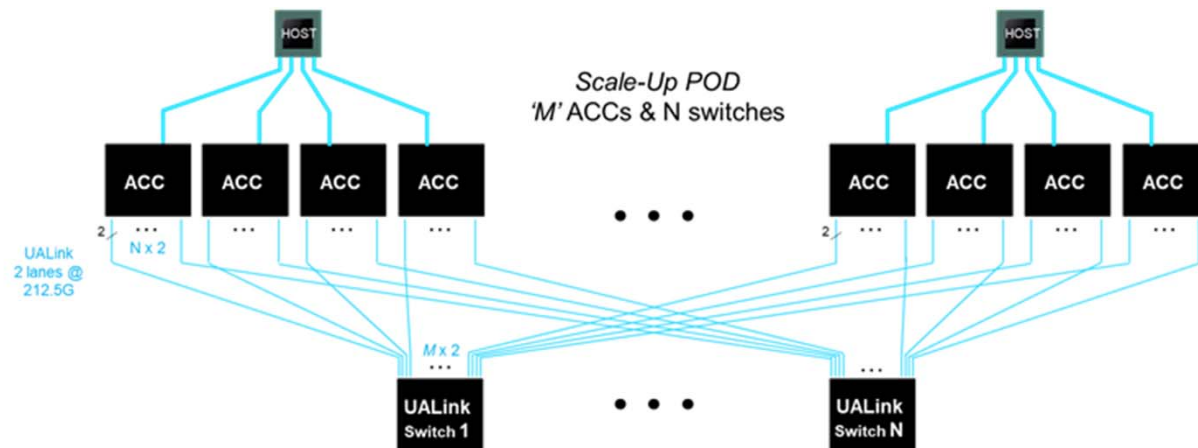
- 640 Byte DL FLIT
 - Flit Header = 3 Bytes
 - Segment Hdr = 5 Bytes
 - CRC = 4 Bytes
 - Efficiency = $628/640 = 98.125\%$
- FEC Code Word = 680 Bytes
 - Higher signaling rate (212.5 GHz) to cover the FEC overhead



Scale-up POD

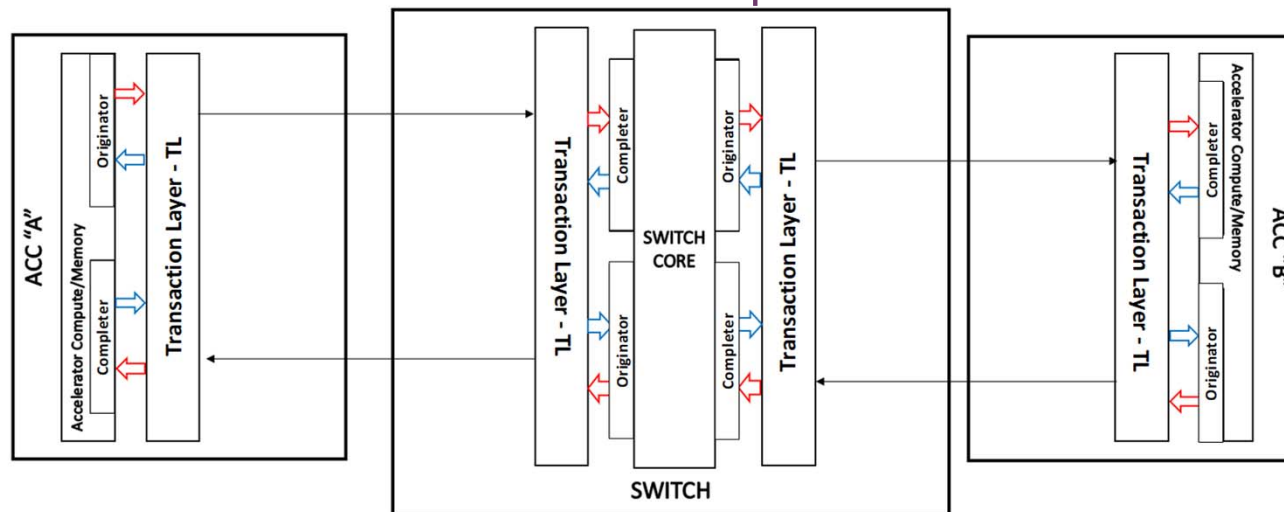


- Single tier switches
- Number of switch planes scaled with bandwidth per accelerator
- Number of Accelerators per POD is limited by lanes per switch
- POD may be configured as many virtual pods
- Virtual POD reconfiguration does not impact each other
- Error in one Virtual POD does not impact another
- Error recovery expected to be contained to a Virtual POD through Port or Station Reset
- Internal Switch Errors may impact the entire POD. Requires application restart



Data Flow

- Accelerators finely interleave (256B) memory channels
 - Maximizes bandwidth to local and peer GPU memory
 - Load/store/atomic memory accesses use small packets
 - Application may communicate with multiple peers simultaneously
- TL packs requests and responses into same FLIT
 - Requests and responses to many destination may be packed together
 - Reduces latency and area
 - TL is a light-weight implementation consuming ~0.3 sqmm in N3 technology



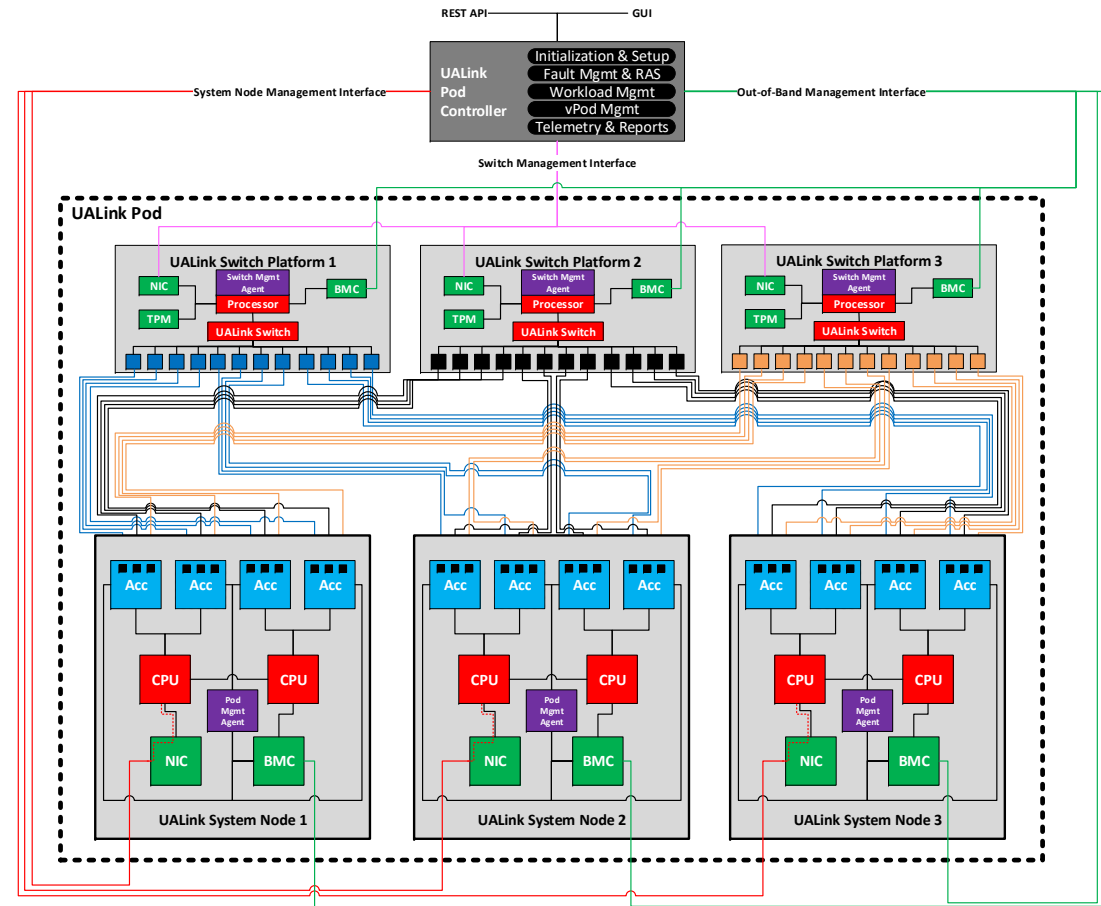


Systems Specifications Conclusion

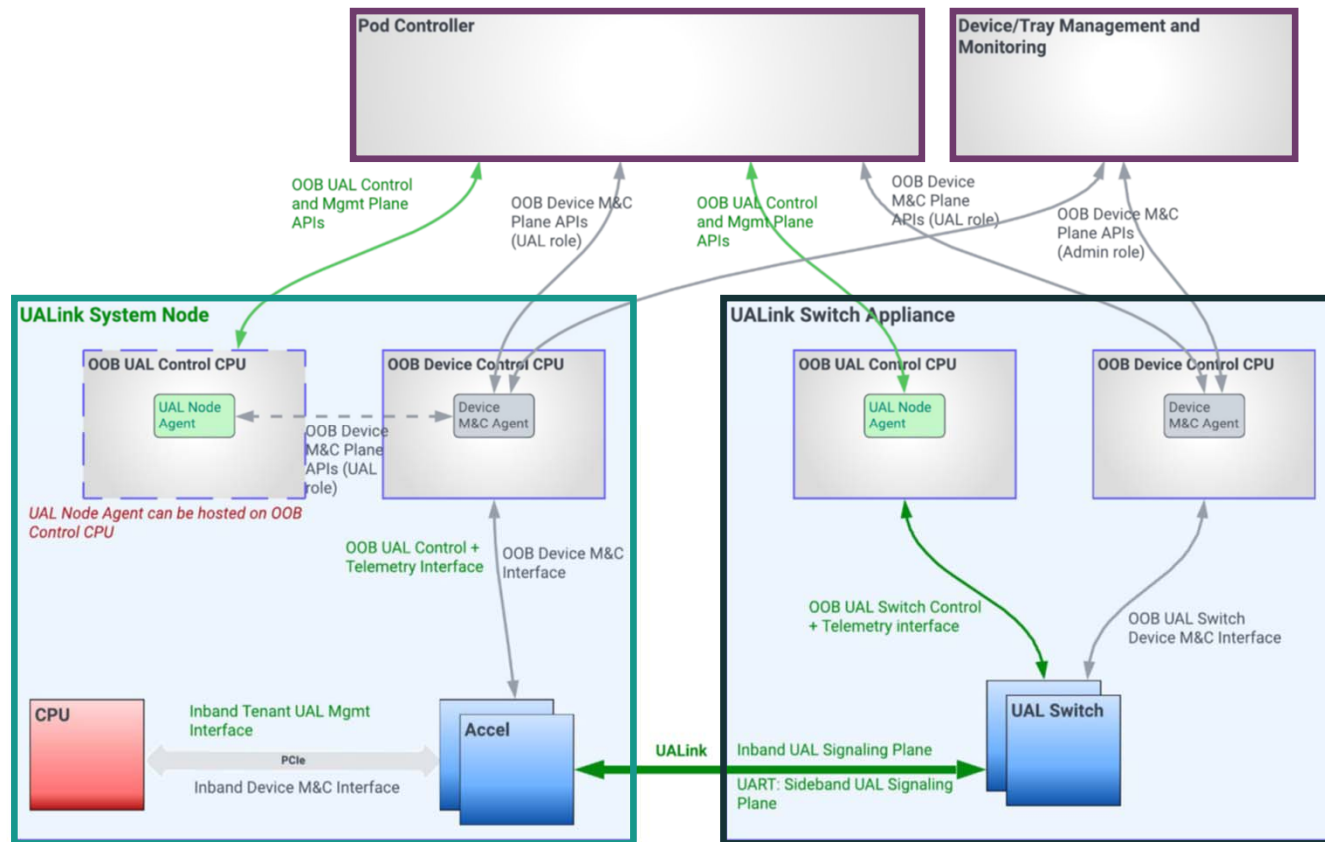
Chris Petersen, Astera Labs
UALink Board Director

Switch & Cluster Management

- Flexible management models for switches
 - Ethernet-like appliance model &
 - Lightweight PCIe-like switch model
- Common work-flows/APIs
- Leverage industry specifications
 - OCP, CPER, etc.
 - For Telemetry, Accelerator management, RAS, etc.



Management Layer

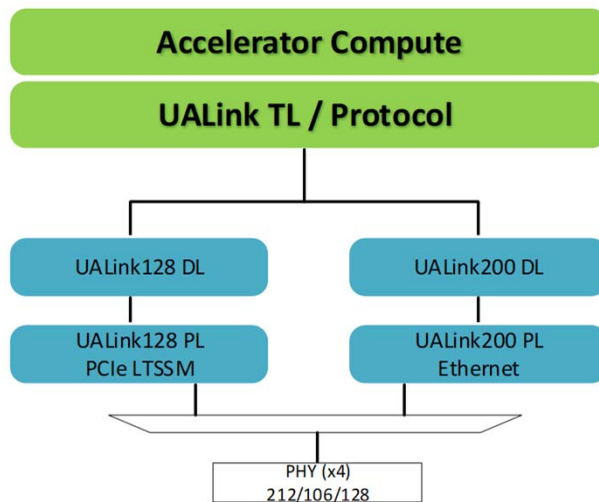


Example for illustration

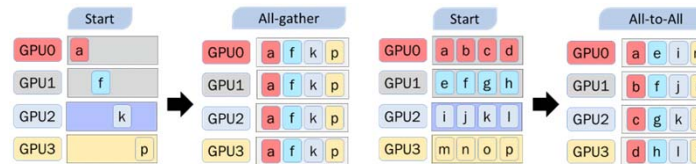
In Progress



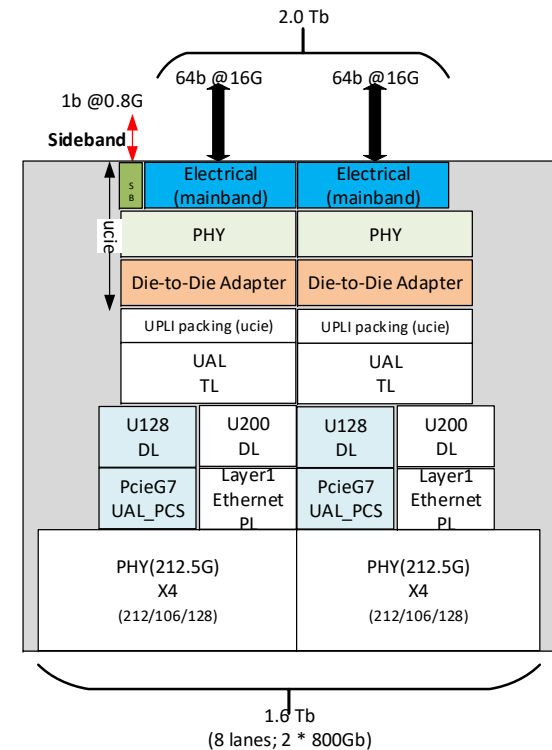
128G DL/PL Specification
Expected release : July 2025



In-Network Collectives (INC) Specification
Expected release : Dec 2025



128G & 200G UCle PHY Chiplet Specification



Summary



- UALink addresses industry demand for a scale-up fabric empowering efficient, scalable AI applications
 - Facilitates direct load/store for AI accelerators
 - Open industry standard enables advanced models across multiple AI accelerators
 - Advances large AI model training & inference
- UALink enables an **efficient, low-latency and high bandwidth** interconnect across hundreds of accelerators within a few racks
- The UALink 200G 1.0 Specification is available for download at: www.ualinkconsortium.org

Thank you!!



Q&A

THANK YOU

