Introducing UALink 200G 1.0 Specification

By: Nathan Kalyanasundharam
AMD Corporate Fellow and UALink Consortium Technical Task Force Co-Chair

Ultra Accelerator Link™ (UALink™) is a new industry standard to enable scale-up interconnects for next generation AI workloads. The UALink 200G 1.0 Specification was initially developed by the UALink Consortium Promoter Group Members from Alibaba, AMD, Apple, Astera Labs, AWS, Cisco, Google, HPE, Intel, Meta, Microsoft, and Synopsys and is now also supported and adopted by more than 70 Contributor and Adopter Members.

Introduction

AI models are rapidly growing, demanding higher compute, memory, and interconnect performance. The cost and complexity of delivering reliable scale-up solutions is a significant burden for the entire industry. Scale-up solutions are critical to distribute AI models across a Pod with 100s of accelerators. There is a growing demand from the industry to establish standards-based scale-up network solutions for training and inference workloads. The UALink Consortium's mission is to establish an open standard to deliver a scalable, performant, resilient and cost-effective networking solution for scale-up connections.

UALink is designed to deliver scale-up solutions, optimized for area, latency and power, featuring

- Improved link efficiency for bidirectional memory access to provide maximum data bandwidth
- Decreased Total Cost of Ownership (TCO) by leveraging existing Ethernet infrastructure including the use of cables, connectors, retimers, and management software
- Reduced software complexity using memory semantics with direct read, write, atomic transactions and maintaining the same ordering model to host attached, local, and remote accelerator memory.

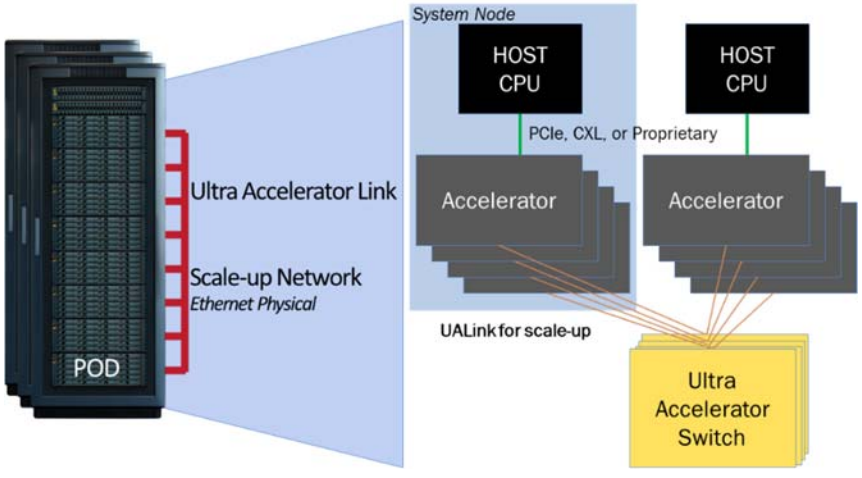

Figure 1: Multi-Node & Multi-Plane Switching in a Scale-Up Rack Infrastructure

UALink enables scalable interconnect performance to match growth in compute, memory bandwidth and capacity. Interconnect performance scalability is met by enabling multi-node systems and a multi-plane switching ecosystem.

The UALink 1.0 specification defines protocols and interfaces for low latency accelerator-to-accelerator communication.

Overview

The UALink 1.0 specification supports a maximum data rate of 200 GT/s per lane. The signaling rate is higher at 212.5 GT/s to account for the bandwidth required by Ethernet Layer 1 for Forward Error Correction Code (FEC) and additional Layer 1 encoding. UALink lanes can be configured into various groupings: a single-lane Link (x1 Link), a dual-lane Link (x2 Link), or a quad-lane Link (x4 Link). A group of four lanes constitutes a Station, offering a maximum bandwidth of 800 Gbps each in transmit (TX) and receive (RX) directions. The number of accelerators and bandwidth allocated to each accelerator can be scaled to meet the demands of AI applications. Figure 2 shares the high-level features and goals for UALink based accelerator systems.

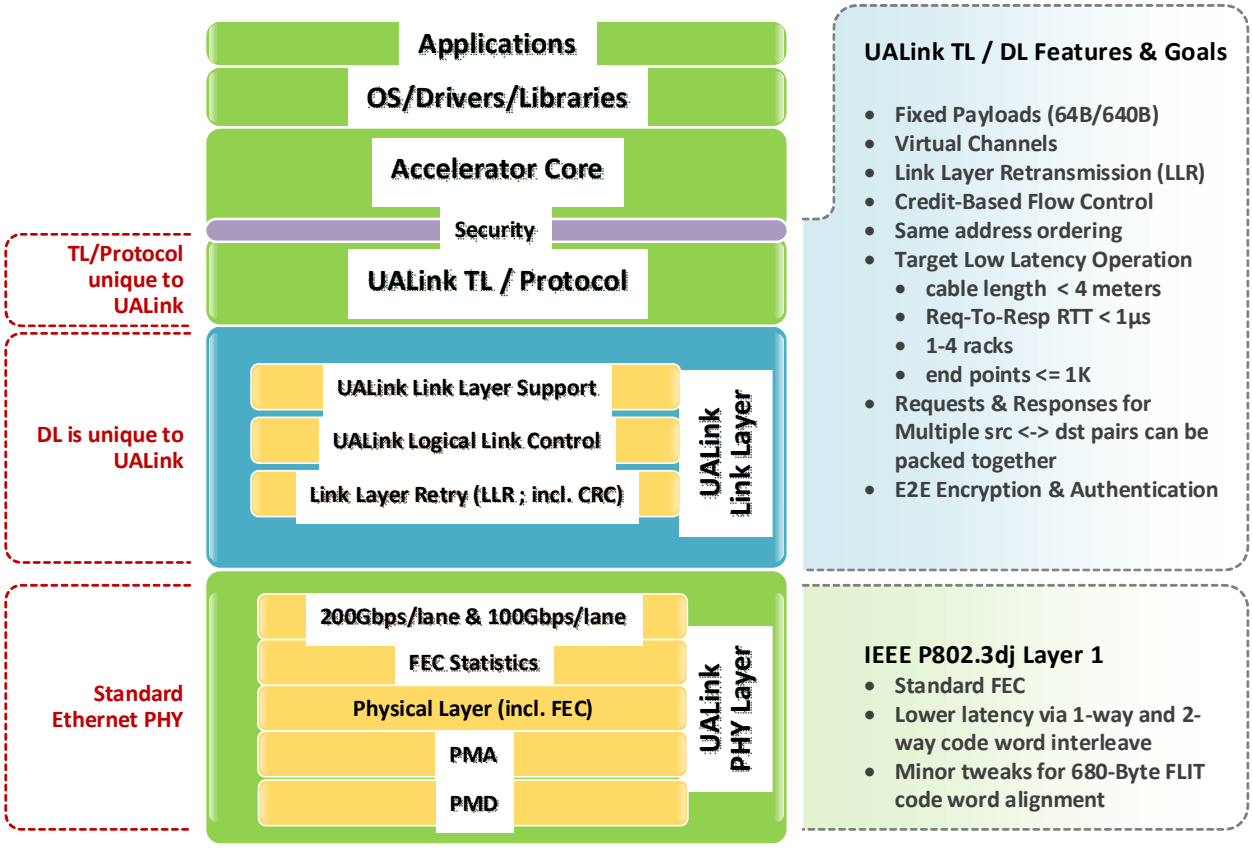

Figure 2: UALink Features & Goals

Figure 3 depicts a system with multiple nodes, each containing a Host processor and four Accelerators. The system includes 'M' Accelerators overall, each with 'N' symmetric Ports distributing traffic evenly. Each System Node is managed by one OS image. UALink Switches (ULS) connect up to 1024 Accelerators or endpoints. Each Accelerator is assigned a unique 10-bit routing identifier. These connected Accelerators form a scale-up Pod. Each UALink Switch port connects to a distinct Accelerator. The number of Switch ports matches the number of connected Accelerators.

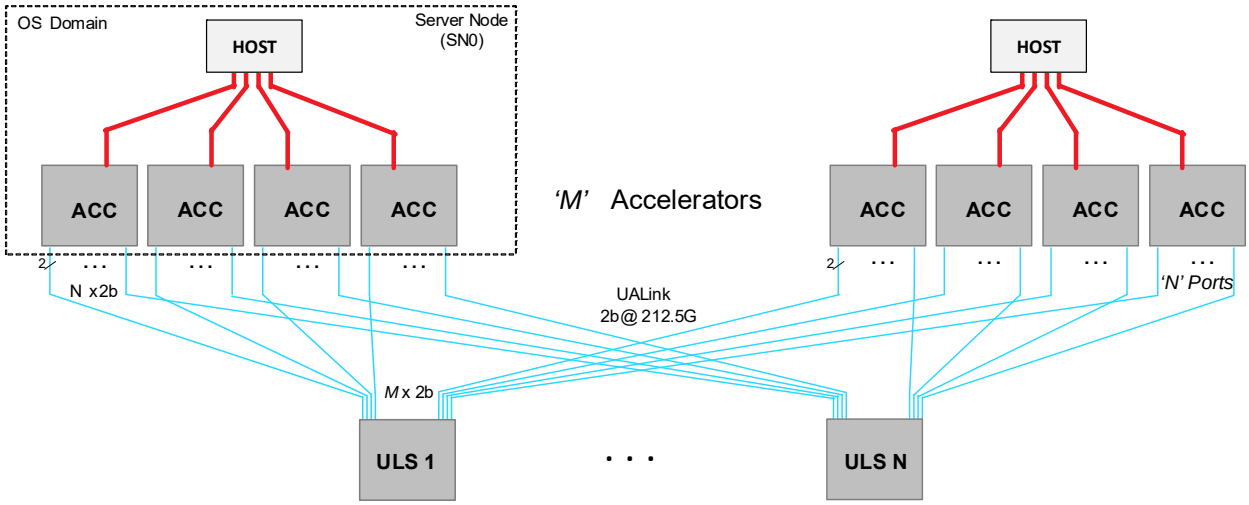

Figure 3: UALink Based Multi-Accelerator System

A Pod may be partitioned into Virtual Pods. A Virtual Pod is a group of one or more Accelerators in the Pod that may communicate amongst themselves but not with any other Accelerator in the Pod. The Pod may be divided into Virtual Pods by partitioning the Switches into non-overlapping subsets of Ports on each Switch. The Ports within a subset can communicate with one another but not with any Port outside the subset. Switches provide the mechanisms to configure partitions. All Accelerators in a Pod have a unique Accelerator ID, regardless of Virtual Pod partitioning.

Link Stack

A UALink stack shown in Figure 4 includes a

- Protocol Layer
- Transaction Layer
- Data Link Layer and
- Physical Layer

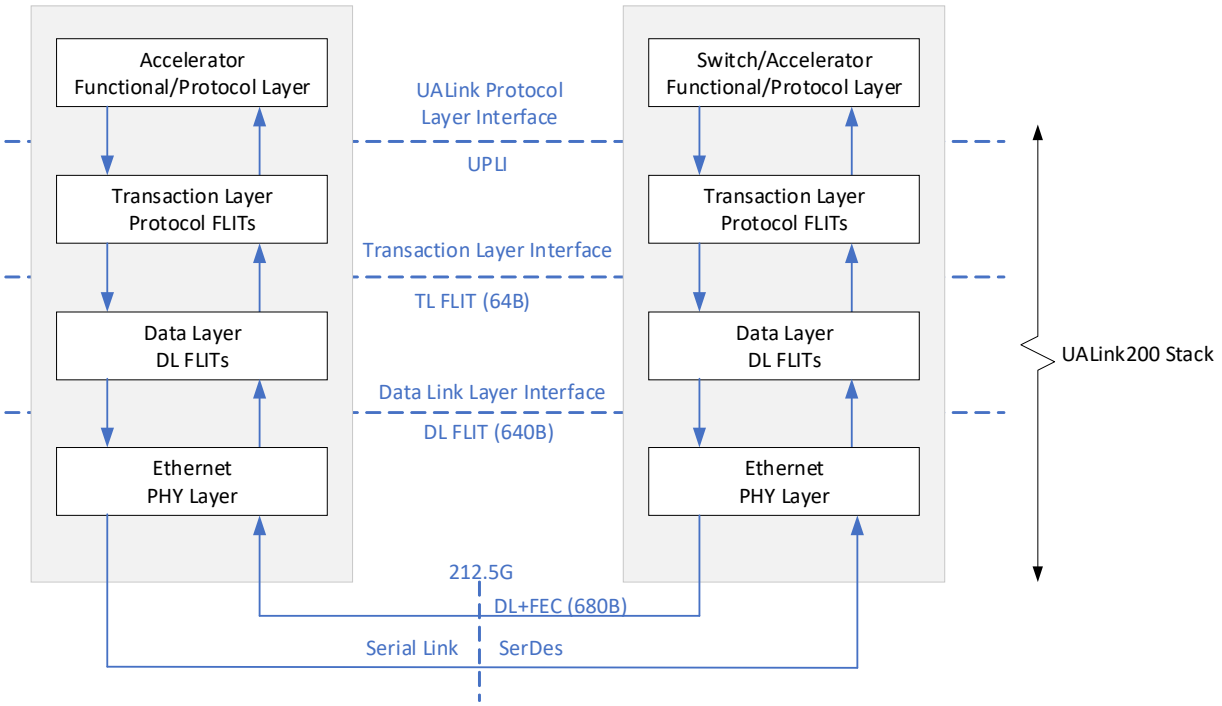

Figure 4: UALink Stack

Messages between accelerators are transmitted over UALink. UALink is a symmetrical protocol that supports the same set of messages and channels in both transmit and receive paths. These messages pass through multiple functional layers of the UALink stack. The UALink Stack consists of an independent UPLI (UALink Protocol Layer Interface) Originator and an independent UPLI Completer, along with logic units to implement the UALink Protocol Layers. The Completer and Originator are connected to a Transport Layer (TL). The TL converts the Protocol Channels into a TL Flit that is passed to the Data Link (DL) Layer. Similarly, the TL receives a TL Flit from the DL that is unpacked into the UPLI Protocol Channels received by the Completer and the Originator. A Completer interface receives requests from peer accelerators and returns responses. An Originator interface initiates requests targeting remote accelerators and receives responses in return.

The UALink DL receives several TL Flits, adds CRC (Cyclic Redundancy Check) protection, a header to the Flit to form a DL Flit, and passes the DL Flit on to the UALink Physical Layer (PL). Similarly, the DL can receive DL Flits from the PL, strips the CRC and header from the Flit, and form TL Flits that are passed on

to the TL. The UALink PL receives DL Flits and produces a code word with FEC encoding that is serialized and transmitted through a serial link. The PL can also receive a serialized code word with FEC from a link partner on a switch or accelerator to perform FEC decoding and convert that into a DL Flit.

Physical Layer (PL)

The UALink PHY is based on 802.3 Ethernet PHY. UALink is defined for one, two, or four serial lanes running at a serial rate of 212.5G (200GBASE-KR1/CR1, 400GBASE-KR2/CR2, 800GBASE-KR4/CR4), as well as a lower speed serial rate option of 106.25G (100GBASE-KR1/CR1, 200GBASE-KR2/CR2, 400GBASE-KR4/CR4).

The * in Figure 5 indicates modifications from IEEE 802.3. The PCS (Physical Coding Sublayer)/PMA (Physical Medium Attachment Interface) operates in additional codeword interleave modes, with reduced interleave, to achieve better FEC latency at the cost of decreased burst error correction. The PMD is unmodified from 802.3. Auto Negotiation and Link Training (AN/LT) is unmodified from 802.3. The 64B/66B encoding is a subset of what 802.3 supports.

The PCS and RS (Reconciliation Layer - an interface between PCA and DL) require additional behavior to synchronize DL Flits to codewords, so that 640-byte Flits from the DL fit exactly into one RS (544, 514) codeword. This will optimize latency and minimize replay Flits. The DL generates a CRC as part of each 640-byte DL Flit.

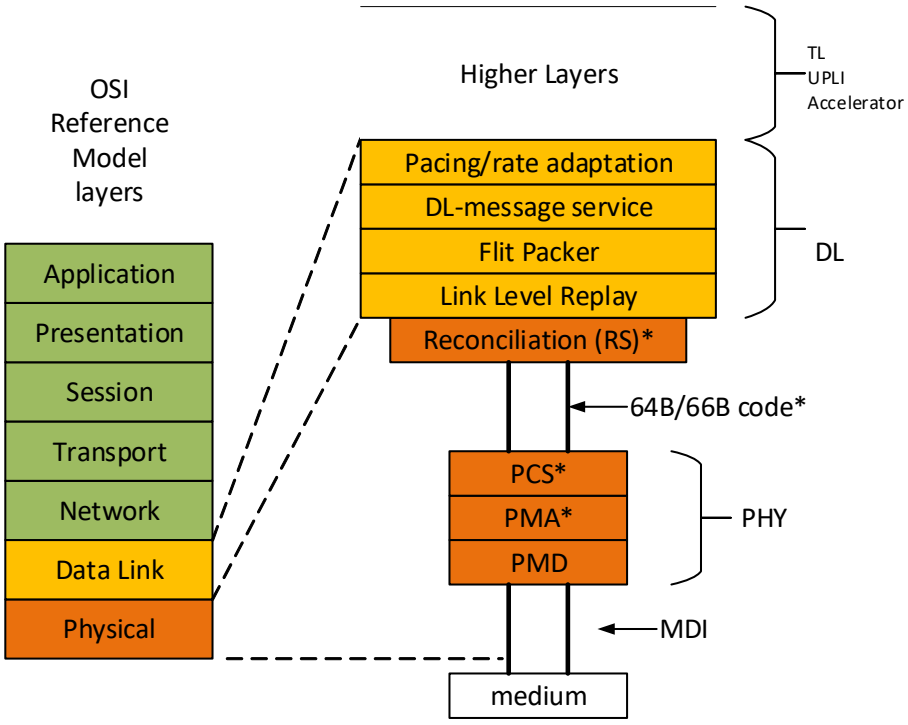

Figure 5: Physical Layer

Data Link Layer (DL)

The Data Link sits between the Transaction layer and the Physical Layer. The Data Link packs 64-byte Flits from the transaction layer into 640 Bytes Flits for the Physical Layer. The Data Link also provides a message service between link partners that originates and terminates at the Data Link layer. The message service is used for advertising the Transaction Layer rate, querying device and port ID on connected Link Partner, and other functions. The message service also provides a UART (Universal Anonymous Receiver Transmitter) style communication between link partners, intended for F/W (Firmware) communications. The Link level replay is provided on a 640 Byte Flit basis. A 32-bit CRC is computed, checked, and included as part of the 640 Byte Flit.

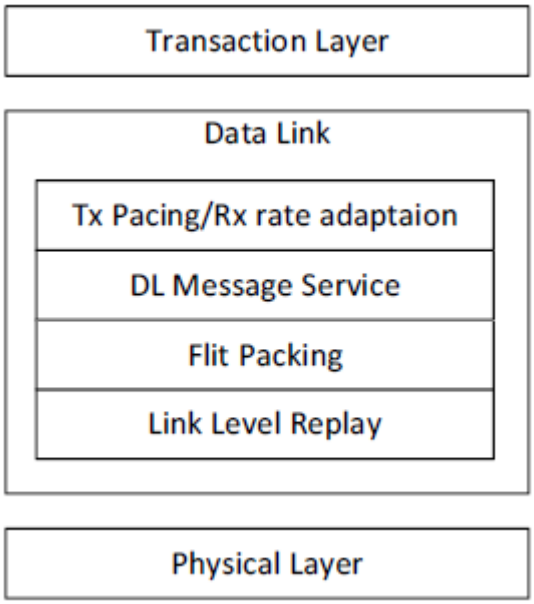

Figure 6: Data Link Layer Block Diagram

Transaction Layer (TL)

The Transaction Layer (TL) is responsible for converting protocol messages from the inbound channels of the two UPLI interfaces (Originator/Completer) into outbound (TX) TL Flits. The TL also converts TL Flits received from the inbound (RX) DL back into UPLI messages on the UPLI interfaces.

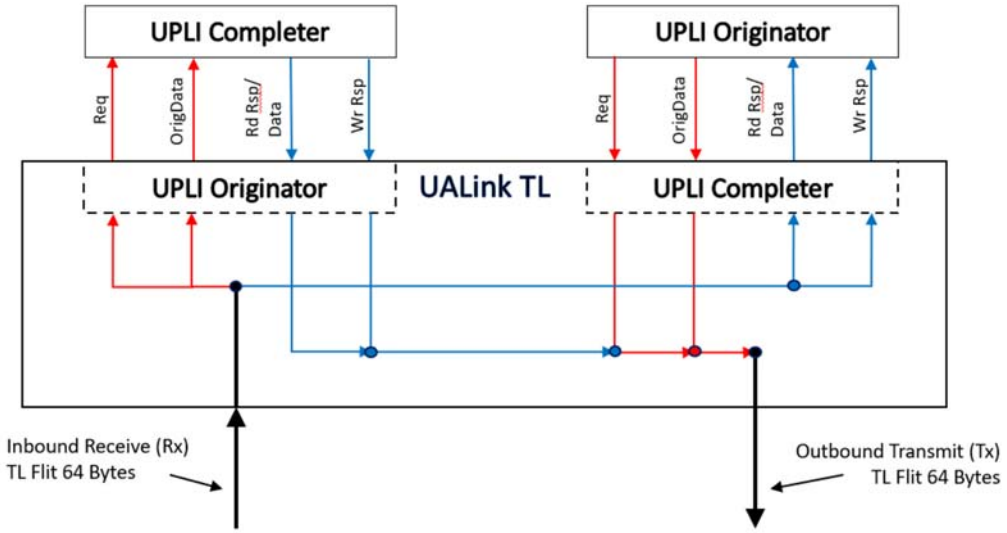

Figure 7: Transaction Layer Interfaces

Figure 7 schematically illustrates the TL Interfaces and their connections between the various channels for the two UPLI interfaces. It also illustrates the relationship to the Tx and Rx TL Flit Channels. Due to the symmetry of the interfaces, the format for both the Receive and Transmit Flits is identical. Each 64-byte Transmit Flit and Receive Flit Channel encodes the information for a UPLI Request Channel, Originator Data Channel, Read Response/Data Channel, and Write Response Channel.

The Transaction Layer supports a streaming address cache to compress addresses. This enables TL to achieve a very high bidirectional protocol efficiency where requests and completions are transferred in both TX and RX directions.

Figure 8 illustrates an example where five write requests, five write completions, and one Flow Control packet are sent over 21 TL Flits achieving an efficiency of 95.2% (20/21). A compressed request header is 8 Bytes, and a compressed write response and flow control each are 4 Bytes.

	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.Data6							

Figure 8: Write requests with max payload (256B) and completions packing

Security

The UALink security feature, referred to as UALinkSec, is intended to protect traffic on a UALink network and switches from a physical adversary; the adversary might be present at the time of the attack or may have placed a device (e.g., an interposer) to snoop or tamper with the UALink traffic. Additionally, in platforms that support Confidential Computing (CC), UALinkSec protects the Tenant data on a UALink network and switches from the infrastructure provider and other Tenants co-located on the same UALink Pod. CC implies that a Trusted Execution Environment (TEE) (e.g., Intel TDX, AMD SEV and ARM CCA) exists on the platform. The TEE under the control of the Tenant is responsible for UALinkSec configuration. When enabled, UALinkSec provides data confidentiality and optional data integrity (including replay protection).

UALinkSec supports encryption and authentication of all the UPLI protocol channels – requests, read responses, and write responses.

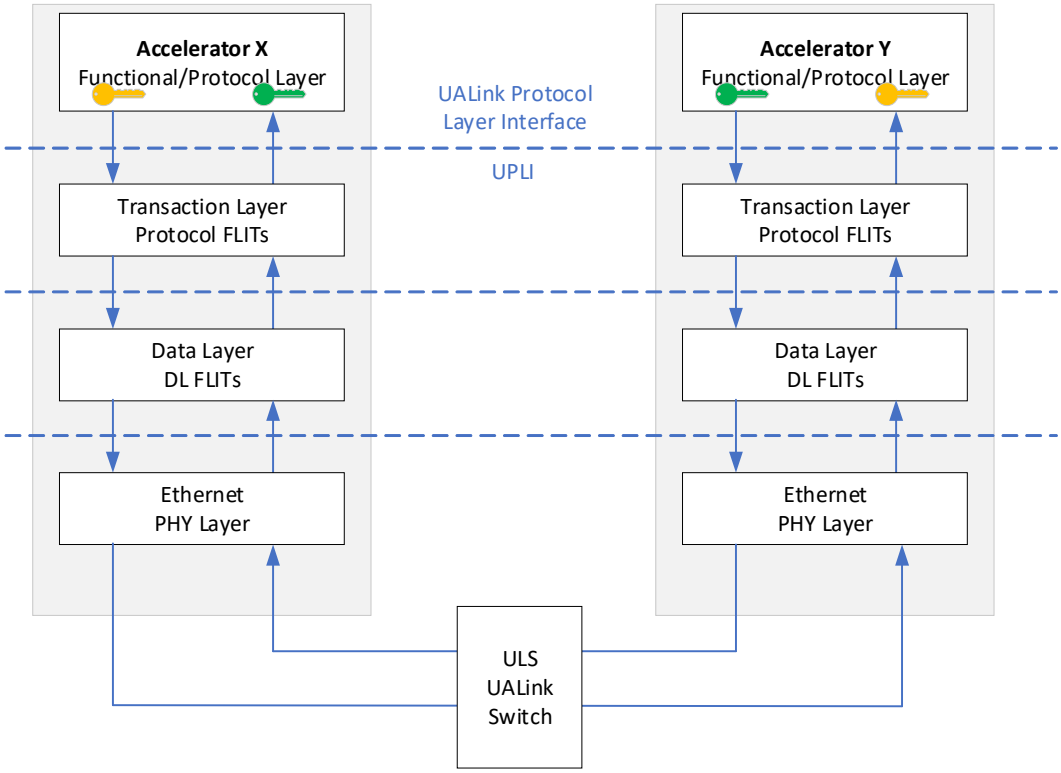

Figure 9: End-to-End encryption and authentication

Manageability

A UALink Pod is comprised of one or more UALink Accelerators connected via the UALink network and managed by a UALink Pod Controller. Each Accelerator is hosted on a UALink System Node, and Accelerator traffic may be routed through the UALink fabric via UALink Switches. Multiple UALink Pods may be connected to create even larger Accelerator clusters; however, inter-Pod communication and control is not specified by the UALink specification.

Figure 10 illustrates a possible UALink Pod set up with Switch and Server Platforms. A System node with four UALink Accelerators (each with three ports), two host CPUs, a NIC, and a BMC is shown. The UALink Switches are responsible for routing traffic between Accelerators and are managed by software/firmware called a Switch Management Agent. Physical Switches are implemented in hardware but may be partitioned into multiple Switches for the purpose of routing and creating Virtual Pods. Often, Physical Switches are hosted on Switch Platforms that run the Switch Management Agent on a processor (such as an x86 CPU or a Baseboard Management Controller). The processor is attached to each Physical Switch via a high-speed interface such as PCIe®, as well as to a network interface for communication with the Pod Controller.

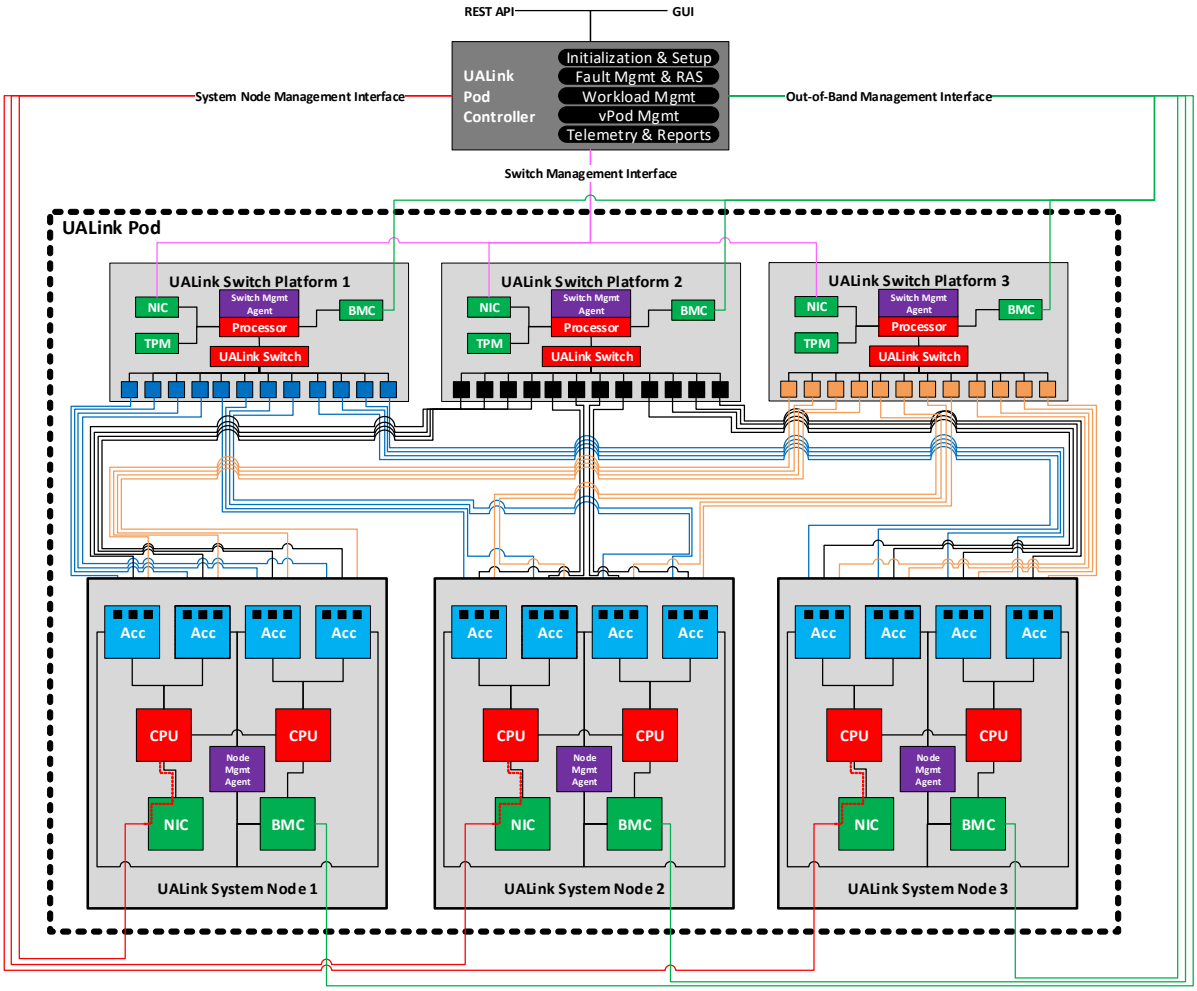

Figure 10: UALink Rack Scale system management interface

Summary

UALink is an open industry standard scale-up interconnect set to transform data center connectivity. Tailored for ultra-high bandwidth and low-latency exchanges among homogeneous accelerated computing devices, UALink is key to unlocking a new epoch of compute efficiency.

UALink is at the vanguard of innovation in the Artificial Intelligence and Machine Learning domains, providing an open ecosystem path to a dedicated accelerator interconnect leveraging the ubiquitous Ethernet ecosystem. By incorporating UALink Switches, accelerators with UALink capability, can expand the scale-up domain, creating ultra-high bandwidth multi-node Accelerator Pods. UALink also enables a simple software model by supporting load/store operations across an entire Pod of up to 1024 accelerators. The technologies developed by the UALink Consortium will have a lasting impact, improving performance, ease of use, and the TCO of demanding AI and HPC applications of the future.

About the Ultra Accelerator Link Consortium

The Ultra Accelerator Link (UALink) Consortium, incorporated in October 2024, is an open industry standard group dedicated to developing the UALink specifications, a high-speed, scale-up accelerator interconnect technology that advances next-generation AI & HPC cluster performance. The consortium is led by a board made up of stalwarts of the industry; Alibaba, AMD, Apple, Astera Labs, AWS, Cisco, Google, HPE, Intel, Meta, Microsoft, and Synopsys. The Consortium develops technical specifications that facilitate breakthrough performance for emerging AI usage models while supporting an open ecosystem for data center accelerators.

For more information on the UALink Consortium, please visit www.UALinkConsortium.org

Interested to Contribute?
Join the UALink Consortium to participate in the technical working groups and influence the direction of the UALink specification. Learn more about UALink membership here.

ULTRA ACCELERATOR LINK™