What Designers Need to Know About UALink for Scalable AI Systems

Author

Diwakar Kumaraswamy
Sr Staff Technical Product Management,
Synopsys

As the computational requirements of AI models rapidly increase—often doubling in scope within mere months—the traditional approaches to connecting accelerators, such as GPUs, are struggling to keep up with the demand for bandwidth, latency, and efficient resource sharing.

Existing interconnect technologies were not originally designed to support tens of thousands of accelerators working in tandem, each requiring rapid access to vast, shared pools of memory. This is particularly evident in scenarios like the pre-training of Llama3, where over 700TB of memory and 16,000 accelerators are necessary. To overcome the limitations of previous interconnects, UALink enables high-speed, low-latency communication directly between accelerators, allowing them to function as a unified system.

By introducing UALink, the industry is paving the way for seamless GPU-to-GPU communication, effective memory pooling, and flexible scaling beyond the constraints of individual hardware boundaries. In doing so, UALink is essential for supporting the next generation of AI models, ensuring that performance and efficiency grow in step with ever-expanding computational ambitions.

This whitepaper will explore the fundamentals and architectural considerations of UALink, focusing on how it enables scalable AI systems through efficient GPU integration, memory pooling, and high-bandwidth communication. Readers will gain insights into the protocol's stack design, transaction and data link layers, and practical implementation strategies, as well as learn about the unique optimizations and features of Synopsys' UALink solution for building large-scale, high-performance AI clusters.

Understanding AI Engine Components and Scaling Strategies

To understand scale up fabrics, let's analyze the common architecture of the AI engines, which consists of scalar and vector units for arithmetic and logic operations, a memory interface for accessing the main memory, and instruction fetch and decode engines, as shown on Figure 1.

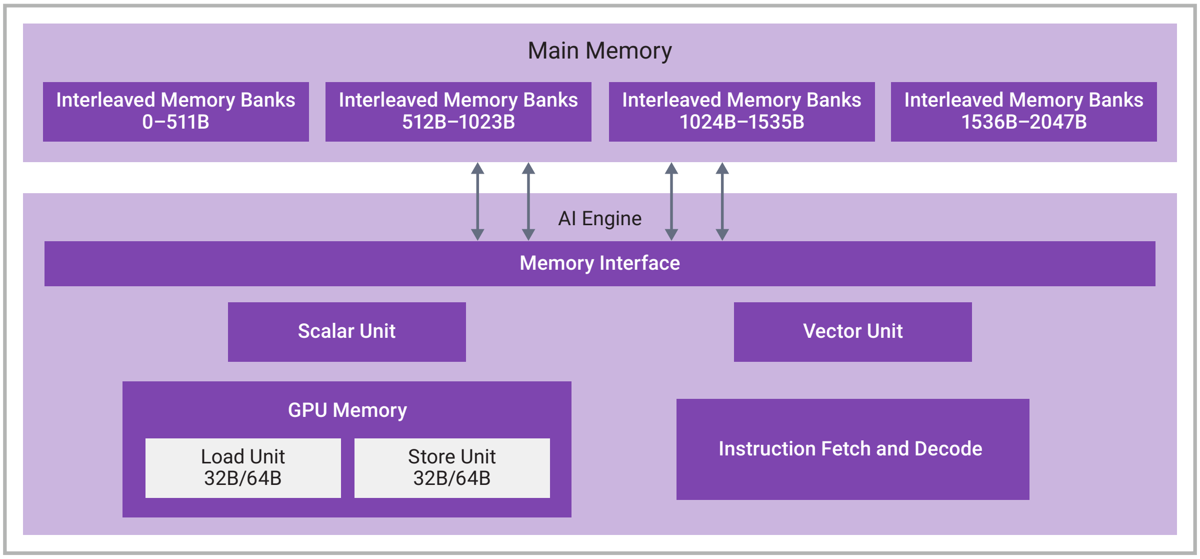
Main Memory
Interleaved Memory Banks 0–511B
Interleaved Memory Banks 512B–1023B
Interleaved Memory Banks 1024B–1535B
Interleaved Memory Banks 1536B–2047B
AI Engine
Memory Interface
Scalar Unit
Vector Unit
GPU Memory
Load Unit 32B/64B
Store Unit 32B/64B
Instruction Fetch and Decode

Figure 1: Typical AI Engine Architecture Featuring Scalar and Vector Units, Memory Interface, and Instruction Fetch/Decode Engines

The AI engine retrieves data from main memory using the Load unit and stores it in registers for processing. Once processed, the Store unit transfers the data back to memory. Load and store instructions define this process supporting 32B or 64B operations, with memory typically interleaved across 256B or 512B boundaries. This is a common workflow for AI engines performing basic computations.

Combining multiple GPUs increases overall computational capacity. If each GPU accesses interleaved memory channels concurrently, memory bandwidth utilization improves. Effective pooling of GPU memory helps address capacity limitations and supports AI system scalability, as illustrated in Figure 2.

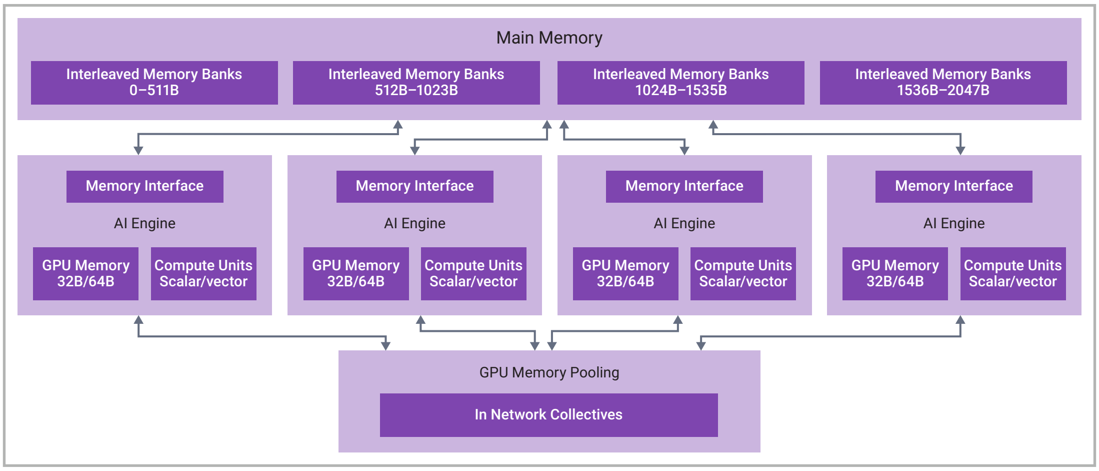
Main Memory
Interleaved Memory Banks 0–511B
Interleaved Memory Banks 512B–1023B
Interleaved Memory Banks 1024B–1535B
Interleaved Memory Banks 1536B–2047B
Memory Interface
AI Engine
GPU Memory 32B/64B
Compute Units Scalar/vector
Memory Interface
AI Engine
GPU Memory 32B/64B
Compute Units Scalar/vector
Memory Interface
AI Engine
GPU Memory 32B/64B
Compute Units Scalar/vector
Memory Interface
AI Engine
GPU Memory 32B/64B
Compute Units Scalar/vector
GPU Memory Pooling
In Network Collectives

Figure 2: Scaling AI Systems Through GPU Integration and Memory Pooling

Minimizing Overhead in GPU Communication

As AI workloads continue to scale, it is essential to establish links between GPUs that support efficient communication with minimal overhead. These interconnects must mimic the load and store semantics familiar to main memory access, ensuring that data movement between GPUs is both rapid and predictable. This approach lays the foundation for high-performance, low-latency operations crucial for large-scale AI deployments to meet these stringent requirements, the UALink Protocol was developed as a purpose-built solution. Designed from the ground up, UALink enables accelerators to exchange data seamlessly, maintaining low latency and high throughput. Its architecture is specifically tailored to support the unique demands of scalable AI systems, focusing on efficient resource sharing and reliable communication.

GPU Organization in Pods and Racks

In typical AI system deployments, GPUs are organized into pods or racks. Within these configurations, clusters of GPUs are interconnected through switches. These switches leverage UALink connections to facilitate fast and effective data exchange, ensuring that each GPU within the cluster can communicate without bottlenecks or excessive overhead to the other GPU. By adopting this approach, AI clusters can achieve optimal performance and scalability, supporting the evolving computational needs of advanced models.

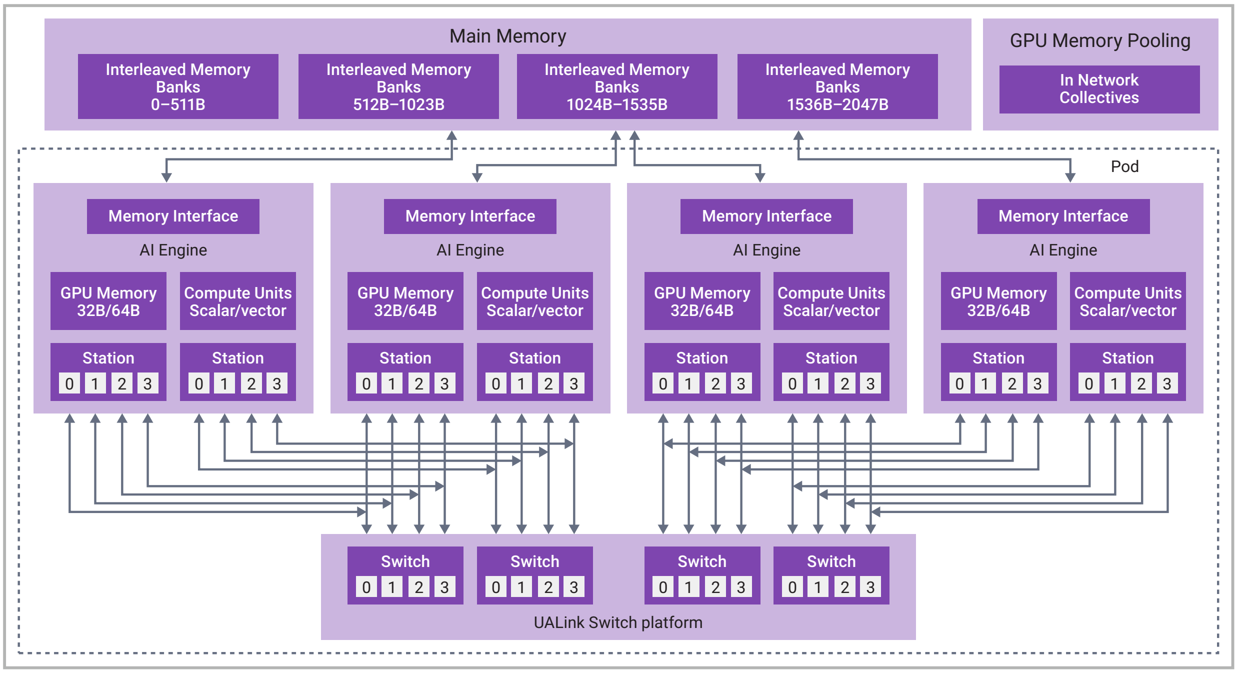
Main Memory
Interleaved Memory Banks 0–511B
Interleaved Memory Banks 512B–1023B
Interleaved Memory Banks 1024B–1535B
Interleaved Memory Banks 1536B–2047B
GPU Memory Pooling
In Network Collectives
Pod
Memory Interface
AI Engine
GPU Memory 32B/64B
Compute Units Scalar/vector
Station 0 1 2 3
Station 0 1 2 3
Switch 0 1 2 3
UALink Switch platform

Figure 3: High-Performance GPU Clusters Connected by UALink

Optimized Data Transfer and Transaction Layer Design in UALink

UALink is engineered with a fixed 64-byte packet size, ensuring efficient data movement between accelerators. This design leverages the fastest SerDes technology currently available—224G—based on the Ethernet protocol. Since the physical layer and SerDes are defined by the IEEE 802.3dj specification, which leaves little room for further optimization, the UALink consortium has concentrated its efforts on refining the upper layers of the protocol stack.

Among these, the Transaction Layer (TL) has been particularly optimized for packing efficiency. The TL is responsible for bundling both requests and responses, and it communicates directly with the accelerator's UPLI interface. Its lightweight implementation and latency-focused design make it highly effective for high-performance AI workloads.

The UPLI interface features a symmetric architecture, consisting of two components: the "Originator" and the "Completer." These components integrate directly with the UALink stack, allowing seamless and efficient communication between accelerators as illustrated in the accompanying figure.

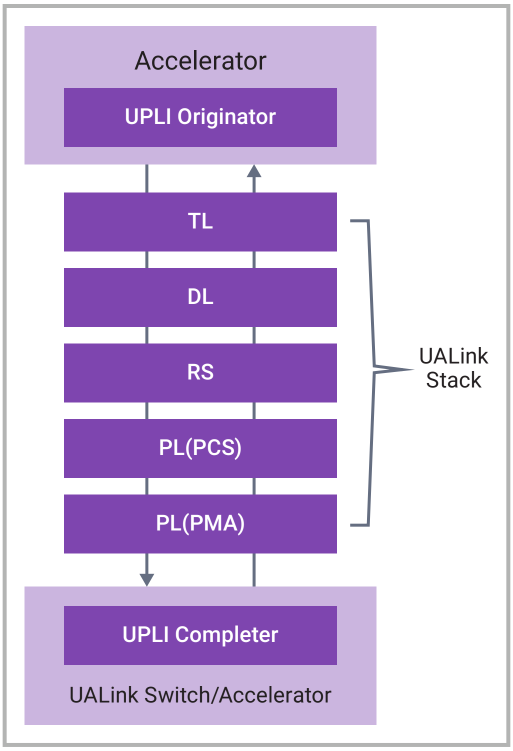

Figure 4: UALink Protocol Stack and Accelerator Interface Overview

UALink Protocol Stack: Layered Architecture and Key Features

The UALink protocol stack is meticulously designed to deliver high-performance, reliable, and secure communication between accelerators in large-scale AI systems. Each layer of the stack plays a specific role, from initiating and managing transactions to ensuring efficient data transfer and robust error handling. The following sections provide an overview of the core protocol layers—UPLI/Protocol Layer, Transaction Layer, and Data Link Layer—highlighting their primary functions and unique optimizations that enable scalable and efficient accelerator interconnects.

Transaction Initiation and Protocol Operations

The Originator device is responsible for initiating transactions within the UALink Protocol by issuing requests such as Read, Write, Atomic operations, UPLI Write Messages, or Vendor Defined Command Requests. Each UPLI port utilizes transaction identification tags, enabling the support of multiple outstanding requests on the interface port. Additionally, the protocol allows for Split Requests and Responses, which enhances interface bandwidth utilization. Also by allowing the requests and responses to be packed together the protocol allows for high efficiency and low latency communication.

Virtual Channels and Source-Ordered Protocol

UALink supports up to four virtual channels allowing independent flow control of the channels with dedicated and pool credits for prioritization of the traffic. Switches in the system use an “ID” field to route transactions from the source accelerator to the destination accelerator. UALink is mostly source ordered protocol, completions for all the requests (read, writes and atomics) can be unordered. For requests targeting the same address from source to destination, switches are required to maintain ordering , ensuring consistency and reliability in data delivery. The exception being, when encryption and authentication is enabled, where the packets within the stream is expected to be in order from source to destination.

Addressing and Security Features

The protocol defines a 57-bit address field, providing implementers with significant flexibility to construct different address spaces tailored to their requirements. Furthermore, UALink supports a comprehensive Confidential Compute solution, delivering end-to-end encryption to safeguard data throughout its journey within the AI system.

Transaction Layer (TL) Flit Structure and Compression Mechanisms

The Transaction Layer (TL) within the UALink protocol is designed with a focus on maximizing efficiency and clarity in data management. Each TL flit is fixed at 64 bytes in size and is segmented into sixteen distinct 4-byte sectors. These sectors are organized into two main portions: the upper half flit, comprising 32 bytes, and the lower half flit, also comprising 32 bytes. This dual-half organization allows the protocol to differentiate the handling of control and data information within each flit.

Control flits are specifically confined to the lower half flit. By isolating control information in this manner, UALink ensures that protocol operations and signaling are managed in a dedicated and streamlined section of the packet. This separation supports efficient protocol management and reduces the potential for contention between control and data traffic.

Conversely, data flits are afforded greater flexibility in their placement. They can reside in either the lower or the upper half flit, depending on the needs of the data transfer at any given moment. This adaptable organization supports the efficient movement of data across the interconnect, ensuring that bandwidth is utilized effectively and that data transfers remain agile to the demands of large-scale AI workloads.

To further enhance packing efficiency, the TL incorporates compression mechanisms for both requests and responses. Compressed requests are reduced to 8 bytes, compared to their uncompressed size of 16 bytes. Similarly, compressed responses are condensed to 4 bytes, in contrast to the 8-byte size of uncompressed responses. This selective compression strategy is supported by an address cache implemented within the TL, which aids in achieving better utilization of the available bandwidth. Notably, while responses can always be compressed, requests are not always eligible for compression and may remain uncompressed in certain scenarios like when there is no cache hit or compression being supported by the link partner

Data Link Layer Data Link Flit Aggregation and Efficiency

Within the UALink protocol, each Transaction Layer (TL) flit—fixed at 64 bytes—is aggregated into a larger Data Link (DL) flit measuring 640 bytes. This aggregation introduces a total overhead of 12 bytes, which consists of a 3-byte header, a 5-byte segment header, and a 4-byte cyclic redundancy check (CRC). As a result, the overall data transfer efficiency achieved at this stage is 98.125%.

Physical Coding Sublayer (PCS) Encoding and Forward Error Correction

After aggregation, the 640-byte DL flit undergoes further processing at the Physical Coding Sublayer (PCS), where it is encoded into a 680-byte forward error correction (FEC) code. This step adds additional bytes specifically for error correction purposes, enhancing the integrity of data transmission across the interconnect.

Reliability and Flow Control Mechanisms

The Data Link layer is instrumental in ensuring reliable and lossless data transmission. It employs a Link Level Retry mechanism to guarantee that data is delivered without errors. Additionally, credit-based flow control is implemented to efficiently manage the data exchange between the Originator and Completer interfaces, helping to optimize bandwidth utilization and prevent congestion.

Additional Service Features

Beyond its core responsibilities, the Data Link layer supports both port-level and station resets to maintain system stability. It also provides a UART-style message service, which enables in-band queries and key exchanges between link partners. These services are managed through dedicated firmware, contributing to the overall robustness and flexibility of the UALink protocol. These DL messages originate and terminate at the DL layer.

Together, these architectural innovations and protocol optimizations make UALink a foundational technology for building the next generation of scalable AI systems. By enabling high-bandwidth, low-latency, and secure communication across thousands of accelerators, UALink addresses the pressing challenges of memory pooling, data movement, and efficient resource sharing in large-scale deployments.

Synopsys UALinkSec Security Module

The protection of data traveling across the UALink interface is paramount due to the sensitive nature of the information being transmitted. Ensuring data security involves implementing robust encryption and authentication mechanisms to safeguard against unauthorized access and data breaches. Encryption is mandatory to maintain confidentiality, while authentication, though optional, enhances the security framework by verifying the integrity and origin of the data. These measures are crucial to prevent data from being intercepted or tampered with during transit.

Synopsys UALinkSec_200 Security Module aligns with the standard and is specifically designed to address the security needs of data traveling across UALink interfaces. This module provides highly efficient encryption and decryption capabilities based on the AES-GCM cryptographic algorithm, ensuring data remains secure throughout its journey. Additionally, it includes key derivation functionality and optional authentication support, allowing for flexible security configurations tailored to specific requirements. By using this security module, which comes pre-validated with the Synopsys UALink controller, data centers can achieve failsafe secure operations, minimizing the risk of data exposure and enhancing overall system integrity.

Synopsys UALink IP Solution

The Synopsys IP UALink solution comprises silicon-proven 224G PHY, Controller, Security, and Verification IP. This offering incorporates the UALinkSec_200 Security Module, the UALink Controller, and a 800G Quad PCS and an advanced four-lane 224G PHY, providing secure high-throughput, low-latency connectivity tailored for AI cluster and memory-semantic fabric applications. The UALink Controller manages protocol layer operations and supports multi-channel aggregation via four Nx66b interfaces that integrate efficiently with the 800G PCS logic.The PCS block features optimized Reed-Solomon Forward Error Correction (RS-FEC) engine which provides the industry best latency profile ensuring robust signal integrity for extended board-level connections.

And our 224G PHY built on silicon-proven technology with wide industry interoperability, is engineered for ultra-low jitter, superior channel performance, and power-efficiency across extended temperature and voltage ranges—enabling next-generation UALink deployments targeting AI and HPC disaggregated architectures. This co-optimized PHY—PCS—controller stack forms a complete, silicon-proven UALink interface IP subsystem, providing customers with a high-performance, low-risk path to first-pass silicon success in large-scale AI network systems. The UALink Controller supports a single x4 link supporting 800G throughput or two bifurcated x2 links like each supporting 400G or 4 bifurcated x1 links each supporting 200G as per the Protocol. And our UALink solution can be used with both Switches and Accelerators to establish the high performance and low latency UALink connectivity.

In addition, Synopsys UALink Verification IP (VIP) delivers comprehensive, protocol-compliant verification for UALink-based interconnects, enabling designers to validate advanced features essential for high-bandwidth, low-latency connectivity in AI, HPC, and data center environments. Developed in close collaboration with industry partners, UALink VIP introduces new interfaces for seamless access to the UALink protocol stack. With robust protocol checks, configurable traffic generation, and integrated coverage models, this solution accelerates verification closure and ensures reliable operation across large-scale AI network systems.

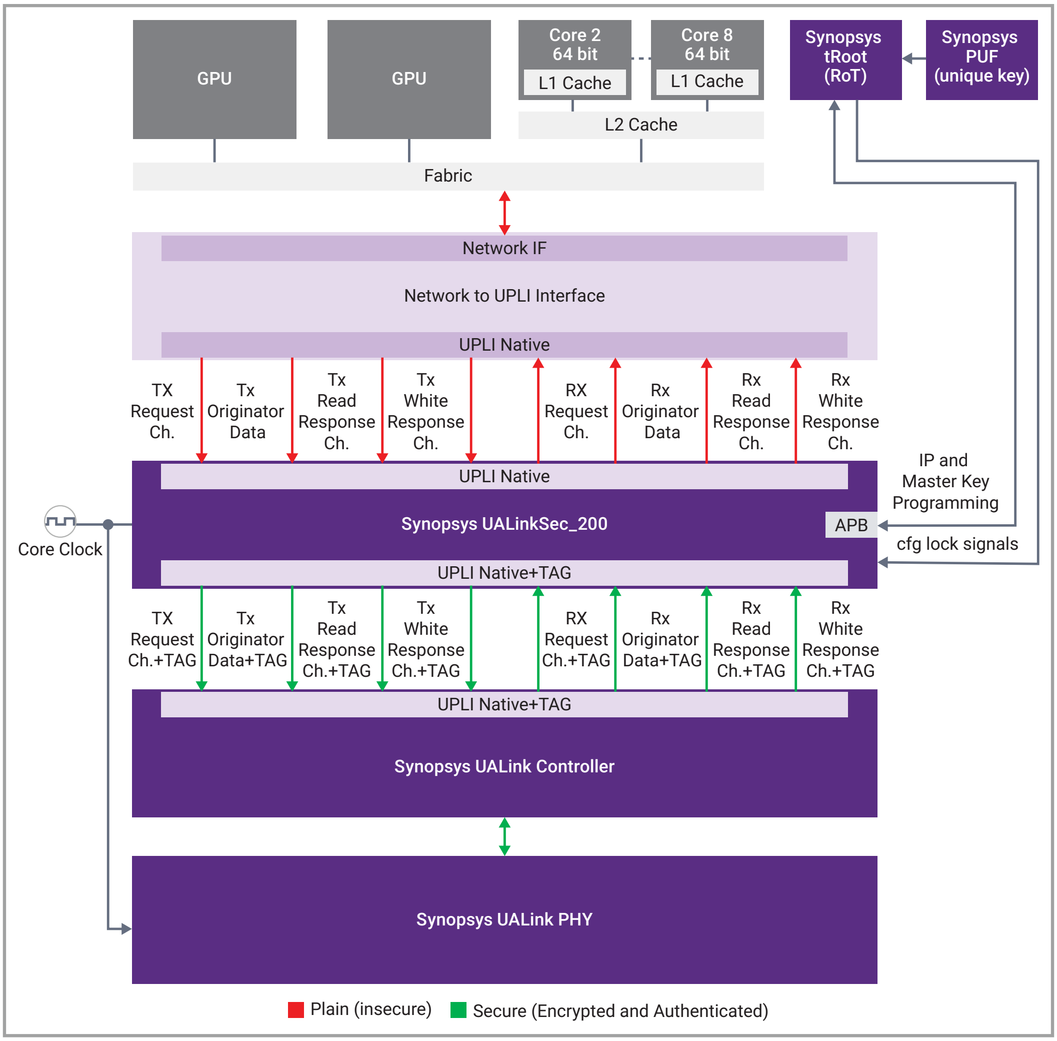

Figure 5: End-to-End UALink IP Solution

Summary

UALink is engineered specifically to address the unique needs of modern, scalable AI infrastructure. Its architecture is optimized for integrating GPUs efficiently, supporting robust memory pooling, and facilitating secure high-bandwidth, low-latency communication across expansive arrays of accelerators. This level of integration ensures that as AI workloads and deployments grow in scale, the interconnect infrastructure remains both agile and capable of meeting performance demands.

At the core of UALink's effectiveness is its comprehensive protocol stack, which includes the UPLI/Protocol Layer, Transaction Layer, and Data Link Layer. Together, these layers enable reliable, flexible, and secure data exchange between system components. The protocol stack is meticulously designed to ensure that data flows seamlessly, supporting the demands of advanced AI systems where reliability and security are paramount.Through its advanced features and purpose-built design, UALink is positioned as a foundational technology for next-generation AI platforms. By supporting efficient GPU integration, scalable resource sharing, and secure communication, UALink empowers organizations to build and deploy advanced AI solutions with confidence, meeting the pressing requirements of today's and tomorrow's large-scale AI applications.

Synopsys leads the industry in providing a complete UALink IP solution, which has been adopted by multiple organizations for large-scale switches and AI accelerators. Featuring silicon-proven PHY, Controller, Security, and Verification IP, Synopsys is establishing the benchmark for future AI platforms. Learn more about how Synopsys UALink IP can accelerate your AI deployments.