

# Overview of and Motivation for the Forthcoming Ultra Ethernet Consortium Specification

## Networking Demands of Modern AI Jobs

Networking is increasingly important for efficient and cost-effective training of AI models. Large Language Models (LLMs) such as GPT-3, Chinchilla, and PALM, as well as recommendation systems like DLRM and DHEN, are trained on clusters of thousands of GPUs.

Training consists of frequent computation and communications phases, where the initiation of the next phase of the training is dependent on the completion of the communication phase across the suite of GPUs. The last message to arrive gates the progress of all GPUs. This *tail latency* – measured by the arrival time of the last message in the communication phase – is a critical metric in system performance.

The size of large models in terms of the number of parameters, entries of embedding tables, and words of context buffers continue to increase. For example, in 2020, GPT-3 was state-of-the-art with 175 billion parameters. Recently, a GPT-4 model was announced with an expected one trillion parameters, while DLRM is at many trillion parameters and is expected to grow. These increasingly large models require increasingly large clusters for training and drive larger messages on the network. When the network underperforms, these expensive clusters are underutilized. The network interconnecting these compute resources must be as efficient and cost-effective as possible.

High-Performance Computing (HPC) jobs are similarly demanding, and there is an increased convergence between the needs of HPC and AI with respect to scale and efficient use of distributed computing resources. While AI workloads are typically extremely bandwidth hungry, HPC also includes workloads that are more latency sensitive.

## The Ethernet Advantage

Currently, many large clusters including hyperscale deployments of GPUs used for AI training are already operating on Ethernet-based IP networks, leveraging their many advantages:

- a broad, multi-vendor ecosystem of interoperable Ethernet switches, NICs, cables, transceivers, optics, management tools and software from many participating parties
- proven addressing and routing scale of IP networks, enabling rack-scale, building-scale, and datacenter-scale networks
- a spectrum of tools for testing, measuring, deploying, and efficiently operating Ethernet networks
- proven history of driving down costs through a competitive ecosystem and economies of scale

- proven ability of the IEEE Ethernet standards to advance rapidly and regularly across many physical and optical layers

We expect these advantages to become table-stakes requirements, and that Ethernet networks will increasingly dominate AI and HPC workloads of all sizes in the future.

## Key Needs of AI and HPC Networks of the Future

Even when considering the advantages of using Ethernet, improvements can and should be made. Networks must evolve to better deliver this unprecedented performance for the increased scale and higher bandwidth of networks of the future. Paramount is the need to have the network support delivery of messages to all participating endpoints as quickly as possible, without long delays for even a few endpoints. “Tail latency” should be minimized.

To achieve low tail latency, the UEC specification offers significant improvements by addressing the following critical networking requirements for the next generation of applications:

- Multi-pathing and packet spraying
- Flexible delivery order
- Modern congestion control mechanisms
- End-to-end telemetry
- Larger scale, stability, and reliability

This last point places an extra burden on all of the previous ones. High-performance systems leave little margin for error, which compounds in a larger network. Determinism and predictability become more difficult as systems grow, necessitating new methods to achieve holistic stability.

In the following sections, as motivation for the solutions proposed by Ultra Ethernet Consortium, we elaborate on each of these needs and show how the currently available technologies have deficiencies that must be addressed. We seek to provide simpler and more efficient remote direct memory access (RDMA) and interconnection for these future workloads.

### Multi-Pathing and Packet Spraying

Traditional Ethernet networking was based on a spanning tree, ensuring one path from A to B to avoid loops in the network. Then came multi-pathing – technologies like Equal-Cost Multipath (ECMP) in which the network attempts to leverage as many links as possible between communicating partners. ECMP typically uses a “flow hash” to send all traffic of a given layer-four flow on one path while mapping different flows to different paths. However, this still confines a high throughput flow to one path. Further, network performance degrades when the multi-pathing technology maps too many flows map to a single network path, and careful management of load balancing is required for best performance. The next phase in the technology evolution is for every flow to simultaneously use all paths to the destination – a technique known as “packet spraying” – achieving a more balanced use of all network paths.

## Flexible Ordering

The rigid packet ordering used by older technologies (e.g., as required by the Verbs API) limits efficiency by preventing out-of-order packet data from being delivered straight from the network to the application buffer, i.e., its final location in host memory. This constraint, along with Go-Back-N packet loss recovery (which forces the re-transmission of up to N packets for a single lost packet), causes under-utilization of the available links and increased tail latencies – inadequate for large-scale AI applications. Ideally, all links are used, and order is enforced only when the AI workload requires it.

Much of the inter-accelerator communication in AI workloads is part of a “collective” communication operation, where All-Reduce and All-to-All are the dominant collective types. Key to their rapid completion is a fast bulk transfer from A to B, where the AI application is only interested in knowing when the last part of a given message has arrived at the destination. Flexible ordering enables this to be done efficiently. It likewise enables the benefits of packet spraying in bandwidth-intensive collective operations by eliminating the need to reorder packets before delivering them to the application. Support for modern APIs that relax the packet-by-packet ordering requirements when application-appropriate is critical to curtail tail latencies.

## AI and HPC Optimized Congestion Control

Network congestion can occur in three places:

- the outgoing link from the sender to the first switch
- the links between the first switch and the last switch
- the final link between the last switch and the receiver

For AI and HPC, congestion on the outgoing link from the sender is primarily controllable through scheduling algorithms on the sending host, which has visibility to all outgoing traffic. Multipath packet spraying described above minimizes hotspots and congestion between the first and last switch by evenly spreading the load on all paths. The final form of congestion – “Incast” – occurs on the last link to the receiver when multiple senders simultaneously send traffic to the same destination; it can occur as part of an “All-to-All” communication mentioned above.

In recent decades, many proposals for addressing congestion have been made (e.g., DCQCN, DCTCP, SWIFT, Timely). None of the current algorithms, however, meet all the needs of a transport protocol optimized for AI, which are:

- ramping quickly to wire rate in a high-speed, low round-trip-time network where there is an uncongested path, without reducing the performance of existing traffic
- managing path congestion in the fabric and on the last hop to the destination
- controlling incast by fairly sharing the final link without resulting in expensive packet loss, retransmission, or increased tail latency

- not requiring tuning and configuration as traffic mix changes, compute nodes evolve, link speeds increase, and networking hardware evolves

A congestion control algorithm for AI workloads of the future must be designed to both support these requirements and to work in conjunction with multipath packet spraying.

## End-to-End Telemetry

These optimized congestion control algorithms are enabled by emerging end-to-end telemetry schemes. Congestion information originating from the network can advise the participants of the location and cause of the congestion. Shortening the congestion signaling path and providing more information to the endpoints allows more responsive congestion control. Whether the sender or receiver schedules the transmission, modern switches can facilitate responsive congestion control by rapidly transferring accurate congestion information to the scheduler or pacer – improving the responsiveness and accuracy of the congestion control algorithm. The result is reduced congestion, fewer dropped packets, and smaller queues - all in service of improved tail latency.

## The Success of RDMA and Its Limitations – a Case to Restart

As AI models increase in size, diversity of communication patterns, and variety of computational methods, it is time to revisit the transport and APIs employed at the core of most AI networks. Generically, remote direct memory access (RDMA) has been a very successful technology for allowing a CPU, GPU, TPU, or other accelerator to transfer data directly from the sender's memory to the receiver's memory. This *zero-copy* approach results in low latency and avoids operating system overheads. Because of this, network technology that supports RDMA is a fundamental component of AI training jobs today.

RDMA over Converged Ethernet, or RoCE, was created to allow the IBTA's (InfiniBand™ Trade Association) transport protocol for RDMA to run on IP and Ethernet networks. That underlying protocol, expressed through the Verbs API, was envisioned at the end of the last century and first standardized by IBTA many years ago. It is now showing its age for modern, highly demanding AI network traffic see [*Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale*, Hoefler et al., in *Computer*, July 2023]. The issue is not with the generic RDMA *principles* of operating system bypass and zero-copy, nor with the use of an Ethernet network, but rather with the current transport protocol services common to both RoCE<sup>1</sup> and InfiniBand.

A single accelerator, at the time of this writing, may integrate multiple terabits of network I/O, and PCIe NICs will soon deliver 800 gigabits per second and beyond – orders of magnitude faster than when RDMA was first envisioned. Tomorrow's more-demanding and higher-speed networks will further test the status quo and require new solutions.

---

<sup>1</sup> Further references of RoCE in this document pertain to RoCEv2

It is common for RoCE to be used with DCQCN as the congestion control algorithm to avoid overrunning links in the network while attempting to ramp its rate quickly. However, DCQCN requires careful manual tuning for performance. DCQCN tuning is sensitive to the latency, speed, and buffering capabilities of the network beneath it, as well as the nature of the workloads communicating over it. A great success of the TCP/IP protocol suite that powers the Internet is that TCP does not need to be tuned for the network and that it “just works.” AI networks of the future need a transport protocol that, like TCP, works “out of the box” for any data center network.

It is well known that, while the RDMA transport used in InfiniBand and RoCE can handle lost packets, it does so very inefficiently. A lost or out-of-order packet results in “Go-Back-N” recovery in which already-received packets are retransmitted, resulting in lower “goodput” and poor efficiency. Network operators frequently run RDMA over a “lossless” network to avoid triggering this behavior. Ethernet can be lossless if configured to generate hop-by-hop backpressure using Priority Flow Control (PFC) from the receiver toward the sender in the event of congestion. So rather than dropping a packet, its transmission is delayed at the previous hop. However, when this backpressure propagates through the network, it can produce “congestion trees” and head-of-line blocking; both of which can cause serious performance degradation at scale.

While large lossless RoCE networks can and have been successfully deployed, they require careful tuning, operation, and monitoring to perform well without triggering these effects. This level of investment and expertise is not available to all network operators and leads to a high TCO. A transport protocol that does not depend on a lossless fabric is needed.

Additionally, RoCE, as well as InfiniBand, uses an API (Verbs) designed for much lower scale – both in terms of bandwidth and number of peers – than what is required by modern AI and HPC jobs and future accelerators with integrated networking. The RC (Reliable Connection) transport mode is not well-suited to efficient hardware offload implementation at high speed which requires a reduced fast-path state. While proprietary attempts have been made to address RC’s limitations, none have been broadly accepted, nor have they fully addressed the limits imposed by its inherent Process to Process ( $P^2$ ) scalability issue. While implementations of RC work on a modest scale, they add endpoint cost and complexity that is burdensome for AI jobs at the scale of tomorrow; a new solution is needed.

Lastly, AI applications transfer huge amounts of data. Legacy RoCE transmits this data as a small number of large flows that must be carefully load-balanced to prevent overloading any individual link, as mentioned above. AI workloads often cannot proceed until *all* flows are successfully delivered, and even one overburdened link throttles the entire computation. Improved load-balancing techniques are essential to improve AI performance.

# Ultra Ethernet Transport (UET): a Protocol for Next-Generation AI and HPC Networks

Ultra Ethernet Consortium's members believe it is time to start afresh and replace the legacy RoCE protocol with Ultra Ethernet Transport, a modern transport protocol designed to deliver the performance that AI and HPC applications require while preserving the advantages of the Ethernet/IP ecosystem.

Two fundamental lessons from the success of TCP/IP and Ethernet are that the *transport protocol* should provide loss recovery and that lossless fabrics are very challenging to operate without triggering head-of-line blocking and congestion spreading. Embracing these principles, the UEC transport builds on the proven path of distributed routing algorithms and endpoint-based reliability and congestion control. The UEC transport protocol advances beyond the status quo by providing the following:

- An open protocol specification designed from the start to run over IP and Ethernet
- Multipath, packet-spraying delivery that fully utilizes the AI network without causing congestion or head-of-line blocking, eliminating the need for centralized load-balancing algorithms and route controllers
- Incast management mechanisms that control fan-in on the final link to the destination host with minimal drop
- Efficient rate control algorithms that allow the transport to quickly ramp to wire-rate while not causing performance loss for competing flows
- APIs for out-of-order packet delivery with optional in-order completion of messages, maximizing concurrency in the network and application, and minimizing message latency
- Scale for networks of the future, with support for 1,000,000 endpoints
- Performance and optimal network utilization without requiring congestion algorithm parameter tuning specific to the network and workloads
- Designed to achieve wire-rate performance on commodity hardware at 800G, 1.6T and faster Ethernet networks of the future

The UEC specification will go beyond the transport layer to define standard semantic layers, improved mechanisms for low-latency delivery, and consistent AI and HPC APIs with standard, multi-vendor support for implementing those APIs over the UEC transport protocol.

## Security for AI and HPC

AI training and inference often occur in hosted networks where job isolation is required. Moreover, AI models are increasingly sensitive and valuable business assets. Recognizing this, the UEC transport incorporates network security by design and can encrypt and authenticate all network traffic sent between computation endpoints in an AI training or inference job. The UEC

transport protocol leverages the proven core techniques for efficient session management, authentication, and confidentiality from modern encryption methods like IPSec and PSP<sup>2</sup>.

As jobs grow, it is necessary to support encryption without ballooning the session state in hosts and network interfaces. In service of this, UET incorporates new key management mechanisms that allow efficient sharing of keys among tens of thousands of compute nodes participating in a job. It is, designed to be efficiently implemented at the high speeds and scales required by AI training and inference.

HPC jobs hosted on large Ethernet networks have similar characteristics and require comparable security mechanisms.

## Further Efforts in UEC - HPC and Beyond

In addition to enabling improved networking for AI, UEC is developing technology to support the network needs of High-Performance Computing (HPC) of the future. Going forward, AI and HPC workload and network requirements are anticipated to increasingly overlap. Hence, we expect the UEC transport protocol to serve the networking demands of both AI and HPC jobs. Recognizing the different sensitivities to bandwidth and latency, the UEC specification will offer two profiles – one optimized for AI and another optimized for HPC.

With speeds and scale increasing, the traditional approach of relying only on end-to-end retry is increasingly burdensome for latency-sensitive workloads. Local error handling at the link layer has proven valuable in scale-out HPC networks, such as those used in exascale systems. The UEC specification provides this capability for Ethernet.

## Summary

AI systems are typically deployed on a network topology with many paths from sender to receiver. It is critical to use *all* lanes of this expensive highway simultaneously and efficiently. To make this happen, scalable and efficient remote memory access, will be required, implemented with packet spraying, flexible ordering, and optimized congestion control algorithms. Additionally, new end-to-end telemetry, scalable security, and AI-optimized APIs will be essential for networks optimized for the unique communication needs of the intense AI computations of tomorrow.

UEC protocols are also designed to support modern HPC workloads, leveraging the same transport mechanism outlined above while preserving broadly used APIs such as MPI and PGAS.

---

<sup>2</sup> See [github.com/google/psp](https://github.com/google/psp)



The founding members of UEC include suppliers and operators of many of the largest AI and HPC networks today. UEC's efforts leverage its members' many years of experience building and operating those networks. The forthcoming UEC draft specification will be open to use as an interoperable basis for AI and HPC networks. The technologies under development in UEC will have a lasting impact, improving the performance, ease of use, and cost of the demanding AI and HPC applications of the future.

For further information, please check [www.ultraethernet.org](http://www.ultraethernet.org).

### **About Ultra Ethernet Consortium**

Ultra Ethernet Consortium brings together companies for industry-wide cooperation on interoperability and to build a complete Ethernet based communication stack architecture, that best matches the rapidly evolving AI/HPC workloads at scale and provides for best-in-class functionality, performance, interoperability and TCO as well as developer and end-user friendliness. UEC is a Joint Development Foundation Projects, LLC Series, an affiliate of the Linux Foundation. The founding members include **AMD, Arista, Broadcom, Cisco, Eviden (an Atos Business), HPE, Intel, Meta, Microsoft**. Learn more at [ultraethernet.org](http://ultraethernet.org).

## DISCLAIMER

THESE MATERIALS ARE PROVIDED "AS IS." The parties expressly disclaim any warranties (express, implied, or otherwise), including implied warranties of merchantability, non-infringement, fitness for a particular purpose, or title, related to the materials. The entire risk as to implementing or otherwise using the materials is assumed by the implementer and user. IN NO EVENT WILL THE PARTIES BE LIABLE TO ANY OTHER PARTY FOR LOST PROFITS OR ANY FORM OF INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES OF ANY CHARACTER FROM ANY CAUSES OF ACTION OF ANY KIND WITH RESPECT TO THIS DELIVERABLE OR ITS GOVERNING AGREEMENT, WHETHER BASED ON BREACH OF CONTRACT, TORT (INCLUDING NEGLIGENCE), OR OTHERWISE, AND WHETHER OR NOT THE OTHER MEMBER HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.