Intersect360 Research White Paper: **UEC 1.0: NEW HIGH-PERFORMANCE STANDARD FOR SCALING HPC-AI**

MARKET DYNAMICS: THE NEED FOR A HIGH-PERFORMANCE STANDARD

For decades, Ethernet has been the networking standard underpinning all of enterprise computing. It connects nodes in scalable clusters, ties storage to computing, and links resources on local area networks. Due to its ubiquity, with a wide range of suppliers and features, Ethernet is the first and often only networking choice. According to an old computing industry adage, there are two types of networking: Ethernet and ether-not.

The only segment with any significant outliers bucking the Ethernet trend has been High Performance Computing (HPC). Owing to the need to chase the highest levels of scalability and performance, many organizations have eschewed the Ethernet standard, instead embracing niche or proprietary interconnect solutions for their largest HPC systems.

Historically, these noteworthy HPC supercomputers have mostly belonged to academic or government research centers, publicly funded institutions whose charters were to push the boundaries of scientific computing. Operating outside of corporate environments, these centers were freer to explore non-standard technologies; in fact, being non-standard could be viewed as a benefit or a goal, in the pursuit of the development of new technologies.

Today, the forefront of supercomputing is no longer driven by national research labs. The advent of AI has brought a wave of massive-scale computing from the private sector. The largest hyperscale companies, including Alibaba, Amazon, ByteDance, Google, Meta, and Microsoft, can spend tens of billions of dollars per year on AI infrastructure, dwarfing the hundreds of millions that might be spent on a single traditional supercomputer.

As private companies race to AI, there is a growing need for scalable, high-performance infrastructure capable of meeting the unprecedented demands of modern HPC and AI workloads. The focus is on maximizing performance, scalability, and efficiency to enable new innovations and operational breakthroughs.

Not Just Ethernet. Ultra Ethernet

The trend toward AI and hyperscale computing has vastly increased the need for high-performance, scalable networking. Data center computing is booming, driven by these segments. Spending on hyperscale AI infrastructure has increased more than sixfold in the past two years, from $18 billion worldwide in 2022 to $121 billion in 2024. Hyperscale data centers now account for over sixty percent of data center infrastructure worldwide.

White paper sponsored by Ultra Ethernet Consortium. Full neutrality statement at https://www.intersect360.com/features/neutrality-statement.
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

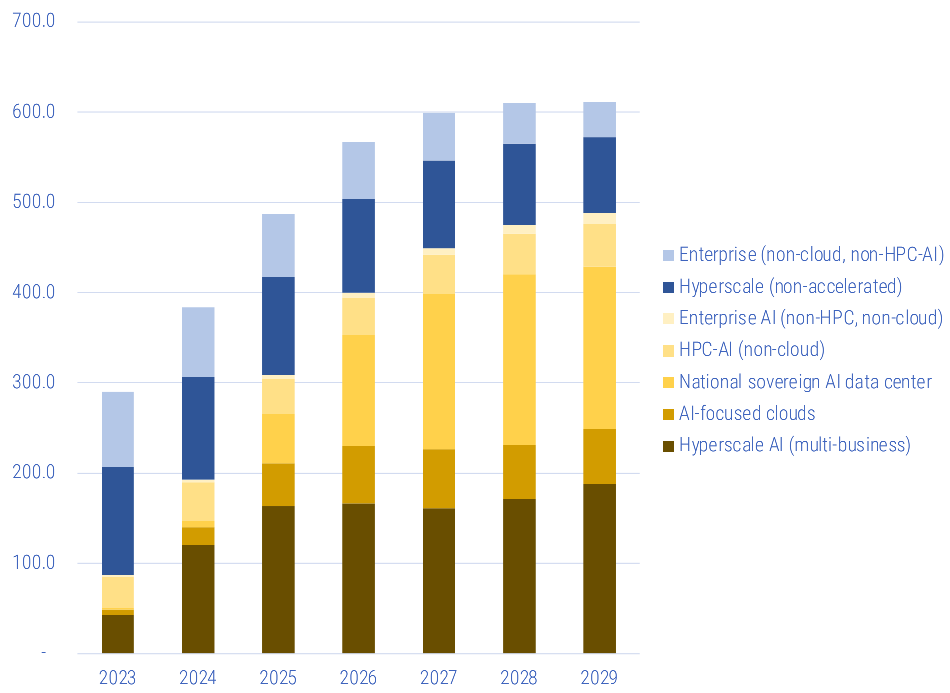
Furthermore, spending on accelerated computing infrastructure will continue to soar, doubling in the next two years. Beyond spending by hyperscale companies, new growth areas like AI clouds and national sovereign AI data centers will emerge as a complement to traditional HPC data centers. (See chart.)

Worldwide Data Center Computing Forecast, Accelerated and Non-Accelerated

2023-24 Actuals, 2025-29 Forecast (in $B).

Accelerated or performance-optimized in shades of yellow and gold; non-Accelerated in shades of blue.

Intersect360 Research, 2025



The effect of this market dynamic is the establishment of large, high-growth data center segments that require the extreme performance and scalability of supercomputing and that also rely almost exclusively on Ethernet. HPC may be a niche segment where a non-standard technology can thrive, but as accelerated computing becomes the norm, Ethernet needs to evolve to embrace new level of capability, without sacrificing on standards.

As private companies race to AI, there is a growing need for scalable, high-performance infrastructure capable of meeting the unprecedented demands of modern HPC and AI workloads.

INTERSECT360 RESEARCH ANALYSIS

Ultra Ethernet Consortium: Addressing the Gap

This surge of activity within the HPC-AI industry has led to increased interest in high-performance networking solutions. For many, the answer to their problems lies within Ethernet-based solutions. However, for Ethernet solutions to be used to their full potential, the industry needs a standardized approach for broader adoption and interoperability.

The Ultra Ethernet Consortium (UEC) is enhancing Ethernet's capabilities through a collaborative, standards-driven approach, with the goal of standardizing a high-performance Ethernet stack purpose-built for the unique demands of HPC and AI. UEC hopes to deliver an open, interoperable, and scalable Ethernet-based fabric that enables data-intensive applications to fully utilize compute and network resources.

There are many working groups within UEC focusing on specific parts of this technology, but the most important deliverable for UEC 1.0 is the specification of a new transport protocol. This new protocol provides the ability to deliver data straight from the network and into application memory and vice versa, without software involvement, in a capability that is known as Remote Direct Memory Access (RDMA). Called the Ultra Ethernet Transport (UET), this protocol incorporates multiple elements that distinguish it from any RDMA protocol that has come before.

Beyond the transport layer, the UEC has a range of working groups, set to deliver innovation not only in the UEC 1.0 protocol, but also in generations to come:

- *Transport Layer:* Develops transport specifications that deliver higher throughput, lower latency, and greater scalability for end-to-end data delivery in AI and HPC networks.
- *Physical Layer:* Focuses on enhancing Ethernet's physical infrastructure—improving performance, reducing latency, and strengthening the foundation for demanding AI and HPC workloads.
- *Link Layer:* Optimizes Ethernet's link layer for efficient, secure, and scalable data communication, extending protocols to support AI and HPC environments.
- *Software Layer:* Creates APIs, specifications, and open-source tools to make Ethernet more flexible, programmable, and adaptable for a wide range of AI and HPC use cases.
- *Storage Working Group:* Integrates high-performance storage services with UEC, ensuring compatibility, security, and seamless management for AI and HPC workloads.
- *Compliance Working Group*: Defines standards and testing to ensure devices and services meet UEC specifications, promoting interoperability and rigorous compliance in AI and HPC networks.
- *Management Working Group:* Develops management models and tools for UEC fabric, enabling efficient monitoring, discovery, and seamless operation across network components.
- *Performance and Debug Working Group:* Establishes benchmarks, metrics, and debug tools to validate UEC performance and reliability, supporting robust deployment in AI and HPC applications.

UEC 1.0 addresses the unique challenges of AI and HPC networks with a scalable, sender-based congestion control model tailored for microsecond-level latency and complex, multi-flow traffic patterns.

Key Stakeholders Shaping the Future of HPC-AI Networking

The UEC is meant to function as a broad industry coalition with the goal of improving performance and interoperability gaps for HPC-AI. As such, UEC leverages the expertise of a diverse set of stakeholders, including processor vendors, system builders, interconnect vendors, cloud providers, and storage companies. Each sector brings with it a unique set of expertise and experience that helps to make UEC a comprehensive and collaborative organization. In fact, Steering Members are particularly influential in setting UEC's technical and strategic direction, while General and Contributor Members are able to provide depth and diversity to this initiative.

These stakeholders have extensive histories within high-performance networking, often having developed proprietary interconnect technologies to meet the demands of HPC. Their participation in UEC reflects a strategic shift – from fragmented, vendor-specific solutions to a unified, open Ethernet-based standard. Thus, previous proprietary interconnect technologies have evolved into today's participation in UEC.

UEC benefits from the collective expertise of organizations that have developed and deployed some of the world's most advanced interconnect technologies for HPC and AI. These legacy technologies have obviously provided foundational expertise for UEC, but it is the participation of hyperscale cloud providers and major AI providers that truly outlines UEC's potential for large-scale, commercial impact. Meta, Microsoft, and Oracle are Steering Members of the Ultra Ethernet Consortium (UEC), with additional participation from General Members such as Alibaba, ByteDance, and Google. These organizations, along with other consortium members, bring experience operating large-scale cloud and AI infrastructure. Their involvement reflects broad industry engagement in UEC's efforts to develop open, high-performance Ethernet standards for HPC and AI workloads, supporting the consortium's goal of expanding relevance beyond traditional high-performance computing environments.

UEC's Technical Approach: Solving Real-World Problems

To get there, UEC is focusing on some important technical innovations that will advance Ethernet to meet the shifting needs of HPC and AI. A good solution to begin this examination is UEC 1.0.

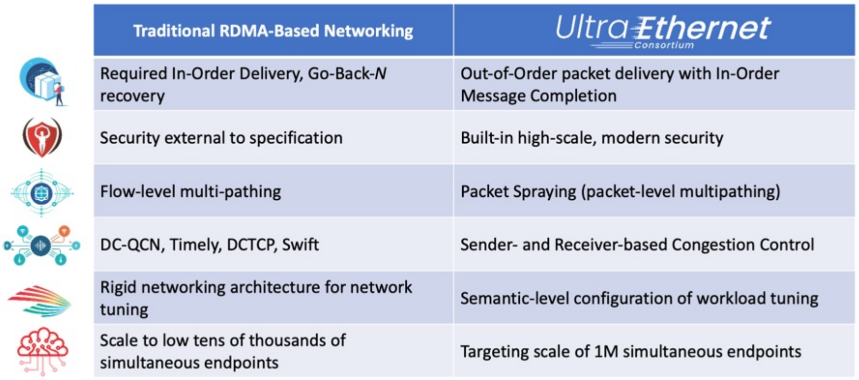
UEC 1.0 defines a set of standards and specifications that reimagine Ethernet for the high-performance needs of HPC and AI. Ethernet has been the first choice for general-purpose networking for years, but it was not originally made for the tightly coupled, high-bandwidth, low-latency communication patterns required by modern HPC and AI. UEC 1.0 is different.

This new approach to Ethernet transport and congestion control is optimized for the way HPC and AI workloads actually behave. One of the key innovations of UEC 1.0 is its ability to align the transport layer with the high demands of parallel applications, particularly in environments with large numbers of nodes and accelerators. Traditional Ethernet was designed for general network traffic, not the high-throughput, low-latency requirements of modern AI and scientific computing. UEC 1.0's approach allows organizations to use existing Ethernet infrastructure while still benefiting from significant performance improvements tailored to these demanding applications.

UEC 1.0 also standardizes new Ethernet's transport protocols to support specialized workloads in HPC-AI environments. By enhancing Ethernet's capabilities for optimized message passing, semantic (i.e., workload-aware) operations, and reduced protocol overhead, UEC 1.0 ensures better network-to-compute efficiency. This minimizes the need for proprietary interconnects and provides a scalable solution that works across a wide range of HPC-AI infrastructures.

Benefits of Ultra Ethernet Networking

Source: Ultra Ethernet Consortium



Traditional RDMA-Based Networking	Ultra Ethernet Consortium
Required In-Order Delivery, Go-Back-*N* recovery	Out-of-Order packet delivery with In-Order Message Completion
Security external to specification	Built-in high-scale, modern security
Flow-level multi-pathing	Packet Spraying (packet-level multipathing)
DC-QCN, Timely, DCTCP, Swift	Sender- and Receiver-based Congestion Control
Rigid networking architecture for network tuning	Semantic-level configuration of workload tuning
Scale to low tens of thousands of simultaneous endpoints	Targeting scale of 1M simultaneous endpoints

Another important part of UEC 1.0 is its congestion management. Standard Ethernet congestion control mechanisms like TCPs weren't designed for the bursty, high-bandwidth, low-latency requirements of HPC and AI. This can cause packet loss, increased latency, and reduced overall throughput.

UEC 1.0 generates a unique solution for congestion management, introducing a sender-based congestion control model that is specifically designed to address the bursty, multi-flow traffic patterns characteristic of AI and HPC workloads. In addition, UEC 1.0 integrates advanced flow control mechanisms that can efficiently handle traffic in large-scale, distributed environments without overwhelming switches and routers. The improved handling of incast events and the ability to perform receiver-side credit allocation ensures that UEC 1.0 can maintain performance during high-demand periods, preventing congestion that could otherwise degrade throughput.

UEC 1.0 addresses the unique challenges of AI and HPC networks with a scalable, sender-based congestion control model tailored for microsecond-level latency and complex, multi-flow traffic patterns. For demanding communication patterns like AllReduce and All-to-All, UEC enhances performance with coordinated multi-path congestion control and optional

UEC 1.0 defines a set of standards and specifications that reimagine Ethernet for the high-performance needs of HPC and AI.

receiver-side credit allocation to prevent incast congestion. UEC also distinguishes itself by handling mixed traffic types—such as single-path and sprayed flows—simultaneously, maintaining effective congestion management across diverse libfabric request scenarios.

UEC also delivers purpose-built transport protocols that are optimized for the performance and scalability demands of HPC and AI. Traditional Ethernet protocols are often too generalized to meet the latency and throughput requirements of parallel applications. This is especially true in systems with heterogeneous accelerators. UEC works to address this by enabling tuned protocol behavior specifically aligned with collective operations and high-bandwidth message flows common in AI training and scientific computing.

By supporting performance-critical operations, UEC 1.0 minimizes protocol overhead and maximizes compute-network efficiency. This is especially beneficial in on-premises environments, where organizations want to maintain tight control over latency-sensitive workflows. UEC's transport-level optimizations are a key part of its differentiated approach to high-performance Ethernet.

Additionally, in environments where multiple workflows, users, and applications compete for shared resources, UEC provides mechanisms for Quality of Service (QoS) to ensure predictable performance. HPC and AI systems demand strict prioritization for critical traffic, particularly during time-sensitive collective operations or when managing large-scale distributed training jobs. UEC enables flow-level control that distinguishes between latency-sensitive and background traffic.

UEC's architecture supports QoS enforcement across both sender and receiver paths, with congestion-aware routing and bandwidth reservation strategies that ensure priority traffic gets the network access it needs. This is especially valuable in enterprise deployments that consolidate workloads across departments or use hybrid traffic patterns. The result is a more stable, efficient network that can scale with the complexity and concurrency of modern AI and HPC use cases.

UEC's solution also works to address the hardware-to-application mismatch that can occur. Traditional Ethernet was not designed with the performance requirements of HPC and AI in mind. Thus, traditional Ethernet can lead to inefficient use of modern accelerators and multi-node systems. UEC 1.0 addresses this hardware-to-application mismatch by optimizing the network stack to align with the specific needs of modern accelerators, such as GPUs, FPGAs, and specialized AI chips. Through improved transport behaviors and advanced congestion management from hardware to application, UEC 1.0 maximizes accelerator utilization, ensuring that compute resources are not underutilized due to network bottlenecks. This end-to-end optimization enables more efficient parallelism and improves overall system throughput.

UEC aligns the capabilities of the network stack with the needs of contemporary HPC and AI applications, especially those involving tightly coupled, latency-sensitive operations. Additionally, UEC utilizes end-to-end optimization. By optimizing transport behavior and congestion management from hardware through to the application layer, UEC 1.0 helps eliminate bottlenecks that reduce the utilization of compute resources. These innovations improve overall system throughput, minimize idle compute cycles, and ensure more predictable performance when working with diverse workloads. UEC also supports standard Ethernet switching hardware, allowing organizations to upgrade performance without a full overhaul.

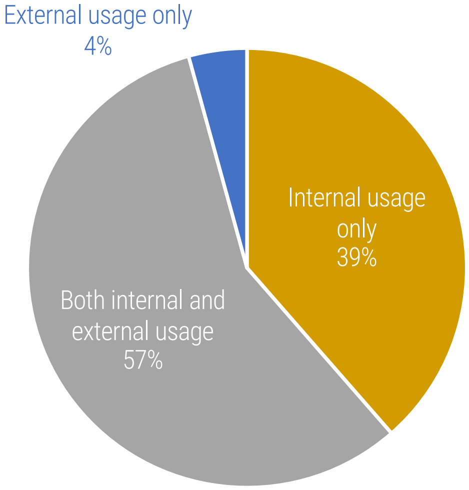
UEC: Use Cases and Deployment

HPC and AI is all about solving problems, and as such it would be impossible to discuss the UEC and UEC 1.0 without mentioning specific use cases for this technology.

In terms of HPC applications, UEC 1.0 has many uses with scientific simulations, data analytics, and computational research. Each of these fields relies on high-bandwidth, low-latency interconnects to coordinate large numbers of tightly coupled compute nodes. UEC works to address these performance challenges in a variety of ways. To begin, UEC 1.0 utilizes sender-based congestion control, which is optimized for microsecond-level latency and bursty traffic. UEC also can leverage on receiver-side credit allocation, which is especially useful during incast scenarios common in collective operations. What's more, UEC's ability to manage parallel processing and systolic multi-flow traffic makes it suitable for the network patterns found in scientific workloads.

Intended Usage of LLMs, Internal vs. External

Intersect360 Research, HPC-AI Software Survey, 2024



UEC is also heavily focused on enabling performance in AI applications. Large Language Model (LLM) training and inference generate large amounts of traffic due to frequent synchronization across accelerators as well as large-scale distributed computation. UEC also offers multi-path congestion control and coordinated packet spraying, improving throughput. Additionally, by providing differentiated QoS, UEC ensures that latency-sensitive operations like gradient aggregation receive priority treatment over background tasks.

The involvement of hyperscale cloud providers is important because it helps UEC-compliant fabrics to be deployed both on-premises and in the cloud. In a survey of HPC-AI users who were investigating LLMs, the majority were anticipating both internal and external usage, implying the need for hybrid cloud deployments. (See chart above.) UEC is compliant with any locality.

CONCLUSION

The Ultra Ethernet Consortium has stepped into a dramatically dynamic computing landscape, where the demands of AI and HPC are outpacing what legacy networking can deliver. To solve this, UEC 1.0 establishes a practical, standards-based foundation that is tailored for the unique demands of HPC and AI workloads. The immediate improvements of UEC 1.0 – including congestion control, semantic routing, and workload-aware protocols – are obvious, but UEC doesn't plan to stop there. It is the consortium's intent to provide iterative refinement aligned with what the industry needs.

Additionally, UEC improves interoperability by uniting diverse vendors and cloud providers under a common standard, thereby reducing vendor lock-in risks. Ethernet's broad ecosystem and existing infrastructure provide massive cost advantages, and therefore work to lower barriers to adoption. What's more, the scalability inherent within Ethernet as well as UEC's protocols are designed to handle massive node counts and data flows typical in both AI training and HPC simulations.

UEC 1.0 is already valuable, but it will also serve as a foundation for future innovation. The modular design of UEC's standards allows for the introduction and incorporation of new technologies as the HPC-AI world changes. UEC 1.0 is an open platform that invites ongoing contributions and encourages innovation from the community rather than being locked into a single vendor's roadmap. Work is already underway for the future of Ultra Ethernet, including new ideas about improved telemetry and congestion control, UE bindings for storage protocols, and in-storage compute.

It is UEC's diversity that provides much of its strength. From cloud hyperscalers to networking experts, the UEC team is deeply entrenched in the HPC-AI industry. This collaboration ensures that standards are grounded in real-world requirements as well as shared operational experience. The open dialogue provided by UEC reduces fragmentation and will hopefully accelerate adoption.

By engaging with this open, collaborative initiative, the community can help shape the future of high-performance networking to meet the evolving needs of tomorrow's workloads.

For more information, visit: https://ultraethernet.org/

White paper sponsored by Ultra Ethernet Consortium. Full neutrality statement at https://www.intersect360.com/features/neutrality-statement.
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com