Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

UALink, Ultra Ethernet, and PCIe: Transforming Next-Generation HPC and AI Workloads

Rajesh Arsid
Edinburgh Napier University, UK

ARTICLE INFO

Received: 19 Feb 2026
Revised: 22 Feb 2026

ABSTRACT

As HPC and AI workloads fundamentally transform data center architectures, the demand is growing for heterogeneous compute infrastructures with unprecedented bandwidth, low latency, and massive parallelism. Legacy interconnect technologies, such as InfiniBand and Ethernet, can no longer address the scale and diversity of the AI training and scientific simulation workloads deployed today. Complementary vendor-neutral, open standards include UALink, Ultra Ethernet, and next-generation PCIe with Compute Express Link (CXL), forming a set of hierarchical interconnect technologies that are optimized for future compute ecosystems. UALink provides multi-terabit throughput at sub-microsecond latency in vendor-neutral, flexible topologies with optimal GPU placement and contention-free configurations. Ultra Ethernet transforms Ethernet networking protocols with deterministic forwarding, advanced congestion control, and hardware-accelerated collective communication primitives to enable commodity Ethernet as a low-latency fabric suitable for scale-out AI use cases while retaining Ethernet protocol compatibility. PCIe evolution, through Compute Express Link (CXL), allows heterogeneous memory architectures with cache-coherent, memory-mapped memory, including Managed DRAM, ReRAM, and persistent non-volatile memory. Managed DRAM with hardware-managed tiering between classes of memory has demonstrated order of magnitude performance improvements compared to software-managed tiered memory systems. These interconnect technologies address green computing issues across the hardware life cycle and enable the topology-aware scheduling, data migration, and resource composition that will be important in future cloud platforms that handle heterogeneous workloads. The success of these convergence architectures will ultimately determine whether cloud-based data center architectures will cope with the computational demands of future AI, natural language processing, and scientific discovery workloads, such as climate modeling.

Keywords: UAlink Accelerator Interconnect, Ultra Ethernet AI Networking, Compute Express Link Memory, Topology-Aware Gpu Scheduling, Heterogeneous Memory Tiering

1. Introduction: Evolving Demands of HPC and AI Workloads

High-Performance Computing (HPC) and AI workloads are driving a transformation in data center architecture. AI training and inference, and scientific simulations require extreme levels of parallelism with ultra-low latency and unprecedented bandwidth to move and manipulate data across thousands of heterogeneous compute nodes. As compute demand grows, there is also an increasing focus on the sustainability of AI infrastructure, with considerations for architecture as well as hardware component environmental impact across their life cycle, beyond compute energy consumption [1]. Customary data center networks (DCNs) are increasingly challenged by the scale and also the diversity of workloads.

As a result of the increasing compute intensity of large AI models, such as LLMs that require distributed training across thousands of computing accelerators, the communication cost in distributed deep learning

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

has become a performance bottleneck, as collective communication operations have been found to consume large fractions of training time. Several studies have shown that as distributed training workloads scale beyond a few hundred nodes, their performance is bottlenecked by network congestion and synchronization overheads. These workloads expose limitations of customary interconnect fabrics along the data center hierarchy. These include limited bisection bandwidth, poor RDMA support, and rigid topologies along all levels of the data center hierarchy. The convergence of the fields of AI and HPC has led to the requirement that a system be capable of supporting both tightly-coupled synchronous communication patterns as well as loosely-coupled asynchronous data movement required in inference workloads and HPC applications that need to transfer petascale amounts of data.

2. The Limitations of Legacy Interconnect Technologies

A lifecycle analysis of data center hardware quantifies the represented carbon impact of interconnect components during manufacturing and materials extraction as the largest contributor to the data center's environmental impact, which must be considered along with operational energy consumption [1]. InfiniBand offers low latency and high throughput, but may suffer from vendor lock-in, limited diversity in its ecosystem, and cost at hyperscale. Due to the protocol overhead and the specialized switching infrastructure required, it has not seen common deployment in cloud service providers, which offer commodity solutions with broad industry support.

Customary Ethernet networks do not meet the low-latency or lossless requirements of distributed AI workloads, training, and high-performance computing (HPC) simulations. In order to fulfill the requirements of training distributed AI workloads, the data center network must scale the bandwidth between the accelerators, minimize the latencies of synchronizing operations, and route packets dynamically and adaptively based on the network traffic patterns. [3] Even with features such as Data Center Bridging and RDMA over Converged Ethernet, standard Ethernet fabrics have been shown to have varying congestion performance, congestion packet loss on collective operations, and poor message latency for synchronizing parameters with small messages, as is required by distributed machine learning. The increase of heterogeneous compute fabric with a diversity of accelerator architectures, such as GPUs with different bandwidth and memory size, tensor processing units focused on different computations of certain neural networks, and reconfigurable field-programmable gate arrays, has exposed the limitations of customary interconnects to provide high-bandwidth communication among heterogeneous compute elements. Modern AI training pipelines also require network fabrics that support multiple classes of traffic (from latency-sensitive gradient updates to bandwidth-heavy checkpointing traffic) while supporting fairness and preventing head-of-line blocking of packets that may otherwise degrade the overall performance of the system.

Interconnect Technology	Key Limitations	Impact on AI/HPC Workloads	Scalability Challenges
InfiniBand	Vendor lock-in and limited ecosystem diversity	Prohibitive cost structures at hyperscale deployments	Specialized switching infrastructure creates adoption barriers
InfiniBand	Protocol complexity requires specialized hardware	Restricted flexibility for cloud service providers	Limited commodity-based solution availability
Conventional Ethernet	Performance variability under network congestion	Disrupts collective operations in distributed training	Inadequate lossless transmission guarantees

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Conventional Ethernet	Inefficient small message handling	Poor parameter synchronization in machine learning	Packet loss scenarios degrade RDMA performance
Legacy Fabrics (General)	Insufficient bisection bandwidth	Cannot support petascale data movement requirements	Rigid topologies fail to adapt to dynamic workloads
Legacy Fabrics (General)	Inadequate RDMA support at scale	Limits tightly-coupled synchronous communication	Fails to support heterogeneous accelerator architectures

Table 1: Limitations of Legacy Interconnect Technologies for AI and HPC Workloads

3. UALink: Open Standard for Accelerator-to-Accelerator Connectivity

The Ultra Accelerator Link (UALink) consortium is an industry consortium defining an open, high-bandwidth interconnect specification to support accelerator-to-accelerator communication, both within servers and across servers. The Ultra Accelerator Link (UALink) Promoter Group announced the publication of the Ultra Accelerator Link (UALink) specification. This is a major advance in creating vendor-neutral standards for AI infrastructure. [5] In distributed AI training workloads, data transfer between GPUsis a bottleneck. In multi-GPU configurations, hardware topology and interconnect topology have been shown considerably affect application performance. Dual-lane NVLink interconnect provides up to 40 GB/s of unidirectional bandwidth between GPUs in the same socket. Each GPU is likewise connected to the socket using two lanes of NVLink [6]. With a compatible memory protocol, shared memory semantics can be made consistent across multiple accelerators, and the software overhead is reduced, as is the use of massively distributed compute resources. The openness of the specification avoids the proprietary lock-in that comes with other high-performance interconnect technologies such as InfiniBand or Intel QuickPath Interconnect.

To scale, UALink supports several topologies such as all-to-all, torus, and fat-tree topologies. System designers can use these topologies to optimize workload mapping to a desired topology. When testing UALink using deep learning workloads, topology-aware placement algorithms can achieve up to a 1.30x speedup over state-of-the-art placements that place GPUs to meet the workload requirement while avoiding interference [6]. For interconnect hardware, UALink seeks to enable sustainable hardware manufacturing and end-of-life hardware disposal [1]. UALink specifies multi-terabit per second targets and sub-microsecond latency requirements to enable computationally efficient training of large models, where distributed gradient synchronizations dominate the training run time when many model parameters are hosted across multiple devices. Performance characterization shows that bad placement strategies can lead to more than 30 percent performance degradation when co-scheduling jobs with high GPU communication requirements in shared environments [6].

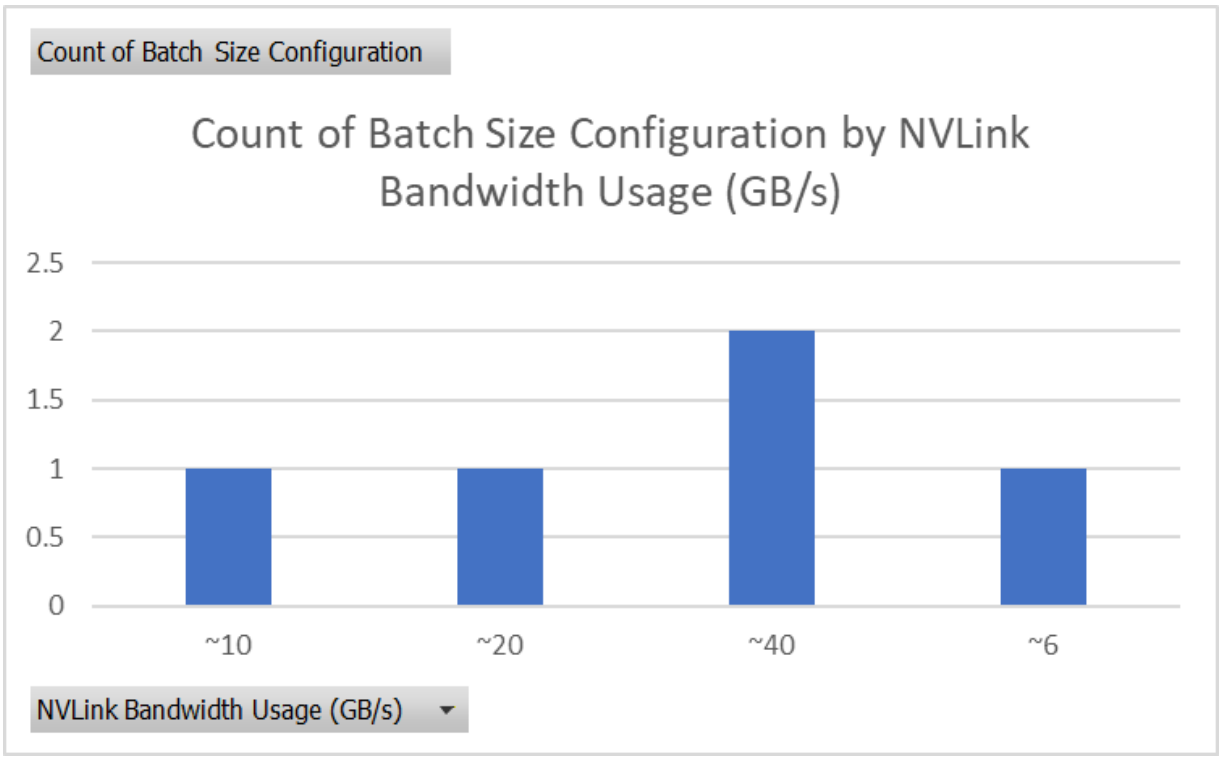
The open specification for UALink has been implemented by multiple vendors to provide system-wide diversity and interoperability found to reduce enterprise lock-in. The specification supports adaptive routing, where a path is dynamically adjusted every time the packet is traversed, and congestion management to address AI workloads' bursty nature. The protocol also supports native hardware acceleration for collective operations such as all-reduce and all-gather, which are often used in distributed machine learning training workloads. Profiling distributed training workloads indicates that when configured with a small batch size, applications will consume almost all of the NVLink bandwidth, approximately 40 GB/s, only during the communications phases of the training workload. Using larger batch sizes, this drops to 6 GB/s because the synchronization phases are less frequent [6]. They also make data-parallel training more efficient, as time for communicating data using software libraries is negligible,

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

and the accelerators spend more time in compute mode rather than idle, waiting for data. The batch size parameter strongly affects the communication-to-computation ratio: smaller batch sizes require more frequent exchange of gradients between the GPUs, and therefore a higher bandwidth for the interconnect [6].

Energy efficiency for optimal signaling and dynamic power management is also a goal of UALink. These are needed to meet the challenge of sustainability for data centre operators, who need to balance demand for higher performance with environmental impact and low cost of ownership. A topology-aware UALink is also helpful to cloud computing providers, where GPUs are one of the biggest capital and operational expenses. In practice, workloads with communication-intensive GPU kernels suffer up to 30% slowdown when co-scheduled on the same physical machine, while other workloads are not affected. The introduction of support for multi-topology networks in the UALink standard represents an opportunity for cloud and enterprise data centers to implement topology-aware scheduling mechanisms in their network hardware and provision interconnect topologies that match the optimal physical GPU placements discovered through workload characterization and performance modeling. With our topology-aware algorithms, we have achieved, on average, 1.30x, 1.28x, and 1.27x improvements in cumulative execution time, over best-fit, first-come-first-served, and standard topology-aware algorithms, respectively, enabling improved training throughput and time-to-solution in ML applications across a range of neural networks and training frameworks [6].



Graph 1: NVLink Bandwidth Usage vs Batch Size [6]

4. Ultra Ethernet: Reimagining Ethernet for AI-Era Data Centers

Ultra Ethernet is a complete evolution of Ethernet networking for AI and HPC workloads that is backward compatible with the existing Ethernet ecosystem. Data center networks suited for AI workloads need to fundamentally change how Ethernet networks deliver quality of service (QoS), network congestion, and

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

collective communications [3]. The Ultra Ethernet Consortium brought together many leading technology companies to define a new series of extensions and enhancements to Ethernet, to evolve Ethernet from a universal networking fabric to a specialized computing fabric, taking advantage of Ethernet's ubiquity, maturity, and low cost, and to provide the same capabilities found in dedicated interconnect fabrics such as InfiniBand. As part of this effort, consideration of the impact of network infrastructure on the environment is being discussed across new dimensions: energy consumption, represented carbon, and circular economy principles over the hardware life cycle [1].

Core features of the Ultra Ethernet protocol include deterministic forwarding for real-time flows with bounded latency, congestion control to avoid throughput collapse for high loads, and lossless delivery of RDMA messages. The performance modeling and implementation of the protocol have shown that distributed training workloads have very diverse traffic patterns, which merge short bursts of synchronization from multiple parallel processes with compute-bound workloads [4]. The standard introduces new and improved quality-of-service (QoS) mechanisms. This allows for fine-grained differentiation between several classes of traffic, such as latency-sensitive control plane messages, bulk data transfers, and background management traffic, enabling heterogeneous communication patterns for training AI applications. Today's AI training pipelines, for example, feature different classes of traffic with different performance characteristics for updating parameters, transferring activations, or writing checkpoints. In the case of Ultra Ethernet, hardware offload support is available for collective communication using all-reduce, broadcast, or scatter-gather operations to move data among nodes in the network switches. This results in reduced host CPU overhead and better performance, because collective operations do not need to go through multiple software layers and are less expensive if implemented in hardware. The specification also allows for rich telemetry and observability features, giving operators observability of the network and its activities, which is instrumental when performance issues occur in distributed training, and the cause of poor performance could be high communication times. Ultra Ethernet is backward-compatible with standard Ethernet, and can be incrementally deployed, letting organizations use existing Ethernet infrastructure while adding functionality and performance as their AI workload requirements grow.

Feature Category	Innovation	Benefit	Application Scenario
Traffic Management	Deterministic forwarding mechanisms	Guaranteed bounded latency for time-sensitive traffic	Latency-critical gradient synchronization operations
Congestion Control	Enhanced algorithms preventing throughput collapse	Maintains performance under high load conditions	Burst traffic during distributed training phases
Quality of Service	Sophisticated multi-tier frameworks	Fine-grained traffic differentiation capabilities	Coexistence of control plane and data plane traffic
Transmission Reliability	Lossless transmission guarantees	Essential for RDMA operations integrity	Parameter updates in distributed machine learning
Hardware Acceleration	Native collective operation support	Reduces host CPU overhead significantly	All-reduce and broadcast primitives for training
Network Observability	Enhanced telemetry and monitoring features	Granular visibility into network behavior	Diagnosing performance anomalies in complex scenarios
Compatibility	Backward compatibility with standard Ethernet	Protects existing infrastructure investments	Incremental deployment in hybrid environments

Table 2: Ultra Ethernet Innovations for AI-Era Data Centers

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5. PCIe Evolution leveraging CXL: Enabling High-Bandwidth Device Interconnection

PCI Express continues to be the predominant interconnect technology in compute nodes connecting processors, accelerators, memory, and storage. Compute Express Link (CXL), an open industry-standard implementation of cache-coherent, byte-addressable interconnect based on the PCIe 5.0 physical layer (PHY), in high-efficiency SerDes technology, connects CPUs with different types of memory with a common load-store semantic [10]. CXL builds on the customary DDR memory architecture, which places memory controllers on the device side instead of the host CPU, allowing for flexible memory expansion without needing to swap out the host CPU and leverage existing CPU architectures. The environmental impacts of interconnect technology include hardware longevity and upgrade cycles. Backward compatibility of CXL allows systems to last longer with new generations of CXL, improving performance. [1]

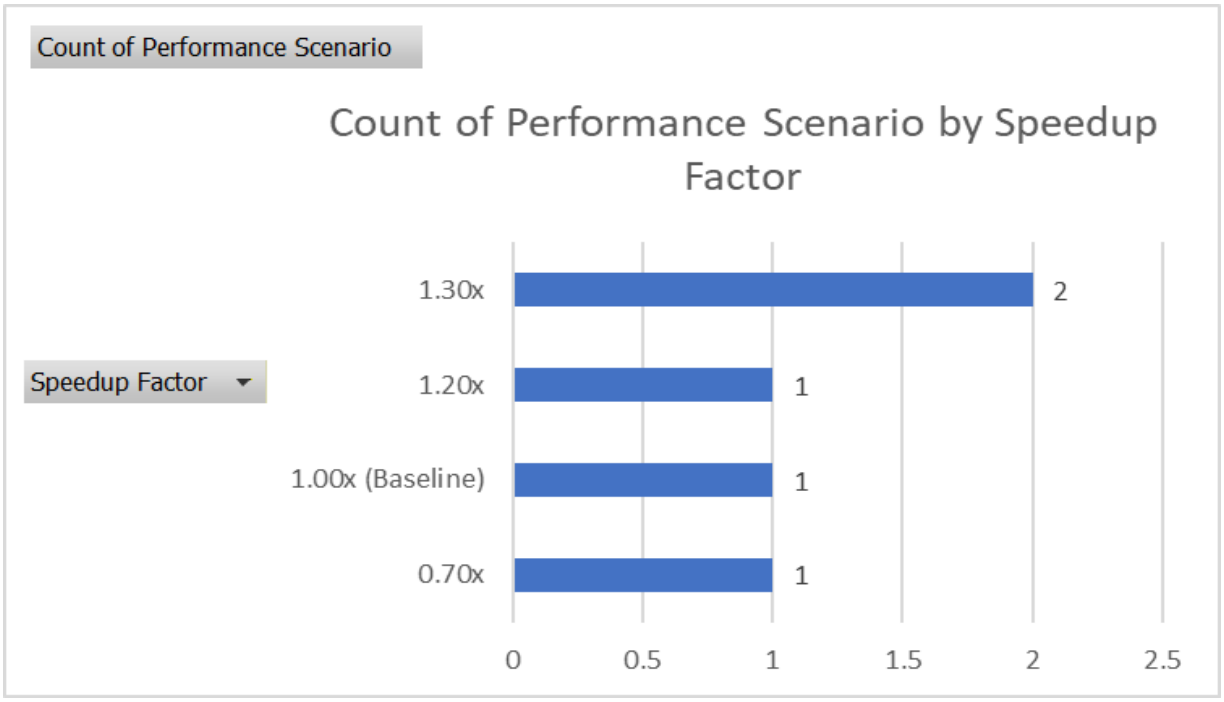
With each generation of CXL, more advanced topologies have become available, leading to the single-level switching CXL switch being introduced with CXL 2.0 that supports one or multiple hosts with one or multiple CXL devices. CXL 3.0 extends to multiple switches on the CXL fabric and introduces Port-Based Routing for inter-switch links, enabling more complex CXL memory device topologies [10]. It has been found that adding a CXL switch can increase the overall latency of the system by around 70 nanoseconds, creating heterogeneous latencies for systems using cascaded switches to connect the CXL memory devices to the host, as some devices will contain more switches than others in the critical path. [10] In addition to increased bandwidth,next-generation interconnect technologies aim to reduce latency with improvements to transaction layer protocols, flow control, and power management to achieve an acceptable balance of performance and power consumption.

CXL-based memory systems open a path to allow new memories such as Managed DRAM, emerging memory technologies such as ReRAM, and heterogeneous storage memories such as 3D-XPoint/Optane, that can have larger capacity and lower cost, lowering the TCO. The studies on memory tiering systems in the CXL extended memory environment have shown performance improvements of 5.1 percent to 16.2 percent compared with customary memory tiering systems through hardware-managed techniques on the device side of the CXL memory systems. With this Compute Express Link technology, heterogeneous memory with different latencies and memory bandwidths can be used by the CPU to access the endpoints. This hierarchical memory management can help increase performance on heterogeneous memory systems [10]. The performance of distributed training with communication patterns could be negatively affected by the overhead of the cache coherency traffic, which was dependent on memory bandwidth and latency [4].

Experiments show that hardware-managed memory tiering systems outperform several existing memory tiering systems by 5.7 percent to 17.6 percent on CXL-enabled systems [10]. PCIe was designed as a superset of fabric interconnects. In disaggregated computing architectures, computational resources, memory, and storage can be composable at fine granularity with respect to the requirements of the workloads utilizing them. It can be particularly applicable to cloud data centers with mixed AI and HPC workloads of variable compute and memory needs. In CXL-enabled experimental systems, hardware-based data migration achieves a 12.9x bandwidth advantage over CPU-based data migration when migrating data between fast and slow memory tiers at a page granularity of 4 kB [10]. Increased error detection and correction capabilities and reliability features, among others, help ensure that data is not corrupted as it flows over the high-speed link. This is important for long-running scientific simulations and training jobs, as bit errors in the data can cause downstream corruption even after long processing times. CXL-extended heterogeneous memory systems can also enable new types of computation with lower data movement overheads to be co-designed and executed through the interaction of computation, storage, and communication subsystems within the CXL-extended heterogeneous memory system.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



Graph 2: Performance Impact of GPU Placement Strategies[10]

Conclusion: Integrated Architectures for Future Computational Ecosystems

The co-development of UALink, Ultra Ethernet, and PCIe with Compute Express Link enables the re-architecture of the data center interconnect fabric with the compute and performance needed for the workloads of AI and HPC spanning servers and data centers. UALink, Ultra Ethernet, and PCIe can be used in tandem to create a hierarchy of communication architectures optimized for heterogeneous workloads, with PCIe and UALink as high-bandwidth, low-latency interconnect for within and between servers, and Ultra Ethernet optimized for rack and data center scale, including distributed workloads. Environmental considerations are relevant to interconnect architecture because they may affect the power consumption, hardware durability, and carbon footprint of a system. Future research topics include collective communication primitives that help alleviate synchronization bottlenecks in distributed training setups, photonic interconnect technologies that can enable higher bandwidth and better energy efficiency compared to electrical signaling, and clever networking solutions leveraging machine learning to optimize routing and resource allocation according to the workload and performance data. Along with newly evolving memory hierarchies such as persistent memory and computational storage devices, these transformations have pushed interconnects to evolve towards an integral part of next-generation architectures that optimize the computation-storage-communications trade-off. With AI applications reaching trillions of parameters and scientific applications evolving beyond exascale, interconnects determine how computation resources can keep working or wait for data, potentially dominating the overall system performance. Distributed training workloads dictate the communication patterns that overwhelm conventional networks and demand flexible interconnects that can switch between different traffic patterns and prioritize different types of messages based on their impact on the training efficiency. UALink, Ultra Ethernet, and high-speed PCIe have the key enabling capabilities to deliver high-performance compute for artificial intelligence and computational science workloads. In addition, architectures for future systems based on these technologies will need to trade off workload

characteristics, scalability, ecosystem maturity, total cost of ownership (TCO), and sustainability over the hardware life cycle to determine how and where data center architectures will address the insatiable demand for compute for natural language processing, computer vision, drug discovery, climate modeling, and machine learning and high-performance computing (HPC)-related workloads to solve social challenges.

References

[1] Fleur Jeanquartier, et al., "Assessing the carbon footprint of language models: Towards sustainability in AI," ScienceDirect, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921344925005476
[2] Lang Xu, et al., "Characterizing Communication Patterns in Distributed Large Language Model Inference," arxiv, 2025. [Online]. Available: https://arxiv.org/abs/2507.14392
[3] Tejas Network, "Rewiring the Future: Data Center Interconnects for the AI Era," 2025. [Online]. Available: https://www.tejasnetworks.com/resource/rewiring-the-future-data-center-interconnects-for-the-ai-era/
[4] Yuxuan Li, et al., "Congestion Control for AI Workloads with Message-Level Signaling," ACM Digital Library, 2025. [Online]. Available: https://dl.acm.org/doi/10.1145/3735358.3735378
[5] Business Wire, "UALink Consortium releases the Ultra Accelerator Link 200G 1.0 specification," 2025. [Online]. Available: https://www.businesswire.com/news/home/20250408050548/en/UALink-Consortium-Releases-the-Ultra-Accelerator-Link-200G-1.0-Specification
[6] Marcelo Amaral, et al., "Topology-aware GPU scheduling for learning workloads in cloud environments," ACM Digital Library, [Online]. Available: https://dl.acm.org/doi/10.1145/3126908.3126933
[7] Erik A. Träff et al., "Simple and efficient GPU accelerated topology optimisation: Codes and applications," ScienceDirect, 2023. https://www.sciencedirect.com/science/article/pii/S0045782523001676
[8] Torsten Hoefler, et al., "Ultra Ethernet's Design Principles and Architectural Innovations," arxiv,2025. [Online]. Available: https://arxiv.org/abs/2508.08906
[9] Dr. Mohiuddin Mazumder and K. Lee, "PCIe 6.0: A high-performance interconnect for storage and networking challenges," Storage Networking Industry Association (SNIA), SDC 2021. [Online]. Available: https://www.snia.org/sites/default/files/2025-05/SNIA-SDC21-Mazumder-PCIe6p0-A-High-Performance-Interconnect-for-Storage-Networking-Challenges.pdf
[10] Yiqi Chen et al., "Exploring Memory Tiering Systems in the CXL Era via FPGA-based Emulation and Device-Side Management," arXiv, 2025. [Online]. Available: https://arxiv.org/html/2502.19233v3