JOURNAL OF INTERNATIONAL CRISIS AND RISK COMMUNICATION RESEARCH
ISSN: 2576-0017
2026, VOL 9, NO 2

The Future Of Next-Generation HPC At Scale: The Transformative Impact Of PCI Express 8.0

Rajesh Arsid

Edinburgh Napier University, UK

Abstract

High-Performance Computing and Artificial Intelligence are getting even closer, at unprecedented scales, building the computational models that exascale and the next generation zettascale architectures will require. PCI Express 8.0 is the next disruptive evolution of the de facto standard for high bandwidth, low latency interconnect that will help heterogeneous computing overcome the very limits that have historically restricted its ascendance. The specification also includes Pulse Amplitude Modulation four-level (PAM4) signaling, improved Forward Error Correction (FEC), and improved protocol stacks, which result in meaningful improvement of the effective data rate. Modern computers adopt more and more diversified accelerators such as Graphics Processing Unit (GPU), Field-Programmable Gate Array (FPGA), or domain-specific accelerated processors with complex interconnect topologies. PCIe 8.0's high bandwidth and low latency empower system architectures to efficiently keep nodes well fed with data to achieve the desired system performance. These system architectures may utilize cache-coherent memory fabrics, disaggregated pools of resources, and different peer-to-peer communication modes to power the next-generation systems enabled by PCIe 8.0. Furthermore, the implementation of such approaches at scale encounters difficulties related to power delivery network design, thermal management, and channel implementation in order to maintain the integrity of the transmitted information. Such implementations have been shown to provide varying benefits to applications such as streaming data processing and machine-learning workloads in distributed systems. In latency-sensitive scientific computing applications, the speed advantage is less impressive, and system integration considerations such as new cooling methods, power management, and signal conditioning become major success factors. As a vendor-neutral and open industry specification, PCIe 8.0 benefits from interoperability and multivendor support across the broad PCI Express ecosystem. The PCIe 8.0 specification will support the next generation of computing systems from the present generation of exascale systems to future systems based on zettascale technology.

Keywords: PCI Express 8.0, High-Performance Computing, Heterogeneous Computing Architectures, Interconnect Bandwidth, Exascale Computing.

1. Introduction: The Evolving Landscape of HPC and AI

HPC and AI are entering an era of potential exascale performance and zettascale data. Already, the convergence of AI and HPC has changed the landscape of computational science, with AI workloads dominating the utilization of HPC infrastructure. According to thorough market research and analysis, the global HPC market accounted for USD 57.00 billion in 2024 and is expected to grow at a CAGR of 7.2% between 2025 and 2030, reaching USD 87.31 billion by 2030, owing to technological advancements and

increasing demand from applications in scientific research, financial modeling, weather forecasting, and molecular dynamics simulations [1]. The North American region held the largest market share of 41.6% in the global market in 2024. The US market is expected to register a CAGR of 7.8% from 2025 to 2030. By component, the servers segment dominated the market in terms of revenue in 2024, accounting for more than 32.0% of the market. By deployment, the on-premise segment led the market in 2024; by end use, the government and defense segment led the market in 2024.

The field of HPC, or supercomputing, overlaps with AI in two ways: AI workloads requiring supercomputer-level capabilities and HPC workloads that benefit from optimizations through AI. For example, in drug discovery, machine learning can be used to predict the force field for molecular dynamics simulations, and in climate modeling, neural networks can be used to improve the spatial resolution of predictions of atmospheric quantities [2]. Specific industries such as aerospace and automotive, pharmaceuticals, and climate science tend to have large-scale computing needs for modeling complex systems and analyzing big data. These industries drive demand for HPC systems, especially where existing computing infrastructures are inadequate for the required modeling and simulations. As seen with many modern AI applications, such as large language models and foundation models, the distributed nature of training on hundreds or thousands of accelerators generates unprecedented amounts of data traffic. The increasing interest in artificial intelligence, machine learning, and big data analytics results in unprecedented levels of data being stored, processed, and generated by organizations, with HPC infrastructure often used to analyze data in real time and create advanced machine learning models. Because of their parallel data processing capabilities, HPC systems have become a strong fit for these types of data-intensive applications in sectors like finance, healthcare, and e-commerce.

The PCI Express 8.0 specification early draft has been released to alleviate these architectural bottlenecks with a theoretical per-lane data rate up to 256 gigatransfers/second, or approximately 512 GB/s each direction in the standard x16 configuration. It is an enabling technology for a new generation of system architectures, allowing concurrent transfers over three key interconnect paths: accelerator-to-accelerator communication, host-to-device data transfers, and storage subsystem access without contention on the interconnect. These features offer higher bandwidth, superior signaling, low protocol overheads, and lower power consumption, making it feasible to create exascale systems within power and thermal budgets. A further driver of the market is the growth of cloud-based high-performance computing (HPC) services, as an increasing number of cloud providers offer access to high-performance, scalable, and cost-effective HPC services. Such services are targeted towards small and medium-sized businesses that may be unable to afford the upfront investment and maintenance costs associated with operating their own supercomputing platforms. This has led to a rise in HPC as a Service, which reduces capital expenditure and can be scaled according to demand across different industry verticals.

2. Architectural Foundations and Technical Specifications

PCI Express defines a layered packet-based interconnect architecture in which communication between components is carried through Transaction Layer Packets (TLPs). A serialized TLP consists of optional prefixes, a header, an optional data payload, and an optional digest. The generic packet organization is illustrated in the specification, where transmission occurs as an ordered stream of bytes over the link, with the lowest-numbered bytes conceptually transferred first. This packet structure forms the foundation for all PCIe transactions, including memory, I/O, configuration, and messaging operations. Two addressing formats are supported for memory requests: 32-bit and 64-bit addressing modes.[3]

To ensure correctness and interoperability, all reserved fields in TLPs must be transmitted as all zeros and ignored by receivers. Packet headers include common fields such as the Fmt[2:0] format field and the Type[4:0] packet type field, which together define the structure and function of each transaction packet. Flow control is a critical architectural component of PCI Express and is implemented through Data Link Layer Packets (DLLPs). All DLLPs include a defined DLLP Type field and a mandatory 16-bit CRC, ensuring the integrity of link-layer control information. The specification defines multiple DLLP categories, including Ack, Nak, InitFC, and UpdateFC packets. These control messages are essential for managing buffer credits and maintaining reliable transmission under heavy transaction loads.

PCIe also supports Scaled Flow Control, where credit scaling factors are encoded using the HdrScale and DataScale fields. The document specifies that scaling values may take encodings of 00b, 01b, 10b, and 11b, corresponding to scaling factors of 1, 1, 4, and 16, respectively. Under scaling, header credits may range up to 2,032, while data credits may scale up to 32,752, depending on the selected factor.[3]
At higher signaling rates, PCI Express introduces stringent electrical compliance constraints. Calibration channels operating at 16.0 GT/s must satisfy return loss limits of ≤ –12 dB below 4 GHz, ≤ –8 dB between 4 GHz and 12 GHz, and ≤ –6 dB between 12 GHz and 16 GHz. For 32.0 GT/s, the return loss mask requires <-–12 dB below 4 GHz, <- –10 dB from 4 GHz to 16 GHz, and<-≤ –6 dB from 16 GHz to 32 GHz.
Calibration channel electrical design is also constrained by resistance limits. The specification states that calibration channels must not have a DC resistance exceeding 7.5 ohms, measured as the sum of the D+ and D− trace resistances. Insertion loss design flexibility is provided through multiple loss options, requiring coverage from at least 2 dB below FHIGH-IL-MIN to 3 dB above FHIGH-IL-MAX, with loss step spacing of 0.5 dB or less.

The physical construction of receiver calibration fixtures is also numerically defined. For example, the 16.0 GT/s Rx Calibration Base Boards include sixteen differential pairs with 85 Ω nominal impedance, and base board insertion loss ranges are specified as 4–11.5 dB (low-loss), 12–19.5 dB (mid-loss), and 20–27.5 dB (high-loss), measured at 8.0 GHz in 0.5 dB steps.

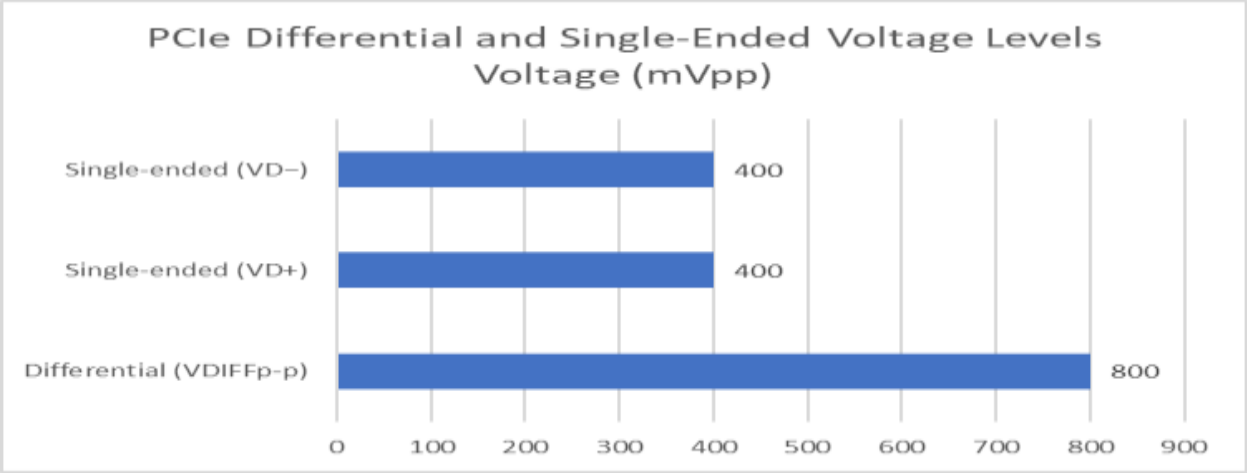
Similarly, the 32.0 GT/s Rx Calibration Base Boards also contain sixteen differential pairs with 85 Ω nominal impedance, with insertion loss ranges of 4.0–11.5 dB, 12.0–19.5 dB, and 20.0–27.5 dB, measured at 16.0 GHz in 0.5 dB steps.

Additional transmitter-side electrical constraints are explicitly defined. All transmitters are required to be AC coupled, with coupling capacitor values ranging from 176 nF (minimum) to 265 nF (maximum). The maximum DC differential transmitter impedance is specified as 120 Ω, and the transmitter short-circuit current must not exceed 90 mA.

Finally, PCIe defines voltage waveform behavior at the physical layer. Differential peak-to-peak voltage is expressed as:

- $V_{DIFFp\text{-}p} = 2 \times \max |V_{D+} - V_{D-}|$

The specification provides an example waveform where the differential voltage swing is approximately 800 mVpp, while each single-ended conductor exhibits approximately 400 mVpp, with a nominal center crossing around 200 mV.



Graph 1: Physical Layer Voltage Swing Specifications [3]

3. Implications for Heterogeneous Computing Architectures

Today's high-performance computers have moved from homogeneous clusters of processors executing the same instruction stream to heterogeneous systems with accelerators for executing specific kernels, owing to limitations in continuing to scale general-purpose computer architecture. In particular, the amount of speedup that could be achieved through instruction-level parallelism has plateaued, and Dennard scaling has ended. Modern computer nodes consist of heterogeneous components: graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs) with a variety of interconnect topologies. The performance of the heterogeneous system largely depends on the data exchange between processing units [5].

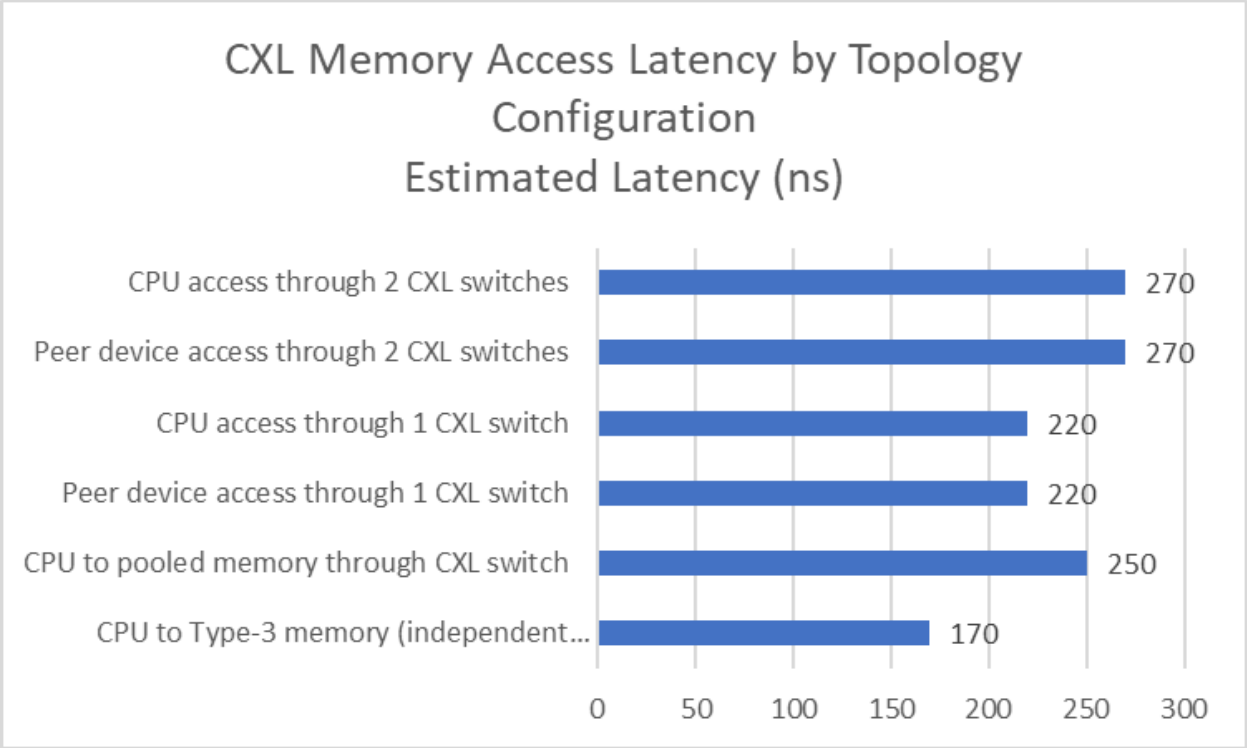
PCIe- and CXL-based interconnects, with their high bandwidth and low latency, are relevant solutions for many of the issues of heterogeneous system architectures. For example, a ×16 link at 32.0 GT/s has a raw bandwidth of 64 GB/s in each direction, and a link at 64.0 GT/s has a raw bandwidth of 128 GB/s in each direction, but protocol overheads mean real payload bandwidth varies depending on workload and traffic mix. A 68-B flit can achieve a CXL.cache read of 56.6 GB/s on a ×16 link; using a 256-B or 128-B latency optimized flit results in 112 GB/s and 104 GB/s, respectively. The estimated rate of write-intensive CXL.cache traffic is around 73.8 GB/s, moderating the added header and protocol overhead to maintain coherence [5].

Hence, the interconnect bandwidth must keep up with the accelerator performance; otherwise, host-to-device and device-to-device interconnects can be a performance bottleneck. On heterogeneous nodes with multiple accelerators, the 128 GB/s per direction raw bandwidth at 64.0 GT/s enables using higher throughput devices effectively, allowing simultaneous transfers between host memory, accelerators, and storage subsystems without severe contention. This is particularly important for distributed workloads, like deep-learning training, which involve frequent synchronization and exchange of parameters between distributed compute elements, and coupled multi-physics simulations, which require a similar kind of exchange of boundary conditions.

Another architectural trend supporting these increased bandwidths is moving to cache-coherent interconnect protocols. For example, Compute Express Link (CXL), which uses the PCIe physical layer, provides a set of coherency semantics to multiple CPUs, accelerators, and pooled memory [6]. Protocol efficiency is key to the usable bandwidth of CXL. CXL.cache and CXL.mem modes have link efficiency (per link utilization) of 0.924 to 0.939 due to 68-B flits, depending on the utilization of the synchronization header. 256-B and 128-B latency-optimized flits have an effective slot efficiency of $15/16 \approx 0.938$. These efficiencies correspond to achievable data rates of 53.5 GB/s of read-only CXL. mem traffic in the 68-B flit mode, in part reflecting the overhead of protocol framing and coherency [6].

Latency characteristics may impact the ability of coherent interconnects to create tightly coupled heterogeneous computing. For estimated latencies, CPU access to independently attached Type-3 memory devices is ~170 ns. Accessing pooled memory through a CXL switch is ~250 ns, peer device access or CPU access through one CXL switch is 220 ns, and peer device access or CPU access through two CXL switches is 270 ns. These latencies are higher than accessing local DRAM but considerably lower than latencies of conventional network-based interconnects, allowing fine-grained synchronization and shared memory programming models on heterogeneous compute elements [6].

Collectively, these metrics show how PCIe and CXL interconnect technologies can be leveraged to build high-performance, scalable heterogeneous computing systems. High raw throughput (64-128 GB/s per direction on ×16 links), high payload throughput (53.5-112 GB/s, protocol and flit format dependent), and low latency for coherent memory and peer-to-peer traffic (170-270 ns) enable such system topologies as memory-attached accelerators, disaggregated memory pools, and coherent device fabrics. PCIe and CXL interconnects are well understood and apply to heterogeneous systems when considering balanced, scalable compute, memory, and I/O architecture.



Graph 2:Estimated Round-Trip Latencies for Heterogeneous Computing Interconnects [6]

4. Scalability and System-Level Integration Challenges

Implementing PCIe 8.0 technology at the system level raises many challenges beyond the specifications relating to the physical and protocol levels. Power consumption is a major concern; the data rate and power dissipation are related with a super-linear relationship, due to the serializer-deserializer (SerDes) circuits used to implement the ultra-high-speed signaling. Effective reductions in bit error rate with higher frequencies would require at least a doubling of the power supplied to the pre-emphasis of the transmitting circuitry, the equalization of the receiving circuitry, and clock recovery circuitry, for each doubling of the signaling data rate (per [7]). Power consumption for SerDes implementations has been numerically estimated. For example, the transition from thirty-two gigabits per second signaling to sixty-four gigabits per second signaling can involve more than double the power per lane. Doubling from there to a full 128 gigabit per second requires nearly a factor of three increase in required current, and scaling for future standards creates meaningful power delivery network (PDN) challenges for clean reference voltages to be provided across multiple high-speed interfaces, with high transient currents of over 50 A for multiple lanes within a x16 interface toggling at the same time.

Thermal management of concentrated high-speed SerDes power dissipation can also be a major issue for system implementers. High-density processing, where four to eight discrete GPUs or other high-wattage processors are contained in a rack, is being used in server platforms to support the high efficiency required in data center operations. These accelerators may expose one or more PCIe 8.0 interfaces, to which the total power dissipation of the interconnect circuitry may exceed a hundred watts in an area a few square centimeters in size [7]. The temperature in the hot spots may be difficult to control with conventional air-cooling solutions without requiring excessive airflow velocities, generating a high acoustic noise level unacceptable in the data center, as well as important fan power consumption. More advanced cooling methods, such as direct liquid cooling (DLC), where liquid-cooled cold plates are thermally bonded to a processor package, two-phase cooling, where refrigerants are evaporated, and experimental immersion cooling, are being used more in other high-performance computing installations using PCIe 8.0.

Longer distance signal integrity is also a problem in rack-scale or multi-rack PCIe networks based on switched fabric topologies. The frequency content of 128 gigabit per second PAM4 signaling is somewhat higher than the 64 gigahertz binary signal, putting it into the region where copper interconnects are typically high-loss and exhibit meaningful dispersion. Due to signal degradation in passive copper cable assemblies, PCIe 8.0 passive copper cables are limited to about one meter in length, with longer lengths exceeding the correction range of PCIe receiver equalization circuits [8]. Active copper cables with signal conditioning electronics can reach two or three meter lengths while still meeting PCIe SI margins. Retiming circuits on active copper assemblies incurs cost and latency penalties, but optical interconnects can be scaled better with respect to distance and electromagnetic interference. Contenders include active optical cables, which can implement PCIe 8.0 signaling over 10+m distances [8]. These cables also have additional latency, power, and system cost due to optical-to-electrical conversion at the cable ends. These factors differentiate it from other architectural choices, such as network-attached storage or Ethernet scale-out interconnects for distributed computing solutions that communicate outside the server.

Table 1: System-Level Integration Challenges for PCIe 8.0 Implementation [7][8]

Challenge Category	Key Issues	Mitigation Approaches
Power Consumption	Super-linear SerDes power scaling; >50 A transient currents	Advanced power delivery; Dynamic management
Thermal Management	>100 W dissipation in small areas	Liquid cooling; Two-phase cooling; Immersion cooling
Signal Integrity	High loss at 128 Gbps; ~1 m passive cable limit	Active cables (2-3 m); Optical cables (10+ m)
Cost and Latency	Retimer latency penalties; Optical conversion costs	Distance-based technology selection

5. Application Performance and Workload Characterization

PCIe 8.0 technology can increase performance depending on the workloads used. Workloads that involve streaming data processing, which dominate most data transfer models in modern HPC and AI workloads, are particularly well-suited to benefit from raw bandwidth. These applications typically involve predominantly one-way data movement and have large transfer sizes and infrequent synchronization requirements. In many applications, video processing workloads that run on GPU-accelerated platforms scale close to the interconnect bandwidth between them. This is because video applications usually read compressed video data from storage or network interfaces into host memory, decode it in a GPU accelerator, process it in the GPU memory, and stream it back over the PCIe interconnect to be encoded and transmitted [9]. Since these data flows are continuous in nature, PCIe bandwidth can be efficiently utilized without incurring additional protocol overhead, nor do such flows set out to deplete the credits allocated to support bursty traffic.

For many existing high-performance computing (HPC) applications with domain decomposition methods, where the domain is divided among multiple processing elements that operate mostly independently on local data, with periodic phases for communication, like exchanging boundary condition information or performing global reductions, the perceived performance impact of PCIe 8.0 will depend on the specific algorithms and implementations. For message-heavy applications, the effects of latency rather than bandwidth are often paramount. At strong scaling, when the problem size is kept constant while the number of processors is increased, the local working set size is reduced, leading to a relatively larger

number of communication events [9]. One such application is molecular dynamics simulation, where many small messages of atomic position and force vector updates are passed between neighboring domain partitions at every timestep. While PCIe 8.0 has built-in protocol optimizations for reducing small message latency, the underlying electrical signaling and multi-layer protocol stack limit the potential for such reductions compared to other specialized and purpose-built low-latency interconnect technologies that have emerged for such usage models.

Machine learning workloads, including the distributed training of large-scale neural networks, could benefit considerably from PCIe 8.0 as their workloads are a mix of compute and communication. Modern distributed training systems typically use data parallelism, where each training batch is split among multiple accelerators executing the computation in parallel. Each accelerator computes the forward and backward passes on its subset of the cluster, after which the gradients across all devices participating in DDP are communicated to reach all other devices, followed by a parameter update [10]. The all-reduce operation dominates the communication cost, which scales with the model parameters, and a timing deadline must be respected to avoid blocking computation. In large language models, the per-iteration gradient tensor is hundreds of gigabytes in size and requires interconnect bandwidth. For example, state-of-the-art language models have over hundreds of billions of parameters. Empirical experiments show that distributed training time decreases with increased interconnect bandwidth. It also enables much larger batch sizes per accelerator without increasing the time per training step (which has been demonstrated in [10]). This is important for training since larger per-device batch sizes give statistically more accurate gradient approximations, and make better use of the accelerator by increasing the overall arithmetic intensity.

Table 2: Application Performance Characteristics with PCIe 8.0 Technology [9][10]

Workload Type	Communication Pattern	Performance Driver	PCIe 8.0 Impact
Streaming Data Processing	Unidirectional, large transfers	Raw bandwidth	High - scales with interconnect bandwidth
Domain Decomposition HPC	Small messages, frequent sync	Message latency	Moderate - limited by protocol stack latency
Distributed ML Training	All-reduce, gradient exchange	Bandwidth and timing	High - enables larger batches, faster training

Conclusion: Strategic Considerations for Next-Generation HPC Infrastructure

PCI Express 8.0 is an important evolutionary addition to the long-running PCI Express interconnect standard. PCI Express 8.0 provides bandwidth capabilities specifically targeted to exascale computing class systems and next-gen artificial intelligence systems. It also provides technical advancements in physical layer signaling, protocol, and error correction that build upon the technical foundation of today's heterogeneous compute systems at the heart of modern high-performance compute strategy. The performance gains across the workload landscape are very modest for latency-sensitive workloads to near-linear scaling for bandwidth-saturating streaming workloads, showing both the value of the specification, as well as the need for workload-aware system configuration and tuning. Deploying and fully leveraging PCIe 8.0 technology is predicated on a system-level view of challenges that could far exceed that of the interconnect specification. The power supply requirements of SerDes and the need for voltage regulation for sensitive analog circuits mean that power delivery networks have high power consumption and must also consider temperature. Hotspotting of high-speed transceiver power leads to

difficult thermal dissipation and trade-offs in efficieny, cost-effective, acoustic, and reliability considerations. Managing these issues through different channel implementations also requires close collaboration among silicon designers, board designers, and system integrators, as well as ensuring that the physical layer margins meet specifications under manufacturing, environmental, and product life conditions. While technically challenging, these issues are manageable by combining best practices from other industries with the ecosystem of tools, components, and expertise developed around the PCIe standard. In addition to the performance, PCIe 8.0 offers multivendor interoperability, vendor choice, and a long-term roadmap for the PCIe ecosystem. PCIe is developed as an open, industry-standard interface in a consensus-based process with semiconductor vendors, system vendors, and end consumers, giving it a clear multivendor interoperability advantage over proprietary interconnects in the industry. This allows a competitive market for improvements and efficiencies to emerge, and minimizes the risk of technology lock-in, maintaining incompatible proprietary technologies. This can make the cost or feasibility of upgrading to a new proprietary platform for infrastructure difficult. In addition, the specification's backward compatibility with prior generations of PCIe can allow modular strategies to be employed where components can be upgraded one at a time. PCIe 8.0 is expected to be a suitable base technology for the entire next generation of compute architectures that could be supported as far out as the late decades of the 21st century, as systems evolve from today's exascale systems to zettascale systems in the future. Its architecture is suitable for implementations from mainstream servers to composable infrastructure to rack-scale computing systems. Workloads are increasingly data-centric, driven by the advent of artificial-intelligence workloads, new scientific instruments generating massive and complex datasets, and a nearly ubiquitous requirement for real-time analytics. PCIe 8.0 provides high-performance interconnect technologies for each of these data-centric workloads. The PCIe 8.0 specification lays the foundation for scientific breakthroughs and technology innovations that will follow.

References

[1] Grand View Research, "High Performance Computing Market (2025 - 2030)," [Online]. Available: https://www.grandviewresearch.com/industry-analysis/high-performance-computing-market

[2] IBM Corporation, "The convergence of HPC and AI: Driving innovation at speed," [Online]. Available: https://www.ibm.com/think/topics/hpc-ai

[3] PCI-SIG, "PCI Express® Base Specification Revision 5.0 Version 1.0," 2022. [Online]. Available: https://picture.iczhiku.com/resource/eetop/SYkDTqhOLhpUTnMx.pdf

[4] Vijay Nagarajan, et al., A Primer on Memory Consistency and Cache Coherence, Second Edition," Springer Nature Link, 2020. [Online]. Available: https://link.springer.com/book/10.1007/978-3-031-01764-3

[5] MDPI Electronics, "New Trends for High-Performance Computing," 2022. [Online]. Available: https://www.mdpi.com/journal/electronics/special_issues/new_trends_HPC

[6] Debendra Das Sharma, et al., "An Introduction to the Compute Express Link (CXL) Interconnect," ACM Digital Library, 2024. [Online]. Available: https://dl.acm.org/doi/full/10.1145/3669900

[7] M. Horowitz et al., "Scaling, Power, and the Future of CMOS," ResearchGate, 2005. [Online]. Available: https://www.researchgate.net/publication/4230730_Scaling_power_and_the_future_of_CMOS

[8] A.K. Kodi; A. Louri, "Design of a high-speed optical interconnect for scalable shared memory multiprocessors," IEEE, 2005. [Online]. Available: https://ieeexplore.ieee.org/document/1375210

[9] John Kim, "Technology-Driven, Highly-Scalable Dragonfly Topology," IEEE, 2008. [Online]. Available: https://ieeexplore.ieee.org/document/4556717

[10] Alexander Sergeev, Mike Del Balso, "Horovod: fast and easy distributed deep learning in TensorFlow,"arXiv, 2018. [Online]. Available: https://arxiv.org/abs/1802.05799