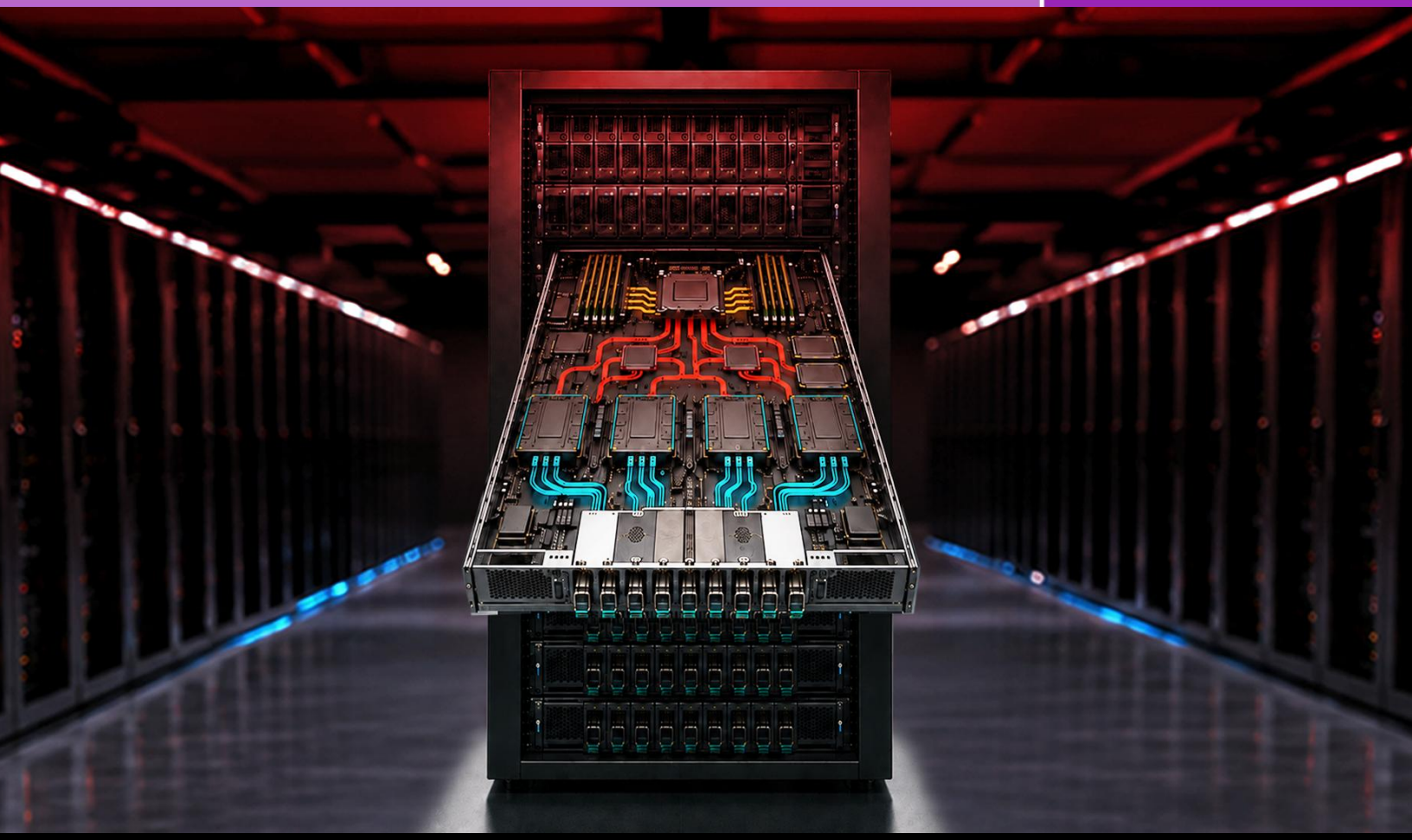
WHITE PAPER



Scaling Up AI Compute

Validating compute I/O, memory, and accelerator fabric interfaces

KEYSIGHT

Moving Compute Data at AI Scale

As artificial intelligence (AI) models grow larger, a major limiting factor for AI infrastructure is how efficiently massive amounts of data can flow within and between compute racks in the data center. Scale-up architectures address this challenge through high-bandwidth, low-latency connections among tightly coupled compute resources, complementing the scale-out networks that connect resources across larger AI deployments.

To support these demands, the industry is advancing high-speed interface technologies across three domains: compute input/output interfaces (I/O), memory, and accelerator fabrics. Each evolves to address different bandwidth, capacity, latency, and power constraints. Together, they help determine how effectively a system uses its available compute. For validation engineers, these advances introduce tighter electrical margins, more sophisticated signaling and training, and greater interoperability demands. Meeting individual interface requirements is essential, but reliable operation also depends on how those interfaces behave within the integrated system.

This paper examines the challenges of scaling up AI infrastructure, the evolution of new compute I/O, memory, and accelerator fabrics, and the growing importance of validation in deploying reliable AI systems. It highlights common engineering challenges and explains how transmitter and receiver testing, margin characterization, and scalable validation strategies help teams keep pace with evolving interface requirements.

Understanding AI Scale-Up

AI infrastructure can expand at several levels, from increasing chip capabilities to connecting compute resources across multiple distinct data centers. Each approach addresses different performance and capacity needs and introduces distinct validation challenges. Understanding these approaches can help clarify the role of scale-up in supporting AI workloads.

AI Infrastructure Scaling Approaches

How organizations scale their AI infrastructure depends on their primary objective: whether to increase compute density, expand capacity, improve resource utilization, or support larger workloads. These approaches are often complementary, and modern AI deployments may combine several scaling strategies. The following overview distinguishes scale-in, scale-up, scale-out, and scale-across by where and how they expand compute capability.

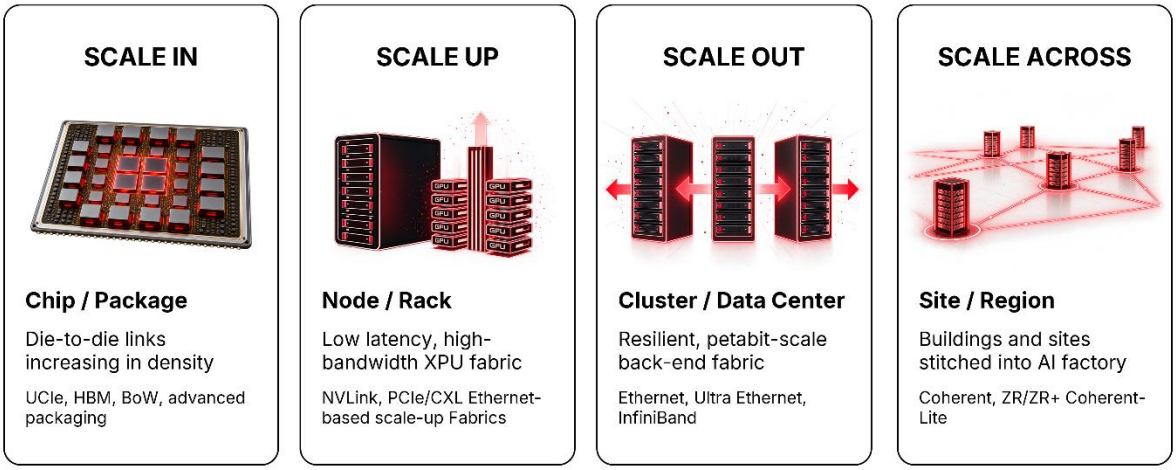


Figure 1. Complementary approaches to scaling AI infrastructure, from chip and package integration to tightly coupled compute systems, networked clusters, and connectivity across sites

Scale In

AI scale-in refers to increasing compute capability within a chip or package through advances in semiconductor design, packaging, and memory integration. Examples include chiplets connected through die-to-die interfaces such as Universal Chiplet Interconnect Express (UCIe) and stacked high-bandwidth memory (HBM) integrated near the processor. These approaches bring compute and memory resources closer together to support greater bandwidth and compute density.

Scale Up

AI scale-up connects processors and accelerators within a tightly coupled compute domain through high-bandwidth, low-latency interfaces. Depending on the architecture, this domain may span a server, a rack, or multiple racks. The goal is to enable participating resources to exchange data efficiently and work together on demanding AI workloads.

Scale Out

AI scale-out expands overall compute capacity by connecting additional servers or compute domains through a network. Technologies such as Ethernet and InfiniBand support communication among these distributed resources for AI training and inference. Scale-out networks complement scale-up connectivity, allowing deployments to grow across larger clusters.

Scale Across

AI scale-across refers to connecting AI infrastructure across geographically distributed data centers. Inter-data-center networking, including coherent optical links, enables organizations to move data and distribute workloads across sites. Distance and network latency influence which workloads can operate efficiently across these locations.

Why Scale Up Matters for AI Innovation

While scale out remains essential for large AI deployments, increasing limitations on power, cooling, and footprint have driven greater demand for scale-up architectures that improve resource utilization within a compute rack. Modern accelerators need rapid access to memory and neighboring resources, making the bandwidth, latency, and reliability of connecting interfaces critical to system performance. For engineering teams, validation must therefore consider both individual interfaces and their behavior within the integrated system, where signal integrity and interoperability help determine how effectively available compute capacity can be used.

The Three Domains of AI Scale-Up

Three interconnected technology domains support AI scale-up: compute I/O, memory, and accelerator fabrics. Each serves a distinct role in moving data through the system, as illustrated in Figure 2.

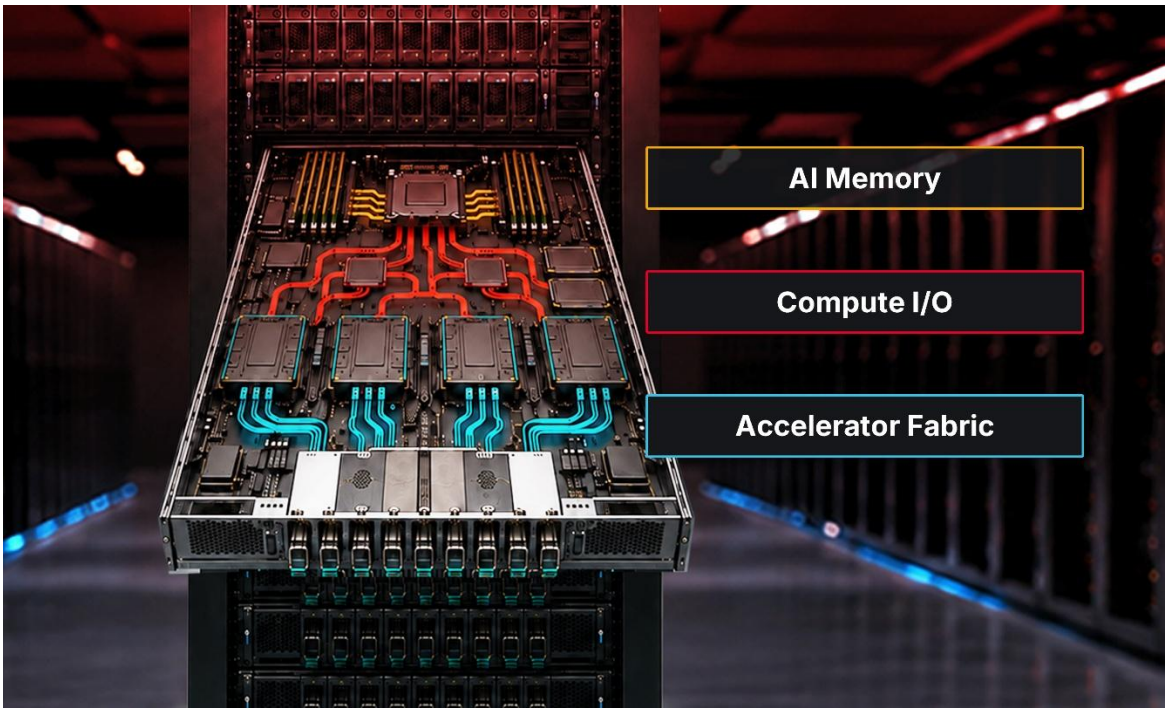


Figure 2. Generalized AI server tray showing memory (yellow), compute I/O (red), and accelerator fabric (blue) traces and cables

Compute I/O

Compute I/O connects processors, accelerators, switches, and other devices within the system. Peripheral Component Interconnect Express (PCIe®) provides high-speed device connectivity, while Compute Express Link (CXL) builds on PCIe physical-layer infrastructure to support coherent caching and memory access.

AI Memory

Memory technologies supply compute resources with the data they need, balancing bandwidth, capacity, power, and packaging requirements. Double Data Rate (DDR) and Low-Power Double Data Rate (LPDDR) serve system memory needs, while Graphics Double Data Rate (GDDR) and high-bandwidth memory (HBM) provide memory bandwidth for graphics and accelerator workloads through different interface and packaging approaches.

Accelerator Fabric

Accelerator fabrics connect multiple accelerators within a scale-up domain, enabling high-bandwidth, low-latency data exchange for shared workloads. Examples include Ultra Accelerator Link (UALink), NVLink, and Ethernet-based scale-up fabrics.

Why New Standards Emerge

Compute I/O, memory, and accelerator fabrics evolve to address different limits on AI performance. Some advances increase transfer rates, while others expand accessible memory or reduce communication overhead. Understanding these drivers helps explain both the diversity of interface technologies and the validation challenges they introduce.

I/O Bandwidth Constraints

Limited I/O bandwidth can restrict data exchange between processors, accelerators, and other devices. The PCIe 7.0 standard addresses this constraint with a transfer rate of 128 GT/s per lane, twice that of PCIe 6.x. CXL addresses a related challenge: connecting compute resources to memory more flexibly. Built on PCIe physical-layer infrastructure, it supports coherent caching and memory access, with expansion and pooling capabilities depending on the specification revision and implementation.

Memory Bandwidth and Access

AI workloads need both sufficient memory capacity and enough bandwidth to keep compute resources supplied with data. Memory technologies address these requirements through different approaches: DDR6 supports greater system memory capacity and bandwidth, LPDDR6 emphasizes power efficiency, and GDDR7 uses PAM3 signaling to increase data transfer rates. HBM combines a wide interface with stacked memory integrated near the processor to deliver high bandwidth. These approaches create different tradeoffs in power, packaging, and signal access.

Accelerator Communication

As more accelerators work on a shared workload, exchanging data among them can limit the benefit of adding compute capacity. NVLink, UALink, and Ethernet-based scale-up approaches such as Ethernet for Scale-Up Networking (ESUN) address this challenge through high-bandwidth, low-latency connectivity. Their different protocols, topologies, and recovery mechanisms introduce distinct requirements for validating reliable communication across the scale-up domain.

Table 1. Selected I/O, memory, and scale-up fabric standards.

Standard	Supported data rate	Role in data center compute
PCIe 7.0	128 GT/s per lane	Connects processors, accelerators, and other high-speed devices
CXL 4.0	128 GT/s per lane	Enables coherent memory access, expansion, and pooling
DDR6	*Pending release*	Supplies system memory for data-intensive compute workloads
LPDDR6	Up to 14.4 Gb/s per pin	Provides power-efficient memory for AI compute systems
GDDR7	Up to 192 GB/s per device	Supplies high-bandwidth memory for graphics and AI accelerators
HBM4	~2 TB/s per stack	Supplies high-bandwidth memory close to compute resources
UALink 2.0	200 Gb/s per lane	Connects accelerators through an open scale-up fabric
NVLink 6	3 TB/s per GPU	Connects GPUs for high-bandwidth scale-up communication
ESUN 1.0	PHY-dependent	Enables Ethernet-based scale-up communication among compute resources

The Validation Challenges of AI Scale-Up

As new standards emerge to address specific bottlenecks, engineering teams must validate more complex interfaces with tighter margins across a wider range of architectures and operating conditions. Many physical-layer challenges are common across standards, including channel loss, shrinking signal timing and noise margins, equalization techniques, and interoperability risks. The most common challenges span interface complexity, signaling margins, architectural diversity, and reliable system operation.

More Interfaces, More Complexity

Modern AI systems combine processors, accelerators, memory interfaces, switches, and fabric connections within the same platform. Each introduces operating modes, training behavior, and interoperability considerations that expand test coverage. Validation teams must account for interactions among these interfaces, including shared power and thermal conditions and simultaneous switching activity. Failures that appear only under combined activity can be harder to reproduce and isolate than those observed on an individual link.

Faster Signaling, Smaller Margins

Higher signaling rates and multilevel modulation increase the demands on electrical channels. PCIe 7.0's higher transfer rate and GDDR7's use of PAM3 illustrate different approaches to increasing bandwidth. These advances can tighten timing and voltage budgets, increasing sensitivity to jitter, noise, crosstalk, channel loss, and equalization behavior. Engineers must understand how these impairments affect signal quality and consume available margin.

Where supported, equalization, adaptive training, and forward error correction help maintain reliable communication, but also introduce validation challenges. A transmitter measurement alone cannot establish how a receiver responds to impaired signals or changing channel conditions. Validation must therefore address both transmitter behavior and receiver tolerance, including the applicable training and error criteria. Additional data rates, operating modes, and equalization settings expand coverage and increase the effort required to characterize the link.

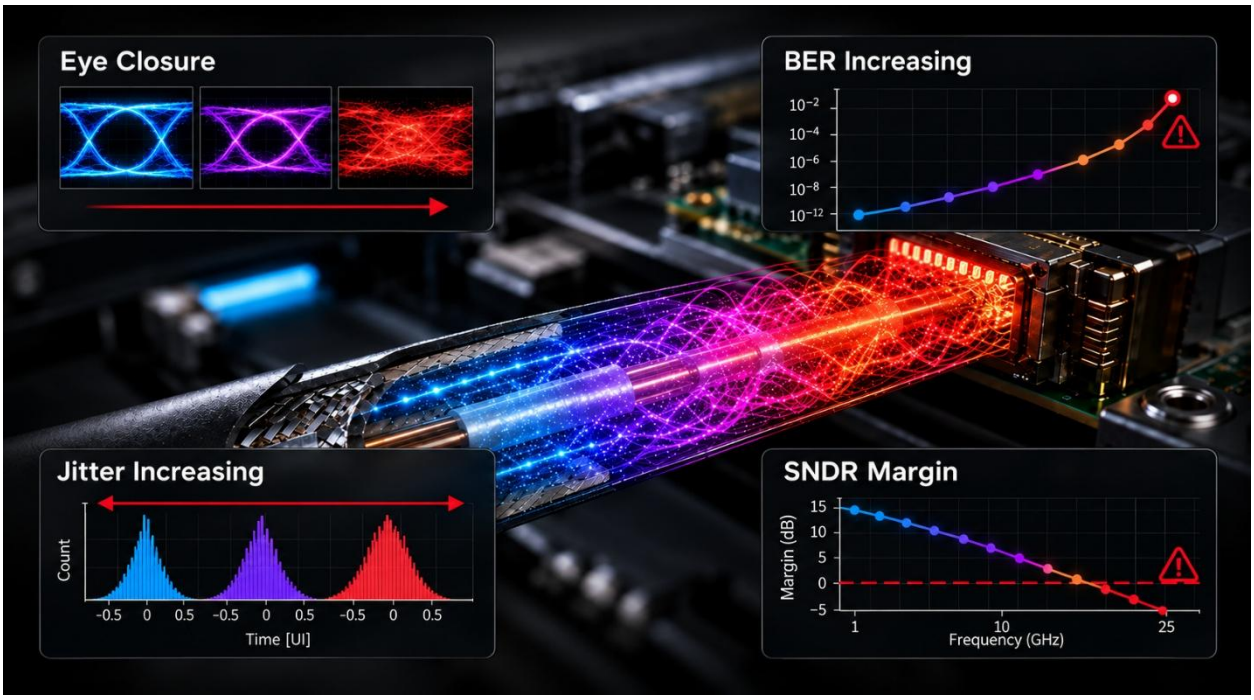


Figure 3. Effects of signal degradation on eye closure, jitter, bit error rate (BER), and signal-to-noise-and-distortion ratio (SNDR) margin

Architectural Diversity

AI compute platforms use different combinations of compute I/O, memory, and accelerator fabrics to meet bandwidth, capacity, latency, power, and packaging requirements. These choices create distinct validation needs. CXL introduces coherency and memory-access behavior beyond the underlying electrical connection, while board-level memory and package-integrated HBM present different signal-access constraints. Accelerator fabrics also differ in topology, protocol behavior, and recovery mechanisms.

Teams supporting multiple architectures must accommodate different modulation schemes, testpoints, and compliance procedures while keeping results consistent and comparable where appropriate. The challenge is to reuse validation workflows without overlooking the requirements specific to each interface, and to adapt those workflows as specifications and implementations evolve.

Compliance, Interoperability, and System Risk

Compliance testing establishes conformance to defined requirements at specified test points and operating conditions. It provides a common baseline across implementations, but does not guarantee interoperability or sufficient operating margin in every deployed configuration. As standards add measurements, modes, and training requirements, teams must maintain consistent testing across products and development cycles.

Devices that pass isolated compliance tests may still encounter problems with particular link partners, traffic patterns, environmental conditions, or topologies. These problems can appear as errors, repeated training, recovery events, or reduced throughput at the link, fabric, or system level. Building confidence in deployment therefore requires connecting compliance results with margin characterization and interoperability testing under representative system conditions.

Building Confidence in AI Compute Validation

Confidence in AI compute systems comes from connecting evidence across compliance, electrical characterization, and system integration. A useful validation strategy establishes whether interfaces meet their requirements, how much operating margin they retain, and how they behave with other devices under representative conditions. Repeatable workflows help teams maintain that evidence as architectures and standards evolve.

Compliance as the Foundation

Compliance testing begins with evaluating transmitters, receivers, and channels against applicable standards-defined requirements. It provides a starting point for interoperability and a consistent basis for assessing implementations.

That baseline must be understood in the context of the specification revision, test procedure, and conditions used. Integration testing extends the assessment to representative link partners, topologies, and workloads, helping identify issues that isolated compliance measurements may not reveal.

Transmitter and Receiver Validation

Reliable communication depends on both the quality of the transmitted signal and the receiver's ability to recover data from it. Transmitter validation characterizes applicable measures of jitter, noise, distortion, and eye quality. Receiver validation evaluates tolerance to defined signal impairments, including the effects of channel loss, jitter, and interference.

As interfaces rely on more sophisticated equalization and adaptive training, receiver behavior becomes critical to understanding link robustness. Evaluating both ends helps engineers distinguish signal-quality limitations from receiver-tolerance or training issues. The measurements and test methods differ by interface, but together they provide a more complete basis for diagnosing link behavior.

Characterize Operating Margin

Compliance may determine conformance to the standard, but margin analysis helps quantify robustness. A device that barely passes compliance tests may behave very differently from one with substantial operating margin. A design with significant electrical margin is generally better positioned to tolerate manufacturing variation, environmental changes, channel impairments, and interoperability differences than one operating close to specification limits.

As timing and voltage margins shrink across modern interfaces, engineering teams increasingly need to understand how much margin exists, where it is consumed, and how it changes across operating conditions. Margin analysis helps quantify design robustness, evaluate interoperability risk, compare implementations, and identify optimization opportunities before issues emerge during system integration.

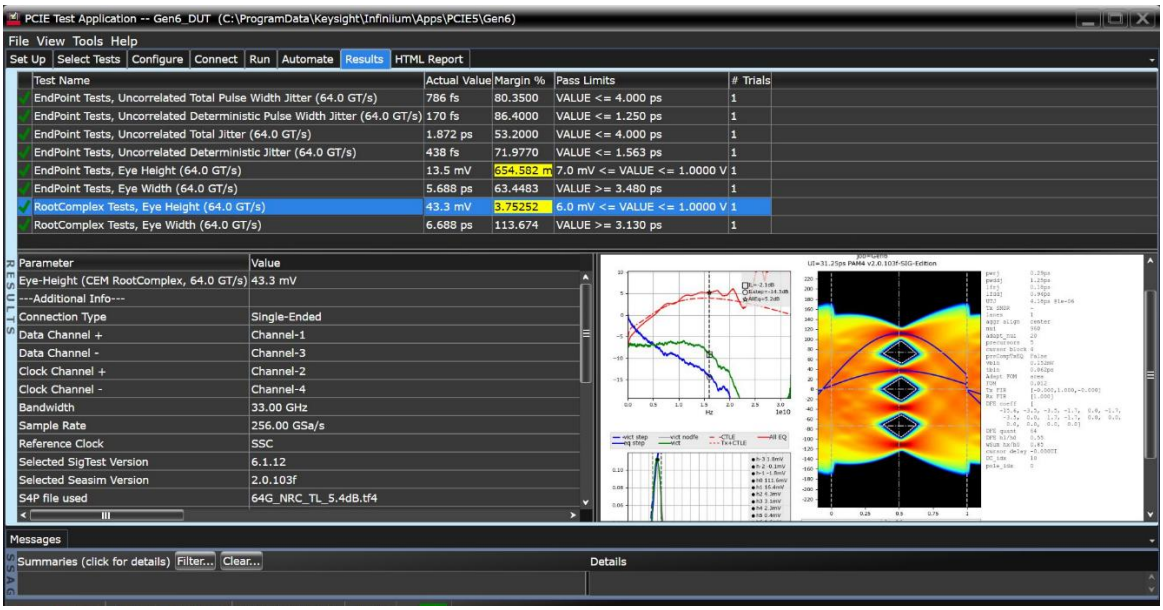


Figure 4. Example PCIe 6.0 compliance report showing measured values, specification limits, and margin

Repeatable Workflows and Automation

Digital interface standards are advancing faster than traditional validation workflows can adapt, making manual test methodologies increasingly difficult to sustain. Each new generation can add new measurements, operating modes, training sequences, calibration requirements, and validation scenarios that expand test effort.

Validation teams must cover more interfaces and operating conditions while keeping results consistent across projects, locations, and product generations. Automation supports this effort by coordinating setup, calibration, test execution, and reporting, reducing manual variation and making broader characterization more practical.

Repeatability also requires traceability. Retaining the relevant specification and software revisions, device configuration, test conditions, and measurement data allows teams to reproduce results and investigate differences between runs. As standards mature or designs change, updated automation workflows make it easier to update coverage and rerun affected tests. This gives engineering teams more time to investigate marginal behavior and resolve integration issues while preserving a consistent basis for validation decisions.

Validation Methods Must Scale with AI Compute

As AI interfaces evolve, validation teams need approaches that can adapt to changing requirements and growing test coverage. Repeatable workflows and automation help teams maintain consistent results across interface generations, device configurations, and operating conditions. Their value extends beyond test efficiency: they provide a dependable basis for assessing margin, investigating interoperability issues, and making integration decisions.

Building confidence requires connecting compliance results with transmitter and receiver characterization and system-level testing. Together, these approaches help engineers understand where operating margin is consumed and whether interfaces perform reliably with representative link partners and workloads.

Across AI compute infrastructure, performance depends on how effectively compute I/O, memory, and accelerator fabrics operate together. Scaling validation capabilities alongside these technologies helps teams identify limitations before deployment, reduce integration risk, and realize the performance that scale-up architectures are designed to deliver.

Explore the accompanying solution briefs for interface-specific approaches to AI compute I/O, memory, and accelerator fabric validation:

- Validating PCIe 7.0 Compliance for AI Compute
- Validating High-Speed Memory for AI Systems
- Validating Ethernet and UALink for AI Scale-Up Fabrics

PCIe® and the PCI Express® are US registered trademarks and/or service marks of PCI-SIG.

As a mission-critical design enablement partner, Keysight connects design, emulation, and test so customers can solve the world's most complex engineering challenges. Accelerate innovation at www.keysight.com.

KEYSIGHT

This information is subject to change without notice. © Keysight Technologies, 2026, Published in USA, September 30, 2026, 7126-1490.EN