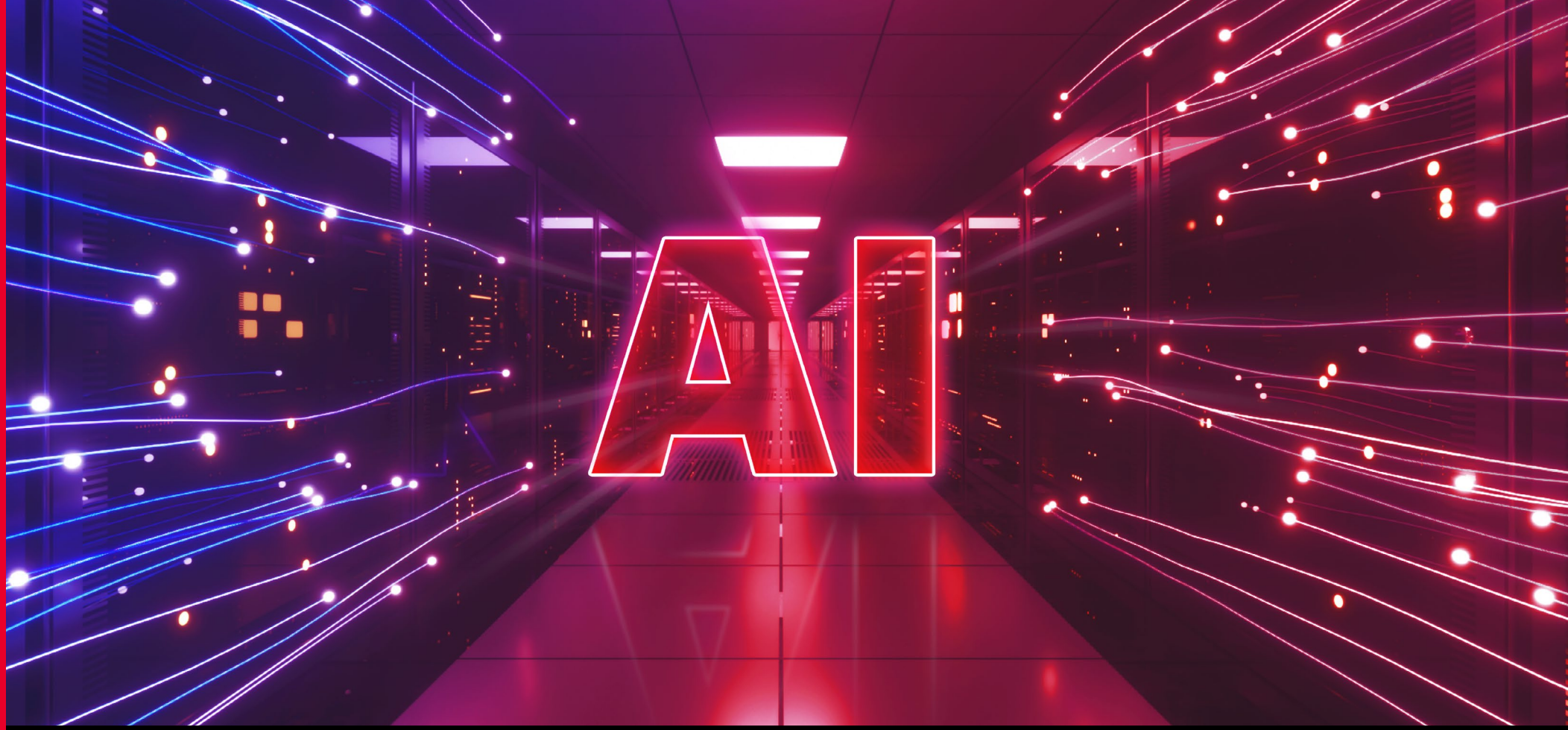


5 Key Strategies to Optimize and Scale AI Data Centers

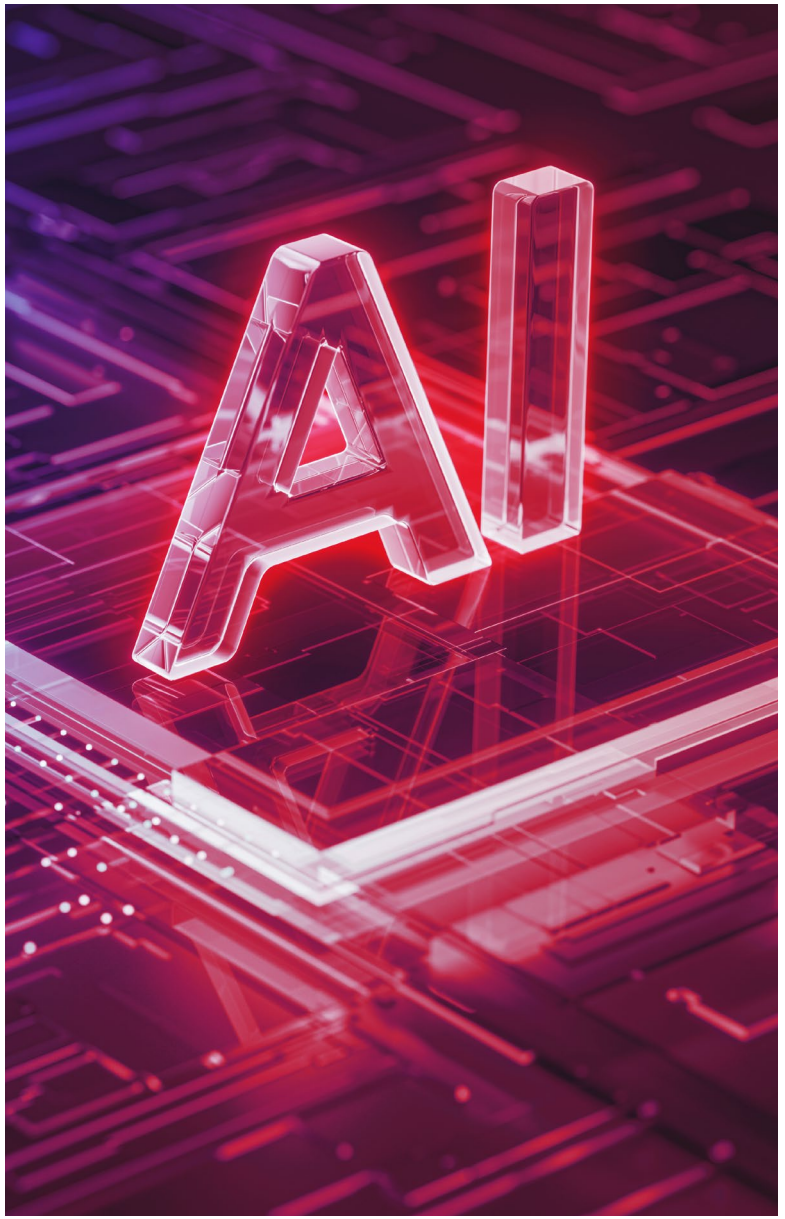
eBook

KEYSIGHT

Executive Summary

Artificial intelligence (AI) is transforming industries and driving innovation. As AI evolves and more participants enter the field, it places increasing demands on data centers to manage unique traffic patterns, dynamic workloads, and relentless performance pressures.

Unlike traditional data centers, AI data centers require seamless integration of components like graphics processing units (GPUs), cables, and servers. Furthermore, component-level validation does not guarantee system-level performance. Even minor integration issues can escalate into critical problems. Traditional validation methods often fail to meet the demands of real-world AI workloads, pushing the boundaries of established industry standards. This eBook outlines practical solutions to optimize AI data center performance for modern applications.



Contents

CHALLENGE

AI Is a Transformative Force

GO TO CHALLENGE >

KEY 1

Scaling AI-ready Data Infrastructure

GO TO KEY 1 >

KEY 2

Designing AI-ready High-speed Digital Compute Products

GO TO KEY 2 >

KEY 3

Optimizing Interconnects for 1.6T and Beyond

GO TO KEY 3 >

KEY 4

Adapting to Higher Network Speeds and Bandwidths

GO TO KEY 4 >

KEY 5

Improving Power Efficiency to Scale AI Workload

GO TO KEY 5 >

CHALLENGE

AI Is a Transformative Force

CHALLENGE

AI Is a Transformative Force

Artificial intelligence (AI) is rapidly emerging as a transformative force across industries, driving innovation, enhancing efficiency, and reshaping business strategies. As organizations increasingly adopt AI technologies to gain a competitive edge, the market is experiencing unprecedented growth.

Projections show that this expansion will have a profound economic impact, fueling both global and domestic markets at an accelerating pace. This economic potential draws more startups and established industry players into the fold. As the number of users and providers continues to grow, the demand for AI-ready computational power has never been greater. With no end in sight to this market's growth, AI data centers must be able to evolve and scale at the rapid pace of AI innovation. Figure 1 represents the projected scale and significant impact of AI's economic growth.

$1.3 Trillion

AI market predicted to reach **$1.3T** by 2030 — up from 86.9 billion in 2022.

Bloomberg

35.7% Growth

AI market projected to **grow at a 35.7%** annual rate from 2024 to 2030.

MarketsandMarkets

21% Increase

Market growth in AI projected to contribute a **21% net increase to US GDP** by 2030.

Forbes

Figure 1. Market growth predictions for AI

AI Is Changing the Way We Design, Deploy, and Validate Data Centers

Meeting growing demands means delivering unprecedented performance, reliability, and efficiency. An AI data center is only as reliable as its weakest link. On the forefront of performance, every chip, cable, interconnect, switch, server, and graphics processing unit (GPU) represents both potential and risk. Individual components must not only have to work independently but also operate cohesively as a system, even under relentless demand.

While industry standards provide a baseline, real-world AI workloads operate far beyond those limits. When transactions occur billions of times per second across thousands of interconnected resources, even the smallest degradation can create bottlenecks that jeopardize performance, effectiveness, and return on investment (ROI), as illustrated in Figure 1.

In AI data centers, these challenges are not exceptions; they are the norm. This eBook examines some of the key challenges AI data centers present and the practical strategies and recommendations to overcome them.

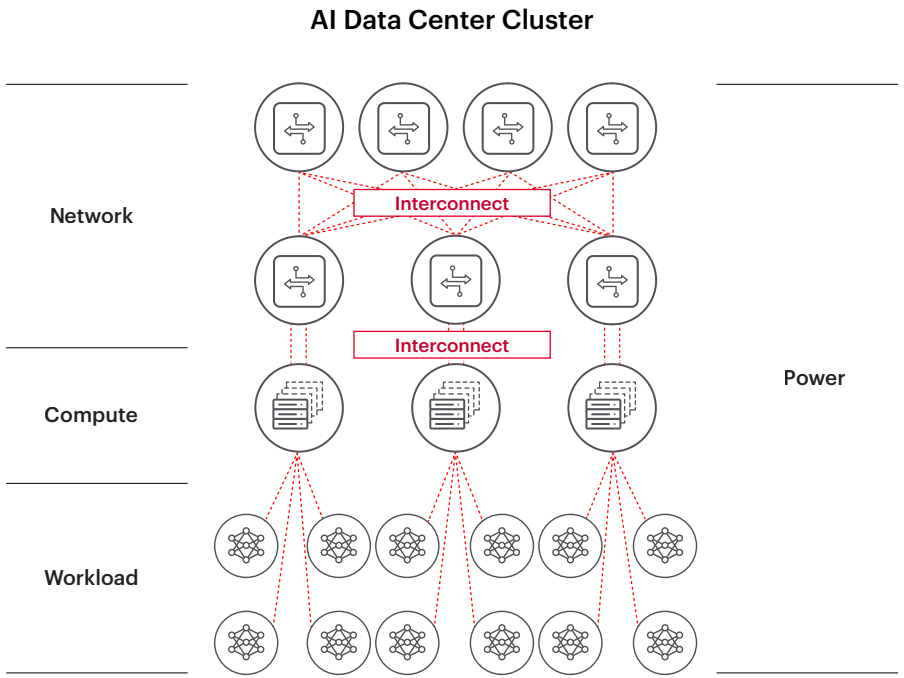


Figure 2. Illustration of an AI data center cluster

KEY 1

Scaling AI-ready Data Centers

KEY 1

Scaling AI-ready Data Centers

The Challenges

Training large language models (LLMs) demands thousands of connected compute nodes, making network performance vital for cost efficiency. However, distributed multi-GPU servers, such as those used in GPT-4, can suffer bandwidth limitations due to slower inter-server links, resulting in resource inefficiencies. In addition, these constant high-scale workloads stress data center components — leading to degradation and unexpected network failures. Figure 3 illustrates that in one study, failure rates reach up to 43.4% for the most resource-intensive LLM training tasks.

That is why it is crucial to identify and strengthen weak data center links before deployment. However, traditional component-level testing struggles to accurately model traffic patterns, workload distributions, and hardware interactions — causing inefficient GPU utilization and elevated operational costs. Cohesive, system-level validation ensures AI training's scalability, reliability, and cost-effectiveness.

> A model like GPT-4 took roughly three months and $100 million to train, yet an estimated $30 million of that was wasted due to graphics processing units (GPUs) sitting idle waiting for data.
>
> Morningstar

The Recommendation

Emulate AI workloads and benchmark performance to gain insights into system efficiency and potential bottlenecks. Figure 4 depicts how replicating AI data center behavior across physical, protocol, and application layers is essential for network architects and equipment manufacturers to optimize data center infrastructure, improve scalability, and ensure reliable system-wide performance.

Streamline AI data center development

Accelerate AI infrastructure design and data center deployment by emulating real-world AI workloads with measurable fidelity using the Keysight AI (KAI) Data Center Builder. With pre-packaged benchmarking applications developed with leading AI operators and infrastructure vendors, you can evaluate how new algorithms, components, and protocols improve AI training performance. The solution supports flexible test engines for comparative benchmarking, including hardware load appliances, software endpoints, and real AI accelerators.

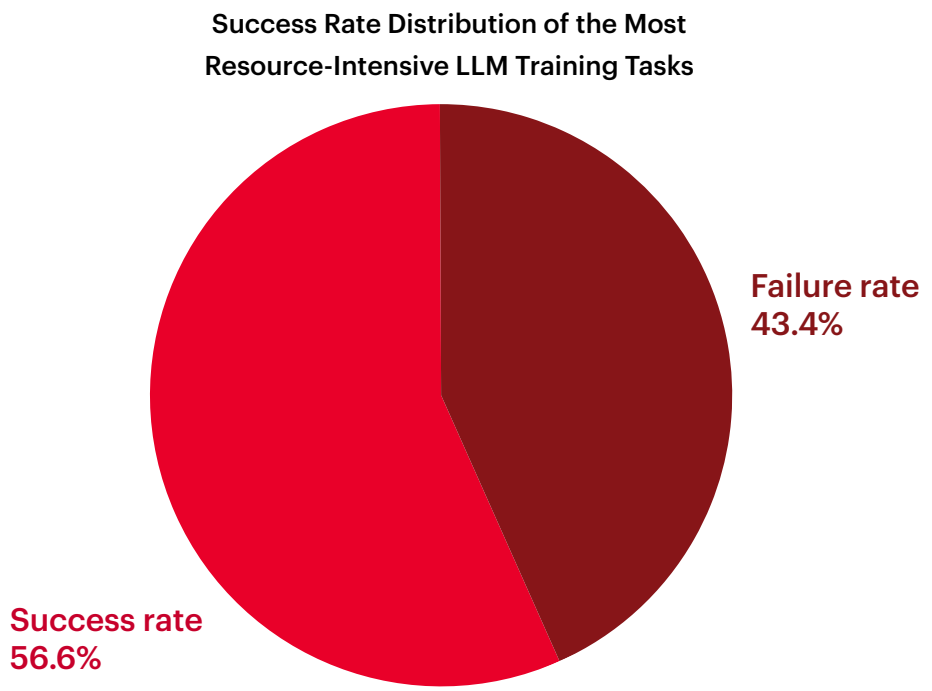
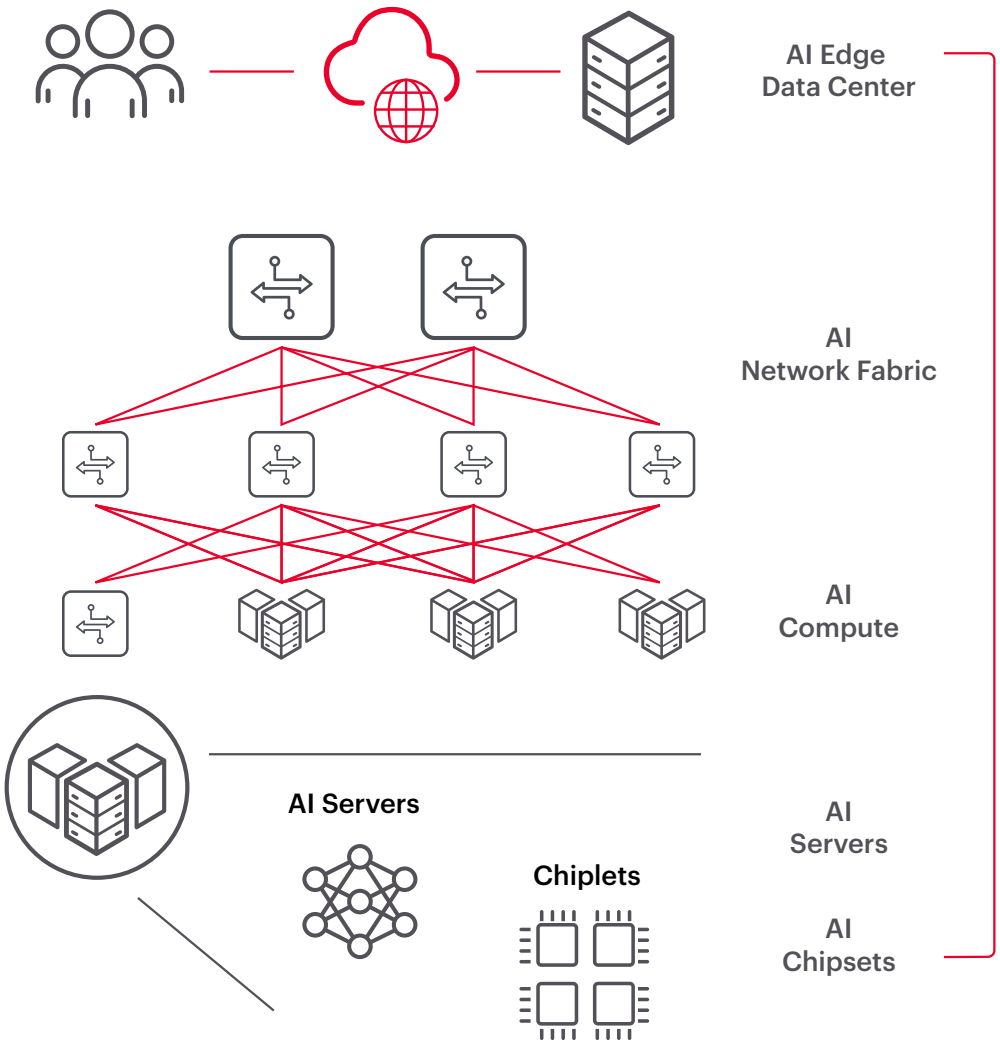


Figure 3. Success Rate Distribution of the Top 5% Most Resource-Intensive Tasks – Failure Rates Reach Up to 43.4%

Source: Unicron: Economizing Self-Healing LLM Training at Scale, Tao He, Xue Li, Zhibin Wang, Kun Qian, Jinabo Xu, Wenyuan Yu, Jingren Zhou, Alibab Group, Nanjing University



KAI Data Center Builder

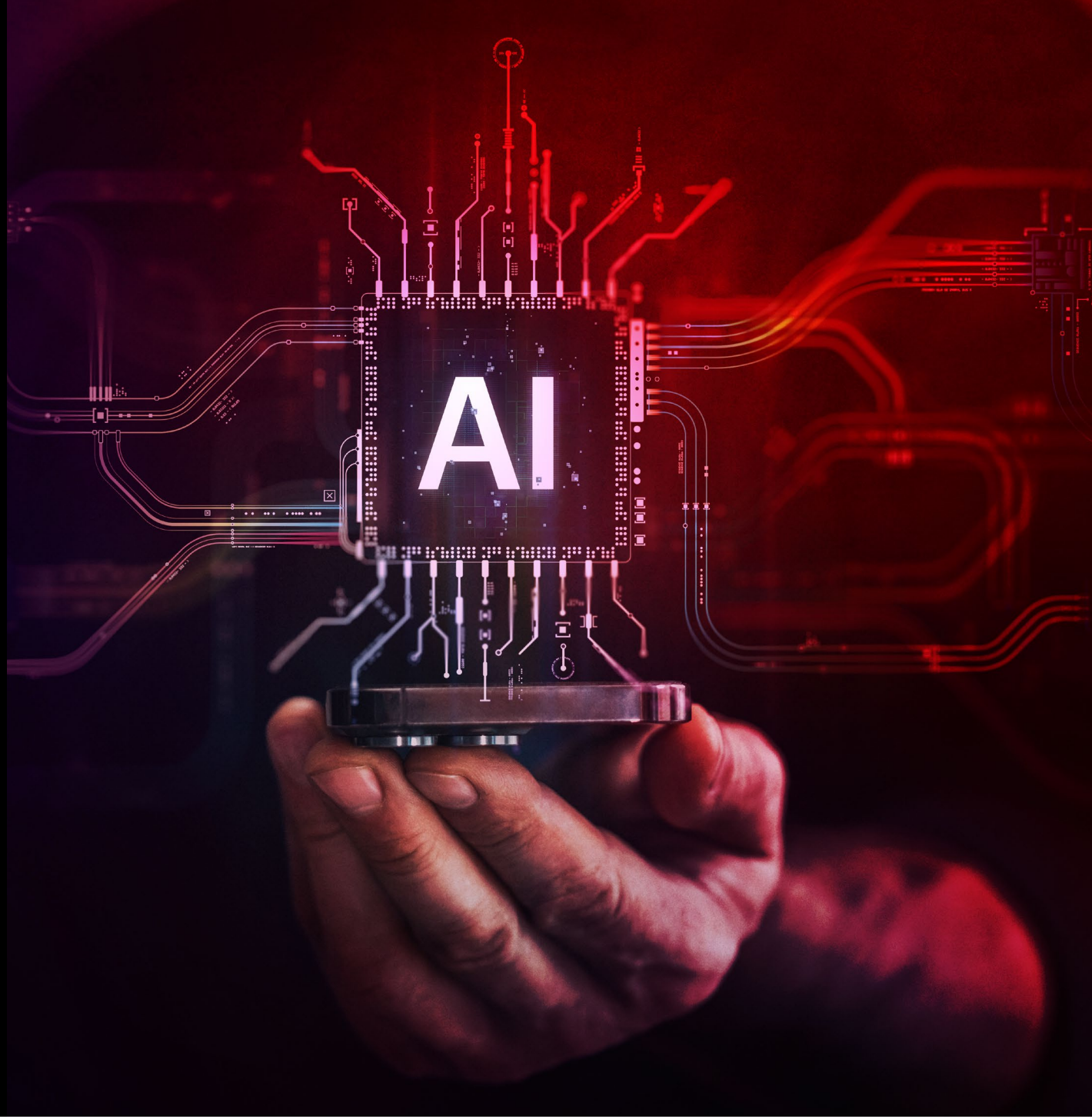
Figure 4. Depiction of AI infrastructure hierarchy — from chiplets to the network's edge

Call to Action

Learn more about KAI Data Center Builder.

KEY 2

Designing AI-ready High-speed Digital Compute Products



KEY 2

Designing AI-ready High-speed Digital Compute Products

The Challenges

The rapid evolution of standards highlighted in Figure 5, such as Ethernet, memory technologies double data rate (DDR), high bandwidth memory (HBM), PCIe®, and Compute Express Link (CXL), demonstrates the increasing demand for higher bandwidth and lower latency driven by AI applications.

For example, Ethernet speeds are set to quadruple from 400G / 800G to 1.6T / 3.2T, while PCIe transitions from Gen5 at 32 GT/s to Gen7 at 128 GT/s — a fourfold increase. These advancements amplify challenges such as noise, jitter, and crosstalk — particularly at higher frequencies and data rates.

Technologies like PAM4 signaling used in Ethernet and PCIe® 6.0 achieve higher data rates but compromise signal-to-noise ratios, necessitating innovations such as forward error correction (FEC) to maintain reliable performance.

> Major AI hyperscalers have started developing their own custom AI chips to reduce the costs of delivering AI-based services and applications.
>
> Gartner

Major AI hyperscalers are also developing custom AI chips to cut costs and optimize workloads, increasing the need for rigorous testing to meet evolving standards. At this pivotal point in technology, companies face challenges and opportunities to overcome existing limitations and drive innovation in AI-centric applications, as illustrated in Figure 5.

The Recommendation

Invest in AI-ready design and validation tools. Designing semiconductors, debugging cutting-edge high-speed digital designs, and meeting — or exceeding — the latest PCIe, DDR, and CXL standards are all crucial tasks for delivering AI-optimized data center infrastructure. That is why semiconductor producers and network equipment manufacturers alike rely on Keysight AI Compute solutions.

Design and simulation software

Ensure reliability with electronic design automation (EDA) software suite for high-speed signal and power integrity simulation and analysis of printed circuit boards (PCBs) and integrated circuits, including signal integrity tools for minimizing electromagnetic errors, designers for DDR and PCIe, and optimizing energy-efficient power delivery.

AI: A Technical Inflection Point
AI pressures for high bandwidth and low latency, offer companies a chance to leapfrog current technologies

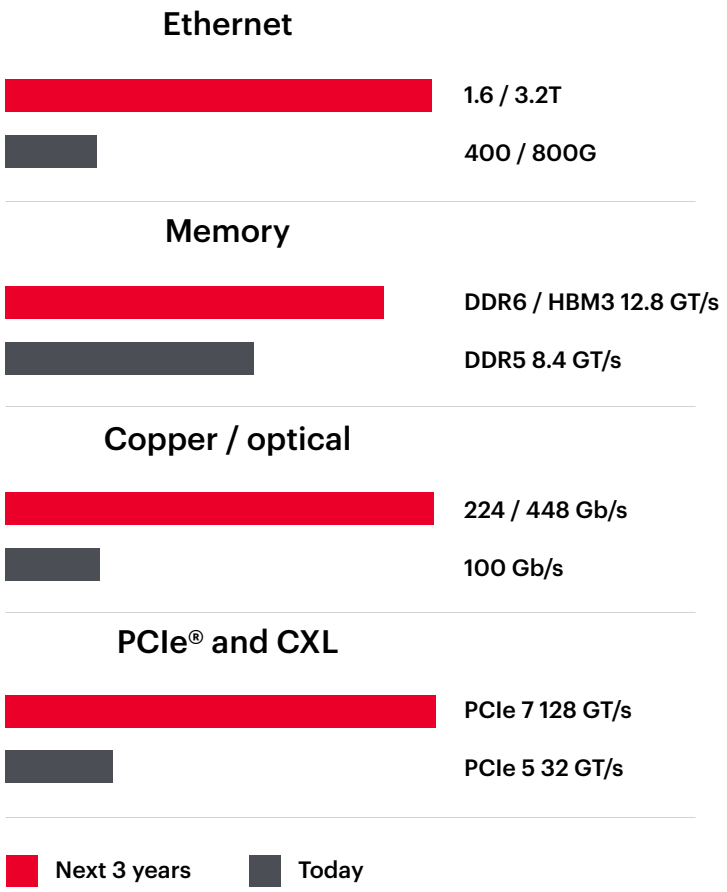


Figure 5. AI technical inflection points

Ultra-high-performance oscilloscopes

Experience up to 110 GHz bandwidth with Keysight Pro+ high-performance oscilloscopes featuring a 10-bit analog-to-digital converter (ADC) for precise measurements and advanced tools like jitter analysis and protocol decoding. Our oscilloscopes are compatible with compliance test software, which automates performance validation for high-speed standards, including PCIe, USB Type-C, and DDR5, as shown in Figure 6.

Call to Action

Discover Keysight AI Compute solutions.

Figure 6. Keysight UXR-B Series oscilloscope showing PAM4 modulated signal, with eye diagram displayed

KEY 3

Optimizing Interconnects for 1.6T and Beyond

KEY 3

Optimizing Interconnects for 1.6T and Beyond

The Challenges

AI data center interconnects also face growing challenges with demand for higher bandwidth and lower latency. As transceiver speeds increase, there is pressure to reduce power consumption and investigate new technologies — such as linear-drive pluggable optics (LPO) and co-packaged optics (CPO). In addition, evolving IEEE and Optical Internetworking Forum (OIF) standards requiring error correction and complex equalizations further complicate design and testing.

For example, 56 GBd PAM4 signals enable 112 Gb/s lanes for 800G networks, doubling to 224 Gb/s lanes for 1.6T networks. As a result, organizations must navigate complex compliance testing across various link configurations.

> By 2029, 1.6T and 3.2T transceivers will make up the bulk of the data center optics market for AI — totaling almost $10 billion.
>
> LightCounting — LightTrends Newsletter

Figure 7 illustrates the 800G / 1.6T design cycle — highlighting critical stages such as channel / interconnect testing, conformance validation, network analysis, and manufacturing verification. Addressing these challenges requires robust simulation, validation, and testing solutions to ensure reliable interconnect performance in AI-driven environments.

The Recommendation

Deploy future-proof test and measurement tools designed for the AI data center, as illustrated in Figure 7. Validating individual high-speed links is the first step toward ensuring system-level data center performance. Network equipment manufacturers and hyperscalers rely on Keysight AI Interconnect solutions to validate performance, compliance, and efficiency.

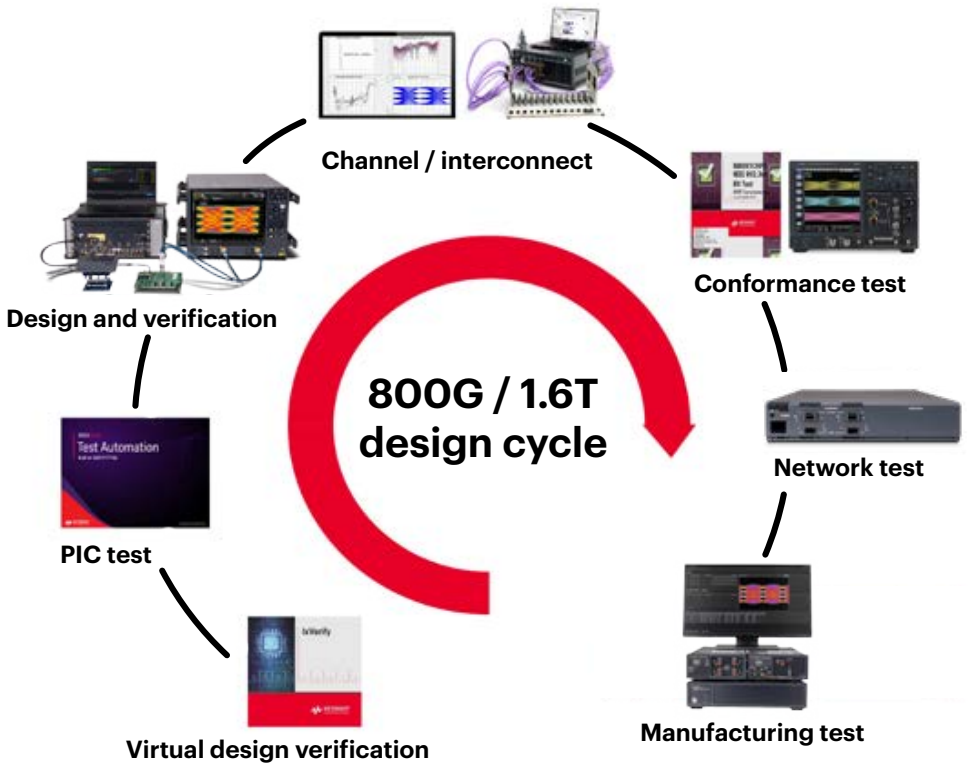


Figure 7. Example of a comprehensive workflow from design and verification to manufacturing and conformance testing for next-generation interconnects

Optical and electrical digital communication analyzer sampling oscilloscopes

Ensure interconnect compliance, validate signal integrity, and optimize optical transceiver performance. Analyze PAM4 signals, measuring TDECQ, jitter, and other specifications to improve transceiver interoperability in 800G / 1.6T systems, as highlighted in Figure 8.

Bit error ratio testers

Confirm physical layer receiver compliance with PCIe, USB, and Ethernet standards with up to 120 Gbaud highly-integrated bit error ratio (BER) test solutions for physical layer characterization, validation, and compliance testing.

BERT, FEC, and packet test solution

Validate high-speed Ethernet technologies like 800G / 1.6T is essential to ensure performance, interoperability, reliability, and scalability in modern network systems. This solution enables BERT, FEC analysis, and Layer 2 protocols across optical transceiver lanes for various configurations, ensuring reliable, high-speed performance in data centers.

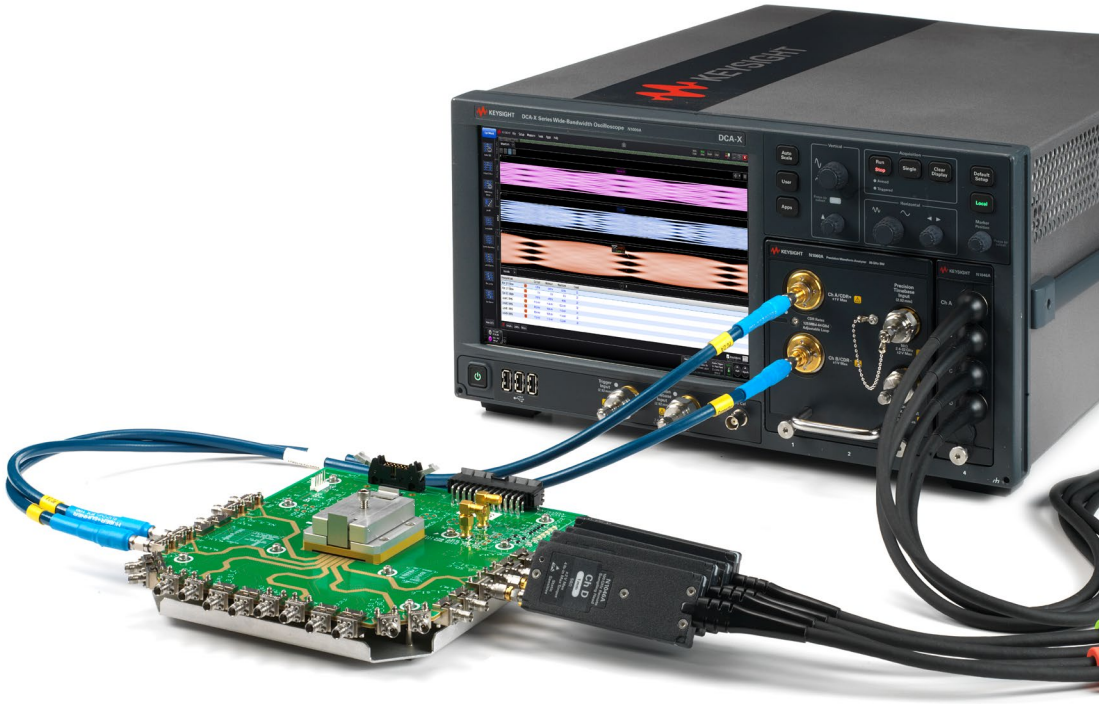


Figure 8. Keysight DCA-X wide-bandwidth sampling oscilloscopes provide accurate and precise measurements of high-speed digital designs from 50 Mb/s to 224 Gb/s.

Call to Action

> Learn more about Keysight AI Interconnect solutions.

KEY 4

Adapting to Higher Network Speeds and Bandwidths

KEY 4

Adapting to Higher Network Speeds and Bandwidths

The Challenges

Growing workloads for AI data centers drive demand for high-performance networks and the infrastructure and equipment that powers them. Achieving speeds of 1.6T and beyond begins at the component level.

Proper protocol support and rigorous testing are necessary to ensure reliability, but that is just the beginning. Optical transceivers require precise 224 Gbps transmission and receiver testing, as well as comprehensive testing at Layer 1.5 to ensure data integrity across striped 8-lane transmissions. Moreover, FEC data needs careful review to ensure optics meet the required specifications for reliable transmission in demanding AI networks.

> Tech giants and beyond are set to spend over $1T on AI CapEx in coming years.
>
> Goldman Sachs

At the system level, avoiding network bottlenecks and maximizing GPU utilization are essential to maintaining efficient AI workflows. For example, as shown in Figure 9, uneven workload completion across GPUs can create bottlenecks, emphasizing the need for precise testing and optimization. Addressing these challenges is vital to eliminating weak links and delivering efficient, reliable AI data center performance.

The Recommendation

Enhance performance, efficiency, and security across AI network environments with a holistic, end-to-end network test strategy. Network traffic generators, application and security testing tools, and distributed traffic emulators can all help pinpoint bottlenecks, inefficiencies, and performance degradations across AI data centers. That is why network architects and AI hypervisors trust Keysight AI Network solutions to maintain peak performance.

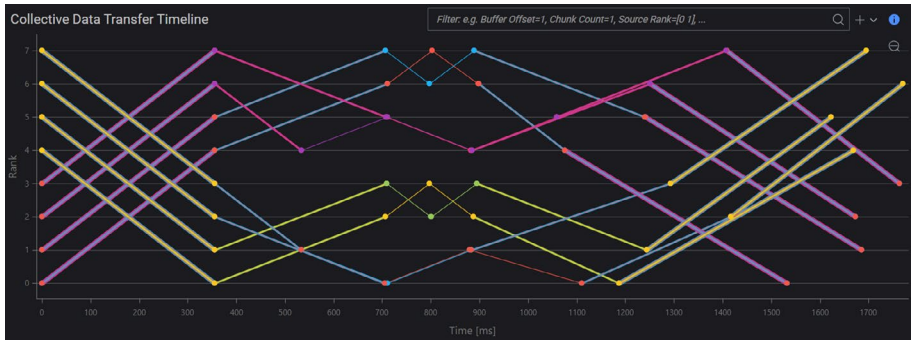


Figure 9. Example of test results of workload progress across eight GPUs using the all-reduce halving doubling algorithm, showing uneven completion times with some GPUs finishing later and adding delays

Validate AI network fabrics efficiently

Accelerate AI infrastructure design and data center deployment by emulating real-world AI workloads with measurable fidelity with the KAI Data Center Builder. With pre-packaged benchmarking applications developed with leading AI operators and infrastructure vendors, you can evaluate how new algorithms, components, and protocols improve AI training performance. The platform supports flexible test engines for comparative benchmarking, including hardware load appliances, software endpoints, and real AI accelerators.

Test the performance and reliability of high-speed network interconnects

Consolidate Layer 1-bit error ratio testing and Layer 2 / 3 switch and router performance validation in a single test platform. The Keysight Interconnect and Network Performance Tester Benchtop 800G is a portable, high-performance Ethernet network emulator. The platform supports 30W optics and offers key capabilities, measurements, and workflows at Layer 1.5. The solution is optimized for validating AI- and ML-ready optical and active cable interconnects, as demonstrated in Figure 10.

Test security and performance

Model dynamic application traffic, user behavior, and threat vectors at scale to ensure robust performance, security, and user experience in complex network environments. Recreate every aspect of a realistic workload across various physical and cloud-based environments. Keysight CyPerf is an instantly scalable, cloud-native network test solution that delivers unprecedented insights into the end-user experience, security posture, and performance bottlenecks of distributed, hybrid networks — including AI / ML cloud platforms.

Call to Action

Explore Keysight AI Network solutions.

Figure 10. Keysight Interconnect and Network Performance Tester 800G testing network protocols

KEY 5

Improving Power Efficiency to Scale AI Workloads

KEY 5

Improving Power Efficiency to Scale AI Workloads

The Challenges

Scaling AI workloads is increasingly limited by power efficiency and thermal constraints. Growing AI model complexity drives exponential energy demands from compute nodes, storage, and networking — often exceeding current power and cooling capacities.

Crosstalk, electromagnetic interference (EMI), and power integrity issues degrade performance, while high-frequency switching in AI accelerators introduces electrical noise. However, thermal constraints might be the most important performance constraint in AI data centers. Overworked cooling systems, which constantly struggle to dissipate heat from dense hardware deployments, not only increase costs but also hinder the scalability of AI workloads.

> AI power demands will surge as high as 652 TWh by 2030 — an 8,050% increase from projected 2024 levels.
>
> Forbes

The Recommendation

Optimize power efficiency and management across all components and infrastructure within AI data centers. The best way to do this is by investing in a comprehensive range of power products like Keysight AI Power solutions — which are designed for precise measurement, reliable performance, and efficient energy management across AI data center applications.

Power analyzer

Combine DC power supplies and loads with the capabilities of a multimeter, an oscilloscope, a waveform generator, and a data logger into a single instrument. Keysight power analyzers enable precise voltage and current measurements, waveform generation, and data logging, as shown in Figure 11.

Oscilloscope with power analysis software, power rail probes

Quickly resolve the root causes of power integrity issues in network equipment. Keysight oscilloscopes enable you to analyze noise, ripple, and crosstalk accurately. Measure power rails and high-speed data signals to reduce crosstalk between high-speed data traces and high-voltage power rails — which cause issues with power efficiency and signal integrity.

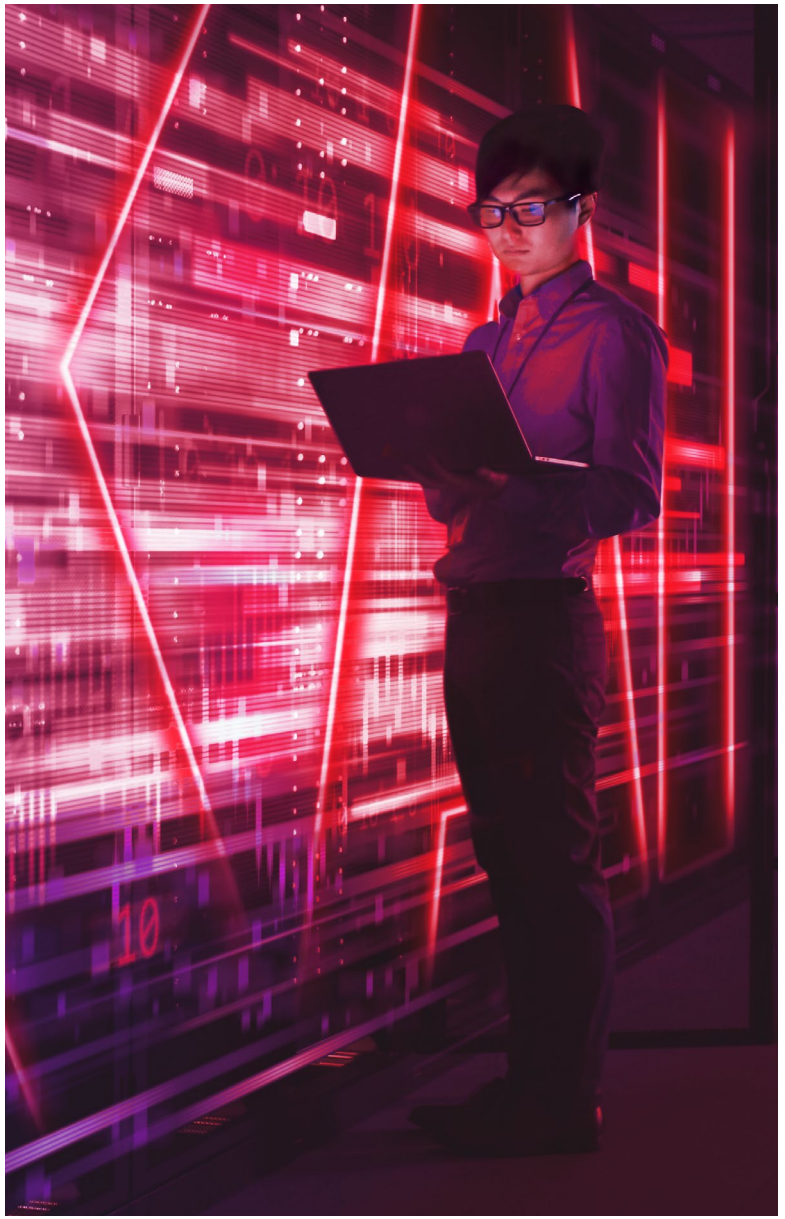
Figure 11. Keysight power analyzer displaying voltage and current waveforms in real time

Power integrity design and simulation software

Optimize power reliability and thermal performance early in the design using electronic design automation (EDA) software. Analyze power delivery networks, predict reliability, and optimize thermal performance, streamlining the power integrity workflow.

Call to Action

> Learn more about Keysight AI Power solutions.



Build AI-ready Data Centers with KAI Solutions

Building data centers capable of managing the demands of AI workloads means validating every component, connection, and configuration. With the stakes — and scale — this high, even the smallest efficiency gain, operational improvement, or performance enhancement can unlock significant returns, reducing disruptions and preventing cascading failures.

AI-ready semiconductors, networks, and data center infrastructure demand AI-ready test and emulation tools. With a full-stack portfolio of emulators and test hardware, KAI solutions make it easy to replicate real-world AI workloads, validate network components, and optimize system-level performance across every layer — from physical hardware to application-level behavior.

Call to Action

- Unlock the full potential of your AI infrastructure.
- Discover state-of-the-art KAI solutions.

> With Keysight, you can emulate anything and optimize everything — ensuring your AI data center meets today's benchmarks and tomorrow's breakthroughs.

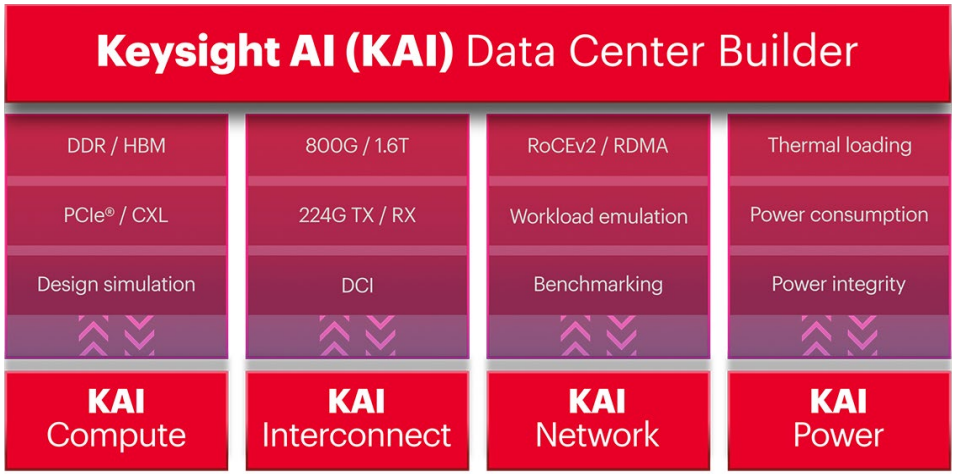


Figure 12. Keysight AI Data Center Solutions

Keysight enables innovators to push the boundaries of engineering by quickly solving design, emulation, and test challenges to create the best product experiences. Start your innovation journey at www.keysight.com.