Beyond the Bottleneck: AI Cluster Networking Report 2026

Trends, Investments & Optimization Strategies to Unlock the Potential of AI Data Centers

eBook

KEYSIGHT

AI at a Crossroads

Artificial intelligence (AI) is the engine of next-generation innovation. However, increasing complexity means increased demand on data center networks. As AI grows into a central component of enterprise strategies, organizations must carefully consider how they design, test, and scale their infrastructure.

This report, based on a global survey conducted by Heavy Reading in collaboration with Keysight Technologies, examines the critical challenges, technology choices, and investment priorities shaping AI cluster networking in 2025.

What emerges is a story of urgency, ingenuity, and redefinition. Together, these findings form a blueprint for navigating the next phase of AI infrastructure: **one built not just on speed, but on a scalable foundation.**

KEY FINDINGS

The Pressure is Mounting

89% of respondents plan to expand or sustain AI infrastructure investments, even as 59% cite budget constraints as a leading concern.

Emulation Becomes the Reality Check

95% of respondents view real-world emulation as critical, but a majority still lack the tools to replicate production-level AI workloads.

Leadership is Being Redefined

The future belongs to those who can combine flexibility, interoperability, and operational maturity to scale AI infrastructure responsibly and strategically.

Networks Are the Next Bottleneck – and the Next Breakthrough

55% of respondents are using 400G and 22% are trialing 1.6T, as emerging AI models demand faster interconnects.

Smarter, Not Just Faster

61% of respondents are focused on optimizing what they already have, leveraging SDN, Ultra Ethernet, and advanced validation workflows over new capital outlays.

Contents

CHAPTER 1

Key Findings from Global Leaders

CHAPTER 1

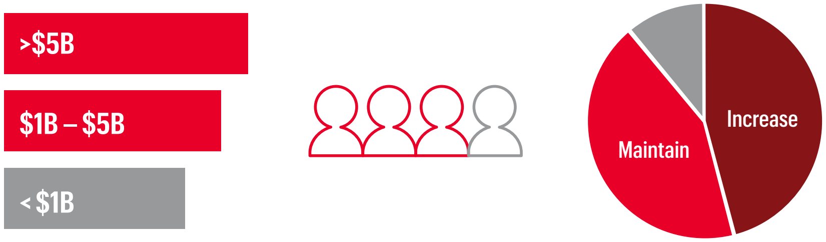
Key Findings from Global Leaders

Conducted from March 13 to April 11, 2025, the survey collected responses from 103 participants — mostly from large telecom and cloud providers. Respondents were primarily senior professionals in network planning, engineering, and IT. Notably, 72% of those surveyed work at companies with over $1 billion in annual revenue.

With nearly 90% of respondents reporting plans to expand or sustain current infrastructure investments, the industry is clearly accelerating toward higher capacity, lower latency, and greater operational efficiency.

However, while operators are gearing up for growth, they also face mounting pressure to extract more from existing network infrastructure.

Large Telecoms	Job Roles	Infrastructure Investment
72%	76%	89%
$1 Billion plus in revenue	Network Planning, Engineering, R&D, Operations, IT	Increase or maintain



Learn More:

Explore insights on scalling, optimizing, and validating AI infrastructure with Keysight.

Explore solutions for AI Data Centers

> We're entering a phase where performance gains must come not only from new infrastructure, but from smarter orchestration of what's already in place.

CHAPTER 2

Smarter AI Scaling Strategies

CHAPTER 2

Smarter AI Scaling Strategies

Overcoming Key Deployment Challenges

As AI adoption accelerates, data center operators are under pressure to scale faster, smarter, and more efficiently. However, several persistent barriers are slowing progress — each requiring strategic attention and innovation.

Together, these challenges highlight the need for more efficient scaling strategies that balance performance, cost, and talent development.

Greatest Challenges in Implementing AI Data Center Infrastructure

Budget constraints – **59% of organizations** are struggling to secure the capital needed to expand AI infrastructure, especially given the high costs of compute, cooling, and power.

Infrastructure limitations – **55% of respondents** highlighted limitations in existing infrastructure, likely due to the lack of flexibility and throughput to handle AI workloads at scale, which creates bottlenecks in performance and capacity.

Shortage of skilled personnel – The demand for AI-savvy engineers, network architects, and data center specialists far exceeds the current talent pool, creating a critical skills gap identified by **51% of respondents**.

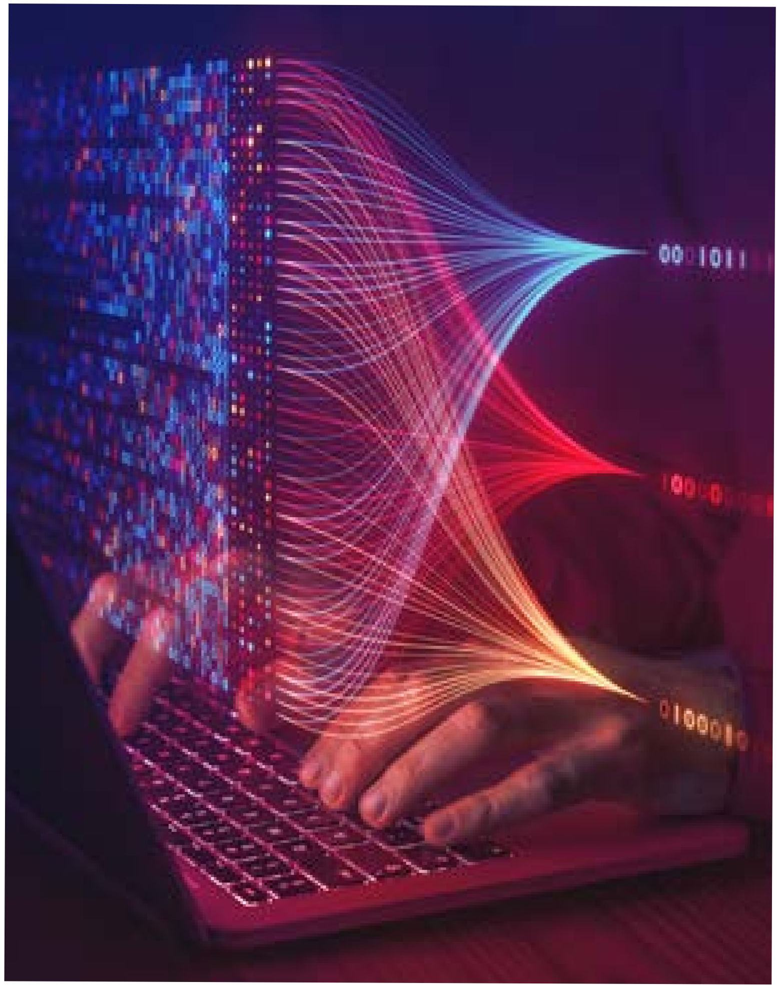
Smarter AI Scaling Strategies

As demand intensifies, operators are adopting flexible strategies to accommodate emerging AI workloads and traffic patterns. Agentic AI, DeepSeek, and Grok 3 are reshaping infrastructure decisions with their need for faster processing and adaptive networking.

Top strategies for expanding AI Data Centers:

- **51%** are planning cloud-based AI integration
- **49%** are prioritizing higher-performance GPUs
- **45%** are investing in new software architectures
- **45%** are upgrading networking hardware

These results reveal a balancing act: organizations are keen to boost computational horsepower while maintaining agility. Agentic AI's iterative reasoning capabilities are particularly driving changes in architectures and interconnect approaches.



Learn More:

Discover how Keysight enables real-world deployment for AI-optimized networks.

Keysight AI Infrastructure eBook

> We're not just scaling capacity; we're redesigning the fabric of AI computing.

CHAPTER 3

Breaking Network Bottlenecks

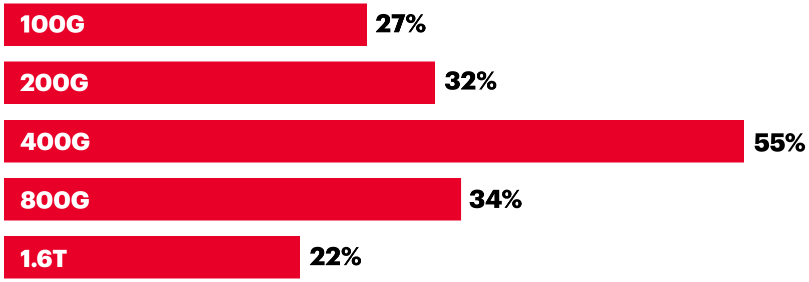
CHAPTER 3

Breaking Network Bottlenecks

As AI models grow in scale and complexity, underlying network fabrics are being pushed to their limits. Operators aren't just maxing out current speeds — they're evaluating next-generation solutions.

Survey results reveal a clear shift toward higher bandwidth. While 400G remains the dominant standard for AI and cloud data centers, adoption of 800G and even 1.6T is rapidly gaining momentum to meet large-scale training and inference demands.

Network speeds deployed in AI / ML infrastructure:



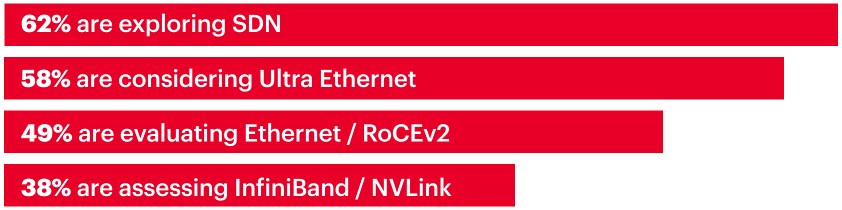
Learn More:

Learn how an arbitrary waveform generator is an enabler of 448 Gbps interfaces for 3.2T.

Watch the 3.2T Demonstration

The data also shows a trend toward Ethernet-based solutions, particularly newer architectures like Ultra Ethernet and software defined networking (SDN). However, high-performance options like InfiniBand / NVLink still have a foothold.

Networking technologies considered for AI / ML workloads:



62% are exploring SDN

58% are considering Ultra Ethernet

49% are evaluating Ethernet / RoCEv2

38% are assessing InfiniBand / NVLink

Ultra Ethernet: Built for AI and High-Performance Computing (HPC)

Ultra Ethernet is gaining traction as a purpose-built interconnect standard for AI and HPC. Unlike traditional Ethernet, Ultra Ethernet redefines the protocol stack to deliver the precision, speed, and scalability AI demands.

- **Full-stack design for AI and HPC**
  Ultra Ethernet is a vertically integrated architecture — optimized from the physical layer to the transport layer for parallelism, synchronization, and deterministic performance.
- **Low-latency, reliable performance for AI**
  With ultra-low latency and zero packet loss, Ultra Ethernet ensures real-time, predictable data delivery at scale through advanced scheduling and traffic shaping.
- **Industry-wide influence and reach**
  Already under evaluation by **58% of operators**, Ultra Ethernet challenges InfiniBand and RoCEv2 with open, standards-based performance and no vendor lock-in — making it a strong contender for next-generation AI data centers.

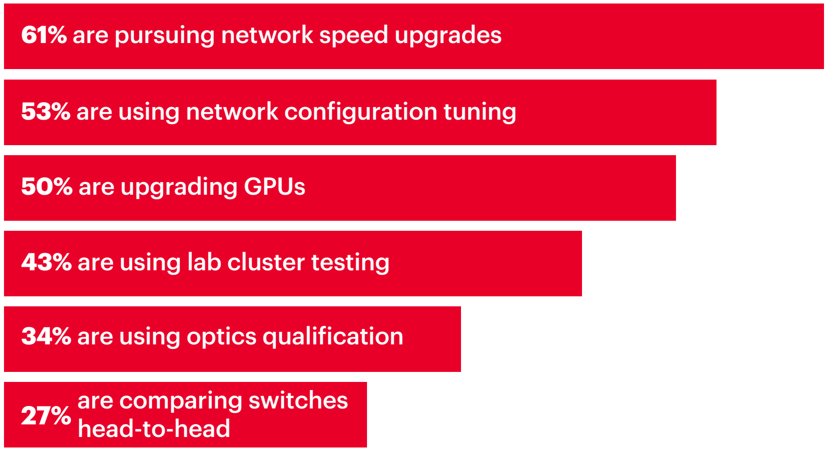
Learn More:

Discover why Ultra Ethernet is being considered by nearly 60% of respondents.

> Discover Scalable Strategies for AI Data Center Clusters

To optimize AI network performance, operators are blending hardware upgrades with software-level fine-tuning. This multi-pronged approach reflects the complexity of scaling AI infrastructure effectively — and the importance of maximizing current infrastructure vs. future investments.

Strategies to optimize AI network performance:



> As AI accelerators multiply, the network is no longer a backbone — it's a bottleneck or a breakthrough.

Test Smarter, Not Harder

Use AI-optimized validation tools to emulate large-scale clusters.

Benchmarking AI Networks White Paper

CHAPTER 4

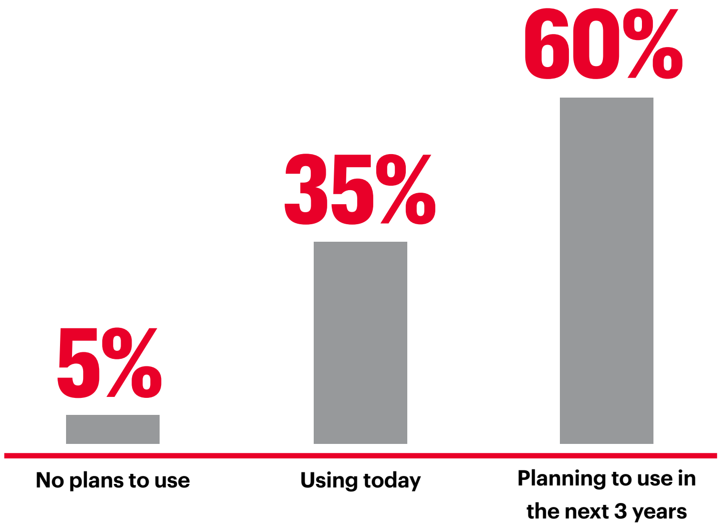
Emulation That Reflects Reality

CHAPTER 4

Emulation That Reflects Reality

As AI data centers increase in capability, complexity, and capacity, real-world traffic emulation is increasingly viewed as essential for validating performance under realistic conditions. It's no longer enough to rely on component-level validation — operators need to simulate system-level actual AI traffic patterns to ensure infrastructure can keep pace.

Organizations use of AI traffic emulation



Why Emulation Matters

1

Validates real-world behavior

Emulation bridges the gap between lab conditions and production realities. By replicating how AI workloads behave across nodes, protocols, and failure conditions, operators gain a clearer view of how their infrastructure performs under pressure.

2

Enables reliable test repeatability

Consistency is key in network validation. Emulation allows operators to recreate the same traffic patterns, error conditions, and synchronization challenges across test cycles — ensuring issues are identified, addressed, and verified with confidence.

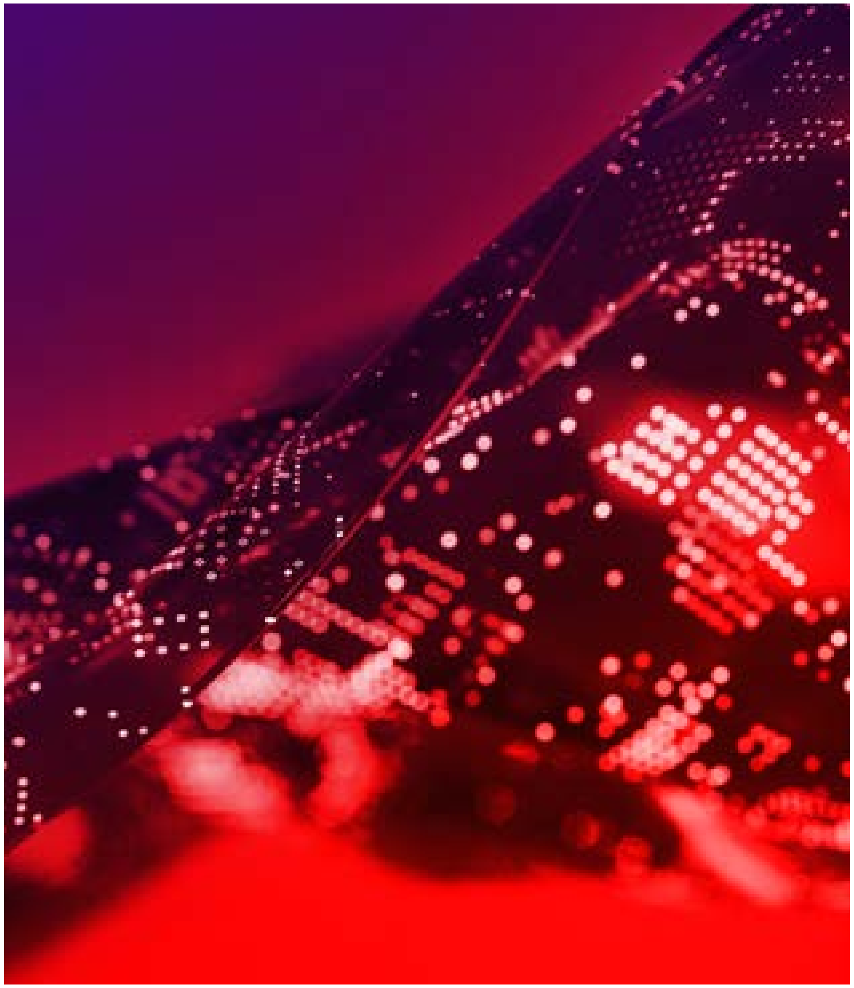
3

Improves pre-deployment confidence

Emulators help identify hidden bottlenecks, incompatibilities, or edge-case failures before scaling or upgrading an AI cluster — reducing the risk of issues in production, shortening rollout timelines, and protecting ROI.

Testing Challenges Hindering Optimization

Despite growing consensus on its importance, operators face major obstacles in scaling AI workload emulation. These results highlight an urgent need for smarter, more consistent solutions that mirror the unpredictable, high-throughput demands of modern AI environments.



Biggest challenges in testing / optimizing AI infrastructure:

- **56%** cite the lack of standardized benchmarking tools & equipments
- **56%** struggle to reproduce real-world AI workloads
- **50%** note a lack of knowledgeable AI network engineers
- **49%** report results skewed by vendor-specific optimizations

Learn More:

Find out how advanced emulators can help ensure scalable and reliable AI networks.

Scalability & Optimization White Paper

> Emulation bridges the gap between theoretical performance and production reality.

CHAPTER 5

Innovation vs. Cost Pressures

Innovation vs. Cost Pressures

AI infrastructure may be expanding rapidly, but financial constraints remain a dominant theme. Operators are clearly signaling a shift in mindset — not just building more, but building smarter. The survey results highlight a strong preference for optimization-driven strategies over capital-intensive upgrades. The top benefits of optimizing AI data center infrastructure reflect this trend:

62%

Work to get more out of infrastructure without additional investment

There is a growing appetite for tools that help fine-tune existing networks. Operators want to squeeze more efficiency and performance out of what they already have — without the delays or costs of physical upgrades.

59%

Improve reliability for workload completion

As AI applications evolve, maintaining consistent performance across varying loads and models is no longer optional. Operators are looking for testing solutions that ensure workloads run predictably and within SLA thresholds — regardless of scale or model behavior.

43%

See optimization as a competitive advantage

Operators increasingly see optimized infrastructure as a strategic differentiator. Improved performance, agility, and responsiveness in AI data centers directly translate to a market advantage, enabling faster innovation and increasing customer value.

41%

Look to reduce power

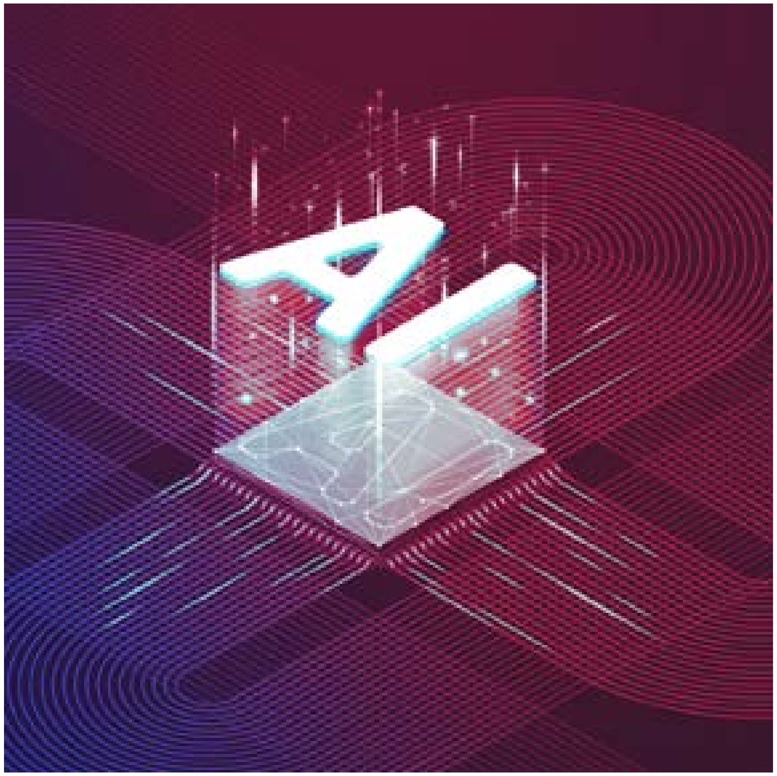
With rising energy costs, minimizing power consumption is a strategic imperative. Operators need tools that can simulate and monitor power draw under real AI loads, identify energy hotspots, and validate sustainable deployment approaches.

Strategic Alignment: Intelligent Infrastructure Testing

The top benefits of optimizing AI infrastructure — infrastructure optimization (62%), workload reliability (59%), competitive advantage (43%), and power efficiency (41%) — highlight a clear shift in strategy. Rather than just investing heavily in new hardware, AI operators appear to be seeking more innovative and sustainable ways to maximize what they already have, recognizing that better performance, agility, and responsiveness can drive faster innovation and greater customer value.

> Performance alone is no longer enough. Success now hinges on resilient, efficient, and sustainable AI networking — enabled by smarter, simulation-driven validation practices.

Together, these preferences signal a new era of intelligent infrastructure testing. Here, the goal isn't just to go faster, but to go further with what you already have. Organizations aren't just looking to scale up — they're focused on scaling smart.



Learn More:

Explore the fundamentals of validating AI networks with real-world demonstrations.

> Register for the bootcamp

CONCLUSION

From Limits to Leverage

The AI infrastructure evolution is well underway, with operators investing in faster, smarter, and more resilient networking strategies. As data center architectures shift to support advanced AI models, innovations like Ultra Ethernet and real-world emulation will be critical to success.

The industry's top priorities — efficiency, reliability, and flexibility — signal a future where performance gains must come from both new investments and more innovative use of existing infrastructure.

Keysight is proud to support this transformation with test solutions designed for the next generation of intelligent, high-performance AI networks. These insights form the basis for evolving standards, smarter design practices, and ultimately, a more scalable AI future.

Discover 5 Proven Strategies to Optimize and Scale AI Data Centers

AI is reshaping every industry — but with it comes complex traffic patterns, volatile workloads, and intense performance demands. This eBook reveals how to overcome these challenges and build AI-optimized data centers that are ready to scale.

Get the eBook

KEYSIGHT

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.