

The rise in AI workloads is impacting datacenter capacities, so new architectures are being created to solve the issues and prepare for the next wave of applications.

Datacenter Scale-Across Networking Architectures for the Next Generation of Applications

February 2026

Written by: Paul Nicholson, Research VP, Cloud and Datacenter Networks, and Peter Rutten, Research VP, Worldwide Infrastructure Research

Evolving challenges in modern datacenter architecture and connectivity

The rapid expansion of AI workloads is placing unprecedented demands on datacenter capacity and driving the evolution of new distributed networking architectures. AI applications are fueling increased datacenter and network buildouts across industries and regions, as organizations strive to support modern workloads and comply with distributed datacenter requirements, including sovereign mandates. These trends are prompting enterprises to refine their strategies, with a particular focus on network infrastructure as a critical enabler.

Despite optimizations for AI workloads, the surge in AI-driven traffic is straining existing datacenter resources, including power, cooling, space, and maintenance, thus intensifying competition for these assets. Power consumption and geographic dispersion are particularly acute challenges, driving multi-datacenter deployments and the need for more efficient datacenter interconnect (DCI) networking solutions (see Figure 1).

AI infrastructure strategies for networking have progressed from a scale-up approach that vertically scales GPUs to operate as integrated systems to a scale-out approach that scales across racks of GPU systems within datacenters horizontally for greater capacity and flexibility. Now, scale-across is emerging as the latest approach, connecting multiple geographically dispersed datacenters into a unified system by using advanced DCI technologies to support distributed, synchronized workloads at scale.

AT A GLANCE

KEY TAKEAWAYS

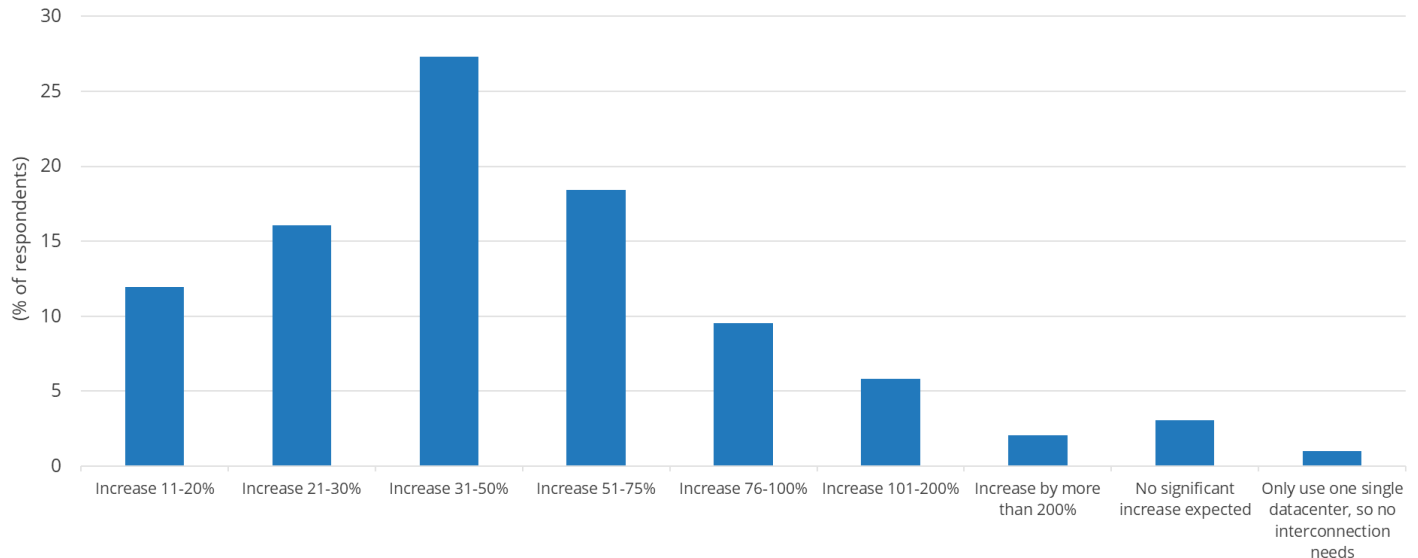
AI and modern workloads are driving datacenter interconnection requirements. IDC's December 2025 *Worldwide AI in Networking Special Report* shows that organizations reported rising bandwidth demands, with inter-datacenter demands being the highest. In detail, in the next year:

- » Nearly all organizations expect bandwidth requirements (89% within and 91% between datacenters) to grow by 11% or more.
- » Regarding the inter-datacenter bandwidth surge, 36% of enterprises anticipate increasing bandwidth between datacenters by over 51%, which is the highest rate of growth observed.
- » Regarding intra-datacenter bandwidth growth, 29% of organizations are planning to increase bandwidth within datacenters by more than 51%, underscoring the need for robust internal network architectures.

Figure 1 details the increase in bandwidth between datacenters. Within the next year, 91% of worldwide datacenters will add 11% or more interconnection bandwidth, with more than half (61%) increasing their bandwidth by 31% or better. Over a third (36%) will increase bandwidth by 51% or more.

FIGURE 1: **Datacenter bandwidth increase between datacenters**

Q Over the next year, how do you expect your organization's bandwidth needs to increase for the interconnection of multiple datacenters because of AI workload deployments?



n = 293

Base = Respondents who indicated that their organization only uses at least one on-premises datacenter. They do not use cloud, hyperscale cloud, or on-premises platforms.

Source: IDC's Worldwide AI in Networking Special Report, December 2025

A key area of industry focus is the development of distributed networks and datacenter architectures that span multiple compute environments, including public cloud providers, on-premises datacenters, and emerging neocloud providers. DCI is becoming central to supporting the scale and complexity of modern AI workloads, with requirements intensifying for high-bandwidth, low-latency, and resilient connectivity. Hyperscalers, neoclouds, enterprises, and communication service providers each play distinct roles in this evolving landscape, with unique DCI needs: Hyperscalers drive large-scale model training and inference; neoclouds can address flexible, geographically dispersed enterprise deployments; enterprises are deploying AI capabilities across new and existing datacenters while also prioritizing data sovereignty and operational efficiency; and communication service providers enable distributed AI and multicloud interconnection. Enterprises deploying AI at scale can also require robust datacenter interconnects to address data gravity, ensuring that large, distributed data sets (often existing) remain accessible for AI processing while maintaining compliance, performance, and cost efficiency across geographically dispersed environments.

Recent advances in silicon, AI workload monitoring, deep-buffer routing platforms, and ZR/ZR+ coherent pluggable optics are contributing to the development of high-performance, scalable, and secure DCI architectures. Improvements in silicon enable higher bandwidth and lower latency, supporting the data transfer needs of both AI and traditional

applications. Enhanced AI workload monitoring provides more detailed telemetry and analytics, which can inform networking resource allocation and network management operations. Deep-buffer routing platforms help manage network congestion and maintain throughput during periods of high traffic. Collectively, these technologies can support the evolving requirements of DCI environments for AI and a range of application types.

Business urgency is increasing as organizations seek to modernize their DCI infrastructure to unlock the full potential of AI, mitigate operational risks, and ensure compliance in distributed environments. Key pain points include resource constraints, network reliability, security, and the ability to scale infrastructure efficiently across diverse environments and connection points, as well as for cluster networks/AI factory locations. Addressing the technologies and strategies needed to build resilient, scalable, and future-ready DCI solutions for the next wave of AI, cloud, and datacenter applications is an important foundation.

Definitions

- » **Datacenter interconnect** is a set of technologies and architectures that enable high-speed, low-latency, and secure connectivity between multiple datacenters. DCI solutions support data replication, workload mobility, disaster recovery, and distributed AI applications by facilitating seamless communication across sites.
- » **Sovereign datacenter** is a datacenter designed to ensure that all data, workloads, and operations remain within a specific nation's jurisdiction, meeting strict data residency and compliance requirements to support digital sovereignty.
- » **Scale-across networking** refers to the architectural approach of connecting multiple, geographically dispersed datacenters or clusters to operate as a unified system for AI workloads. Scale-across architecture enables distributed workloads, synchronous operations, and resource sharing across locations, often leveraging advanced DCI technologies.
- » **Scale-out networking** is the process of increasing system capacity by interconnecting multiple racks of AI workload resources horizontally, frequently within AI factories. Scale-out architectures are designed for high-performance networking, allowing organizations to handle growing workloads by expanding the number of resources rather than upgrading individual components.
- » **Scale-up networking** involves enhancing the capacity of an AI system by using a high-speed networking backbone in a rack to vertically connect multiple GPUs and associated resources to ensure high-speed, nonblocking communications. Scale-up architecture is typically used for demanding, resource-intensive workloads that benefit from larger, high-performance processing capabilities but may face limitations in terms of scalability and fault tolerance as compared with scale-out architecture.

Strategic value of modern interconnected application infrastructure in the AI era

As AI workloads proliferate and datacenter architectures evolve, organizations are adopting modern approaches to application infrastructure. The need for flexibility, optimization, and repeatability across DCI environments, in addition to compliance requirements, is driving this shift as sovereign datacenter requirements become common.

The following highlights the value of embracing advanced infrastructure strategies as the architecture matures:

- » **Adoption of advanced interconnect architecture:** Modern silicon, optics, and deep-buffer technologies enable scalable, high-bandwidth connections between geographically dispersed datacenters. These advancements support both AI and traditional workloads, allowing organizations to efficiently manage data movement and resource allocation across multiple sites.
- » **Enhanced performance for AI and data-intensive applications:** Low-latency, high-throughput interconnects are essential for synchronous operations across datacenters, accelerating AI training, inference, and other data-intensive processes. As AI applications become more prevalent (and increase in size/parameters), the demand for robust interconnect solutions will grow, ensuring that organizations can meet performance requirements at scale.
- » **Improved resiliency and reliability:** Modern DCI platforms incorporate deep buffering and proactive congestion control, which help absorb network failures and maintain service continuity. This fault tolerance is particularly critical for AI workloads, where downtime can disrupt operations and impact business outcomes. Enhanced reliability also supports repeatability and consistency in application performance and for AI GPU utilization.
- » **Strengthened security and compliance:** Integrating advanced encryption, hardware root of trust, and post-quantum cryptography (PQC) into interconnect solutions ensures secure data transfer and integrity. These features support compliance with data sovereignty and privacy regulations, which are increasingly important as organizations operate across more jurisdictions and cloud environments.
- » **Operational efficiency and cost savings:** Energy-efficient interconnect platforms, optimized hardware, and intelligent software reduce power consumption and lower operational costs. This efficiency enables organizations to scale infrastructure sustainably, reallocating saved power to support additional GPU systems or other high-demand resources. Cost-effective datacenter operations also provide greater flexibility and access to capacity as needs evolve.
- » **Strategic opportunities for infrastructure planning:** The adoption of new standards, protocols, and repeatable architecture creates opportunities for organizations to reevaluate and optimize their networking infrastructure. By leveraging flexible, scalable solutions, enterprises can better align IT strategy with business objectives, adapt to emerging AI requirements, and future proof their operations against ongoing technological changes.

Modern DCI solutions are evolving beyond traditional scale-out models to support geographically dispersed AI clusters, address power and cooling constraints, and enable low-latency, synchronous operations. As AI training and inference use cases expand, the benefits of a modernized application infrastructure become increasingly clear, thus positioning organizations to capitalize on the next wave of AI, cloud, and datacenter innovation with increased infrastructure agility.

Considering Cisco

Cisco has introduced new DCI solutions designed to address the evolving challenges of modern application infrastructure. Central to this approach is the shift toward job-centric AI networking, where the network is treated as a critical component of the AI compute engine, optimized for the demands of distributed AI workloads and traditional applications.

As datacenter requirements and interconnection drive up WAN traffic and the demand for scale-across solutions to facilitate more favorable power and real estate conditions, Cisco's offerings are positioned to support large enterprises, including those in finance, healthcare, and other data-intensive sectors.

Cisco has engineered its technology portfolio for a broad range of use cases. While hyperscalers require robust support for large-scale AI model training and inference, large enterprises can also benefit from the same high-performance infrastructure for both AI and traditional applications. Cisco's solutions seek to enable organizations to address emerging needs as they expand datacenter footprints and interconnect geographically dispersed sites.

Product highlights

- » **Cisco Silicon One P200:** This solution offers high bandwidth and deep buffering for scalable low-latency DCI, optimized for both AI and traditional workloads. The P200 powers a 51.2Tbps scale-across system, supporting the throughput and reliability that modern AI/GPU clusters require. Its architecture enables deployment in a fixed, modular, and disaggregated chassis, providing flexibility for diverse network designs.
- » **Cisco 8000 and N9000 Fixed and Modular:** The P200-powered, deep-buffer systems offer 800G connectivity for scale-across universal spine deployments, with significant improvements in power efficiency, space utilization, and operational simplicity as compared with previous generations. These systems are designed to absorb massive traffic surges from AI training and inference, enabling consistent performance and reliability.
- » **Cisco 800G ZR/ZR+ coherent pluggable optics:** This product offers feature-rich and high-performance 800G coherent pluggable optics that can transmit wavelengths using Dense Wavelength-Division Multiplexing (DWDM) over 1,000km while consuming under 30W. It expands Cisco's 400G ZR/ZR+ portfolio in DCI, metro, and regional applications.

Key features and capabilities

- » **Flexible operating system support:** Cisco offers the choice of NX-OS, ACI, IOS-XR, or SONiC to meet diverse operational and application requirements.
- » **Scalable, programmable networking:** Silicon programmability and SDK support allow for future function expansion and complex logic execution without packet recycling, supporting evolving datacenter needs.
- » **Compact form factor:** The compact designs enable space efficiency, less power usage, and flexible deployment.
- » **Power efficiency:** The systems' high capacity with optimized power consumption aids in scaling distributed AI and traditional workloads.
- » **Deep-buffer routing:** This capability is a cornerstone of job-centric networking, as it maintains reliable workload distribution and supports the high-throughput demands of AI/GPU systems by absorbing dynamic traffic shifts to ensure job completion and efficiency.
- » **Integrated security:** Hardware-based security features — including ACLs, line-rate encryption (MACsec and IPsec), PQC, tamper-resistant root of trust, and authenticated data plane software — for end-to-end data and network protection throughout the product life cycle further enhance system integrity.

Cisco's DCI solutions provide a foundation for companies seeking to modernize their infrastructure for the AI era.

The breadth of Cisco's portfolio (e.g., fixed, modular, and disaggregated architectures), combined with platform consistency and adaptability, enables the company to provide solutions to meet customers' requirements depending on their readiness and deployment journey. Cisco's ongoing investment in silicon, optics, and software offers a future road map to customers for Cisco's unified platform.

Challenges

With the rapid changes in modern applications, Cisco needs to continually evaluate the new standards, the new architectures, and the pressures on the availability of datacenter power and overall capacity. This will ensure that the company provides solutions that continue to address customers' needs and (ideally) allow customers to pivot and adjust strategies, if/as needed, because of the pace of innovation. However, all vendors face this situation in emerging technological environments.

If Cisco continues to introduce new solutions to assist customers, it will be well positioned to address these challenges for customers in a rapidly evolving landscape.

Conclusion

The ongoing evolution of datacenter networking architectures is being shaped by the rapid growth of AI workloads, the increasing demands for distributed high-performance connectivity, and the need to address regulatory and operational requirements. Scale-across architecture and advanced DCI solutions are emerging as critical enablers for supporting both AI-driven and traditional applications across geographically dispersed environments. Technological advances in silicon, workload monitoring, and deep-buffer routing are contributing to more scalable, resilient, and efficient DCI platforms.

As organizations modernize their infrastructure, they must balance performance, reliability, security, and operational efficiency while remaining adaptable to new standards and deployment models. Vendors are responding with a range of solutions designed to address these evolving requirements, though the pace of innovation and resource constraints will continue to present challenges. Ultimately, the ability to deploy flexible, future-ready DCI architecture will be central to supporting the next generation of applications.

About the analysts



Paul Nicholson, Research Vice President, Cloud and Datacenter Networks

Paul Nicholson is IDC's Research VP of Cloud and Datacenter Networks. He provides thought leadership and actionable insights on cloud and datacenter networking markets. His deep understanding of these markets — and their product road maps, competitive differentiation, and go-to-market strategies — enables him to provide informed guidance to vendors, cloud providers, enterprise IT buyers, and practitioners.



Peter Rutten, Research Vice President, Worldwide Infrastructure Research

Peter Rutten is Research VP within IDC's worldwide infrastructure research organization and global research lead for the performance-intensive computing (PIC) practice. IDC's PIC coverage includes research on high-performance computing (HPC), AI and GenAI, big data and analytics (BDA) and quantum computing (QC) infrastructure stacks, deployments, solutions, workloads, and use cases.

MESSAGE FROM THE SPONSOR

AI networking has evolved from delivering "speeds and feeds" to becoming an intelligent, programmable fabric. While scale-across architectures remain essential for connecting distributed AI clusters, performance limitations are increasingly emerging *within* the AI fabric itself. A new approach is required to support the next generation of massive AI workloads and power-hungry clusters at greater scale.

Addressing this challenge calls for a shift to job-centric AI networking — an architecture purpose-built for the unique demands of AI workloads. In this model, the network becomes a critical component of the AI compute engine, directly influencing the efficiency, reliability, and delivery of training and inference workloads.

Cisco's foundational approach enables the network and AI cluster to operate as a single, unified system. This integration is designed to optimize performance, increase intelligence and programmability, and deliver more efficient and future-proof AI infrastructure at scale.

Learn more here: www.cisco.com/site/us/en/solutions/artificial-intelligence/ai-networking-in-data-center/index.html



The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Custom Solutions produced this publication. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis that IDC independently conducted and published, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. This IDC material is licensed for external use, and in no way does the use or publication of IDC research indicate IDC's endorsement of the sponsor's or licensee's products or strategies.

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries. IDC's analysis and insight help IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives.

©2026 IDC. Reproduction is forbidden unless authorized. All rights reserved. [CCPA](#)

IDC Research, Inc.

One Beacon Street
Suite 33100
Boston, MA 02108, USA
T 508.872.8200
F 508.935.4015

blogs.idc.com
www.idc.com