IDC

SPOTLIGHT

Sponsored by: Lumen

As AI becomes distributed across multiple geographic locations, the critical path for AI performance has shifted from the local processor to the cross-site network fabric. Organizations must adopt highly deterministic purpose-built interconnects capable of ensuring synchronized, low-latency execution across multi-metro datacenters.

Networking as a Competitive Advantage for Neocloud Providers

May 2026

Written by: Dave McCarthy, Program Vice President, Cloud & Infrastructure Services

At a glance

Key stats

- IDC estimates worldwide AI infrastructure spending will grow to roughly $740 billion in 2029, representing a 2024–2029 CAGR of 37.9%.
- By 2029, at least 40% of advanced GPU needs will have been met by specialized AI cloud providers offering true cloud features, flexible pricing, APIs, and software services.

What's important

Competitive advantage will no longer be determined solely by who has the most silicon, but by who possesses the most intelligently engineered multisite network fabric to orchestrate that silicon around the globe.

Introduction

The unprecedented surge in AI adoption has catalyzed a fundamental paradigm shift throughout the global technology ecosystem, decisively redefining foundational infrastructure requirements. For the past decade, cloud and enterprise architectures have been optimized for general-purpose compute within centralized datacenters. However, the commercialization of generative AI and massive-scale foundation models has exposed the limitations of these localized environments. AI workloads require massive synchronized parallel processing, and the sheer scale of modern clusters means they can no longer be contained within a single facility. This fundamentally alters how datacenters must be connected across geographic regions.

The financial implications of this architectural shift are staggering. Organizations are rapidly reallocating capital to secure the necessary hardware to participate in the AI economy. According to recent market intelligence, IDC estimates that worldwide AI infrastructure spending will grow to roughly $740 billion in 2029, representing a 2024–2029 CAGR of 37.9%. While this explosive investment initially targeted raw compute — specifically, high-performance GPUs — the physical realities of power constraints, cooling requirements, and sheer real estate limitations mean monolithic gigawatt-scale datacenters are increasingly difficult to permit and construct. Consequently, AI infrastructure is becoming heavily distributed across multiple sites. This spatial distribution places extraordinary new demands on wide area networks (WANs), shifting the critical path of AI performance from the local processor to the long-haul fabric that connects disparate datacenters.

To capitalize on the hundreds of billions of dollars being invested, network architectures must scale capacity rapidly and predictably to keep pace with workloads that span cities and regions. In the context of large language models, a single training run or massive inference operation can span tens of thousands of GPUs fractured across multiple server nodes in different physical facilities. These sites must exchange massive volumes of parameter data and gradients almost instantaneously; consequently, underperforming network links can delay global training synchronization and stall the

entire pipeline. Therefore, predictable deterministic connectivity between datacenters helps improve efficiency, reduce idle time, and accelerate returns on infrastructure investments.

To address these logistical realities without sacrificing workload integrity, reliable interconnects are required to link sites and regions with consistent performance as platforms expand to meet power, space, and availability demands. When an organization operates a distributed cluster of GPUs costing tens of millions of dollars, even microsecond-level delays across a multi-metro optical link translate directly to massive financial waste. The wide-area network is no longer just a pipeline between datacenters; it is the central nervous system of geographically distributed AI.

The evolution of "neoclouds"

As AI infrastructure demands outgrew traditional enterprise IT and single-site capacities, legacy hyperscalers often proved too rigid or supply-constrained to meet the agile, distributed needs of AI-native organizations. This market gap led to the creation of neoclouds, which IDC considers a new type of specialized cloud service provider engineered from the ground up explicitly for accelerated computing workloads. Many started strictly as single-site GPU-as-a-service providers of bare metal hardware. However, as cluster sizes grew beyond the power envelope of individual facilities, these providers had to evolve.

Recognizing early architectural bottlenecks, neoclouds first integrated high-performance storage to keep expensive GPUs busy. Now, the market is moving to its next critical phase: addressing the cross-site networking needs of highly distributed AI. Having optimized local compute and storage, the long-haul network fabric is now the frontier of competitive differentiation.

This multisite evolution is essential for low-latency inference, which is rapidly outpacing training in overall market demand. The rise of autonomous AI agents further accelerates this urgency. Because agentic workflows trigger multiple sequential or parallel backend inference calls (often routing requests to different specialized models hosted in different datacenters), the latency of the interconnect between these sites becomes critical. Even minor delays across the wide area network cumulatively cripple performance and ruin the end-user experience. Shifting inference economics increasingly rewards proximity, meaning a centralized single-region GPU farm simply cannot compete on either latency or unit economics in this highly distributed environment.

Robust cross-site networking within the neocloud ecosystem is also becoming a prerequisite for multicloud connectivity. Modern enterprises refuse walled gardens; they often store governed datasets in established clouds but rely on globally dispersed neocloud facilities for cost-efficient model training and inference. Bridging these distinct physical environments requires secure, high-bandwidth, and highly deterministic multi-metro routing.

The transformation of neoclouds into sophisticated, fully integrated distributed AI platforms is accelerating rapidly. Underscoring this shift, IDC predicts that, by 2029, at least 40% of advanced GPU needs will be met by specialized AI cloud providers offering true cloud features, flexible pricing, APIs, and software services — unlike GPU-only providers. The ultimate winners will be those who successfully master the complexities of wide area and cross-site networking to deliver a seamless and globally unified experience.

Designing for distributed AI

To realize the commercial and operational potential of next-generation AI, infrastructure design must undergo a philosophical shift, with cross-site networking as the key to enabling distributed AI. Because foundational models possess

hundreds of billions of parameters, the computational load and associated power requirements vastly exceed any single datacenter's capacity. Workloads must be distributed across vast geographically dispersed fleets of processors and continuously synchronized. In this architecture, the datacenter interconnect (DCI) acts as the backplane for a distributed supercomputing architecture, fusing isolated facilities into a cohesive entity.

As AI platforms scale across regions, the long-haul network transitions from a technical detail to a core business driver that shapes performance consistency and customer trust. For infrastructure providers and enterprise AI divisions, the ability to guarantee service-level agreements (SLAs) for training times and inference latency depends entirely on the reliability of the interconnects linking their sites. Without predictable job completion times across geographically fractured clusters, securing enterprise contracts becomes impossible; thus, wide area network integrity directly dictates commercial viability.

The root of many scaling failures is applying traditional enterprise WAN principles to AI workloads. Legacy WANs, designed for statistical multiplexing and best-effort delivery of standard internet traffic, routinely suffer from latency variance, jitter, and unpredictable routing. When scaling AI clusters across sites, these legacy constraints show up as idle GPUs, inconsistent runtimes, and severe SLA risks.

During the synchronization phase of multisite model training, geographically separated GPUs simultaneously attempt to transmit massive blocks of gradient data across the long-haul link at a scale legacy IP backbones were never engineered to handle. If the DCI cannot support these massive micro-bursts of traffic seamlessly, it leads to dropped packets and severe tail-latency. Because synchronous training dictates that the entire global cluster must wait for the slowest data transfer, this cross-site "straggler effect" halts the entire multi-million-dollar deployment.

The definitive solution is implementing wide area networks purpose-built for scale. This requires abandoning best-effort internet routing and shared oversubscribed long-haul topologies. Deterministic connectivity provides a scalable foundation for cross-site traffic and multi-metro expansion. Because server-to-server traffic now regularly spans cities, the multisite fabric must provide predictable, uniform latency between any two nodes, regardless of their physical placement in different facilities.

To maintain the illusion of a unified compute cluster, despite the massive power requirements forcing them to fracture across multiple physical locations, highly reliable DCIs with massive optical bandwidth, such as dense wavelength-division multiplexing (DWDM), are required. The most complex and critical challenge facing AI network architects today is engineering this multi-metro deterministic connectivity to ensure the unavoidable latency introduced by the speed of light across physical distance is perfectly stabilized and doesn't break synchronization.

Considering Lumen

Lumen offers flexible connectivity options designed to support GPU platform growth, helping customers scale capacity, manage performance, and expand with greater confidence.

High-capacity connectivity where GPUs run

Bring bandwidth to where compute lives, without redesigning the network as demand grows:

- Optical capacity (100G/400G today, with a future road map for 800G) to support high-bandwidth connectivity between GPU clusters, datacenters, and metros

- Private inter-datacenter connectivity for low-latency deterministic traffic across distributed training and inference environments
- Elastic capacity scaling to better align bandwidth growth with GPU utilization and customer demand

Performance-isolated network services to reduce variability

Performance-isolated network services reduce variability by isolating AI traffic and engineering the network for consistent performance under load:

- Dedicated SLA-backed connectivity designed to support GPU-intensive workloads
- High-throughput transport (up to 400 Gbps) optimized for large-scale AI data movement
- Multipath diversity to support resilient multi-datacenter architectures
- Built-in resiliency and security controls to help reduce variability, protect data in motion, and support consistent SLA delivery

Architectural flexibility as platforms mature

Choose the level of control that fits today, with a clear path as the platform evolves:

- Lumen Wavelengths RapidRoutes for pre-engineered, validated capacity on select point-to-point routes, with a 20-day SLA on qualifying connections to help accelerate site and region turn-up
- Managed Wavelength services for rapid scale and simplified operations, providing high-capacity deterministic transport between sites without the operational overhead of owning the network
- Managed Optical Fiber Networks (MOFN) for increased control and performance isolation, extending private optical connectivity across datacenters and cloud regions while reducing operational burden
- Dark Fiber to support organizations with the scale and utilization density to justify full ownership and operational control, supporting long-term platform economics for organizations that want to engineer and operate their own optical layer
- Network as a service (NaaS) to support dynamic capacity needs and agile connectivity, helping enable on-demand bandwidth changes and cloud-like consumption as traffic patterns shift

Challenges

Although the demand for cross-site AI networking is immense, Lumen faces go-to-market hurdles regarding customer adoption. Organizations accustomed to the frictionless integrated provisioning of public cloud environments may hesitate to introduce a third-party telecommunications provider into their highly orchestrated AI stacks due to perceived architectural complexity.

To overcome these adoption barriers and capture the AI connectivity market, Lumen is aggressively pivoting its commercial model to mirror cloud-native consumption. By launching purpose-built solutions such as RapidRoutes and transitioning to a NaaS model, the company aims to strip away traditional telecom friction. This allows AI infrastructure

providers and enterprises to dynamically allocate massive optical bandwidth via APIs, matching the speed and ease of cloud provisioning.

Conclusion

The commercialization and global scaling of AI represent the most significant forcing function in enterprise infrastructure since the advent of public cloud computing. As evidenced by projected infrastructure spending, the market is aggressively capitalizing on the promise of generative and agentic AI. However, this massive capital expenditure will yield suboptimal returns if it remains narrowly focused on raw, localized computational power. The true bottleneck in the modern AI ecosystem has decisively shifted from the local processor to the cross-site network fabric.

The true bottleneck in the modern AI ecosystem has decisively shifted from the local processor to the cross-site network fabric.

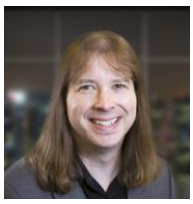
As power and space constraints force training clusters to fracture across regions, and as global inference demands escalate to support autonomous agentic workflows, traditional wide area network architectures simply cannot cope. The latency, jitter, and packet loss inherent in legacy long-haul designs result in catastrophic financial inefficiencies by leaving incredibly expensive GPUs idle while waiting for data to traverse between cities. To unlock the full potential of distributed AI, organizations and service providers must pivot to purpose-built, highly deterministic, and massive-bandwidth optical DCI architectures capable of seamless scaling across multi-metro geographical regions.

The explosive evolution of neoclouds illustrates the market's demand for holistic solutions. Moving beyond isolated single-site GPU offerings, these specialized providers are aggressively integrating advanced cross-datacenter networking capabilities to deliver true full-stack AI performance on a global scale. As the industry matures, competitive advantage will no longer be determined solely by who has the most silicon, but by who possesses the most intelligently engineered optical multisite network fabric to orchestrate that silicon around the globe.

As neocloud platforms scale, selecting the right networking partner becomes a critical infrastructure decision. To future-proof growth, consider the following:

- Pressure-test interconnect architecture against tomorrow's demands. Evaluate whether the current setup can handle increasing capacity requirements and identify routes where deterministic capacity is already constraining growth.
- Assess the partner's ability to scale. Prioritize providers that offer proven optical capacity scaling, broad route diversity and global footprint, and fast turn-ups that keep pace with your platform expansion without sacrificing reliability.
- Scrutinize commercial models and SLAs. Ensure the networking partner offers flexible, growth-aligned commercial structures alongside robust service level agreements that provide performance guarantees and delivery certainty.

About the Analyst



Dave McCarthy, *Program Vice President, Cloud & Infrastructure Services*

Dave McCarthy leads IDC's cloud and infrastructure services global research subdomain, focusing on cloud infrastructure and its related adoption strategies: public, private, hybrid, multicloud, distributed, sovereign, and edge. Benefiting both technology suppliers and IT decision-makers, Dave's insights explore how cloud and infrastructure services provide the foundation for both general-purpose and AI workloads, enabling organizations to innovate faster, create new revenue streams, and achieve competitive advantages.

MESSAGE FROM THE SPONSOR

Lumen delivers the network foundation that specialized cloud providers need to support the surge in GPU capacity demand, combining high-capacity optical transport, deterministic performance, and broad metro and long-haul reach. With one of the largest fiber networks in North America, Lumen enables fast, predictable connectivity between GPU clusters, datacenters, and cloud environments, helping providers reduce time-to-capacity and maintain consistent SLAs as platforms expand. Flexible deployment models allow neoclouds to align network investment to each stage of growth, turning network infrastructure into a lever for performance, efficiency, and sustained revenue expansion.

Learn more about Lumen solutions for neoclouds: https://www.lumen.com/en-us/solutions/neocloud.html

IDC Custom Solutions

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

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