

**451 Research Discovery Report**

May 2025

# Architecting AI at scale

Networking lessons from industry leaders

**S&P Global**  
Market Intelligence

in partnership with  
**verizon**  
business

# Table of contents

<b>Introduction</b>	<b>3</b>
Methodology	3
Figure 1: Demographic breakdown of participants	3
<b>Executive summary</b>	<b>4</b>
<b>Insight 1: Networking planning should anticipate needs, not just react to current use cases</b>	<b>5</b>
Figure 2: Functional areas with the most advanced AI/ML use cases	6
Figure 3: AI application areas in manufacturing, utilities and healthcare/life sciences	7
<b>Insight 2: Leading organizations leverage diverse infrastructure for AI, and networking strategies must reflect this</b>	<b>8</b>
Figure 4: Multicloud orchestration emerging as a critical infrastructure need	10
<b>Insight 3: Latency, bandwidth and availability shape network strategies for AI frontrunners</b>	<b>11</b>
Figure 5: Key capabilities of networking/network security infrastructure technology partners	12
<b>Insight 4: AI workload management benefits from network segmentation</b>	<b>13</b>
Figure 6: Effectiveness of infrastructure practices for AI workloads	14
<b>Implications</b>	<b>15</b>
Insight 1: Networking planning should anticipate needs, not just react to current use cases.	15
Insight 2: Leading organizations leverage diverse infrastructure for AI; networking strategies must reflect this.	16
Insight 3: Latency, bandwidth and availability shape network strategies for AI frontrunners.	16
Insight 4: AI workload management benefits from network segmentation.	16
<b>About the author</b>	<b>17</b>

# Introduction

The appeal of generative AI has attracted the interest of executives, who regard it as a transformative technology for enhancing efficiency and securing a competitive edge. Substantial funding and executive endorsement have enabled swift experimentation with AI. The primary challenge now is transitioning from these experimental phases to full-scale operational deployment. This report is the first in a three-part series that seeks advice from executives at leading organizations that have successfully implemented AI in production at scale, focusing on the best practices and insights they have gained.

This report examines networking. It specifically explores how AI leaders have designed their networking strategies to successfully implement AI initiatives, as well as some of the challenges faced during this process. The second report will explore AI’s security implications, and the third will discuss general best practices that executives identified over the course of the study.

## Methodology

The insights presented in this report are derived from comprehensive interviews and peer group discussions with senior executives responsible for AI initiatives at organizations that have successfully implemented AI at scale. These interactions included 15 in-depth interviews with AI and IT infrastructure decision-makers, as well as an executive discussion board with 30 participants. All participants in the study were specifically involved in managing and implementing their organizations’ IT infrastructure for AI workloads. Participants were based in the US, UK, Singapore, Australia, Sweden, Germany, Denmark and Japan. This report was developed by S&P Global Market Intelligence 451 Research and commissioned by Verizon.

Figure 1: Demographic breakdown of participants

Executive discussion board				One-to-one interviews			
Region	#	Industry	#	Region	#	Industry	#
North America	10	Healthcare/life sciences	8	North America	5	Retail/wholesale	3
Europe	10	Financial services	8	Europe	5	Healthcare/life sciences	4
Asia-Pacific	10	Manufacturing	6	Asia-Pacific	5	Financial services	5
		Retail/wholesale	3			Manufacturing	2
		Utilities	1			Utilities	1
		Other	4				

Source: S&P Global Market Intelligence 451 Research AI at Scale study commissioned by Verizon.



# Executive summary

Organizations have a clear appetite to engage with AI, and those leading the way are investing in an array of both generative and pattern-recognition models. Executives and practitioners participating in the study were keenly aware of the opportunity the technology holds, particularly for efficiency and cost-reduction, and in some instances top-line growth.

**“Our key business drivers for scaling AI are to improve and enhance the customer experience, operation efficiency and strengthen risk management.”**

**Head of architecture/cybersecurity**

Banking, 20,000-50,000 employees, Singapore

AI use cases mentioned by study participants commonly centered on optimizing functional tasks, such as demand forecasting, compliance and process automation. Additionally, enhanced data insights and support for improved decision-making appear to be continual drivers for many AI initiatives. This emphasis reflects data from [451 Research's Voice of the Enterprise: AI & Machine Learning, Use Cases 2025 survey](#), which found that among the enterprise objectives assessed, generative AI was most consistent in delivering operational efficiency.

Many respondents' AI appetites extend beyond internal capabilities to externally facing applications. Notably, numerous participants are investing in AI to enhance customer service interactions. Other AI use cases that participants highlighted include streamlining sales processes, implementing personalized marketing strategies and driving product innovation.

**“AI-driven automation minimises manual effort, reducing operational costs while improving accuracy and speed.”**

**Principal for AI and cloud**

Professional services, >50,000 employees, UK

A clear refrain across the study was that infrastructure presents a major obstacle to achieving these objectives, and compute capacity (e.g., GPU availability) is not the only relevant resource limitation. As organizations meaningfully embark on their AI journeys, they realize they need enhanced networking resources. The vast majority (90%) of executive discussion board participants expect changes to their networking infrastructure in the next 12 to 24 months, with 71% anticipating moderate or significant upgrades. When asked about the biggest design mistakes they had encountered in their AI initiatives, many respondents pointed to challenges with network architecture.

**“[Our design mistake was] Not taking into consideration latency and bandwidth requirements. Large models require terabytes for that transfer.”**

**Head of architecture/cybersecurity**

Banking, 20,000-50,000 employees, Singapore

**“Running LLMs within [our] current infrastructure would not be possible due to a lack of uncongested bandwidth to support RoCEv2 alone.”**

**Senior infrastructure/network security architect**

Medical devices, 1,000-5,000 employees, Denmark

This report examines four key insights from participants regarding networking strategies:

- **Insight 1:** Networking planning should anticipate needs, not just react to current use cases.
- **Insight 2:** Leading organizations leverage diverse infrastructure for AI, and networking strategies must reflect this.
- **Insight 3:** Latency, bandwidth and availability shape network strategies for AI frontrunners.
- **Insight 4:** AI workload management benefits from network segmentation.

## Insight 1: Networking planning should anticipate needs, not just react to current use cases

As AI projects evolve from the pilot phase to full-scale production, the increased data transfer demands from inference, and in some projects additional training, can significantly strain network resources. Study participants agreed that ensuring effective communication between nodes requires a robust network infrastructure. Many noted that this challenge was overlooked by executives during initial AI experimentation. In the fast-paced environment of generative AI, where organizations are under pressure to deliver projects rapidly, many embarked on initiatives without adequately anticipating future network requirements. The AI trajectory of many businesses is set to add further pressure.

**“Once you start automating processes and talk to the LLM more often, your latencies, your hardware, your GPUs, everything will increase. I think in the matter of the next few years, the majority of AI will be agentic, which means more chatting, more going back and forth between the model and the workflow.”**

**Senior GenAI data scientist**

Financial services, >100,000 employees, US

Respondents from North America generally viewed their AI project deployment processes as more advanced compared to other regions. On a scale from 1–5, North American organizations on average scored the maturity of their deployment processes as 3.0, outstripping European (2.1) and Asia-Pacific respondents (2.2). The challenge of scaling up network planning may be felt more keenly by less advanced organizations. However, even in the US, no respondents felt their processes were as advanced as they could be.

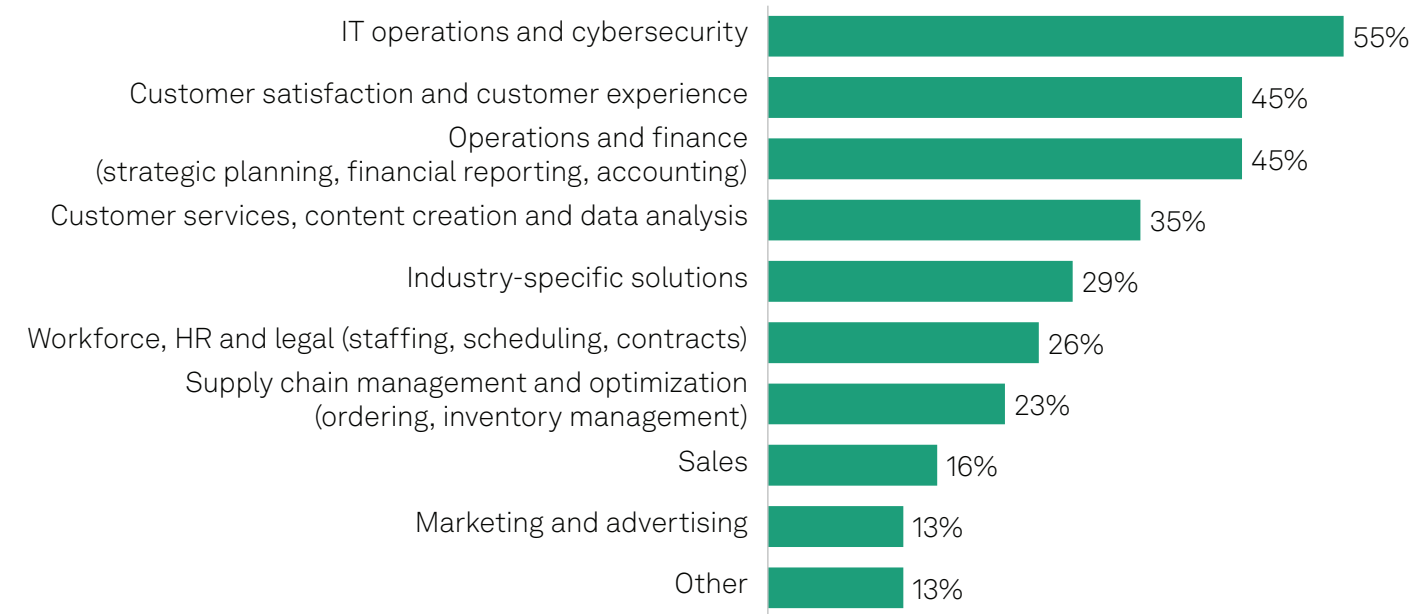
Study participants recognize that the next wave of AI investments will significantly increase pressure on networking resources. This recognition stems from the acknowledgment that many emerging high-impact AI applications will require low latency or edge computing.

**“Real-time sales data analytics; the S&OP [sales and operations planning] and warehouse AI projects are scheduled to be completed in 6-12 months. The experiences with them will inform our future cloud vs. on-prem strategy ... AI edge computing will most certainly be a consideration once we move into production optimization.”**

**Head of cybersecurity and infrastructure**  
Machinery manufacturing, 1,000-5,000 employees, Germany

When asked about advanced AI applications, many participants identified IT operations and cybersecurity, customer satisfaction and experience, and operations and finance as key areas of investment. In many cases, applications in these functional areas appear likely to push the limits of legacy infrastructure. For instance, a conversational interface designed to enhance customer experience may demand higher levels of responsiveness compared to an internally facing application. One study participant highlighted the near five-second processing time of a voice conversation AI service they aimed to deliver, which forced them to explore new configurations and training methods to improve response times.

**Figure 2: Functional areas with the most advanced AI/ML use cases**



Q. Looking at the functional areas below, please indicate which three areas contain your most advanced AI/ML use cases?  
Base: Executive discussion board participants (n=30).  
Source: S&P Global Market Intelligence 451 Research AI at Scale study commissioned by Verizon.

While some respondents focused on broad horizontal use cases, the most challenging AI applications from a networking perspective were often industry-specific. This was particularly true among respondents from manufacturing, utilities and healthcare/ life sciences sectors, where pressure to meet precise requirements is shaping AI infrastructure strategies.

**“Clinical AI for sepsis and stroke triage needs compute power at the edge to work fast. Images for localized GPU compute to process large image datasets without having to use the cloud; ambient listening and transcription in exam rooms.”**

**Senior director of IT**  
Healthcare/life sciences, US

**“Low-latency use cases involving IoT integration for real-time quality control. Quality control systems use AI vision ... Real-time transportation optimization projects involving integration of multiple data sources.”**

**IT director**  
Auto manufacturing, >50,000 employees, US

**Figure 3: AI application areas in manufacturing, utilities and healthcare/life sciences**

Manufacturing	Utilities	Healthcare/life sciences
<ul style="list-style-type: none"> <li>• IoT, including product monitoring/predictive maintenance</li> <li>• Voice of the customer analytics</li> <li>• Real-time quality control/production vision systems</li> <li>• Real-time transportation systems</li> <li>• Real-time sales analytics/sales and operations planning</li> <li>• Warehousing systems</li> <li>• Manufacturing process optimization/OEE</li> <li>• Product innovation</li> </ul>	<ul style="list-style-type: none"> <li>• IoT on supervisory control and data acquisition (SCADA) systems</li> <li>• Water pipe anomaly detection</li> <li>• Water plant data streaming</li> </ul>	<ul style="list-style-type: none"> <li>• Medical triage</li> <li>• Image processing</li> <li>• Medical transcription</li> <li>• Medical device monitoring</li> <li>• Patient behavior monitoring</li> <li>• Care needs assessment</li> <li>• Various vision systems</li> <li>• Remote patient monitoring</li> <li>• Medication monitoring</li> <li>• Drug safety</li> <li>• Robotic nursing assistant</li> <li>• Extended reality systems (e.g., AR, VR)</li> <li>• R&amp;D data analysis</li> </ul>

Source: S&P Global Market Intelligence 451 Research AI at Scale study commissioned by Verizon.

Study participants repeatedly emphasized the advantages of forward planning, with many wishing they had developed a more future-oriented vision at an earlier stage. A common recommendation is to ensure that organizations account for scaling up and deployment during the pilot and training phases of AI projects.

**“Focus ... on model deployment as much as model training.”**

**Head of architecture/cybersecurity**  
Banking, 20,000-50,000 employees, Singapore

**“It is essential to plan with operations in mind and to clarify the non-functional requirements of the production environment — such as performance, availability and security — before starting the PoC.”**

**Director of technology**  
Professional services, 1,000-5,000 employees, Japan

**“Size your GPU/CPU/storage/memory for 18 months in the future. 10gbs [10 gigabits per second] should be the minimum speed for the network, end to end.”**

**VP of IT architecture and infrastructure**  
Healthcare, 20,000-50,000 employees, US

Many study participants noted project delays or derailments due to inadequate up-front planning regarding network architectures. Several pointed to outdated technical infrastructure and a failure to align with technological advancements as placing significant limitations on where AI could be applied and causing serious issues with data movement. This lack of forward thinking often resulted in projects encountering issues while scaling up, at which point upgrading became more expensive, slower and technically challenging. Networking bottlenecks should be anticipated and addressed during the initial planning stages.

**“Response times that posed no issues in the development environment became significantly delayed — by tens of minutes — in the production environment due to a large volume of data flowing through the WAN, which saturated the available bandwidth.”**

**IT strategy director**

Electronics manufacturing, >50,000 employees, Japan

**“At one point, we had a data overflow due to poor planning. We accumulated massive amounts of data and had to do pinpoint problem solving, ‘putting out fires,’ which we look back upon with regret now. Scalability is important to avoid data overflow.”**

**Head of IT systems/R&D**

Manufacturing, 20,000-50,000 employees, Japan

**“Build more than you think you need. Starting down the AI/ML pipeline and then having to stop to upgrade/add more physical equipment will end up costing more down the line than over-purchasing up front.”**

**Director of engineering infrastructure**

Financial services, >50,000 employees, US

## Insight 2: Leading organizations leverage diverse infrastructure for AI, and networking strategies must reflect this

While many organizations initially experimented with generative AI in the public cloud, their workloads started to span various infrastructure venues as they advanced in their journeys. Data preparation, training and inference are distributed across locations from edge to near-edge to core. This complexity is reflected in [451 Research's Voice of the Enterprise: AI & Machine Learning, Infrastructure 2024 survey](#), in which respondents reported using a broad spectrum of venues for inference, including hyperscaler public cloud (61%), computing devices (46%), network operator infrastructure (42%), proximate datacenter environments (41%), stand-alone systems (39%) and specialist AI clouds (32%).

This architectural complexity is important because although some respondents had a simple networking strategy centered on their AI initiatives residing with a single cloud provider, this was the exception rather than the rule. Due to a combination of legacy investments, privacy desires, latency and cost considerations, as organizations increased the scope of their AI projects and introduced new use cases, architectural setups expanded as well. Many experienced survey participants noted that cloud was not always the answer to their AI workload requirements.



**“Depending on workload and use case, we choose whether cloud or on-prem is the best approach for us. Cost considerations are a prime driver for our decision-making.”**

**Head of cybersecurity and infrastructure**

Machinery manufacturing, 1,000-5,000 employees, Germany

**“Edge computing is significantly able to reduce ingress/egress costs of the cloud and latency.”**

**Head of IT infrastructure and security**

Retail, 1,000-5,000 employees, Germany

The majority (54%) of respondents from organizations with more than 20,000 employees are building in hybrid environments comprising a mix of on-premises and cloud infrastructure. Just one participant in the executive discussion board described their current approach as primarily on-premises — a financial services respondent who said this strategy was required for “conforming to our numerous industry certifications that allow us to operate.”

**“We take a hybrid approach: AI training is carried out in the cloud, while real-time inference at the point of service is performed on-premises. This allows us to balance cost and service-level requirements effectively ... For processing sensitive data — such as confidential or personal information that cannot be sent externally — we often rely on on-prem infrastructure to ensure data security.”**

**Head of advanced technology lab**

Retail, >50,000 employees, Japan

Companies that did not account for inter-environment connectivity, routing complexity, internal network needs and bandwidth implications found themselves struggling as workloads scaled and data volumes grew. When asked about their biggest concerns and challenges, companies repeatedly cited the need to move data between environments.

**“Our biggest concern right now is the cost of pushing large datasets from on-prem to Azure and mitigating costs for it.”**

**Senior infrastructure manager**

Healthcare, 20,000-50,000 employees, US

**“We run into training slowdowns when we move large datasets across zones, and inference at remote sites is an issue because of unstable or limited connection options.”**

**Senior director of IT**

Healthcare/life sciences, 10,000-20,000 employees, US

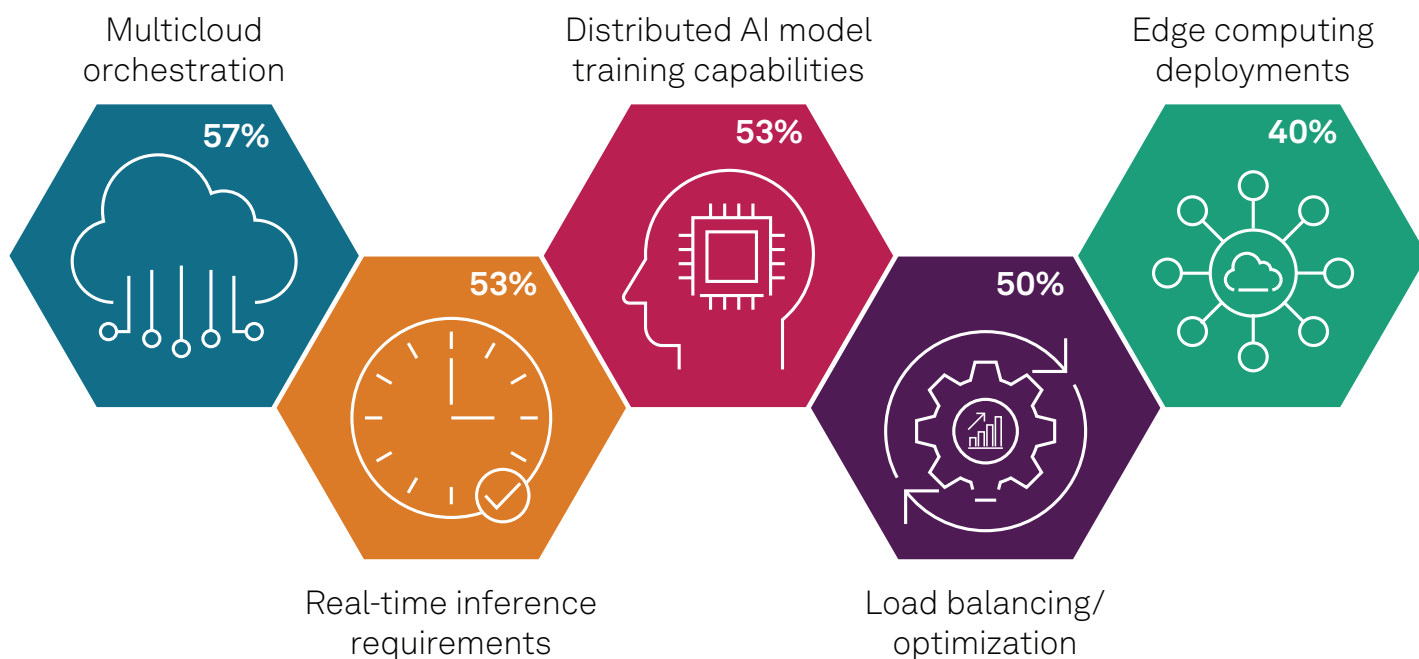
**“Some data is exchanged globally via our company’s private network. With the increasing use of AI, there is a strong need to significantly expand bandwidth.”**

**IT strategy director**

Electronics manufacturing, >50,000 employees, Japan

The scale of this challenge for advanced organizations is evident in their emerging infrastructure needs. When identifying the most critical infrastructure requirements for their AI initiatives, survey participants cited distributed AI multicloud orchestration, followed closely by distributed model training and real-time inference. Respondents also noted the need to address data locality, efficient data transfer, maintenance of consistent model versions and orchestration.

**Figure 4: Multicloud orchestration emerging as a critical infrastructure need**



Q. Looking forward, which of the following, emerging infrastructure needs are most critical for your AI/ML initiatives?

Base: Executive discussion board participants (n= 30).

Source: S&P Global Market Intelligence 451 Research AI at Scale study commissioned by Verizon.

Several AI leaders emphasized the importance of establishing a clearer framework for determining where specific workloads should run. One executive highlighted the need to involve and inform AI application developers and data scientists about this framework to ensure its effective implementation.

**“We are trying to have a strategy ... we are trying to build a process ... to guide the different data science teams, to recommend, saying, ‘Hey, this use case is better for cloud versus this use case is better for on-prem.’ And they need our certification to go and deploy ... when to use batch processing, real-time inferencing and things like that.”**

**Senior director of AI products and platforms**  
Financial services, 20,000-50,000 employees, US

# Insight 3: Latency, bandwidth and availability shape network strategies for AI frontrunners

As organizations scale up AI applications, the need for low-latency, high-bandwidth infrastructure emerges as a key foundation for network planning. Many participants emphasized the need to shape networking strategies to efficiently move high volumes of data and support real-time processing. Architectural advice consistently focused on a trio of factors: availability, latency and bandwidth.

**“Aim for high bandwidth and low latency.”**

**Head of infrastructure services**

Banking, 20,000-50,000 employees, Australia

Organizations that have successfully delivered AI initiatives at scale have already invested to ensure strong practices for network availability. When asked to score their effectiveness on a scale of 1-5, 92% scored themselves as a 4 or 5, higher than any other infrastructure practice assessed. Despite this level of advancement, the area is deemed sufficiently critical that most identified a need for continued investment.

**“Network availability assures the stream of accurate real-time data for analytics that needs to be taken into account on decision-making.”**

**Director of infrastructure and cloud**

Gaming, 1,000-5,000 employees, Sweden

**“Establishing high-speed, low-latency connections by using dedicated lines to connect with on-premises systems and other cloud environments.”**

**IT strategy director**

Electronics manufacturing, >50,000 employees, Japan

**“Design a scalable infrastructure with high-speed and low-latency networks. Optimize data flows with distributed storage and caching.”**

**Director of cloud engineering**

Insurance, 20,000-50,000 employees, US

The emphasis on latency, availability and bandwidth is crucial because respondents cited numerous case studies where networking limitations either derailed projects or continue to constrain AI efforts.

**“Network is hindering our ability to deploy more functionality because of latency issues and the size of the connections.”**

**Senior director of IT ops and strategy**

Chemicals manufacturing, 1,000-5,000 employees, US

“At times, legacy segmentation and sometimes lack of real-time data flow control create bottleneck for scaling our real-time, cross-border AI application.”

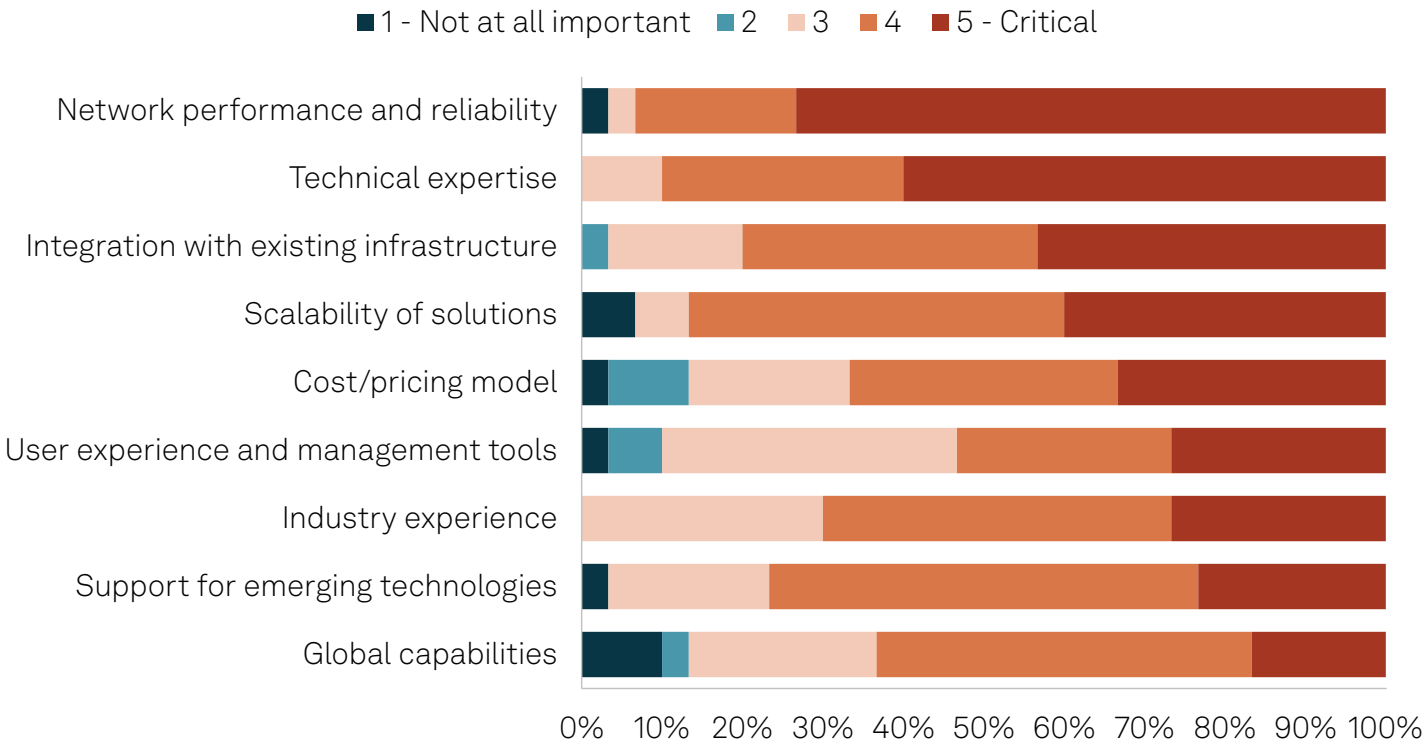
**Head of architecture/cybersecurity**  
Banking, 20,000-50,000 employees, Singapore

“Optimizing data transfer — specifically improving bandwidth and latency — is a key area to consider ... Reason: In AI/ML workloads, large volumes of data are transferred to the cloud, and bandwidth can easily become a bottleneck.”

**Director of technology**  
Professional services, 1,000-5,000 employees, Japan

Participants identified networking performance and reliability as the most crucial factors to consider when evaluating technology partners for networking and security. Participants emphasized this area even more than cost, industry experience and user experience (see Figure 5). This reinforces a central theme across the study: An effective AI infrastructure strategy must have networking performance at its core.

**Figure 5: Key capabilities of networking/network security infrastructure technology partners**



Q. When evaluating networking/networking security infrastructure technology partners for your AI/ML initiatives, please rate the importance of the following (1=Not at all important; 5=Critical).  
Base: Executive discussion board participants (n= 30).  
Source: S&P Global Market Intelligence 451 Research AI at Scale study commissioned by Verizon.

# Insight 4: AI workload management benefits from network segmentation

Survey participants emphasized the importance of establishing dedicated subnetworks specifically for AI-capable hardware. When asked about infrastructural best practices for AI workloads, multiple respondents identified network segmentation — creating isolated network segments with tailored routing policies for AI workloads.

**“Segregation per environment and ML-AI life cycle, so training versus operations.”**

**VP of architecture**

Professional services, 1,000-5,000 employees, UK

**“Create zone layers architecture.”**

**Global director of infrastructure and operations**

Automotive/aerospace manufacturing, 10,000-20,000 employees, Germany

**“For on-prem, we are putting up on the network uplift, and create a separate subnet for AI/ML-capable hardware, and have their specific routes.”**

**Cloud services and hosting lead**

Water utilities, 1,000-5,000 employees, Australia

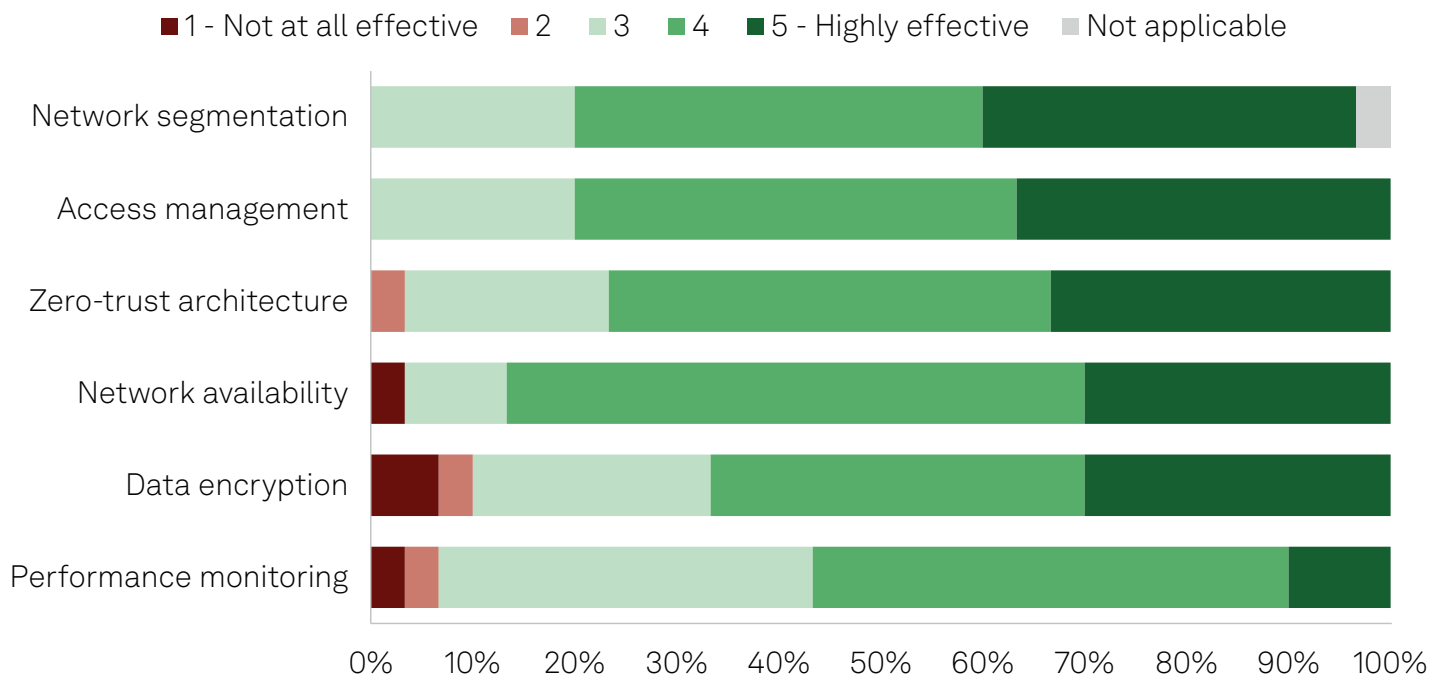
Two primary factors are driving this push toward network segmentation. First, survey participants noted that high-impact AI applications typically generate significant network traffic. By isolating this traffic, organizations can exercise greater control, prevent network congestion and ensure that AI/ML processes do not interfere with other operations. The performance of GPU clusters involved in AI training is highly dependent on network throughput and latency. Ensuring that non-AI traffic is excluded can maximize network performance.

Second, network segmentation can limit the exposure of AI workloads by adding layers of security specifically configured for AI workloads, thereby safeguarding sensitive data. AI models and their concentration of critical insights make them significant targets for attackers. The layered isolation that segmentation provides is a key security benefit. This topic will be explored in greater detail in our upcoming report on security.

Organizations that have successfully delivered AI projects at scale invested heavily in network segmentation (see Figure 6).



Figure 6: Effectiveness of infrastructure practices for AI workloads



Q. Please rate the effectiveness of the following infrastructure practices for your AI/ML workloads (1=Not at all Effective, 5=Highly Effective).  
Base: Executive discussion board participants (n=30).  
Source: S&P Global Market Intelligence 451 Research AI at Scale study commissioned by Verizon.

**“We have a segmented, zero-trust networking architecture with a subnet for training, inference and data pipelines.”**  
**Head of architecture/cybersecurity**  
Banking, 20,000-50,000 employees, Singapore

**“Network segmentation is a bank-wide direction ... It’s managed by a centralized team with very strong SRE processes in place.”**  
**Head of architecture/cybersecurity**  
Banking, 20,000-50,000 employees, Singapore

Maturity in network segmentation does not always correlate with participation in industries that are associated with data privacy concerns. On average, healthcare and life sciences respondents reported less mature network segmentation strategies than other industries. Several respondents from the financial services sector also identified room for improvement.

The emphasis on network segmentation is particularly notable due to the array of challenges that complicate segmentation efforts. One pertinent challenge is cost. When executive discussion board participants were asked about the top three challenges they faced with AI, 58% cited cost-related issues, and this number increases when considering ROI concerns. Additionally, the study illustrated that infrastructure teams are often stretched thin. The burden of deploying and managing segmentation could further strain networking and security teams. Network automation can help, but it's a capability that enterprises have struggled to master.

Additionally, many participants cited the complexity of legacy architecture as a barrier to AI enablement. This complexity has implications for the integration requirements that a network segmentation strategy must address. It may also impact the ability to maintain business continuity while segmenting the network. It is particularly noteworthy that, despite all these challenges, a broad sample of study participants emphasized the importance of segmentation.

## Implications

Few executives question whether AI can offer value to their organization; the bigger question is how to leverage the technology effectively. In 451 Research's Voice of the Enterprise: AI & Machine Learning, Use Cases 2025 study, just 5% of respondents said generative AI was not in use at their organization, with a further 11% saying it was used only informally on an individual basis. However, just 27% have deployed generative AI across their organization, with the bulk of organizations still experimenting with the technology or attempting to scale up smaller deployments.

What is hindering many organizations from scaling their AI initiatives? Participants in this study who have successfully implemented AI at scale identify networking infrastructure as a significant potential bottleneck in the transition. As organizations develop a networking strategy to support full-scale operational deployment, they should keep four leading practices at the forefront of their planning:

### **Insight 1: Networking planning should anticipate needs, not just react to current use cases.**

Executives should avoid the trap of reactive measures and focus instead on proactive network planning. Rather than focusing on immediate needs for pilots and model training, organizations should anticipate the needs of AI initiatives in production, seeking to estimate resource requirements in advance. This may require greater clarification of non-functional requirements such as performance, availability and security before starting proofs of concept. It will also involve planning excess capacity to accommodate growth and avoid project delays.

## **Insight 2: Leading organizations leverage diverse infrastructure for AI; networking strategies must reflect this.**

While early experimentation may take place in the cloud, delivering AI projects into production will often require different infrastructure options. Networking strategy must be flexible to address planning for workloads that span various environments, from edge to core, and anticipate the need for inter-environment connectivity. Modern infrastructure is commonly hybrid, and network investments must reflect that fact. Establishing clear frameworks for workload placement and involving AI application developers and data scientists in that decision-making can mitigate some of this architectural complexity.

## **Insight 3: Latency, bandwidth and availability shape network strategies for AI frontrunners.**

To address the high volumes of data, as well as increasingly real-time processing requirements associated with AI, executives should emphasize low-latency, high-bandwidth networks as a foundational element of network planning. It is advisable to focus on the trio of availability, latency and bandwidth when designing network architectures, potentially looking to further optimize data flows using distributed storage and caching. Proactive, continuous investment in network availability can help organizations avoid project derailments.

## **Insight 4: AI workload management benefits from network segmentation.**

Key strategies to optimize AI workloads include establishing dedicated network segments for AI clusters and ensuring adequate capacity for data and model movement across on-premises and cloud infrastructure. Organizations can also create isolated network segments with tailored routing policies to manage significant network traffic and prevent congestion, taking into consideration different environments and AI/ML life-cycle stages, such as training versus operations.

This report is the first in a three-part series on insights gleaned from AI leaders based on their experiences launching and scaling up successful AI initiatives. The [second report](#) in this series covers security, and [the third](#) examines overall best practices.



### **About Verizon Business**

Verizon Business is all about helping organizations like yours succeed in today's dynamic digital environment. We provide essential network solutions that support and enhance business operations and empower how millions of people live, work, and play every day. This research was commissioned to provide insights into the challenges of AI deployment, share peer experiences, and clarify the strategic choices facing IT leaders. Our aim is to offer a deeper understanding of these intricate dynamics, especially the critical importance of network and security infrastructure. We believe these insights can help your organization to scale AI confidently, maximize its potential, and consistently stay ahead in this rapidly evolving landscape.

# About the author



## Alex Johnston

### Senior Research Analyst

Alex Johnston is a senior research analyst on the 451 Research Data, AI & Analytics team at S&P Global Market Intelligence. He focuses on emerging technologies and how they can be applied in business contexts. Alex's primary coverage areas are artificial intelligence, distributed ledger technology, event stream processing and data marketplaces. Alex's recent areas of concentration include monitoring the emerging generative AI market, tracking the evolution in blockchain use cases and investigating real-time architectures.

## About this report

A Discovery report is a study based on primary research survey data that assesses the market dynamics of a key enterprise technology segment through the lens of the “on the ground” experience and opinions of real practitioners — what they are doing, and why they are doing it.

## About S&P Global Market Intelligence

At S&P Global Market Intelligence, we understand the importance of accurate, deep and insightful information. Our team of experts delivers unrivaled insights and leading data and technology solutions, partnering with customers to expand their perspective, operate with confidence, and make decisions with conviction.

S&P Global Market Intelligence is a division of S&P Global (NYSE: SPGI). S&P Global is the world's foremost provider of credit ratings, benchmarks, analytics and workflow solutions in the global capital, commodity and automotive markets. With every one of our offerings, we help many of the world's leading organizations navigate the economic landscape so they can plan for tomorrow, today. For more information, visit [www.spglobal.com/marketintelligence](https://www.spglobal.com/marketintelligence).

## CONTACTS

**Americas:** +1 800 447 2273

**Japan:** +81 3 6262 1887

**Asia-Pacific:** +60 4 291 3600

**Europe, Middle East, Africa:** +44 (0) 134 432 8300

[www.spglobal.com/marketintelligence](http://www.spglobal.com/marketintelligence)

[www.spglobal.com/en/enterprise/about/contact-us.html](http://www.spglobal.com/en/enterprise/about/contact-us.html)

Copyright © 2025 by S&P Global Market Intelligence, a division of S&P Global Inc. All rights reserved.

These materials have been prepared solely for information purposes based upon information generally available to the public and from sources believed to be reliable. No content (including index data, ratings, credit-related analyses and data, research, model, software or other application or output therefrom) or any part thereof (Content) may be modified, reverse engineered, reproduced or distributed in any form by any means, or stored in a database or retrieval system, without the prior written permission of S&P Global Market Intelligence or its affiliates (collectively S&P Global). The Content shall not be used for any unlawful or unauthorized purposes. S&P Global and any third-party providers (collectively S&P Global Parties) do not guarantee the accuracy, completeness, timeliness or availability of the Content. S&P Global Parties are not responsible for any errors or omissions, regardless of the cause, for the results obtained from the use of the Content. THE CONTENT IS PROVIDED ON "AS IS" BASIS. S&P GLOBAL PARTIES DISCLAIM ANY AND ALL EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, ANY WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE OR USE, FREEDOM FROM BUGS, SOFTWARE ERRORS OR DEFECTS, THAT THE CONTENT'S FUNCTIONING WILL BE UNINTERRUPTED OR THAT THE CONTENT WILL OPERATE WITH ANY SOFTWARE OR HARDWARE CONFIGURATION. In no event shall S&P Global Parties be liable to any party for any direct, indirect, incidental, exemplary, compensatory, punitive, special or consequential damages, costs, expenses, legal fees, or losses (including, without limitation, lost income or lost profits and opportunity costs or losses caused by negligence) in connection with any use of the Content even if advised of the possibility of such damages.

S&P Global Market Intelligence's opinions, quotes and credit-related and other analyses are statements of opinion as of the date they are expressed and not statements of fact or recommendations to purchase, hold, or sell any securities or to make any investment decisions, and do not address the suitability of any security. S&P Global Market Intelligence may provide index data. Direct investment in an index is not possible. Exposure to an asset class represented by an index is available through investable instruments based on that index. S&P Global Market Intelligence assumes no obligation to update the Content following publication in any form or format. The Content should not be relied on and is not a substitute for the skill, judgment and experience of the user, its management, employees, advisors and/or clients when making investment and other business decisions. S&P Global keeps certain activities of its divisions separate from each other to preserve the independence and objectivity of their respective activities. As a result, certain divisions of S&P Global may have information that is not available to other S&P Global divisions. S&P Global has established policies and procedures to maintain the confidentiality of certain nonpublic information received in connection with each analytical process.

S&P Global may receive compensation for its ratings and certain analyses, normally from issuers or underwriters of securities or from obligors. S&P Global reserves the right to disseminate its opinions and analyses. S&P Global's public ratings and analyses are made available on its websites, [www.standardandpoors.com](http://www.standardandpoors.com) (free of charge) and [www.ratingsdirect.com](http://www.ratingsdirect.com) (subscription), and may be distributed through other means, including via S&P Global publications and third-party redistributors. Additional information about our ratings fees is available at [www.standardandpoors.com/usratingsfees](http://www.standardandpoors.com/usratingsfees).