ENABLING AI WORKLOADS WITH AN EDGE-TO-CLOUD NETWORK

Successful AI provisioning depends on a nuanced understanding of how and where the network can play a role. Don’t let network bottlenecks hold back your AI progress. Here’s how.

At a glance

- Today's IT leaders feel confident that their network infrastructure is set up to support the imminent demands of soaring AI traffic.
- However, less than half admit to having a full understanding of what those networking demands might actually be—hinting at a worrying infrastructure disconnect.
- Next-generation AI architectures require a flexible network that can deliver the right combination of high-performance and low-latency connectivity to ensure the fastest provisioning whether the AI ask is training, inferencing, or tuning. It is a mix you can't afford to get wrong.

Over the last decade, businesses have come to appreciate the network's role beyond its connectivity offering: it has been vital in supporting the ever-evolving digital demands of today's workplaces—from enabling a **work-from-everywhere** culture to unlocking the increasing variety of IoT use cases deployed across businesses. Fast forward to 2024, and we're seeing enterprise adoption of AI gaining pace rapidly, giving rise to an important question. Are current enterprise networks offering effective connections—from the cloud right to the edge—in support of the growing compute-intensive AI workloads?

Well yes, 93% of IT leaders believe their network infrastructure is already set up to support AI traffic. This was one of the findings from the **Architect an AI Advantage**[1] survey that HPE commissioned from Sapio Research earlier this year.

However, the **network only ranked fifth in IT leaders' priority investment list for supporting AI efforts**. And perhaps more concerningly, when we dug a little deeper, we found that less than half of respondents admitted to fully understanding the nuanced needs for networking across the full AI lifecycle.

In overestimating their readiness, IT leaders may not yet have given their network the appropriate level of consideration or investment that it needs—which could lead to inadequate performance, security disconnects, stalled progress, and more. The network has an important role to play in long-term AI success, meaning that it's crucial to have the right network strategy in place, especially to support activity from edge to cloud.

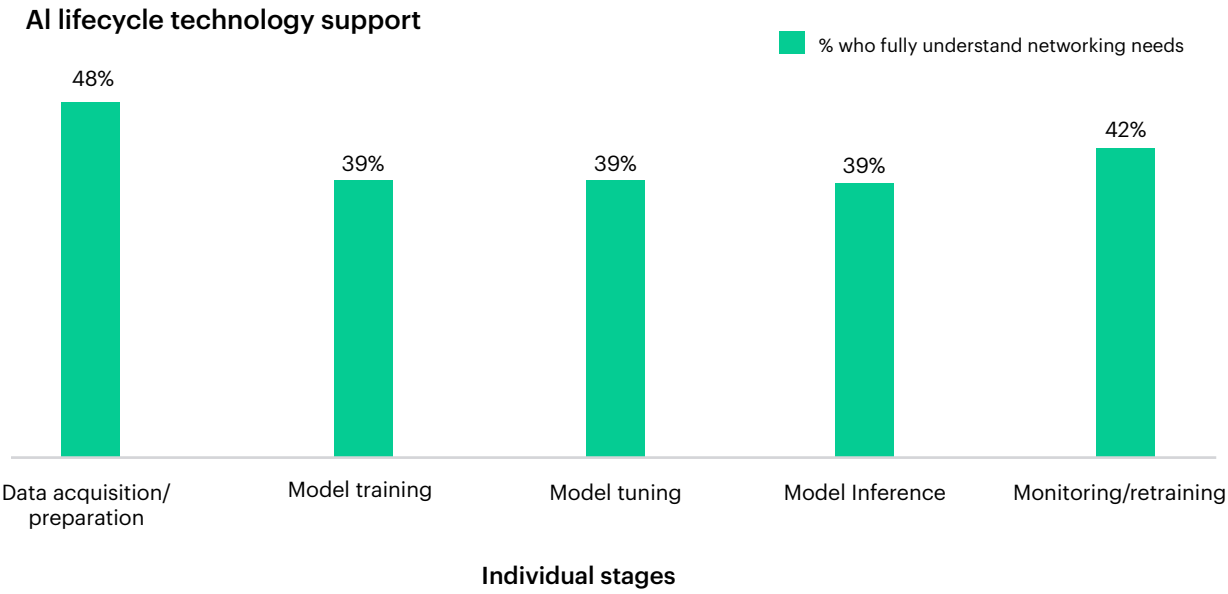


Figure 1. Percentage of IT decision-makers who fully understand networking needs across the AI lifecycle

[1] "Architect an AI Advantage," Sapio Research, January 2024

The network must operate seamlessly from the edge to the cloud to continuously deliver a steady and coordinated flow of data at the right time to GPUs and CPUs.

The network wears many hats

The role of the network will change depending on which of the **three main phases**—data acquisition, model training and tuning, and inferencing—**it is supporting within the AI lifecycle**.

Only 48% of surveyed IT leaders said they had a good understanding of the network requirements to support phase one's data acquisition and preparation—21% had limited or no understanding. For AI model inferencing, training, and tuning, less than 40% of IT leaders knew what was needed from the network for these parts of the AI lifecycle.

A nuanced understanding is critical if businesses are to prevent the network from becoming a bottleneck in AI processes. Let's take a closer look at the network's role across each phase in the AI lifecycle.

1. Data acquisition phase

Data is the lifeblood of any AI project—for any AI system to make smart and accurate decisions, it first needs access to large and diverse datasets to learn from. It is therefore essential to **control data to provide a much safer, more precise, and more effective AI solution**. This is a mission for the network: to capture, secure, and transport data quickly and easily from any source to any destination—across edge, data center, and cloud. But, to succeed in its mission, the network must have an integrated connectivity fabric with a single operational services interface running from edge to cloud.

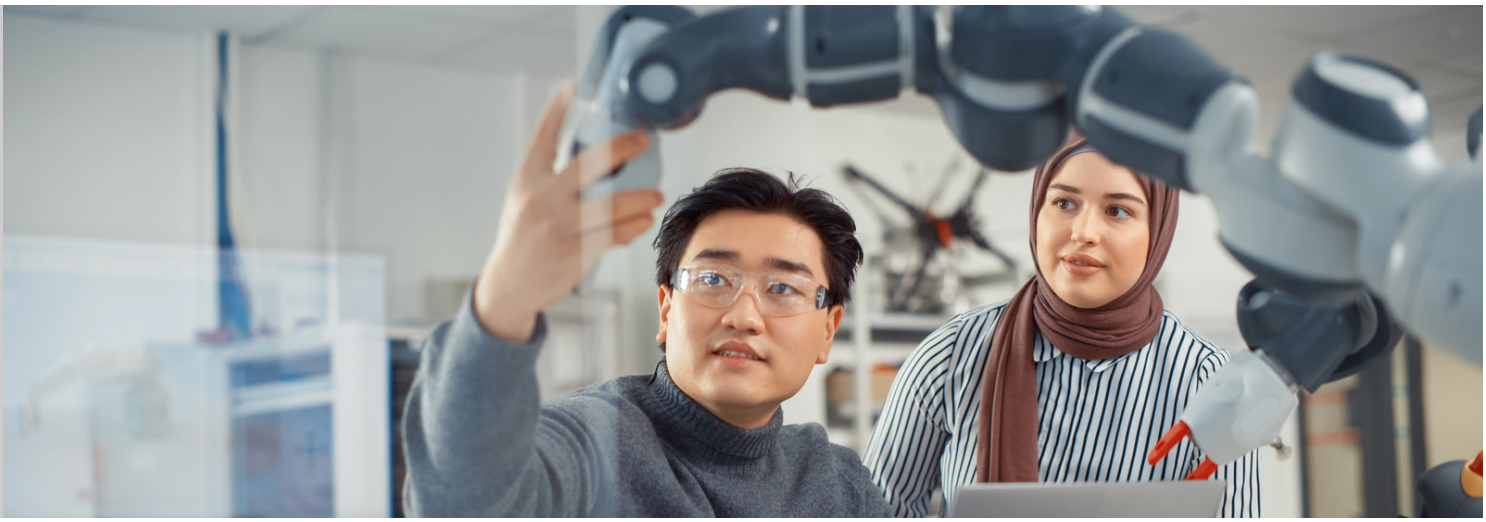
Access to the edge—the point at which people and devices connect—is particularly important here because it's where an organization generates and consumes data, without it first needing to travel through the cloud or data center for processing. The edge is playing host to a growing number of IoT devices as well, and these offer AI a rich source of data for training and inferencing, which the network should support. How? By **acting as an onramp for quick delivery of high-quality, insight-rich data that is captured and processed right at the source (the edge) and then transported to wherever it is needed**—whether the data lake for AI training or operationally for AI inferencing.

2. Model training and tuning phase

Once data has been captured, it must be leveraged to train the algorithm. Training and retraining AI models on huge amounts of data is an intense activity, requiring heavy horsepower matched with low-latency, high-performance network interconnects for hundreds or thousands of GPUs. Simply put, the **network needs to be up to the task of delivering optimized, power-efficient, and predictable performance** for training and tuning across multitenant workload environments.

3. AI inferencing phase

This phase involves deploying newly trained AI models to wherever the business has chosen to run them (that is, on-prem, at the edge, or in the cloud) so that the fresh data that's generated there can be analyzed and acted upon quickly to unlock business value. Inference workloads tend to need GPU acceleration for large models although some applications can rely on CPUs alone. Either way, the network must provide proper connectivity performance; without this support, there will be issues with latency and losslessness, which will dramatically affect the model's efficacy.



Key network necessities

No matter the lifecycle phase, AI workloads must be supported by a network providing high-performance connectivity. How can the enterprise ensure its AI-related network strategy is sound? By giving due consideration to **key network imperatives, which include broad-based infrastructure, unified visibility and control, automation, and "integral security**."

1. Broad-based infrastructure

To satisfy the many different AI use cases, an enterprise should implement the **broadest set of connectivity** options so that the network can collect all data no matter where it's generated. Options should stretch beyond just wired, to include Wi-Fi with full integrated support for connecting and managing IoT devices and private 5G that extends wireless LAN use cases—both can and should be used simultaneously. Extended Wi-Fi options like Wi-Fi 6E and Wi-Fi 7 offer access to the 6 GHz band for more capacity, wider channels, and less interference—meaning stronger connectivity for demanding AI applications leveraging IoT data.

2. Unified visibility and control

Implementing a broad set of connectivity options can easily lead to siloed infrastructure. Enterprises need a unified **edge-to-cloud network**, which will enable them to apply consistent security policies, centralize management, and understand the status of the network from a single point of visibility and control.

3. Automation

At an enterprise level, the network should enable zero-touch service management with a single toolset—letting you **automate the management of physical infrastructure across devices and locations**. By simplifying and standardizing all network operations through intelligent automation, businesses can lower operational costs and increase user satisfaction.

4. Integral security

Today's networks must have security built in as opposed to security bolted on. Integrated security will help an enterprise mitigate the growing threats landscape—something that 94% of IT leaders say will have been made worse by AI—with greater success. The bottom line is that keeping bad actors away from data, applications, and infrastructure—and meeting compliance requirements—demands a **security-centric network that leverages Zero Trust and SASE** principles to ensure that the wrong workloads don't get to the wrong places and that access is restricted to only those with the right credentials. Security must be prioritized from the outset too because the network has a defining role to play within the data acquisition phase of the AI lifecycle: to secure all the data collected from the start, so it doesn't get corrupted or tampered with. Beyond protecting the model itself, the right network helps ensure that AI is trained on sound data and that AI outputs down the line are rooted in a reliable source, with inferencing generating credible business insights.

Beyond these necessities, enterprises should prioritize flexibility, too. AI is not static—it is a living solution, and the network must be able to respond to this both in its provisioning and perhaps even its financing (that is, considering network as a service).

While the network is critical for enabling AI, many network capabilities can also be enabled by AI.

Symbiotic relationship: AI and the network

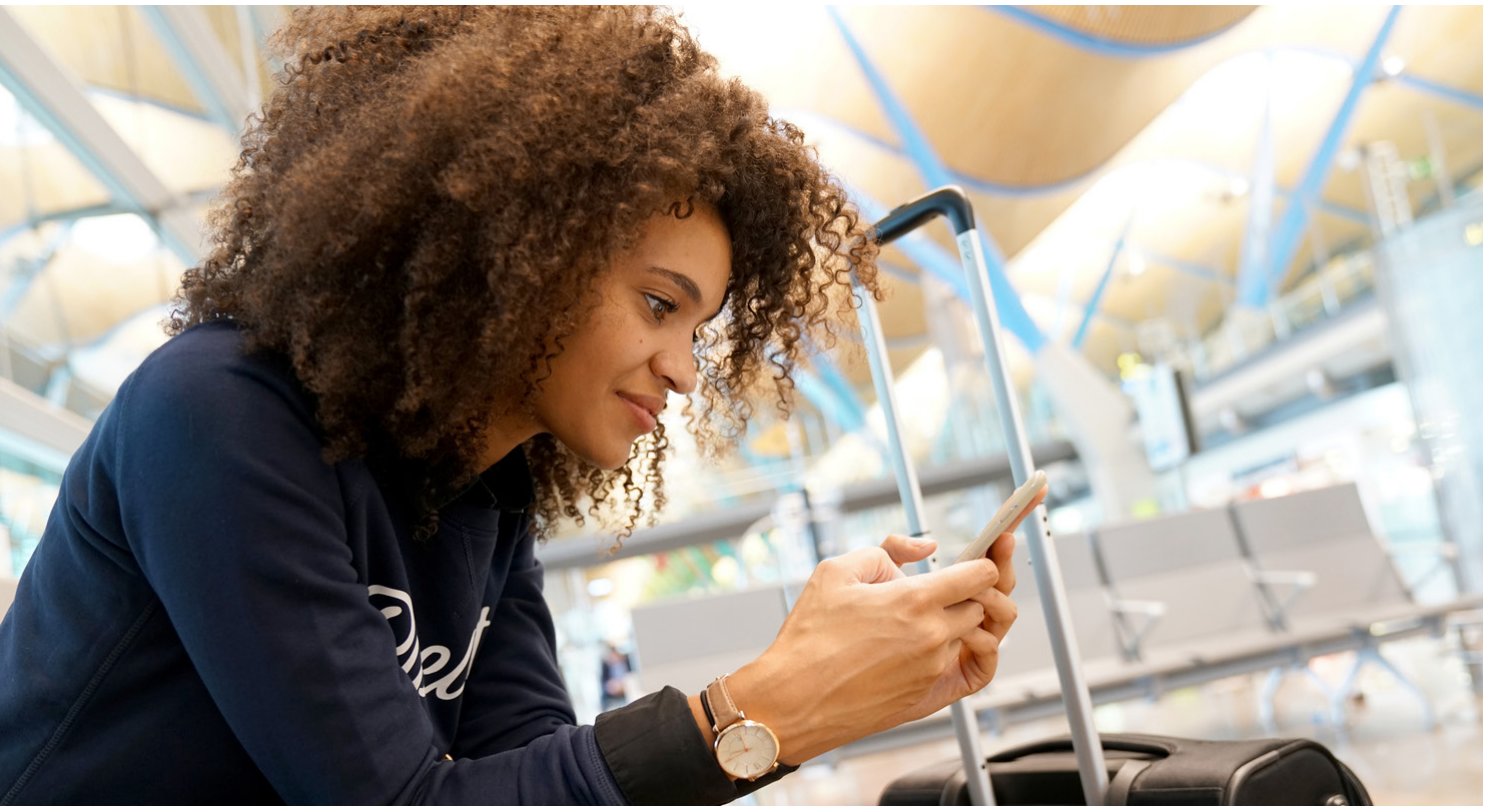
Enterprise networks are growing more diverse and distributed—to the point where managing and protecting them with traditional, manual techniques is becoming almost impossible. Networks are only as good as the people overseeing them and errors are inevitable when operators are stretched and overwhelmed. This is where AI flips into support mode.

Organizations around the world are embracing AI as a force multiplier for improved productivity of IT operations—and networking has taken a leadership position in this use case. AI for IT operations (known as AIOps) combines Big Data, analytics, large language models (LLMs), and machine learning to make life easier for networking teams. It offers 24x7 monitoring, automatically identifies issues before they impact the business, determines root causes, and delivers troubleshooting and optimization guidance in real time. Used this way, AI helps to save time and money, improve security, and boost the efficiency of the team and network.

Networking for AI

No matter the AI use case, two things are certain: modern AI applications need to be supported by a high-bandwidth, low-latency, secure, and scalable network. And this network must operate effectively across the full AI lifecycle—from data capture and aggregation to high-performance model training and edge-to-cloud inference deployment.

Given the network-boosting benefits that AI can also deliver, the two technologies have an important future together, where one cannot perform at its best (or at all) without the other. Success in AI means embracing these interlinked technologies simultaneously. HPE Aruba Networking can help you get started.



About HPE

HPE is the edge-to-cloud company that helps organizations accelerate outcomes by unlocking value from all of their data, everywhere. Built on decades of reimagining the future and innovating to advance the way people live and work, HPE delivers unique, open, and intelligent technology solutions, with a consistent experience across all clouds and edges, to help customers develop new business models, engage in new ways, and increase operational performance.

Learn more at

HPE.com/edge

Visit HPE.com

Chat now

a50010954ENW, Rev. 1

HEWLETT PACKARD ENTERPRISE

hpe.com