Where Should AI Live?

The AI Infrastructure Report

Ninety-four percent of organizations have regrets about their initial AI infrastructure decisions. New research from 1,201 IT leaders shows what they would do differently.

By Scott Sinclair, *Practice Director,*
and Jim Frey, *Principal Analyst*
Omdia

August 2026

Table of contents

Introduction

While most organizations have an AI infrastructure plan, research shows that 94% of them have regrets. Organizations are learning the hard way that AI is not a single workload but a broad and diverse mix of applications that span multiple use cases, models, and datasets. To support the diversity of AI, organizations need hybrid infrastructure. Having a mix of cloud, on-premises data centers, and edge infrastructure provides vital flexibility, enabling IT decision makers to optimize specific AI workloads and maximize the optimal mix of key metrics, including performance, cost, data security, and compliance. Deployment flexibility does not guarantee success, though.

Early deployments have brought missteps and cost surprises, leading to retrenchment, missed goals, and even repatriation.

The specific needs for AI training, fine tuning, and inference vary for organizations and across use cases, and, as such, different deployment locations offer distinct advantages. On-premises wins on cost, security, and control for the AI work that matters most: fine-tuning and training on proprietary data. That advantage is about to drive a serious wave of investment in data centers and edge environments. This investment will address the realization that server and network resources become the new bottleneck. To gain insight into these trends related to AI infrastructure decisions, Cisco commissioned Omdia to execute a survey, along with five in-depth interviews, of 1,201 IT decision makers with purchasing authority for data center infrastructure and influence over hosting decisions for AI workloads. All respondents have custom-trained or fine-tuned models in deployment today. The data tells a consistent story: *Hybrid is here to stay, on-premises is undervalued, and most organizations have already paid the price for getting this wrong.*

96%
of organizations currently utilize a hybrid approach to AI workload distribution.

54%
The most common primary objectives for hybrid AI deployments are workload-specific optimization **(54%)**, scalability and flexibility **(53%)**, and cost efficiency and performance **(51%)**.

94%
of organizations have at least some regrets regarding their initial AI infrastructure strategy.

54%
of organizations confirm that on-premises infrastructure is more cost-effective than the public cloud for training workloads.

82%
of organizations agree that their AI infrastructure needs to keep pace with security threats and regulatory changes.

82%
agree that controlling who can access their AI systems and data is a significant factor in data center infrastructure decisions.

76%
of organizations agree that data sovereignty requirements influence their AI infrastructure decisions.

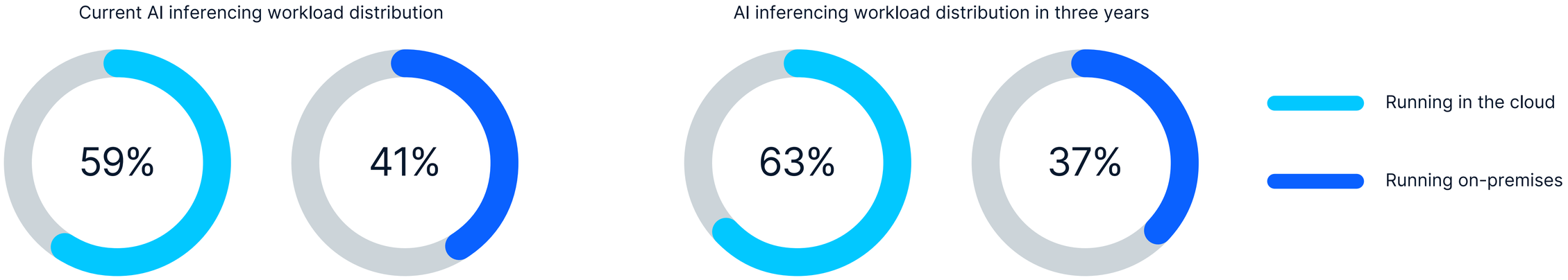
Back to Contents

The journey to hybrid

Hybrid is not a phase—it is the plan

Ninety-six percent of organizations already run AI across a mix of cloud, on-premises, and edge infrastructure. That is not a transitional state; it is where AI infrastructure lives, and it is where it is likely going to stay. The data shows a slight expected shift toward public cloud, but the overall hybrid model stays dominant—and it is being actively rebalanced. In fact, 60% of organizations also shared they have already moved AI workloads from the cloud back to on-premises, a clear signal that early deployment decisions are being corrected.

Hybrid AI workload distribution expected to remain the planned approach throughout the next three years



Back to Contents

Workload fit is the whole game—and many teams are still figuring that out

Flexibility is the right starting point. But it does not deliver results on its own. What actually works is matching each workload to the environment where it runs best. More than half of organizations building hybrid AI are already prioritizing this approach. A manufacturing CIO with over 10,000 employees was direct about this: Anything that needs a real-time decision stays on-premises, but everything else can go to the cloud.

Scalability and cost efficiency are not separate priorities from workload placement—they are the outcome of getting it right. Once workloads live where they belong, efficiency follows. The bigger takeaway: It is not just where an organization's data lives that should decide where AI runs. It is what that AI is being asked to do.

Top reasons for adopting a hybrid strategy for AI

Optimize workloads based on their specific needs

Improve scalability and flexibility

Balance cost efficiency and performance

Meet data sovereignty and compliance requirements

Leverage existing on-premises investments while utilizing cloud capabilities

Reduce latency for real-time applications

Ensure redundancy and disaster recovery

Back to Contents

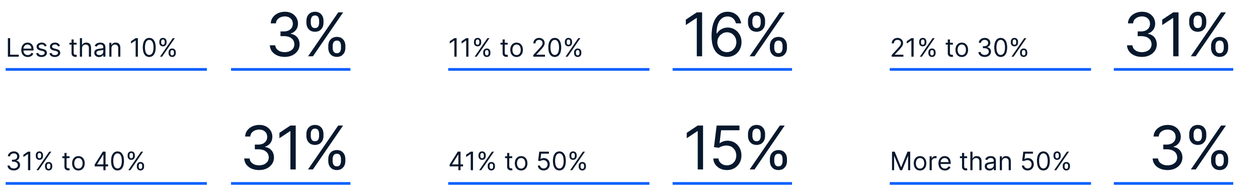
Custom models usually stay close to home, but everything else is up for debate

Enterprise leaders draw a clear line between two types of AI work. Custom models built on proprietary data—engineering drawings, sensitive client files, product designs—almost all stay on-premises. These are the areas where organizations have real domain expertise and where handing over control is not an option. Pre-trained models are a different story. For standardized functions—security, orchestration, threat detection, customer-facing tools—organizations go hybrid because off-the-shelf solutions get them there faster.

The impact: Rising cloud costs

On-premises is cheaper for training, but the full picture is more nuanced. The public cloud has an obvious appeal. It is fast to start, and it gets AI workloads running without major upfront investment. But the costs catch up quickly. Organizations see cloud spending jump an average of 30% after deploying AI, and some are seeing increases of 40–50% or more. More than half (54%) now say on-premises is more cost-effective for training.

Percentage of cloud cost increase for AI workloads after deployment



> As an organization that is new to AI, we are just finding out what the reality is in terms of cost. We are finding out that the more we use, the more it costs."
>
> *— C-level executive in construction/ engineering with over 10,000 employees*

The data problem is real, and executives are not sugarcoating it:

> We work with some very sensitive clients, so we were not very comfortable taking open models to the cloud. We decided to do it in-house. And second, we wanted control over our environment, and, as most of our applications are running on-premises, we decided that it is better for us to make that investment and train our own model and host it in our own data centers...But we would definitely use a [hybrid cloud approach] for areas where we can find specialized models and in areas that are not very sensitive in terms of our data...Security is the primary driver for us to go with our own models."
>
> *— C-level executive in manufacturing with over 10,000 employees*

> The ugly truth for any company [is that] you're probably sitting on more unstructured data than anybody wants to believe. How on earth can you train that model?"
>
> *— C-level executive in healthcare with over 1,500 employees*

Back to Contents

The response: Scale back

Whether from a lack of planning or unanticipated growth, the impact of cloud costs has been direct and immediate, and AI projects themselves have been one of the clear casualties. Over half (52%) of organizations had to scale back AI initiatives because of cost. While this outcome may end up resulting in more efficiencies in the long run, it may also bring delays and missed milestones and objectives.

What organizations wish they had done sooner

Nearly all organizations (94%) have regrets about their original AI infrastructure strategy. When asked what they would do differently, two answers came up most frequently: do a more rigorous cost analysis before deploying (34%) and invest in on-premises infrastructure earlier (33%). Both point to the same problem: Unpredictable cloud costs hit harder than most expected. Organizations that put on-premises infrastructure in place earlier are better positioned to cap that risk before it compounds.

Then there is the pace problem. Over a quarter of organizations (28%) say they wish they had pushed back harder on how fast leadership wanted to move. That is a difficult position to be in—and an expensive one. When the pressure to move fast outpaces the team's ability to get it right, the cost of fixing it later is almost always higher than slowing down would have been.

Common lessons from early missteps with AI deployments

Complete more rigorous cost analysis before committing	34%	Invest in our own infrastructure earlier	33%
Push back harder on the pace of AI adoption	28%	Move to the cloud more aggressively	21%
Keep more workloads on our own infrastructure from the start	13%	Nothing - we made all the right calls	6%

Insights from decision makers on lessons learned

Executives have indicated that, before implementing AI, organizations need to evaluate their existing infrastructure, understand what changes are required, and know if their data will be useful enough for their model. A lack of preparation can lead to significant loss of time and money. The organizations that got burned tend to say the same thing:

> "You really have to be specific with what you want to do with AI. Everybody says let's just get AI, and we can do this or that. They bring in vendors and they start doing all kinds of things and spending money, but it's not doing anything for the company."

— C-level executive in construction/ engineering with over 10,000 employees

> "We've done cost analysis, and we realized that the cost advantage of moving AI workloads to the cloud is not that significant in the long run."

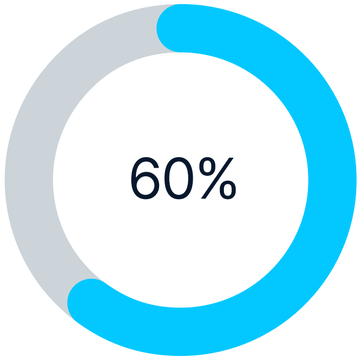
— C-level executive in manufacturing with over 10,000 employees

Back to Contents

The rebalance

Most organizations regret their initial AI infrastructure decisions. This is what happened.

Running AI across hybrid environments is harder than it looks. IT teams are juggling cost, performance, data location, privacy, compliance, and sovereignty all at the same time, across cloud, data center, and edge environments. Most organizations got the balance wrong the first time. The evidence is in the numbers. Nearly two-thirds of organizations (65%) agree that the actual cost of AI has been significant. Organizations desire to retain high-bandwidth, proprietary workloads on-premises within private cloud environments to maintain total control over spending and provide a more predictable cost structure compared to variable public cloud costs.



of organizations have already moved AI workloads from the cloud back to on-premises, which is indicative of lessons learned from the deployment decisions of early AI workloads.

Rethinking the deployment mix

The cost advantages of on-premises deployment are most clear for training or fine-tuning, while, for AI inference, the economics are closer to parity. As a result, organizations need to be more nuanced in decision making, looking beyond just upfront and ongoing costs and including other criteria for workload placement.

For training AI models, large enterprise organizations (5,000 employees or more) indicate, on average, that on-premises infrastructure is approximately 12% less expensive than public cloud infrastructure, and midmarket organizations (500–4,999 employees) indicate ~8% savings.

For fine tuning AI models, large enterprise organizations indicate, on average, that on-premises infrastructure is approximately 7% less expensive than public cloud infrastructure, and midmarket organizations indicate 2–3% savings.

For AI inference, large enterprise organizations indicate, on average, that on-premises infrastructure is approximately 5% less expensive than public cloud infrastructure, and midmarket organizations indicate on-premises infrastructure is ~2% more expensive.

Regarding data gravity and egress costs, 69% of organizations report that the cost of moving data (egress and data movement) is directly impacting their AI workload placement, forcing compute to move where the data already lives.

Back to Contents

Securing AI workloads

The security factor is not to be overlooked. Given that AI systems regularly operate on proprietary data and have their own specific vulnerabilities and risks, the ability to establish and enforce protections around AI infrastructure is an essential pillar of planning and architecture. Some of the most pronounced evidence of security's direct influence over AI workload placement includes the following:

82% of organizations agree that their AI infrastructure needs to keep pace with security threats and regulatory changes.

82% agree that controlling who can access their AI systems and data is a significant factor in data center infrastructure decisions.

60% agree that sending sensitive data to the cloud poses significant security risks for their organization.

Data sovereignty drives AI infrastructure decisions

AI commonly runs on private, sensitive data, and that data cannot always go just anywhere. Most organizations feel this constraint directly: 76% say data sovereignty shapes their infrastructure decisions, and 70% say regulations dictate where their data can be stored. For many sensitive workloads, on-premises is not the preferred option. It is the only viable one.

When it comes to meeting data regulatory and residency requirements, respondents indicated that on-premises deployments are more likely to be advantageous overall (49% versus 42% for the public cloud). In addition, organizations' need to use air-gapped infrastructure to meet data sovereignty requirements reinforces the importance of on-premises infrastructure as AI environments expand to incorporate additional use cases and datasets.

76% of organizations agree that data sovereignty requirements influence their AI infrastructure decisions.

70% of organizations agree that regulations dictate where they can store data.

64% of organizations identify the need to use air-gapped infrastructure to meet data sovereignty requirements.

These findings suggest the combined cost, security, and compliance advantages strengthen the motivations to deploy private or custom AI workloads on-premises.

Insights from decision makers on AI deployment

> When you're talking specifically about the engineering space, we have [AI workloads] on-premises because they are proprietary… and we want to keep them close to us…[For] proprietary workloads in training and post-production inference, we need to have high bandwidth and to maintain a manageable cost structure. We would put those high-bandwidth workloads on-premises in the private cloud. We have total control of the spend."

— Senior IT manager in manufacturing with over 150,000 employees

Back to Contents

The pending capacity crunch

The clock is running: Half of organizations expect a server bottleneck within 12 months

As AI usage increases and the business expands into new AI use cases, on-premises investments must keep pace. Demands to scale compute for AI have led 52% of organizations to expect server infrastructure bottlenecks within 12 months, reinforcing the need for robust, internal data center capacity. Teams must prioritize accelerating the deployment of new server infrastructure, while also helping to ensure that every new investment is optimized and fully utilized to maximize the return on investment it provides to the organization.

Over half of organizations expect their server environment to become a bottleneck for AI in a year or less

52%

It is already a bottleneck or will be within the next year

31%

Within the next one to two years

16%

More than 2 years or don’t expect it to become a bottleneck

Server infrastructure

Back to Contents

The network is the next bottleneck—and two-thirds of organizations already see it coming

Scale a hybrid AI environment fast enough, and the network becomes the thing that breaks it. Two-thirds of organizations (67%) expect their network to hit its limits as an AI bottleneck within 12 months. And the pressure is only going to intensify, as edge computing and deskside AI are creating exponentially more endpoints, all demanding reliable, high-bandwidth connections. Research indicates that 41% of organizations have already deployed edge inferencing workloads, with an additional 56% currently in the process or planning to deploy edge inferencing workloads within the next 12 months. These distributed AI deployments mean organizations not only have to move massive datasets across their networks but also ensure consistent performance across hundreds or thousands of distributed computing nodes simultaneously.

Two-thirds of organizations expect their network infrastructure to become a bottleneck for AI in a year or less

67%

It is already a bottleneck or will be within the next year

21%

Within the next one to two years

12%

More than 2 years or don’t expect it to become a bottleneck

Network infrastructure

Back to Contents

The edge is leading the way

As the usage of AI expands with the influx of new use cases, edge infrastructure provides a strategic opportunity by enabling AI inferencing workloads to be deployed closer to business operations, such as near the factory floor for manufacturing or closer to the customer at retail locations. To capture the business potential offered by AI at the edge, 97% of organizations are deploying or expect to deploy edge inferencing workloads within the next 12 months, representing a 2.4x increase from current numbers.

97%

of organizations are deploying or expect to deploy edge inferencing workloads within the next 12 months, representing a 2.4x increase from current numbers.

Consistent deployment of AI inferencing at the edge over the next 12 months

41%
Already deployed and operational

43%
Currently in the process of deployment

13%
Planning to deploy within the next 12 months

2%
Exploring the concept but no concrete plans yet

0%
No plans to deploy edge inferencing

Back to Contents

The on-premises spending wave is here

Maximizing compute and GPU utilization is essential to optimizing AI economics. Achieving that goal requires investments across the entire data lifecycle and pipeline, especially the network, to ensure it does not become the bottleneck that impedes returns on overall AI technology and infrastructure investments. To this end, as organizations scale server investments, networking investments must scale accordingly.

98%
of organizations anticipate an increase in network infrastructure spending, with 38% planning to increase spending by 50% or more.

95%
of organizations anticipate an increase in server infrastructure spending, with 40% planning to increase spending by 50% or more.

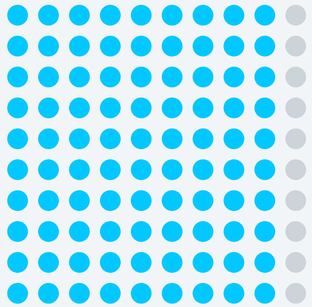
The vast majority expect to scale server and networking investments for AI in the next 12 months

	Network infrastructure	Server infrastructure
Increase more than 75%	6%	3%
Increase 61% to 75%	12%	12%
Increase 50% to 60%	20%	25%
Increase 26% to 49%	25%	26%
Increase 10% to 25%	27%	21%
Increase less than 10%	8%	8%
Stay the same	2%	4%
Expected to decrease	0%	1%

Back to Contents

Building from scratch fades as integrated solutions grow

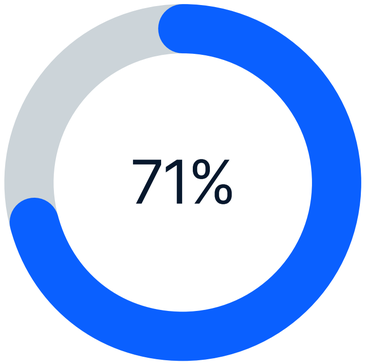
To help accelerate the necessary surge in on-premises server and networking capacity and overcome bottlenecks, integrated AI solutions, such as AI factories, have risen in popularity as a viable option to simplify the procurement, deployment, and management of AI infrastructure. This integration reduces operational risk and accelerates deployment via pre-validated, optimized, and supported solutions.



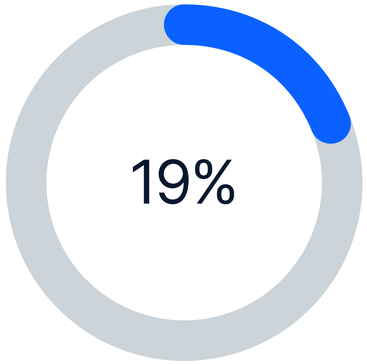
90%

of organizations are using or planning to use integrated AI solutions for managing on-premises workloads in the next 12 months, proving that these options are quickly becoming accepted alternatives to cloud and self-built on-premises approaches.

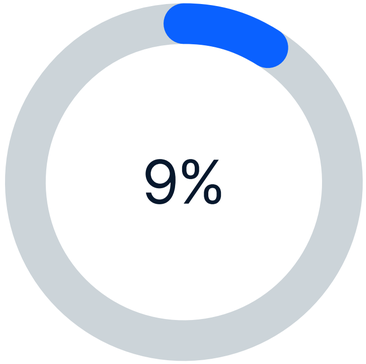
Nine in ten organizations use or plan to use integrated AI solutions on-premises



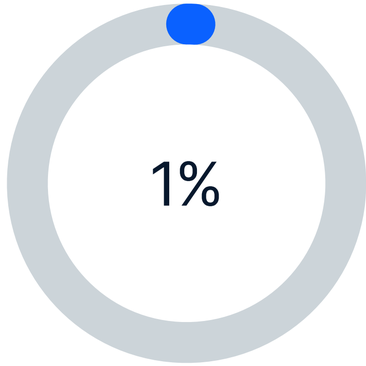
Yes, we currently use integrated AI solutions



No, but we plan to implement integrated AI solutions within the next 12 months



No, but we are in the early stages of exploring integrated AI solutions



No, and we do not plan to implement integrated AI solutions

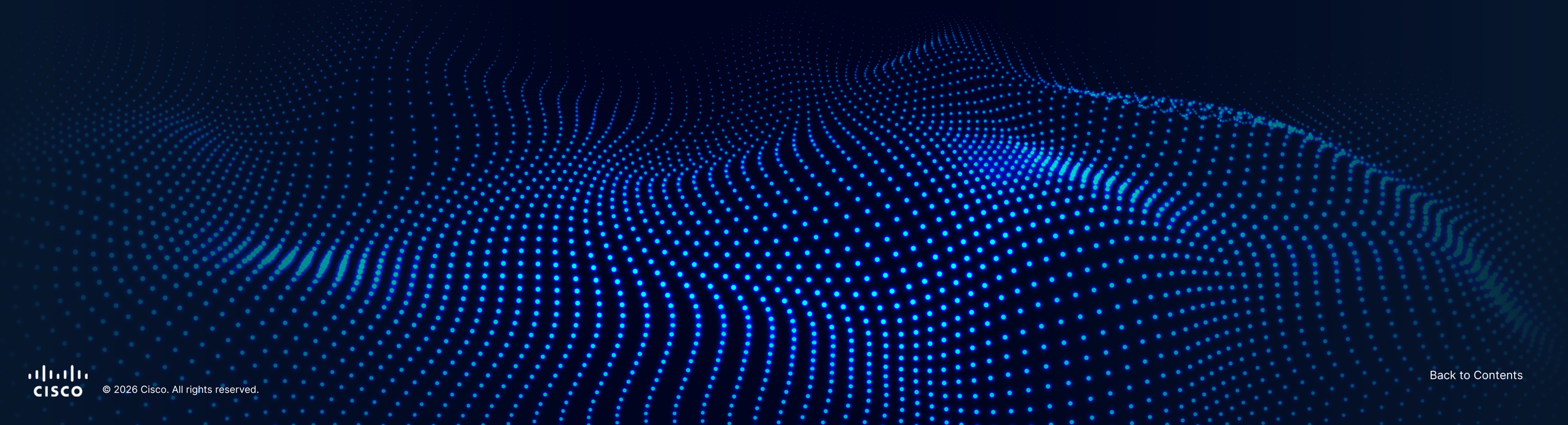
Back to Contents

Getting AI infrastructure right

Investment in AI infrastructure is set to increase across both cloud and on-premises environments, with our research predicting a slight shift toward the cloud. But that shift is not the whole story, as the data shows 60% of organizations have moved workloads from the cloud back to on-premises. Long-term deployment models will be use-case dependent. Priorities for each specific training, fine-tuning, or inferencing workload will ultimately determine where it is best hosted.

For on-premises investments, the data shows the trajectory is clear: The rise of both data center and edge use cases will be constrained by the network. As organizations scale server capacity to keep pace with AI demand, the network risks becoming the bottleneck that limits returns—unless it scales in parallel.

Getting AI infrastructure right is not just about location decisions; it is a workload placement discipline. But this means that, even as AI matures, the complexity challenge will not recede. Flexibility is not a nice-to-have. It is the whole game. And real flexibility only works with a network that is distributed, modern, and built for AI from the ground up.



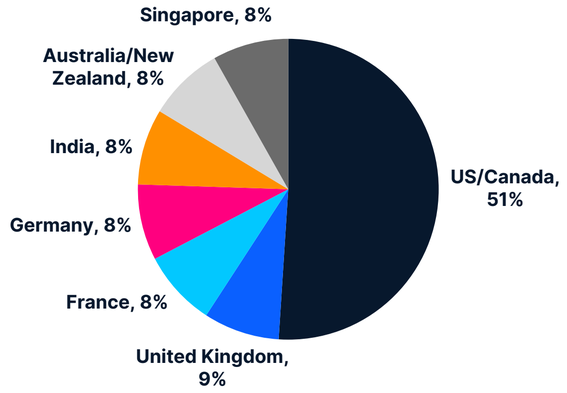
Back to Contents

Research methodology and respondent demographics

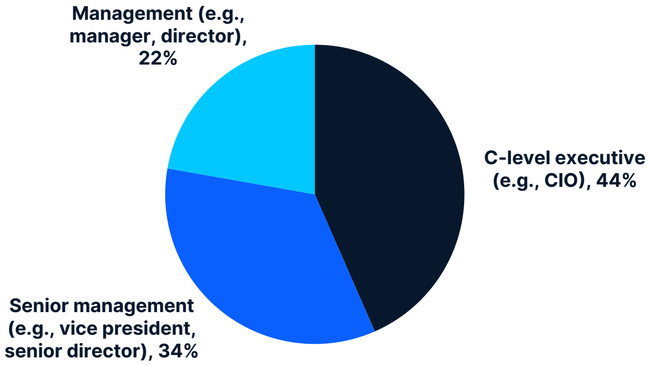
Cisco commissioned Omdia to conduct a comprehensive online survey fielded between March 27, 2026, and April 15, 2026. The research included 1,201 qualified respondents from midmarket organizations with 500 to 4,999 employees (51%) and enterprise organizations with 5,000-plus employees (49%) across North America (50%), EMEA (25%), and APAC (25%). Respondents are IT decision makers with purchasing authority for data center infrastructure (compute, storage, networking) and have influence over hosting decisions for AI workloads (i.e., cloud versus on-premises). The survey also included five in-depth interviews with key decision makers regarding their AI workload placement strategy. All respondents in this survey have AI inferencing that uses custom-trained or fine-tuned models.

The margin of error is +/- 4 percentage points at the 95% confidence level. Note: Totals in figures and tables throughout this eBook may not add up to 100% due to rounding or organizations choosing more than one answer to select questions.

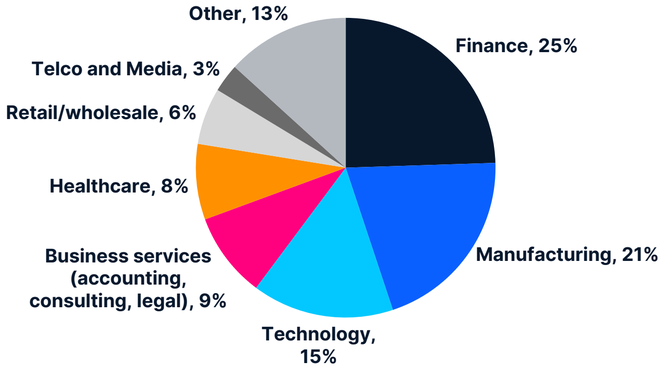
Respondents by region
(Percent of respondents, n = 1,201)



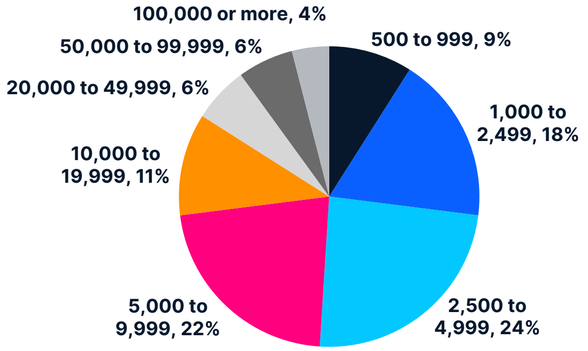
Which of the following best describes your current job title/level?
(Percent of respondents, n = 1,201)



What is your organization's primary industry?
(Percent of respondents, n = 1,201)



How many total employees does your organization have worldwide?
(Percent of respondents, n = 1,201)



Back to Contents

About Cisco

Cisco is the critical infrastructure for the AI era. By uniquely combining the power of the network with security, observability, and collaboration, Cisco powers AI-ready data centers, future-proofed workplaces, and digital resilience.

Learn more about how Cisco partners with organizations to build the secure, high-throughput, low-latency infrastructure that AI demands, from core to edge.

Learn More