Private Cloud Outlook 2026

The AI Tipping Point

vmware by Broadcom

Contents

As We See It:
AI Is Rewriting the Cloud Playbook

In 2026, enterprise AI has moved beyond pilots and into production workflows. That transition is changing how organizations think about cloud economics, infrastructure architecture, security, and day-to-day IT operations.

For many enterprises, the first wave of AI experimentation happened in public cloud. That made sense. Public cloud gave teams fast access to tools, capacity, and services when use cases were still being tested. But AI workloads in production are different. They introduce sustained compute demand, sensitive data flows, new governance requirements, and performance expectations that can quickly expose the limits of a purely public cloud approach.

A year ago Broadcom's inaugural Private Cloud Outlook 2025 report captured a "cloud reset" taking place among enterprise IT organizations. The report uncovered a growing preference for private cloud, driven in part by its advantages in security and cost predictability.

In this second annual report, we examine how enterprise IT cloud strategy has evolved over the past year—and the growing value of private cloud in the AI era. Like its predecessor, the Private Cloud Outlook 2026 draws on a global survey of 1,800 senior IT decision-makers across the Americas, Europe, and Asia-Pacific.

The findings point to four developments that we believe will define enterprise IT strategy in the years ahead:

Private cloud is the preferred platform for production AI.

Geopolitics are reshaping enterprise IT.

The public cloud cost reckoning has arrived.

AI complexity is widening the IT skills gap.

vmware by Broadcom

Private cloud is the preferred platform for production AI.

As AI moves from experimentation to production, enterprises need infrastructure with predictable performance, tighter governance, stronger data control, and more manageable economics. Private cloud is increasingly where those requirements come together.

Fifty-six percent of organizations surveyed are running or planning to run production inferencing in a private cloud.

Public cloud use for production inference fell 15 percentage points year over year to 41%.

Our prediction:
The three Cs—cost, complexity, and control—will make private cloud the default platform for production AI.

Geopolitics are reshaping enterprise IT.

Cloud strategy must consider where data resides, who can access it, and which legal or regulatory regimes apply. **Eighty-six percent of IT leaders say geopolitical and regulatory factors are affecting their IT strategy and operations.** Data sovereignty and residency requirements have become the top geopolitical concern, cited by 54% of respondents.

Our prediction:
Data sovereignty will rank alongside cybersecurity as a board-level infrastructure priority.

The public cloud cost reckoning has arrived.

The economics of public cloud become more difficult to manage as workloads scale, and AI intensifies that challenge. Sustained inference, GPU demand, data movement, and bandwidth charges are making cost predictability a central factor in workload placement.

Cost has overtaken security as the top public cloud concern, and **97% of IT leaders believe some portion of their public cloud spend is wasted.**

Our prediction:
AI workloads will expose public cloud cost structures as unsustainable for many production-scale deployments.

AI complexity is widening the IT skills gap.

AI infrastructure, cloud security operations, and Kubernetes operations now require specialized skills that many enterprises cannot hire fast enough. Professional services can help close the near-term gap, but the long-term answer also depends on reducing operational complexity. **Eighty-one percent of enterprises now either fully outsource or use professional services for cloud-related needs.**

Our prediction:
Enterprises that simplify their platform stack will close the AI skills gap faster.

What’s Changed in 2026

Private cloud is now the preferred platform for production AI.

The 2026 findings show a decisive change in enterprise AI deployment: private cloud has moved to the center of production strategy.

In 2025, many organizations were still testing generative AI use cases, evaluating infrastructure options, and weighing where AI workloads should run. One year later, the experimentation phase is giving way to production reality. The question has become more practical: where can enterprises run AI reliably, securely, and economically at production scale?

Increasingly, the answer is private cloud.

More than half of enterprises surveyed, 56%, are now running or planning to run AI inference on private cloud. Over the same period, public cloud usage for these workloads dropped sharply, falling 15 percentage points in a single year, from 56% to 41%.

That change marks an important turning point. Public cloud remains a critical part of enterprise IT strategy, especially for experimentation, elastic capacity, and access to specialized services.

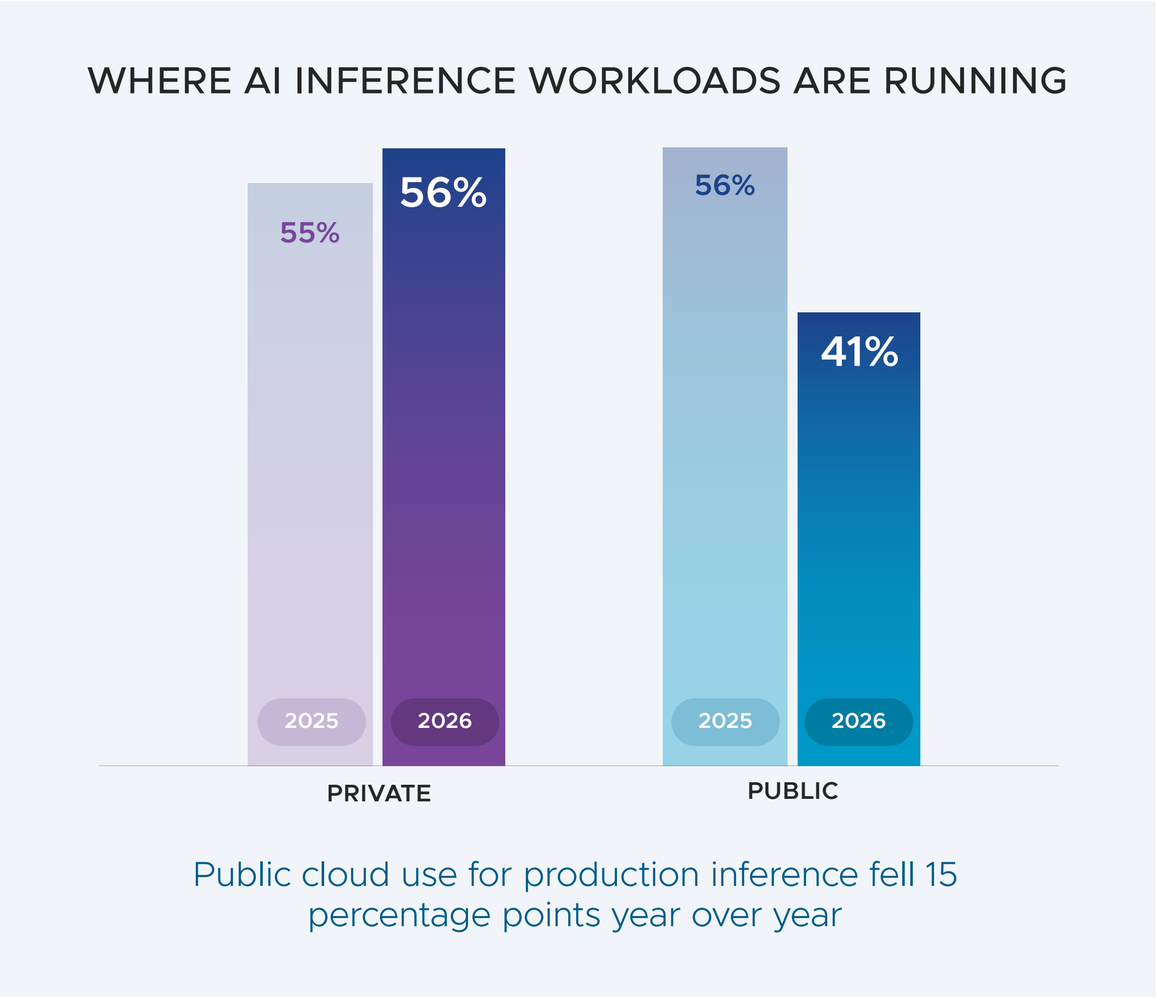


Figure 1: Where AI-based applications or workloads currently run n=1800

But as AI moves closer to production, organizations are putting greater weight on the operational requirements that private cloud is designed to address: cost predictability, data control, security, and performance.

The investment data points in the same direction. Enterprises are increasing their cloud investment overall, and more organizations plan to expand both private and public cloud spend over the next three years. But private cloud momentum is rising faster.

As inference becomes embedded in applications, workflows, and business processes, it requires predictable performance and tighter governance than many organizations can achieve through a purely public cloud model.

Key findings include:

Intent to increase private cloud spend over three years rose from **51% to 72%.**

Priority to build new private cloud workloads climbed from **53% to 58%.**

Private cloud investment growth is expanding at **more than twice the rate of public cloud growth.**

Private cloud leads for high-stakes workloads.

The move toward private cloud is not limited to AI inference. The 2026 findings show a broader pattern in how enterprises are placing their most demanding workloads.

In 2025, cost and security concerns were the top drivers for repatriating workloads from public cloud to private cloud. Those factors still matter. But in 2026, performance and business criticality are also taking a more prominent role, especially as AI workloads become more closely tied to core applications, sensitive data, and real-time decision-making.

For many high-stakes workloads, private cloud leads by a wide margin:

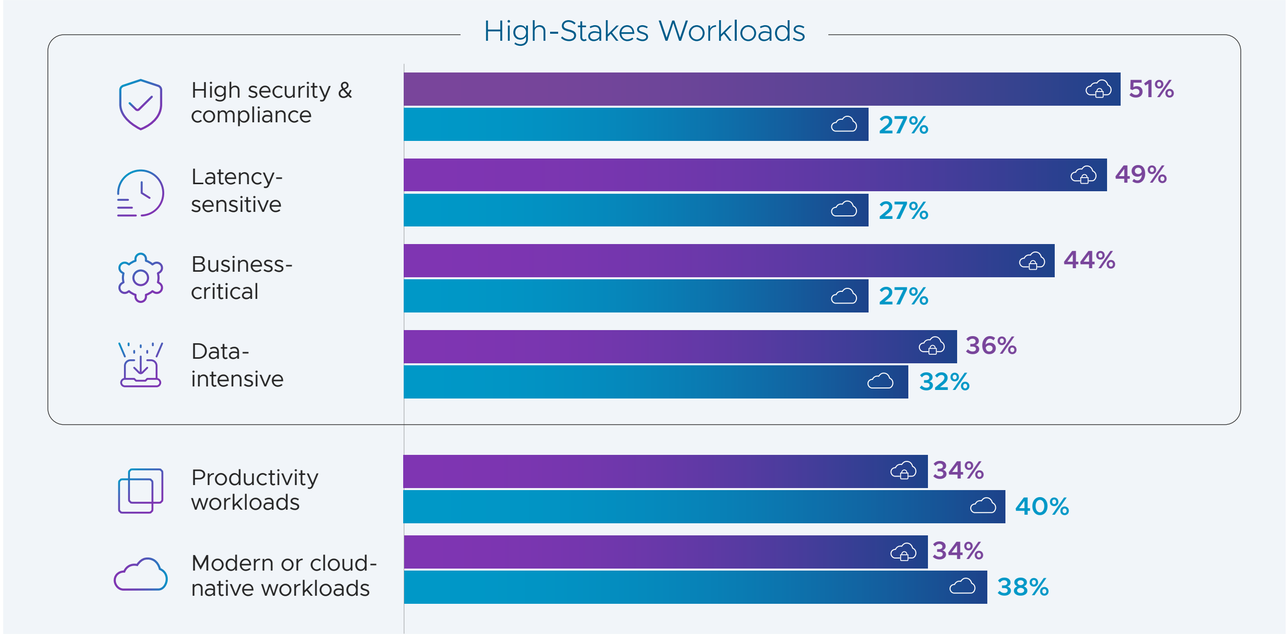


Figure 2: Preferred cloud environment for workload types n=1800

Figure 3: How workload placement is managed across its mix of public and private cloud n=1736

The divide is widest where risk is highest.
When workloads carry higher requirements for security, compliance, latency, or business continuity, enterprises are more likely to choose private cloud. Public cloud remains competitive for data-intensive workloads, but private cloud still has an edge.

Most enterprise IT organizations are approaching these decisions carefully. Seventy percent now have either a central team or standardized policies, tooling, or automation to manage workload placement. That suggests a more disciplined approach to cloud strategy: one based less on default assumptions and more on the specific requirements of each workload.

vmware by Broadcom

Repatriation is accelerating.

Repatriation from public cloud to private cloud continues its trajectory. In 2025, more than two-thirds of enterprises were considering repatriating workloads, and 35% had already done so. In 2026, those figures moved sharply higher.

Now, 50% of enterprises have already repatriated some workloads—a 15-point jump in one year—and 33% percent are considering repatriation.

The reasons behind that acceleration tell an important story:

- **Security and compliance** remain the top driver, cited by 51% of organizations
- **Cost predictability** has become the second biggest driver, cited by 39% of organizations
- **Performance** is now a top-three driver, also cited by 39% of organizations

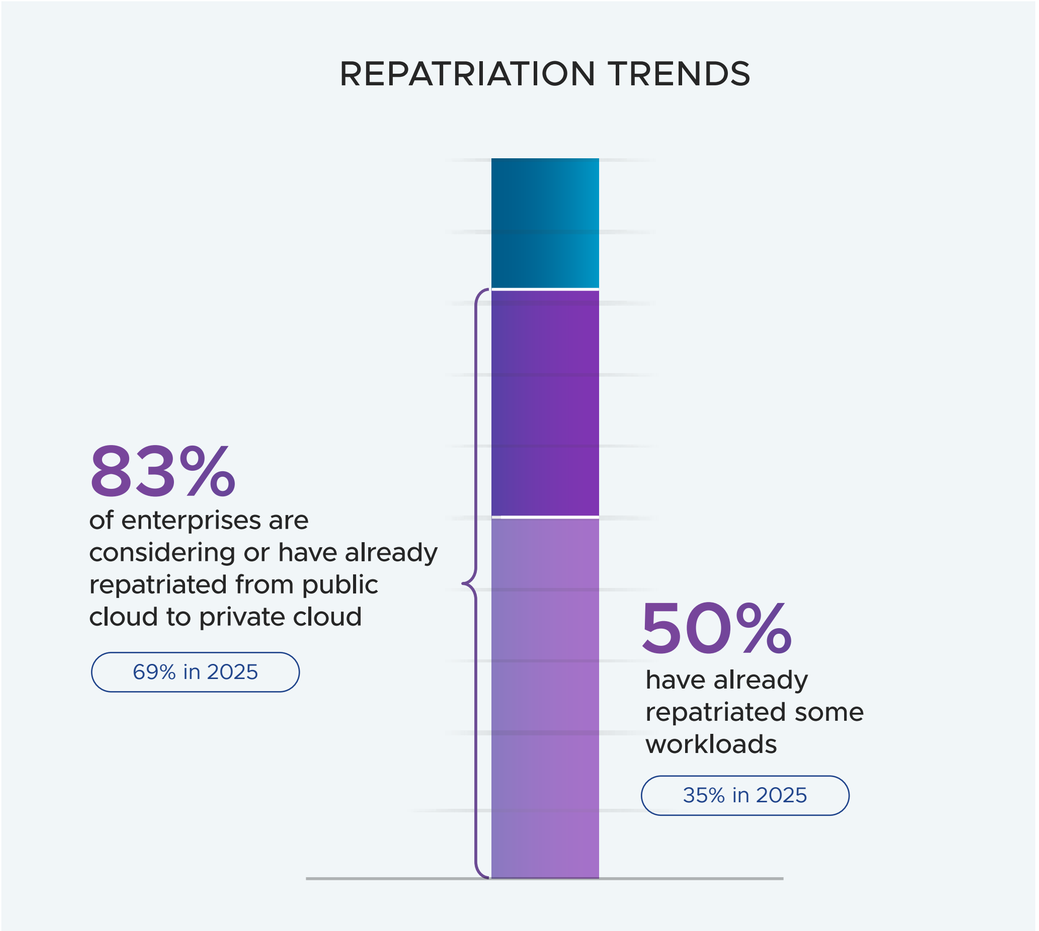


Figure 4: Have you repatriated any workloads from public cloud to private n=1800

vmware by Broadcom

What types of workloads are moving?

The workloads being repatriated are exactly the kinds of workloads where control matters most: those with high security or compliance requirements, data-intensive workloads, and business-critical applications.

AI also appears as a distinct repatriation category for the first time. Forty-three percent of organizations repatriating workloads are moving AI training, LLMs, and inference from public cloud to private cloud—a category that did not exist in the 2025 study.

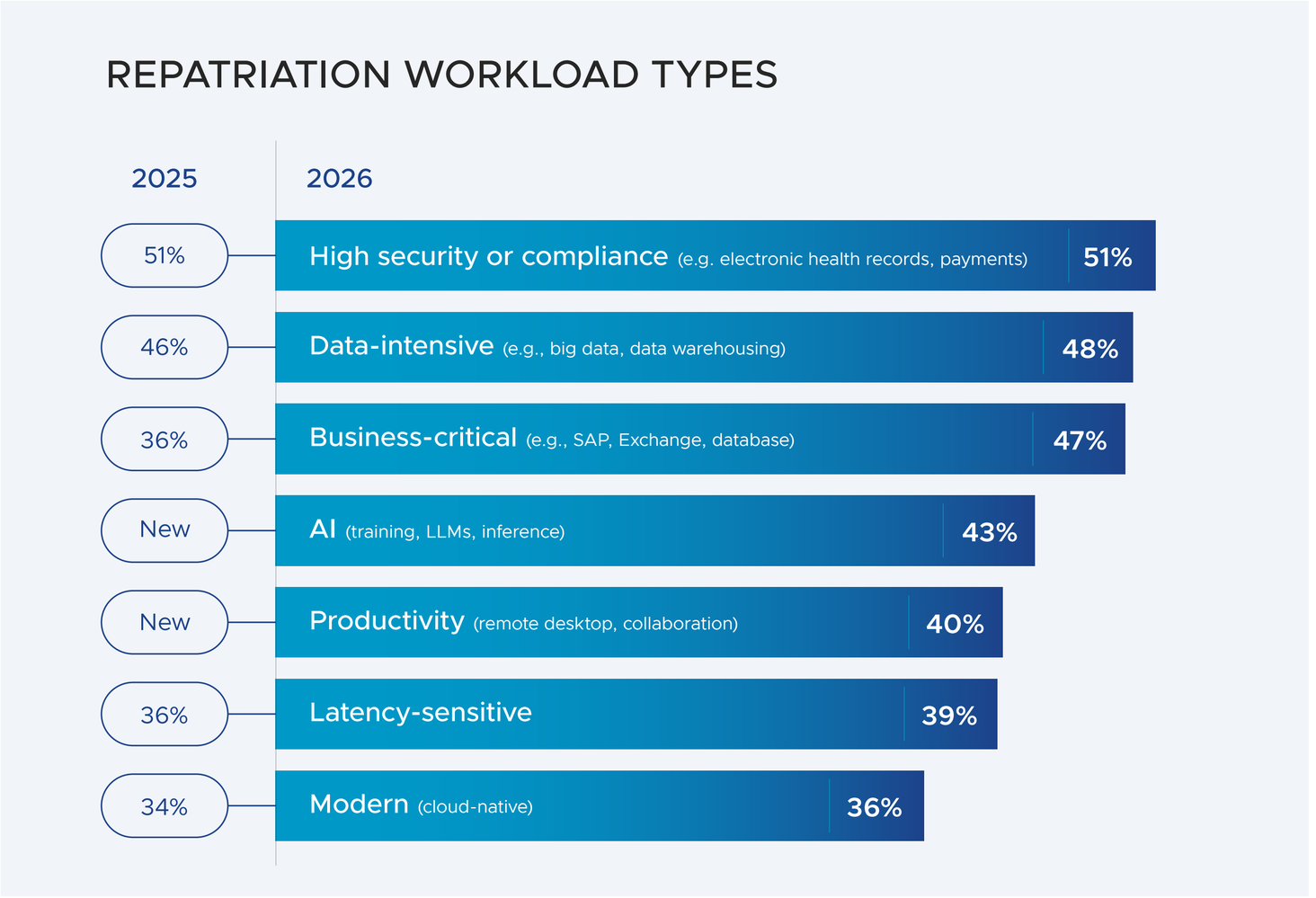


Figure 5: Types of workloads organizations repatriated/considered repatriating from public to private n-1491

The Catalysts for Change

Enterprises that spent the last two years exploring AI in public cloud are now confronting the hard reality of AI at production scale. Cost predictability, security, governance, performance, and operational control all matter more when AI is embedded in business-critical applications and workflows. In this environment, AI is amplifying the three Cs of infrastructure decision-making: **cost, complexity, and control.** Increasingly, private cloud is where enterprises see the clearest path to managing all three.

vmware by Broadcom

The Public Cloud Cost Reckoning

For the first time, cost has overtaken security as the top public cloud concern. In 2025, 26% of respondents cited cost management as a leading challenge. In 2026, that figure rose to 31%, making cost the number-one issue enterprises face with public cloud.

Nearly all IT leaders surveyed, 97%, believe some portion of their public cloud spend is wasted. More than half, 52%, say that waste exceeds 25%. Put plainly, many enterprises believe a quarter or more of their public cloud budget is delivering no business value.

AI is intensifying the problem. The workloads consume more compute, storage, and bandwidth, while GPU pricing, data movement fees, and unpredictable usage patterns make costs harder to forecast. **Sixty-two percent of IT leaders say they are very or extremely concerned about generative and agentic AI infrastructure costs.**

The issue is not simply that public cloud is expensive. It is that the relationship between cost and value is becoming harder to manage.

This tension helps explain why cost predictability has become a leading reason for repatriation. Enterprises that have run workloads in public cloud at scale are encountering costs that are difficult to forecast and harder to control. As AI inference adds new layers of GPU and compute demand, the financial case for private cloud has become considerably stronger.

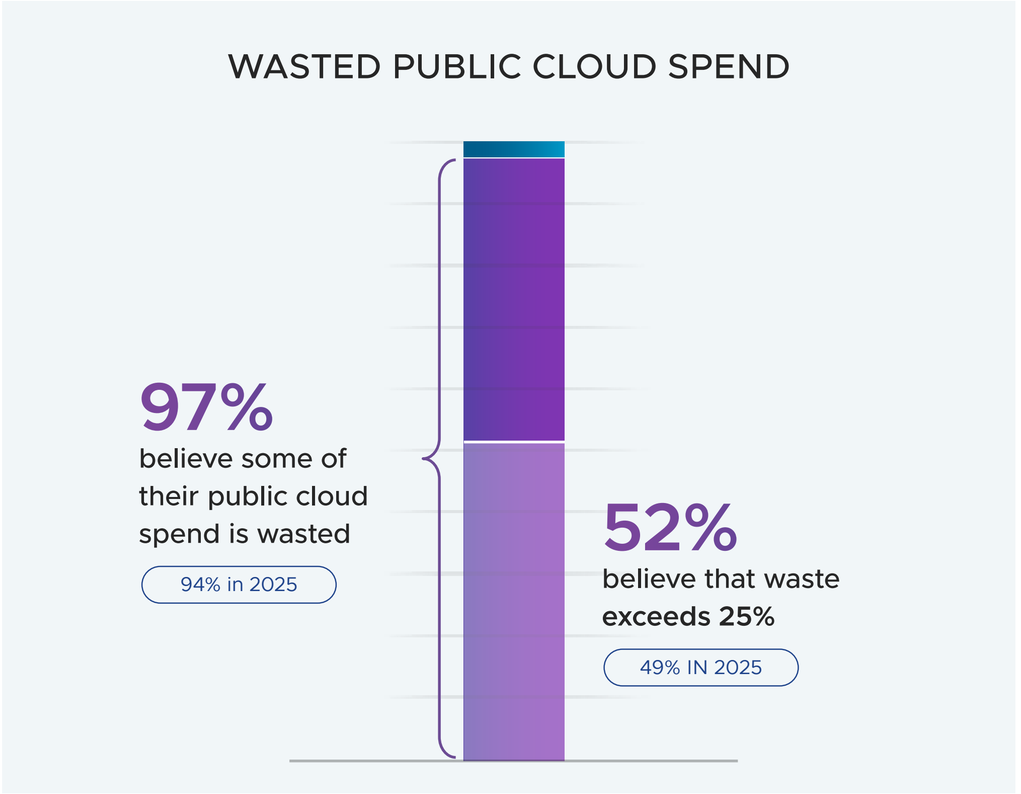


Figure 6: How much of your public cloud spend is believed to be “wasted”? n=1736

> Actual costs of cloud-based AI infrastructure were higher due to hidden fees and bandwidth charges.
>
> – C-Level Executive, Energy and Utilities, Australia

vmware by Broadcom

The Non-Negotiables: Security, Sovereignty, and Compliance

Cost may be rising fast as a public cloud concern, but security remains the key determinant in where critical workloads are placed. When asked to choose a single factor for workload placement, 32% of respondents chose security and compliance—ahead of other considerations including cost, performance, and scalability.

AI raises the stakes. The biggest new requirements introduced by AI are data protection and privacy, cited by 37% of respondents, followed closely by security and control at 36%. These requirements go directly to the strengths enterprises already associate with private cloud: greater oversight of sensitive data and stronger control over access and policy.

This strengthens the case for private cloud as a production AI platform. AI systems are often connected to proprietary data, regulated data, customer records, operational systems, and business-critical workflows. This makes risk management a key factor in decisions about workload placement.

The calculus is also being shaped by forces outside the data center. Four out of five IT leaders say geopolitics and regulatory factors are now affecting their IT strategy and operations. For global enterprises, workload placement is often tied to where data resides, who can access it, and which legal or regulatory regimes apply.

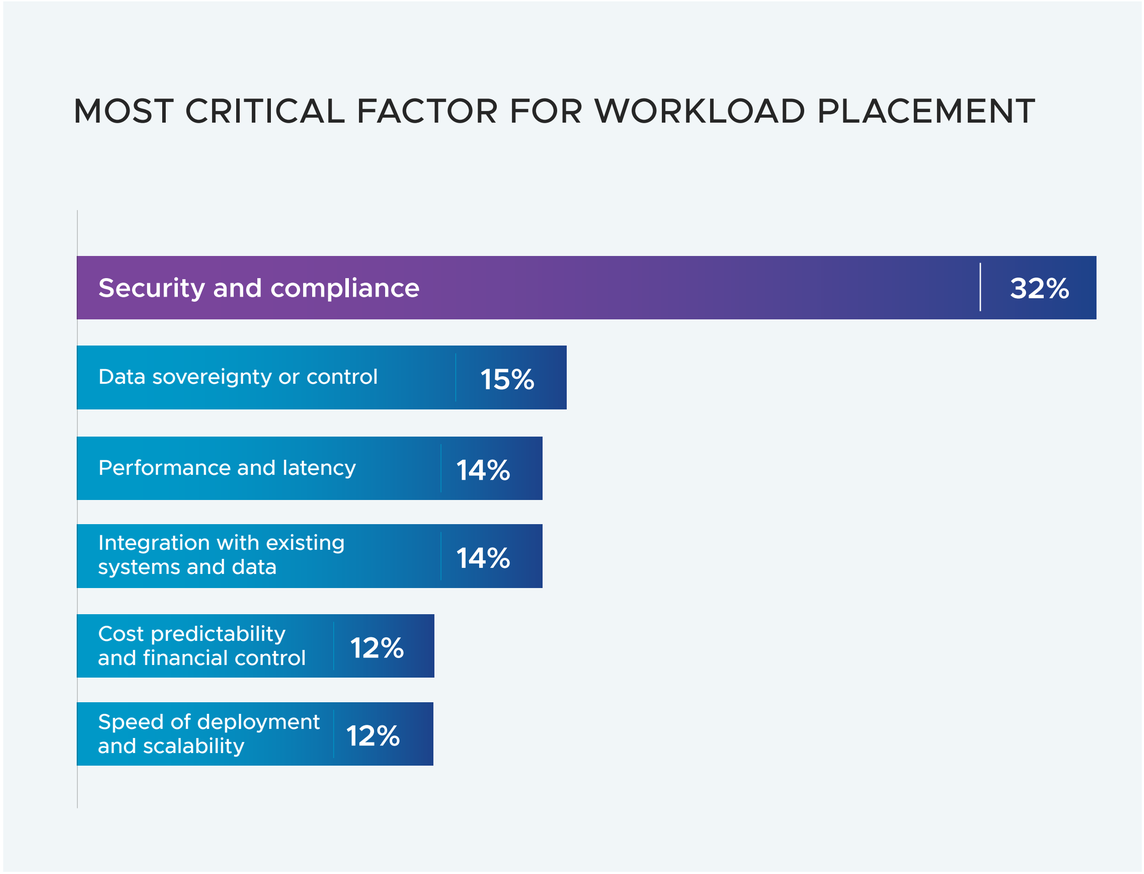


Figure 7: Mission-critical workload placement decisions involve trade-offs n=1800

Figure 8: Extent geo-political or regulatory factors affecting IT strategy n=1800

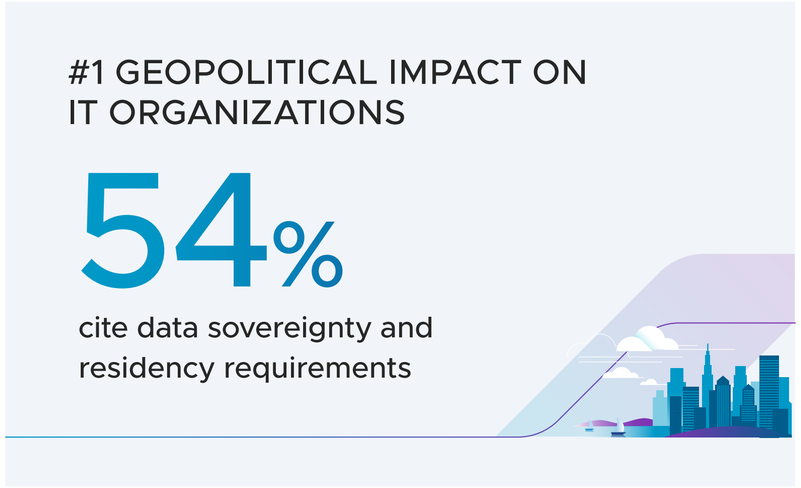


Figure 9: Geo-political or regulatory factors impacting IT decisions today n=1544

In fact, data sovereignty has overtaken traditional compliance as the top geopolitical factor influencing IT strategy:

54% cite data sovereignty and residency requirements.

51% cite jurisdiction-specific compliance requirements.

This distinction matters. Compliance is about meeting rules. Sovereignty is about control: where data lives, how it is governed, and how much authority an organization retains over its digital operations. For enterprises navigating AI, regulation, and geopolitical uncertainty at the same time, private cloud offers a more direct path to maintaining that control.

> I thought compliance would be a minor checklist item, but governance requirements reshaped our entire architecture.
>
> – C-Level Suite Executive, Healthcare, United States

The IT Skills Gap

In addition to changing infrastructure requirements, AI is exposing a gap in the skills required to operate that infrastructure at scale.

The top skills gap cited by enterprise IT leaders is AI infrastructure and operations, named by 40% of respondents. Close behind are cloud security operations at 38% and Kubernetes operations at 37%.

The challenge is that the talent market cannot fill these roles quickly enough. This helps explain why professional services remain so widely used. Eighty-one percent of enterprises now either fully outsource or use professional services for cloud-related needs. For many organizations, this is a rational response to a market where demand for specialized cloud and AI skills is growing faster than supply.

But services alone are not a long-term operating model. The skills gap also underscores the importance of a modern private cloud platform. A private cloud that unifies infrastructure, security, and Kubernetes operations on a single platform can reduce the breadth of skills IT teams need to run complex environments. It can also help organizations standardize operations, apply policy consistently, and scale expertise.

As AI workloads become part of everyday enterprise operations, organizations will need more than additional specialists. They will need platforms that help existing teams operate advanced infrastructure more consistently and efficiently.

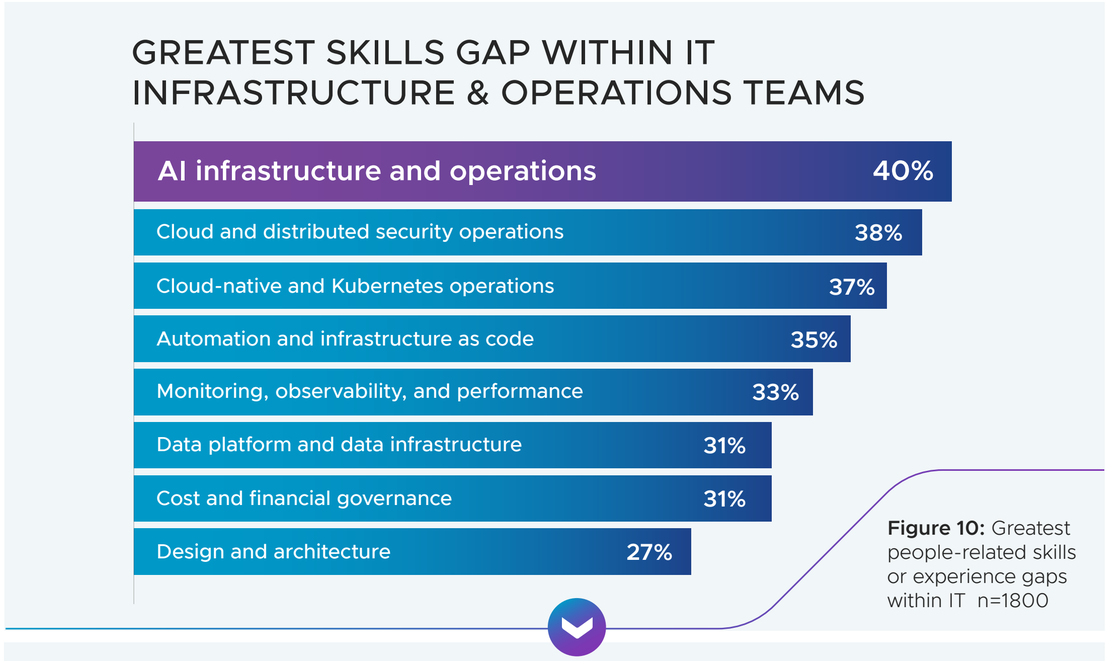


Figure 10: Greatest people-related skills or experience gaps within IT n=1800

81% of enterprises fully outsource or use professional services for cloud-related needs.

Figure 11: Organizations using external professional services n=1800

> The complexity of the AI tools exceeded our team's expertise, requiring extensive training and external consulting.
>
> – VP, Financial Services, Japan

vmware by Broadcom

Recommendations for IT Leaders

AI at scale raises the stakes for every infrastructure decision. The findings in this report show that enterprises are already adjusting: moving more AI workloads to private cloud, scrutinizing public cloud costs, and putting greater emphasis on governance, sovereignty, and operational control.

The path forward starts with three priorities: place AI workloads where control matters most, regain command of cloud economics, and build governance into the infrastructure from the start.

Lead with AI on private cloud.

Reclaim cloud economics.

Govern AI before it governs you.

vmware by Broadcom

Lead with AI on private cloud.

Prioritize private cloud for AI workloads, especially those tied to sensitive data, regulated environments, latency-sensitive use cases, or business-critical applications. Private cloud offers more control over cost, performance, and data.

IT leaders should also reduce unnecessary platform sprawl. Many organizations now run a mix of virtual machines, containers, Kubernetes environments, AI tooling, and LLM platforms across multiple clouds and operating models. That fragmentation increases cost and complexity while stretching already limited teams.

A more disciplined approach means:

- Unifying AI, container, and VM operations where possible
- Reducing redundant tooling with a unified private cloud
- Aligning workload placement with security, performance, cost, and governance requirements

Reclaim cloud economics.

AI changes the cost equation. Inference drives up demand for compute, storage, networking, and GPU capacity, while data movement and unpredictable usage patterns can make public cloud costs harder to forecast.

IT leaders should take an active role in addressing public cloud waste, improving visibility into actual workload costs, and identifying where private cloud can provide better predictability and efficiency.

Key actions include:

- Measuring public cloud waste and identifying workloads that no longer justify their current placement
- Modeling AI infrastructure costs across public and private cloud scenarios
- Planning for repatriation of public cloud AI workloads where cost, governance, or performance requirements have changed

Govern AI before it governs you.

AI governance must move earlier in the infrastructure decision process. Organizations need to know where data is moving, where models are running, who has access, and which policies apply.

This becomes even more important as agentic AI emerges. AI agents will take actions, trigger workflows, access systems, and operate across applications.

IT leaders should prepare now by:

- Auditing cross-border data flows
- Repatriating regulated or sovereignty-sensitive workloads where needed
- Reviewing workload placement against regulatory requirements
- Establishing governance policies for AI models, agents, data access, and outputs
- Monitoring, controlling, and auditing AI systems across environments

Modern Private Cloud for the Workloads That Matter Most

The findings in this report underscore the pressures that every IT leader is feeling at once: rising AI demand, growing geopolitical and regulatory scrutiny, and tighter infrastructure budgets.

VMware Cloud Foundation (VCF) is engineered to address these market pressures. It provides a consistent private cloud platform to run enterprise AI at the lowest cost per workload, under enterprise sovereignty, without compromising security or performance.

VMware Cloud Foundation (VCF) can help IT leaders:

Manage cloud costs by increasing workload density, reducing hardware footprint, and getting more value from existing CPUs, GPUs, and storage.

Maximize advanced compute by pooling high-value accelerators and allocating them based on business priorities.

Strengthen governance and resilience by giving teams more control over where workloads and data run, with integrated security, networking, protection, and recovery capabilities.

Simplify operations across environments by unifying VMs, containers, Kubernetes operations, and AI workloads on a single platform and consistent operating model.

AI has brought enterprise cloud strategy to a decision point. Organizations can keep adding tools, capacity, and people, or they can optimize their infrastructure to run AI at scale.

The 2026 findings make a strong case for the second path. Enterprises that can control cost, protect data, meet sovereignty requirements, and operate efficiently will be better positioned to turn AI ambition into business value.

Appendix

Audience profile

1,800 senior IT decision-makers worldwide across small, medium, and large enterprises, with a majority of participants (69%) from large enterprises of 5,000+ employees

600 participants each from North America, Europe, and APJ

- 400 United States
- 200 Brazil
- 200 United Kingdom
- 200 France
- 200 Germany
- 200 India
- 200 Japan
- 200 Australia

- Director level or above with direct responsibility for or influence over IT infrastructure and cloud strategy

- Responsibility spans cloud infrastructure, security, networking, platform engineering, or related disciplines

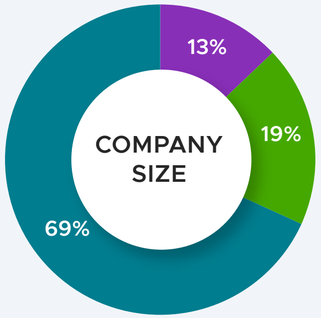
- Industries represented included life sciences/ pharmaceutical (14%), financial services (18%), public sector (18%), healthcare (20%), and others (30%)

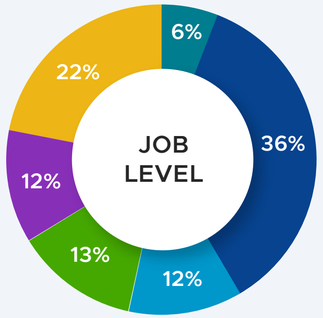
Methodology

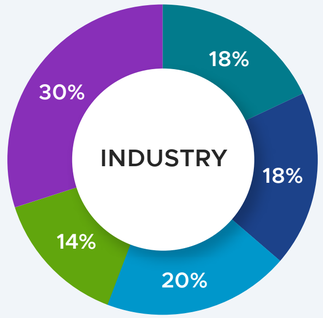
Broadcom partnered with market research firm Radius Tech to uncover insights into private cloud and how it is shaping today's cloud strategies.

Sample:
Web-based survey (20 mins)

Date Range:
February 11 – March 13, 2026







vmware by Broadcom

Appendix (continued)

Definitions Used

- **Public cloud:** shared infrastructure run by a third-party provider (e.g., AWS, Azure, Google Cloud); excludes SaaS

- **Private cloud:** dedicated infrastructure for a single organization, potentially owned or managed by the organization, a third party, or both (e.g., VMware Cloud Foundation, Red Hat OpenShift)

vmware® by Broadcom

Item No: private-cloud-outlook-2026 6/26