Analyst Connection, sponsored by:

Broadcom

Moving Enterprise AI from Pilot to Production on a Private AI Cloud

Questions posed by:

Broadcom

Answers by:

Jevin Jensen

Research Vice President, Infrastructure and Operations

Published:

November 2026

Enterprises are reconsidering the private cloud as they deploy new AI workloads. Technology leaders need to address token cost, sovereignty, performance, and agility when setting their AI deployment strategy. For many enterprises, the gap between an impressive pilot and successful production remains. The predictable total cost of ownership may be elusive. Technology leaders struggle to deliver the promised returns to the business. Enterprises are seeking a solution they can afford, trust, and govern at scale. IDC research shows that the obstacles are rarely limited to the models themselves. Concerns extend to sovereignty over sensitive data, the volatile economics of tokens and inference, and public cloud bills that grow faster than the value they deliver.

That's why IT leaders are taking a fresh look at private cloud. Most enterprises continue to invest in private cloud infrastructure. Some of the benefits include running AI closer to their proprietary data, achieving better cost predictability, and meeting regulatory requirements. This Analyst Connection examines the barriers to scaling AI and how a private AI cloud can help enterprises get back on track with costs and value delivery. The following questions were posed by Broadcom to Jevin Jensen, research vice president, Infrastructure and Operations, at IDC.

Q. IDC data shows strong enterprise AI adoption, yet many organizations struggle to move from pilot to production. What are the top 3 systemic barriers preventing scale, and why do traditional public cloud approaches often fall short for enterprise AI?

AI experimentation is no longer in question as nearly every enterprise is actively piloting AI use cases or testing AI agents today. However, when we examine why organizations hit a wall when attempting to transition these pilots into full production scale, three systemic barriers stand out:

- **Security and data sovereignty:**

  IT leaders and CTOs tell IDC that data sovereignty is their top concern. Enterprises identify data security, sovereignty, and privacy compliance as the single most significant risk in their current data platform. Some tech leaders also worry that extraterritorial legislation could compel their cloud providers to hand over data. AI raises these stakes, as model fine-tuning, retrieval-augmented generation (RAG) integration, and direct tool integrations via standards like the Model Context Protocol (MCP) grant AI agents deep access to an organization's most sensitive data.

- **Unpredictable token economics:**

  The paradox of enterprise AI is that token prices and total spend are moving in opposite directions. IDC expects the cost per agent action to fall roughly 95% between 2025 and 2030, yet total token delivery cost worldwide is on track to climb exponentially. Volume and sophistication outpace price declines. Enterprises already feel it with budgets being squeezed and missed ROIs jeopardizing other IT projects. Most tech leaders lack real-time dashboards to track token consumption. Because AI usage is increasing faster than token prices are falling, total budgets required for AI are rising.

- **Escalating public cloud costs:**

  In IDC's August 2026 *Intelligent Cloud Survey,* 89% of organizations reported unexpected AI-related costs on public cloud or SaaS after go-live. The culprits are the costs teams often overlook: variable inference demand, GPU scarcity pricing, retraining cycles, forecasts for API calls, hyperscaler networking charges, and the steady growth of storage.

  Public cloud is known for its agility and on-demand capacity. But sustained, scaled AI production affects the math. Two common surprises are consumption-based billing for continuous inference and egress charges for moving large data sets. These can erode

business margins. Enterprises increasingly need a strategy to ensure ROI and improve forecast accuracy.

Q. If public cloud isn't the only answer for production AI, what architecture should enterprises pursue to address these barriers?

Many enterprises are considering a private AI cloud, which involves running AI models, data pipelines, and inference on infrastructure they control, built on cloud operating principles. The data supports this shift. According to IDC's August 2026 *Intelligent Cloud Survey,* 91% of organizations plan to increase private cloud investment in 2027. The architecture matters as much as the location. Enterprises must continually contend with implementation complexity, fragmented tools, and poor life-cycle integration. The answer isn't a bespoke private cloud that requires significant investment and specialized skills. It's a unified, software-defined platform that supports traditional virtual machines (VMs), containerized applications, and AI workloads under a single operating model.

Private cloud integrates compute, storage, networking, operations, automation, Kubernetes runtime, and private AI services into a single platform. Private cloud customers are free to leverage open source models, vector databases, and Kubernetes-native toolchains while retaining control over their private infrastructure's economics and security. This strategy allows enterprises to support the rapid rise of open-weight models. Depending on the enterprise use case, the performance of open-weight models can be adequate and may nearly match proprietary frontier models, while ensuring sensitive data never leaves the organization and operates at a lower cost. Enterprises can fine-tune, audit, and optimize the models for their own workloads and costs without depending on a third-party API. Companies often need to run multiple open source and commercial models. This "model as a service" capability gives them agility and a competitive advantage. The goal isn't to abandon the public cloud. It's to build an architecture that gives enterprises agility, cost, and scalability options for each AI workload.

Q. Data sovereignty is a board-level concern. How can a private AI approach, specifically using private cloud, allow

developers to leverage powerful LLMs while ensuring strict enterprise data privacy?

Sovereignty has moved from a compliance checkbox to a strategic mandate for a growing number of enterprises. Enterprises with digital sovereignty requirements will migrate sensitive workloads to new or private cloud environments to reduce risk and increase autonomy. Tech leaders at multinationals will boost investment in sovereign-ready cloud and data localization at a faster rate than the IT economy as a whole. Private digital infrastructure is the ultimate sovereign data environment, but only if it's architected with isolation and governance from the start.

A private AI cloud delivers true sovereignty by isolating AI workloads both physically and logically. In a private cloud, IT can provision AI environments in dedicated, secure namespaces and orchestrate deployments within isolated virtual private clouds (VPCs). Fine-tuned models, RAG databases, and proprietary training data remain inside the corporate firewall. Data science teams can work with open source LLMs without intellectual property bleeding across internal departments or externally. Open source, open-weight models also improve portability, which IDC sees as key to sovereignty and to avoiding vendor lock-in.

IDC recommends that organizations:

- Define tiers for data sensitivity (public, internal, confidential, restricted) and map which AI use cases require private infrastructure.
- Prioritize private cloud deployment for workloads that involve personally identifiable information (PII), financial data, or IP-sensitive processes.
- Secure the entire AI stack, from model weights to training data, with a zero trust architecture and encryption.
- Automate data trust tasks such as classification, tagging, and policy enforcement so governance scales.

Q. The CapEx for AI accelerators and the hidden operational costs of tokenomics can easily derail IT budgets. How does private cloud help organizations rein in these unpredictable AI consumption costs?

You can't control AI tokenomics without controlling resource utilization. IDC found that FinOps teams that don't bring AI into scope across public cloud, private cloud, and SaaS will see 30%

higher costs for AI projects. At the same time, 85.9% of organizations rate improving agent cost management as a critical or high priority for the next 12 months (see *Crossing the Agent Economy's Cost Governance Gap: A Challenge for Tech Leaders,* IDC #US54890226, August 2026).

Private cloud changes the cost equation in two ways. First, it adds predictability. A large share of on-premises hardware and software costs can be capitalized and amortized over the asset's useful life. This approach avoids usage-based fluctuations in compute, storage, and networking charges that can lead to surprise bills.

Second, it increases utilization. Expensive accelerators only pay off when they're busy. A private cloud enables organizations to pool heterogeneous hardware across GPU and CPU types. Another way is multitenant model sharing. It lets many business units or teams serve requests from the same loaded model and GPU pool, so expensive hardware stays busy rather than sitting idle between tenants' sporadic requests. Higher utilization through batching and shared infrastructure spreads fixed costs across more tokens, which lowers the cost per token for every tenant. Policy-driven quotas and lease times for AI environments, combined with visibility into vGPU consumption, keep accelerators fully utilized. This approach prevents the hoarding of AI resources and stranded capacity that plague manual deployments.

This aligns with what enterprises tell us works. A top strategy for enterprises seeking to control AI costs is to automate AI portfolio monitoring, life-cycle management, and reporting. Pair that automation with a FinOps practice that proactively reviews AI projects, and IT can hold teams accountable for meeting total cost of ownership (TCO) targets.

Q. What are the TCO benefits of running AI on a unified private cloud platform compared with relying solely on public cloud AI services?

The TCO advantage of private AI is strongest for sustained, predictable inference workloads. Public clouds charge consumption-based rates for every token generated and every gigabyte transferred, which is ideal for bursty experimentation but expensive for always-on production services. Public cloud has not always delivered lower costs over time, while private AI infrastructure offers predictable costs, avoids vendor lock-in, and eliminates growing transaction charges.

The second TCO lever is operational. IDC consistently finds that Global 2000 enterprises run four or five separate tools for monitoring, observability, and automation. IDC has advocated for standardizing and consolidating multiple IT automation solutions, which reduces training and maintenance costs while increasing automation adoption. Enterprises using a private AI cloud

gain a single automated operating model for VMs, Kubernetes clusters, and AI workloads, thereby migrating away from overlapping third-party products.

Life-cycle and fleet management is the third lever. Automated patching, upgrades, and configuration management reduce configuration drift, lower operational risk, and free IT teams to focus on higher-value work. The result is cloud-like developer and operations velocity at a lower long-term cost.

To capture these benefits, IDC recommends that organizations model AI TCO up front, including tokens, GPUs, networking, storage, egress, and ongoing model retraining. Organizations should also plan for the skills and processes needed to operate a private AI cloud effectively. Enterprises that pair a unified private cloud platform with disciplined FinOps for AI will be best positioned to scale AI from pilot to production with predictable economics.

About the IDC analyst



Jevin Jensen

Research Vice President, Infrastructure and Operations

Jevin Jensen is research vice president, Intelligent CloudOps Market service at IDC, where he covers infrastructure as code/GitOps infrastructure automation, cloud cost transparency, DevOps, hybrid/public/multicloud management platforms, and edge management.

More about Jevin Jensen

Message from the sponsor

VMware Cloud Foundation (VCF) is a unified private cloud platform designed to run modern AI and existing workloads with a consistent cloud operating model, comprehensive automation, and enhanced security. Building on this platform, VMware Private AI Cloud gives organizations a production-ready path to securely build, run, and govern inference workloads, agentic applications, and traditional enterprise workloads together on a single private cloud. Offering diverse hardware, model, and accelerator choices, it enables enterprises to scale AI cost-effectively, operate more securely, and innovate rapidly.

Learn more about VMware Cloud Foundation and the AI Services built into the platform.

Learn how a university system with over 100,000 students scaled GenAI securely, cost-effectively, and with existing IT resources.

IDC Custom Solutions

IDC Custom Solutions produced this publication. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis that IDC independently conducted and published, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies.

This IDC material is licensed for external use, and in no way does the use or publication of IDC research indicate IDC's endorsement of the sponsor's or licensee's products or strategies.

IDC Research, Inc.
One Beacon Street, Suite 33100, MA 02108, USA
T +1 508 872 8200

blogs.idc.com | LinkedIn @IDC | www.idc.com

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries.