Broadcom's Private Cloud Outlook 2026 Reveals an AI Tipping Point as Production Inference Shifts Decisively to Private Cloud

June 9, 2026

Cost, Complexity, and Control Drive Infrastructure Decision-Making as Security Concerns Around AI Mount and Public Cloud Waste Reaches a Critical Threshold

PALO ALTO, Calif., June 09, 2026 (GLOBE NEWSWIRE) -- The AI experimentation phase is over, and the private cloud is where enterprise AI workloads are being deployed for security and scale. That is the central finding of the *Private Cloud Outlook 2026*, a new report from Broadcom Inc. (NASDAQ: AVGO), a global technology leader that designs, develops, and supplies semiconductor and infrastructure software solutions.

Where last year's report documented a deliberate "cloud reset" toward balance between public and private cloud, 2026 marks an acceleration into a full AI tipping point. The shift is being shaped by three forces — costs, complexity, and control — that public cloud environments are increasingly failing to address for production AI at scale. Key findings from the report include:

- 56% of enterprises are running or planning to run production AI inferencing on private cloud, while public cloud use for the same workloads dropped 15 percentage points year over year — from 56% to 41%;
- The biggest new demands being placed on enterprise IT by AI are data protection and privacy (37%) and security and control (36%);
- For the first time, cost has overtaken security as the number one public cloud concern, rising from 26% in 2025 to 31% in 2026;
- 97% of IT leaders believe some of their public cloud spend is wasted, and 52% estimate that waste exceeds 25% of their total public cloud budget;
- 83% of enterprises are considering the repatriation of workloads from public to private cloud, and 50% have already done so —with cost predictability now jumping to second biggest driver for repatriation, cited by 39% of organizations;
- Four out of five IT leaders say geopolitics are now affecting their IT strategy and operations, and for the first time, data sovereignty and residency requirements (54%) have overtaken jurisdiction-specific compliance (51%) as the leading geopolitical factor shaping infrastructure decisions.

“As enterprises move from pilots to running AI at production scale, infrastructure and operational costs spike, security gaps surface, and complexity compounds. The research is clear: enterprises increasingly prefer private cloud for production AI,” said Prashanth Shenoy, vice president of marketing, VMware Cloud Foundation Division at Broadcom.

View the full report here.

The Inference Shift: Production AI Finds Its Home on Private Cloud

The defining data point of this year's report is the scale and speed of the shift in where enterprises are running AI workloads. While public cloud remains a viable environment for AI pilots and model training experiments, the economics of running inference at scale tell a different story. Fifty-six percent of enterprises are running or planning to run production inferencing on private cloud, compared to just 41% on public cloud — a reversal from last year's near-parity. The drop of 15 percentage points in public cloud's share of production AI workloads in a single year is among the most dramatic shifts in this year’s report.

The reason is straightforward. As IT leaders described it in the survey data: public cloud is too expensive and insufficiently governed for AI workloads at scale. For pilots and training, some agility trade-offs may be acceptable. But when organizations need to scale, the cost and governance requirements drive workloads back home. Sixty-two percent of IT leaders report being very or extremely concerned about the infrastructure costs of running generative and agentic AI, while 36% say AI is actively driving new requirements for data protection, privacy, security controls, and risk management.

The Sovereignty Mandate: Geopolitics Reshapes Infrastructure Strategy

Geopolitics has entered the infrastructure conversation in 2026. Four out of five IT leaders now report geopolitics are directly affecting their IT strategy and operations. Data sovereignty has moved from a compliance checkbox to a board-level priority, and data sovereignty and residency requirements (54%) have overtaken jurisdiction-specific compliance (51%) as the leading geopolitical factor shaping infrastructure decisions. Industries with high security and compliance requirements — financial services, public sector, healthcare, and life sciences — are at the leading edge of this shift. For these organizations, the combination of AI-driven data volumes, cross-border data governance complexity, and the rising cost and governance burden of public cloud is making a compelling case for private cloud infrastructure that helps keep sensitive data under organizational control.

The Cost Reckoning: Public Cloud Economics Are Breaking Down

Cost has now become the defining concern about public cloud, overtaking security as the number one public cloud challenge, jumping from 26% in 2025 to 31% in 2026. The waste figures are striking: 97% of IT leaders believe some portion of their public cloud spend is wasted, and 52% believe that waste exceeds 25% of their total public cloud budget.

These economics are directly translating into repatriation activity. Eighty-three percent of enterprises are now considering moving workloads from public to private cloud, and 50% have already repatriated at least some workloads. The top three drivers of repatriation are security and compliance

(51%), cost predictability (39%) and performance (39%). The explosive rise of cost predictability as a repatriation driver is one of the most significant year-over-year changes in the report, underscoring how severely public cloud economics have deteriorated in the AI era.

Against this backdrop, private cloud investment intent is accelerating. Private cloud spend intent is growing at twice the rate of public cloud — up 21 points versus 10 points over a three-year outlook — and 58% of IT leaders now name building new workloads on private cloud as a top priority, up from 53% one year ago.

Survey Methodology

The *Private Cloud Outlook 2026* is based on a global survey conducted by Radius Tech in partnership with Broadcom. The survey was fielded in February–March 2026 and included 1,800 senior IT decision-makers at enterprise organizations (1,000 or more employees) across eight countries in North America, Europe, and Asia-Pacific. The report was published in June 2026.

Additional Resources

- Learn more about VMware Cloud Foundation
- Follow VMware Cloud Foundation social channels on LinkedIn, X, formerly known as Twitter and YouTube

About Broadcom
Broadcom Inc. (NASDAQ: AVGO) is a technology leader that designs, develops, and supplies semiconductors and infrastructure software for global organizations' complex, mission-critical needs. Broadcom combines long-term R&D investment with superb execution to deliver the best technology, at scale. Broadcom is a Delaware corporation headquartered in Palo Alto, CA. For more information, visit www.broadcom.com.

Roger T. Fortier
VCF Division, Broadcom
roger.fortier@broadcom.com
+1.408.348.1569