Private Cloud Emerges as the Preferred Platform for Production AI

Enterprise AI deployment strategies shifted dramatically in the past year, with private cloud now the clear leader for production AI workloads and public cloud starting to fall out of favor.

Broadcom's second annual Private Cloud Outlook 2026 reveals more than half of organizations surveyed **(56%)** are running or planning to run production inferencing in a private cloud. Public cloud use for production inference fell **15%** year over year to **41%.** Additionally, **62%** of IT leaders reported being very or extremely concerned about generative and agentic AI infrastructure costs while **36%** report AI is driving new requirements for data protection, privacy, security controls and risk management.

These findings from the Private Cloud Outlook 2026 study—the second annual research from Broadcom examining private cloud trends — reveal how the three C's of AI economics (costs, complexity, and control) are reshaping infrastructure decisions.

Methodology

The Private Cloud Outlook 2026 data is based on a global survey conducted by Radius Tech in partnership with Broadcom. Radius Tech surveyed 1,800 senior IT decision-makers at enterprise organizations (1000+ employees) in February - March 2026 across 8 countries in North America, Europe, and Asia-Pacific.

Where enterprises plan to run inference/production workloads

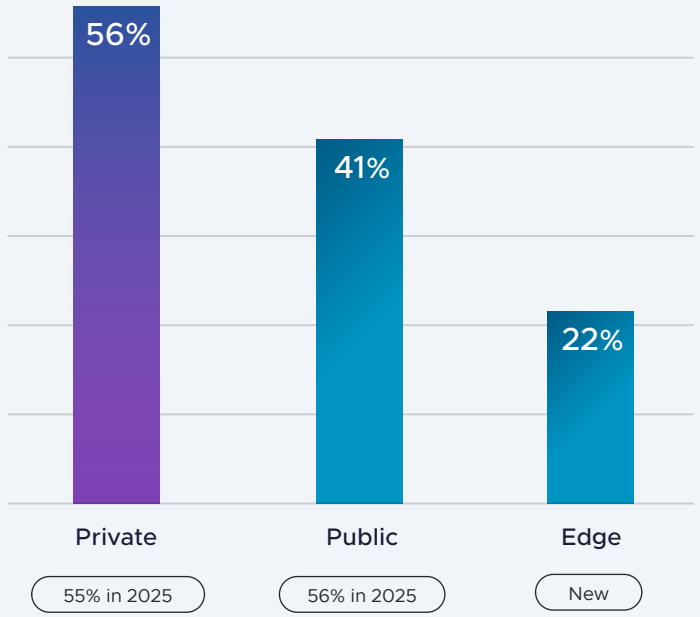


Figure 1: Where AI-based inferencing (production) applications or workloads currently run or expected to run, n=1800

Security is top priority for AI

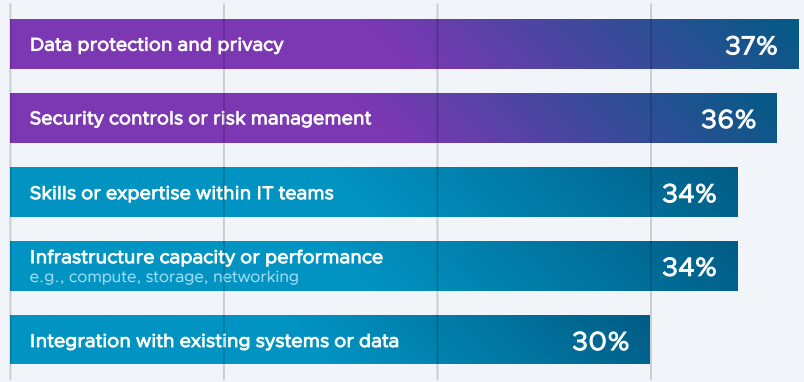
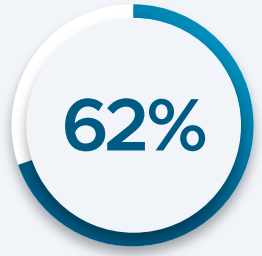


Figure 2: Where is AI creating the biggest new requirements for IT, n=1800



of IT leaders very or extremely concerned about Gen AI / Agentic AI infrastructure costs

Figure 3: Concerns about infrastructure costs for generative/agentic AI, n=1800

vmware by Broadcom