Compute

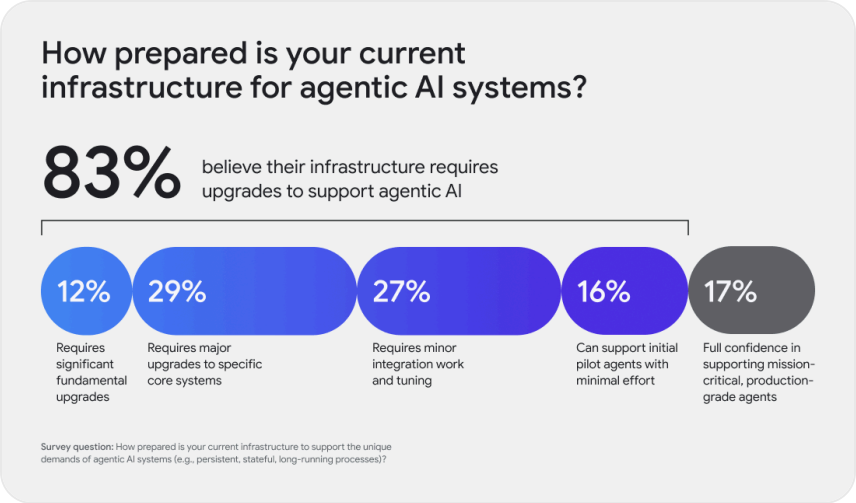
Report: 83% of organizations need to upgrade their infrastructure to support agentic AI

July 7, 2026

Drew Bradstock
Sr. Director, Product, Orchestration & Kubernetes

For years, enterprise AI has been synonymous with conversational AI — the customer service bots and digital assistants we interact with every day. But today, the market has shifted. We've officially moved from moving from AI that answers through simple chats, to AI that takes action, automated workflows, and executes complex tasks on its own. While this unlocks entirely new use cases, there's a catch: it places significant stress on the underlying infrastructure we've relied on in the past.

We recently surveyed more than 1,400 senior IT leaders for our State of AI Infrastructure report, and a resounding pattern emerged: the gap between AI ambition and infrastructure reality is widening. In fact, **83% of organizations say they require infrastructure upgrades** to support production-grade agentic AI.



Why? Because yesterday's infrastructure simply wasn't built for agents that act autonomously.

In this blog, we lay out the core insights from our research on how leading organizations are

rethinking their infrastructure to build resilient, fluid foundations. For more details and depth, we encourage you to download and read the full report.

Escape the “inference tax” with fluid compute

Agentic workloads introduce a new level of scale, where a single prompt can trigger hundreds of downstream actions, requiring massive context windows to be held in memory. Trying to run these continuous reasoning loops on legacy architecture is financially unsustainable. In fact, 62% of leaders are seeing a significant inference tax driven by data egress fees, storage bloat, and idle specialized hardware. Furthermore, 81% cite operational complexity as a hidden cost of scaling AI.

To fix this, organizations need fluid compute — the ability to dynamically match the right silicon to the right task while minimizing operational overheads.

- **For heavy training**: Compute accelerators like our new TPU 8t deliver tremendous scale to train the world's most sophisticated models.
- **For low-latency inference:** The TPU 8i, meanwhile, was purpose-built to maximize on-chip memory, so agents can think and react in real-time.
- **For orchestration**: General-purpose compute powered by CPUs is emerging as a critical component for driving AI control plane

operations. Using highly efficient, Arm-based processors like Google Axion, organizations can cost-effectively run reinforcement learning simulations and orchestrate agents.

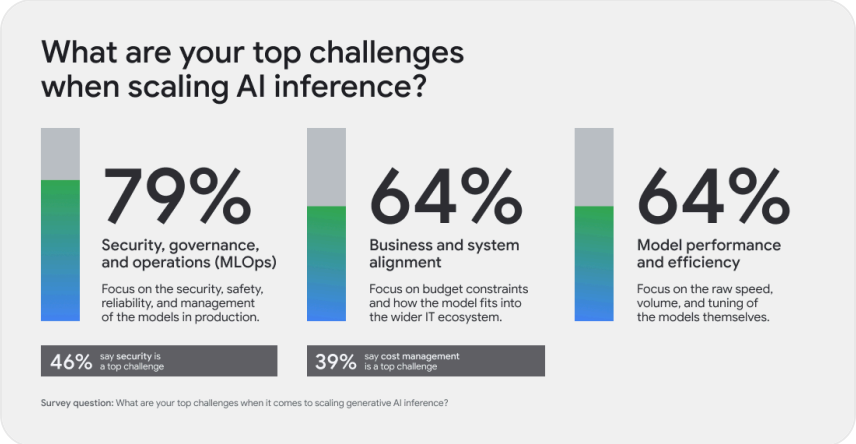
Managing agent sprawl with centralized governance

Agents are designed to act autonomously — reading emails, querying databases, and executing workflows across your business. But as agentic AI scales, organizations are facing a new challenge: agent sprawl. How do you manage thousands of autonomous agents scattered across diverse platforms, without losing visibility and control?

It's no surprise that 79% of tech leaders cite security, governance, and MLOps as their top challenge to scaling inference. In the agentic era, you need a mature governance strategy before you can innovate. This entails creating a centralized control plane that provides a single system of record for agent permissions, identity, and workflows.

Instead of patching together disparate tools, leading enterprises are relying on solutions like Agent Gateway to enforce enterprise-grade governance. Agent Gateway gives you the visibility you need to see exactly how agents are sharing data. It lets you define precise read/write scopes and maintain full audit trails of every interaction, and it provides human-in-the-loop oversight for when an agent needs approval before taking a critical action. This

drive for unified, straightforward governance explains why 78% of organizations now source their gen AI solutions directly from their primary cloud partner — a 30 point increase from 2025.



A unified data layer

Agents perform reasoning, meaning they constantly run heavy queries across your organization. If your data is fragmented across silos, your AI is effectively flying blind.

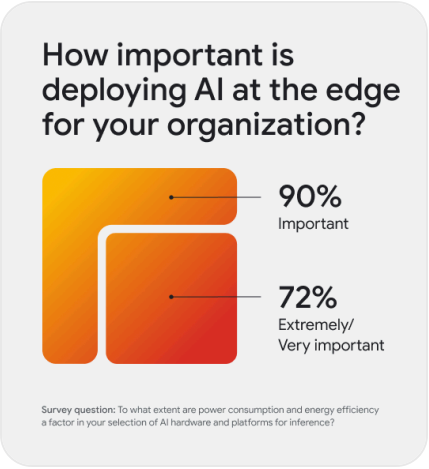
To move from managing disconnected data to gathering unique and actionable business context, leaders are adopting a unified data layer. Using tools like Smart Storage — which automatically annotates unstructured data to make it searchable — and the Cross-Cloud Lakehouse, agents can natively read and understand data no matter where it lives, without needing custom pipelines or duplicated data.

Hybrid multicloud and digital sovereignty

The debate between public cloud and local computing is settled: hybrid is the destination. In fact, 52% of organizations now use a hybrid multicloud architecture.

For technology leaders, this shift is largely driven by digital sovereignty and data gravity. Indeed, 48% of leaders are prioritizing infrastructure with strict data residency controls. You need the flexibility to run AI where it complies with shifting local laws. Whether that's leveraging the public cloud for broad compute, or bringing foundational models entirely on-premises via Google Distributed Cloud for air-gapped isolation, modern infrastructure must adapt to geopolitical realities, not the other way around.

AI at the edge



For technology leaders and infrastructure architects, relying on a strictly centralized cloud topology to process every agentic interaction is not a viable strategy. A staggering 90% of organizations now rank edge deployment as important for AI initiatives, with 72% describing it as extremely or very important.

Moving AI to the edge solves three issues:

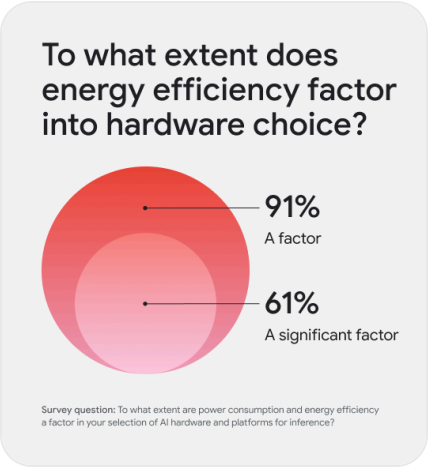
- **The latency bottleneck:** Real-time agents — especially those that rely on voice, video, or financial trading algorithms — can't afford the microsecond gap of a round-trip to a distant data center.
- **Operational resilience:** If an internet connection drops, business can't stop. Edge deployment ensures that agents running in manufacturing plants, retail stores, or hospitals can continue functioning autonomously.
- **Sustaining cost-efficiency:** Running always-on, continuous reasoning in the cloud is expensive. By utilizing highly optimized models on edge devices (like smartphones, IoT devices, or local warehouse servers), organizations shift the compute burden locally, drastically cutting variable per-token costs.

Breaking through the energy wall

Energy consumption used to be a sustainability metric reserved for annual reports. Today, it plays a crucial operational role. 91% of leaders now factor power consumption into their hardware selection, with 61% rating it as a primary or significant factor.

For technology leaders, power consumption presents a three-fold barrier to growth:

- **Grid scarcity:** You simply cannot buy more power in certain regions, heavily limiting how much compute infrastructure can be provisioned.
- **Regulatory compliance:** Energy efficiency is now a strict legal prerequisite to operate. For example, in Germany, new data centers must achieve a Power Usage Effectiveness (PUE) of 1.2 or lower. And Ireland now mandates that large data centers provide 100% on-site dispatchable generation to match their grid draw.
- **Infrastructure economics**: Inefficient power envelopes drastically inflate the Total Cost of Ownership (TCO) of AI deployments. Accommodating high-power hardware requires massive capital expenditure (CapEx) for advanced cooling architectures, specialized rack designs, and facility upgrades.



To address the energy wall, technology leaders must treat energy as a strategic asset. One of the focus areas for optimization must shift to performance-per-watt. This is why co-designed silicon is becoming so important. For example, our new TPU 8t delivers nearly three times the performance of the prior generation while being up to twice as energy-efficient.

Unified, AI-optimized infrastructure

Ultimately, you cannot solve the challenges of tomorrow's agentic systems with yesterday's architecture. When engineering teams are forced to manually integrate heterogeneous compute, storage, and networking layers, organizations incur high operational overhead just to ensure basic interoperability.

To innovate quickly and cost-effectively, technology leaders are therefore moving toward holistic, unified systems.

This is the philosophy behind Google Cloud's AI Hypercomputer. It's an architecture where every layer is co-designed and co-engineered to work together. The custom silicon (TPUs, GPUs, CPUs) isn't designed in a silo; it's engineered alongside the ultra-high-bandwidth networking (Virgo Network), the storage (Managed Lustre, Hyperdisk), and the software orchestration layer (GKE).

Bridging the digital and physical worlds

When you embrace this co-designed, holistic approach, the results go far. With this level of scalable, fluid intelligence operating at the edge, we're entering the era of physical AI. A new generation of autonomous robots can sense,

simulate, and navigate the physical world, practicing tasks millions of times in digital twin simulations on Google Cloud before they ever set foot in the real world. From performing complex industrial inspections to capturing cinematic videography, AI is now solving tangible problems in the real world.

Your blueprint for agentic AI

Adapting your infrastructure to meet the demands that agentic applications place on your systems will help you move from pilot to production. The organizations set to thrive in 2026 are embracing a unified foundation that is cost-efficient, resilient at the edge, optimized for autonomous action — and governed by default.

Ready to start? Download the State of AI Infrastructure report to explore the data behind our findings, and discover how your peers are already building for success.

Posted in Compute—AI & Machine Learning

Related articles

Compute

Three lessons in accelerating foundation model upgrades

By Radhika Mani • 4-minute read

Compute

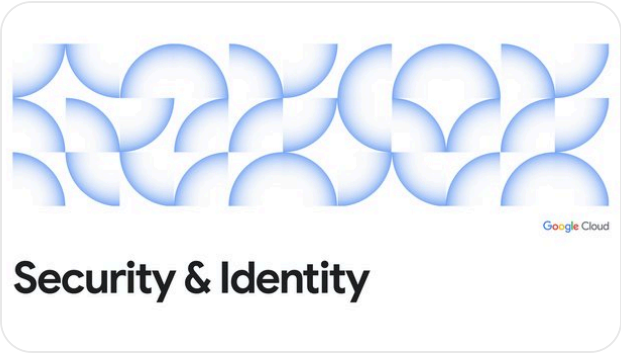
C4N, now GA: Delivering cloud's highest per vCPU network and block storage I/O for x86 workloads

By Parinda Gandhi • 10-minute read

AI infrastructure

Google Cloud named Leader in the 2026 Gartner® Magic Quadrant™ for AI Infrastructure

By Mark Lohmeyer • 5-minute read



Security & Identity

Verifiable, private AI: Google Cloud expands Confidential Computing frontiers

By Sam Lugani • 8-minute read

Follow us