

McKinsey Technology

Reimagining tech infrastructure for (and with) agentic AI

Scaling agentic AI requires turning unstructured data into governed, reusable assets that systems can interpret and trust. Data leaders can start by building shared foundations and enforcing standards.

by Arnaud Tournesac, Arun Gundurao, Ling Lau, and Pankaj Sachdeva



IT infrastructure is entering a new phase as [AI agents](#) increasingly orchestrate, govern, and scale work across the enterprise. Infrastructure no longer acts solely as a support function or control layer; it becomes the backbone of an orchestrated system that determines how effectively organizations capture value from agentic AI (see sidebar “What are AI agents?”).

What are AI agents?

AI agents are autonomous software entities designed to achieve specific goals, execute tasks independently, and make [real-time decisions](#). They can operate on their own or collaborate within an “agentic mesh,” a network where agents coordinate seamlessly with other agents, tools, and transactional systems. By combining multiple AI technologies, these agents can be orchestrated in a modular way, making it possible to handle even highly complex workflows from end to end.

With AI workloads expanding, however, IT infrastructure costs are projected to increase two to three times by 2030 while budgets remain flat.¹ [McKinsey’s latest survey on the state of AI](#) finds that while 62 percent of organizations are experimenting with or piloting AI agents, scaling remains low. In any given business function, no more than 10 percent of respondents say their organizations are scaling AI agents.

For chief technology officers (CTOs), this creates an urgent dual challenge: Upgrade infrastructure so it is fast, scalable, and reliable enough to support agentic AI, while using agentic AI itself to contain the rising cost of doing so (see sidebar “The evolution of IT infrastructure”). Companies must move quickly, given more than one-third of high performers are committing more than 20 percent of their digital budgets to AI.²

¹ Projections based on McKinsey analysis of industry reports and Gartner data.

² “[The state of AI in 2025: Agents, innovation, and transformation](#),” McKinsey, November 5, 2025.

The evolution of IT infrastructure

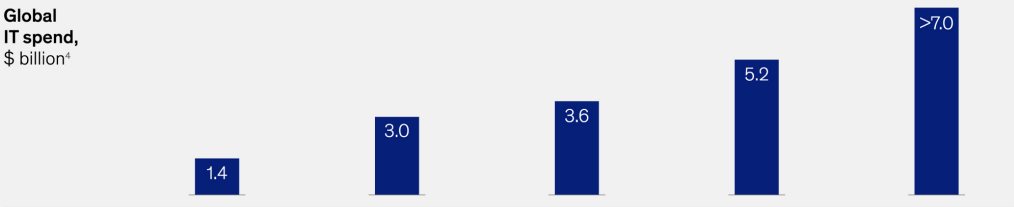
Over the past few decades, IT infrastructure has changed radically multiple times—each shift unlocking greater efficiency, scalability, and resiliency. While mainframe and on-premises systems dominated before 2000, cloudification took hold in the 2010s followed by first enterprise adoption of gen AI and agentic AI today (exhibit). This evolution has also been marked by shifts in compute architecture—from CPU-centric environments to GPU acceleration, and now toward hybrid and specialized accelerators optimized for AI workloads. Now, we stand on the verge of a new technological breakout: [the agentic shift](#).

Exhibit

Agentic AI is transforming IT infrastructure operations into a new era.

Agentic AI opportunities by domain

	Mainframe and on-premises computing (pre-2000s)	Global internet (dot-com era) (2000s)	Cloudification and rise of public cloud (2010s)	Emergence of AI into enterprises (2020s)	Rise of agentic AI ecosystems (2030)
IT infrastructure	Centralized, monolithic compute tightly coupled with storage and apps	Siloed infrastructure with limited reuse, manually provisioned	Infrastructure defined per request, limited automation and standardization	Modular, consumable infrastructure (IaC) ¹ with automated provisioning	Intent-based, autonomous infrastructure orchestrated by AI agents
Operating model	Manual, central IT provisioning, terminal-based access	Tickets, emails, human-driven provisioning	Tickets, human-driven provisioning, self-service portals	APIs, CLI, ² IaC templates; self-service portals	NLP ³ , business intent, AI agents interpret and act
Tech stack	Mainframe (on-premises only)	On-premises (virtual machines, physical servers)	Hybrid (on-premises and public cloud)	Distributed (cloud, edge, on-premises)	Intent-driven (agentic platform, cloud, edge, on-premises)
Pace of change	Years	Years	Months	Weeks	Days



¹Infrastructure as code.
²Command-line interface.
³Natural language processing.
⁴Calculated as (S&P 500 IT sector total return index) ÷ (S&P 500 total return index) to illustrate pace of spend change instead of flat spend.
Source: eMarketer; Gartner; S&P Capital IQ

McKinsey & Company

Agentic AI can fundamentally reshape how infrastructure is provisioned, managed, and optimized, with intent-driven connectivity, autonomous operations, and minimal human oversight. Our experience indicates that agentic AI can enable automation of 60 to 80 percent of routine infrastructure work over time, translating to a 20 to 40 percent run-rate cost reduction in initial deployments, with further gains as adoption scales. It will, however, require CTOs to deliver on a raft of needs such as automation, environment simplification, operating model redesign, and active cost governance, while simultaneously improving resilience and delivery speed.

To support this shift, leading organizations are beginning to redesign infrastructure applying principles more commonly associated with architecture, such as modularity, composability, and orchestration.

Three pressure points for infrastructure

As companies look to scale their agentic AI programs, infrastructure leaders face three structural pressures:

- *Infrastructure must run materially faster and at scale.* Innovation in agentic AI is flourishing but often in silos, creating fragmentation that slows the ability to reuse agents and scale. As a result, [less than 10 percent of agentic programs reach meaningful scale](#). At the same time, demands are increasing as developers work faster and as the need to coordinate agents, tools, and data across environments increases. Environments designed for ticket-based workflows cannot sustain this throughput.
- *Non-labor costs are rising rapidly as AI workloads expand.* Rapid growth in compute and storage demand (both on-premises and cloud), amplified by agentic AI, is driving a projected two- to threefold increase in IT infrastructure costs by 2030.³ At the same time, infrastructure budgets are expected to remain relatively flat.
- *Outage risk carries greater financial consequence than ever before.* Resilience has become inseparable from brand reputation, security, and enterprise risk. Systems are growing more complex as they adapt to the needs of agentic AI, creating more points of failure and making observability and control significantly harder to maintain.

What agent-ready infrastructure looks like

Infrastructure has historically been built primarily for human-led operations. In the agentic era, that is no longer sufficient. At scale, this requires more than incremental upgrades. Infrastructure must evolve toward a more modular, “mesh-like” design, where agents, tools, and enterprise systems are connected through a shared orchestration layer. This enables coordination across domains while maintaining control and reuse.

³ Projections based on McKinsey analysis of industry reports and Gartner data.

To achieve this vision, four foundational capabilities are critical:

- *Repeatable and executable actions through secure APIs.* Repeatable actions must be accessible as code with embedded policy checks.
- *Reliable operational data.* Clear sources of truth for assets, dependencies, ownership, logs, and metrics reduce ambiguity and enable safe automation. Imperfect data should not prevent progress. Many high-value use cases can be piloted even in environments with inconsistent database fidelity or fragmented repositories.
- *Embedded controls and agent governance.* Permission models must define what agents are allowed to do and under what conditions, with clear digital identity, ownership, and accountability for every agent. All actions must be logged, traceable, and auditable, with enforcement of policies across environments. High-impact actions require human approval, supported by supervisory mechanisms to pause or override automated behavior.
- *Agent life cycle management, interoperability, and context.* Organizations need a clear inventory of deployed agents, defined scope for each one, performance tracking, and life cycle management. As agents scale, teams must also manage cost and resource consumption explicitly, including monitoring inference usage and execution patterns to avoid unexpected cost spikes. Agents increasingly operate across systems and platforms, requiring interoperable control planes and integration patterns. They also depend on a structured understanding of the IT estate, including dependencies, ownership, and known failure modes, to operate safely and make informed decisions.

Building on this foundation, leading organizations are converging on a more mesh-like approach to infrastructure, where agents, platforms, and systems are interconnected through a shared orchestration layer. In practice, this introduces a set of design principles that shape how infrastructure behaves at scale:

- *Composability:* Infrastructure components, agents, and tools can be reused across workflows without rework.
- *Decoupling:* Execution, orchestration, and data layers are separated to improve scalability and flexibility.
- *Vendor flexibility:* Components can evolve independently, reducing lock-in and preserving optionality.
- *Governed autonomy:* Agents operate within defined policies, with clear accountability and escalation paths.

Most organizations already run platforms such as ServiceNow, cloud management tools, network controllers, and observability stacks, many of which are adding AI capabilities. The

strategic decision for enterprises is not whether to replace these systems, but how to integrate them into a coherent backbone that enables cross-domain coordination and reuse.

Architecturally, enterprises should retain flexibility in how agents are built and deployed, whether through hyperscaler-native services, leading model providers, or enterprise-hosted models optimized for cost and data sensitivity.

Agentic AI can create the greatest value in five areas

Organizations that successfully adapt their infrastructure for agentic AI focus on a set of high-value domains where automation, simplification, and operating model redesign can create near-term impact. These five areas stand out: service desk, observability and IT service management (ITSM), network operations, hosting operations, and active cost and contract management (exhibit).

Service desk

Service desk is the largest and “quickest-to-value” area, accounting for 20 to 30 percent of total infrastructure labor spend. High ticket volumes, standardized workflows, and predictable resolution paths make this area especially well suited for agentic automation.

Exhibit

Agentic AI can unlock significant value across infrastructure, with the largest impact concentrated in five core domains.

Agentic AI opportunities by domain

			Average share of infrastructure spend, %	Gross potential savings, ¹ %
Labor	Service desk	Use AI agents within service desk to reduce ticket demand, remove manual triage, and automate resolution of repeatable issues	 20–30	 25–45
	Observability and ITSM ²	Apply AI agents across ITSM and observability workflows to reduce incident frequency, shorten diagnosis time, and remove manual coordination	 20–30	 20–40
	Network operations	Use automation, AI agents, and declarative infrastructure (IaC ³) across network to reduce manual activities	 10–20	 20–40
	Hosting operations	Use automation, AI agents, and IaC across cloud and hosting operations to reduce manual run and provisioning activities	 15–25	 20–40
Non-labor	Active cost and contract management	AI-enabled IT asset right-sizing aligning applications, infrastructure, and licenses to actual usage; vendor and contract optimization using agentic monitoring and predictive analytics to support negotiation, consolidation, and ongoing contract decisions	 40–60	 5–15

¹Potential unlock expressed as share of potential automation of tasks within the area. Savings shown are not additive, since baseline amounts differ and overlap in some cases. Realized automation benefits vary by maturity: Advanced organizations see incremental uplift, while those earlier in adoption may achieve larger step change gains.

²Information technology service management.

³Infrastructure as code.

AI agents can autonomously resolve routine high-volume requests such as password resets and account unlocks, while guiding structured ticket intake through self-service interactions. More complex issues can be escalated to humans under clearly defined governance. Agents can also fulfill standard service requests, such as access provisioning, license assignment, and group membership changes, without manual intervention. Organizations can experience 25 to 45 percent savings along with improved service-level-agreement adherence, always-on support, and better employee experience.

In one example, the IT service desk of a multinational enterprise embedded agents across its support model, aiming to boost productivity, enhance user experience, and reduce resolution times for approximately 450,000 tickets annually. The organization redesigned customer journeys and workflows to enable agent-led resolution, prioritizing an AI intake bot, an agent using interactive-voice-response technology, and a proactive infrastructure monitoring capability. The transformation resulted in up to 80 percent of requests being automated, 50 percent of service agent capacity redeployed to higher-value activities, and a customer satisfaction score of 4.8 out of 5.

Observability and IT service management (ITSM)

Observability, ITSM, and infrastructure operations (including both network and hosting operations) together account for 45 to 75 percent of total infrastructure labor spend, reflecting both platform engineering and operational-response activities. Engineers spend a disproportionate share of time responding to alerts as well as manually testing and deploying fixes, limiting their capacity to proactively manage risk.

Agentic AI automates both diagnosis and resolution. Agents continuously correlate logs, metrics, configuration data, and change histories to identify weak signals, anomalies, and emerging failure patterns. When incidents occur, agents automatically provide relevant context, identify root causes, and execute predefined remediation activities within guardrails. Common issues can be resolved autonomously, while engineers engage selectively on complex cases. (See sidebar “Agentic AI in action: Responding to an alarm” for an illustrative example of this workflow.)

Agentic AI in action: Responding to an alarm

When an observability alarm triggers a SEV1 (severity level 1) incident, an incident manager agent initiates a structured triage workflow aligned to information technology infrastructure library (ITIL) principles.

Behind the scenes, multiple domain agents spanning network, infrastructure, application, change history, and ticketing systems launch parallel investigations within defined boundaries. Each agent queries its own tools and data sources, correlating logs, configuration management database (CMDB) records, recent changes, and prior incidents.

Domain-specific agents (for example, network, application, or infrastructure) each generate and test hypotheses within their scope. An orchestrator agent synthesizes these inputs to determine the most likely root cause and defines a remediation plan with ordered steps and validation checks, alongside stakeholder communications, with execution gated by predefined approval thresholds. Low-risk actions can be executed autonomously, while high-impact changes, such as customer-facing communications or production rollbacks, require human approval.

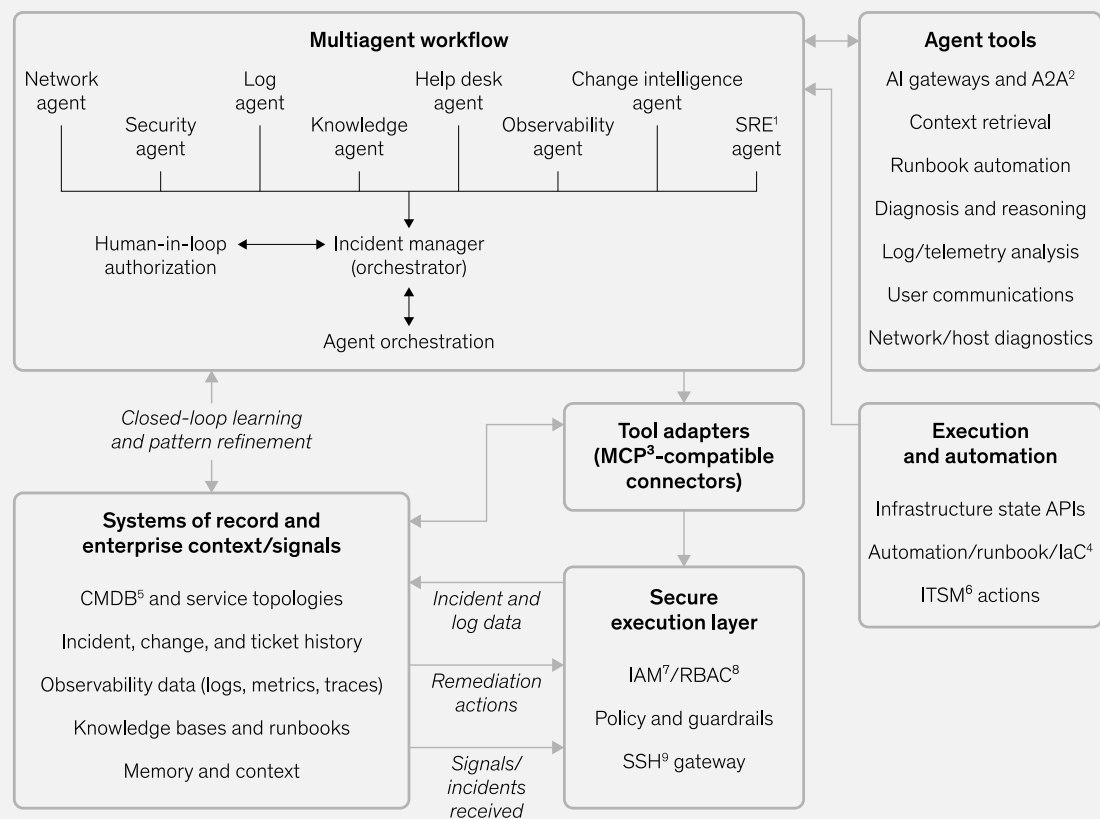
In parallel, the incident manager agent converts incident data into dashboards and summaries for leadership to track autonomous resolution rates, mean time to repair (MTTR), recurrence, and business impact. Reporting is embedded within the agentic workflow and reduces coordination overhead. This workflow is reflected in the multiagent architecture, where domain agents, orchestration, and execution layers operate in a coordinated loop—the kind of foundation required to support scaled agentic operations (exhibit).

The result is a shift from ad hoc firefighting to consistent, rapid, and scalable execution. The aspiration is “ZeroOps,” where agents manage routine workflows and humans focus on systemic improvement.

Exhibit

Multiagent workflows require a robust underlying architecture with a secure execution layer.

Architecture for a multiagent incident response system illustrative



¹Site reliability engineering. ²Agent-to-agent communication. ³Model context protocol. ⁴Infrastructure as code. ⁵Configuration management database. ⁶Information technology service management. ⁷Identity and access management. ⁸Role-based access control. ⁹Secure shell.

Network operations

Network operations account for approximately 10 to 20 percent of total infrastructure labor spend. Traditionally, enterprise networks were designed for relatively stable traffic patterns and human-led troubleshooting. Network teams operate reactively, with limited ability to proactively intervene upstream of issues.

Agentic AI instead enables real-time, intent-driven management. Agents can ingest and correlate traffic patterns, configuration states, logs, and change histories to detect congestion, abnormal traffic, and emerging risks, and translate higher-level intent into governed network actions. They can autonomously triage network events and execute repeated, routine changes (such as firewall/VLAN updates) in response to policy triggers. By reducing manual efforts in investigation and execution, savings of 20 to 40 percent can be realized in initial deployments, with significantly higher automation potential over time.

One example of this transition at scale is Deutsche Telekom's agentic network implementation, the "RAN Guardian agent."⁴ Operating in the context of network events and exceptional situations, agents actively monitor mobile-network performance, assist in troubleshooting, and optimize solutions.

Hosting operations

Hosting operations, which include on-premises, DevOps (software development and IT operations), and cloud compute and storage activities, account for approximately 15 to 25 percent of total infrastructure labor spend. Hosting operations remain dominated by repetitive life cycle activities in capacity management, patching, and environment provisioning that are still coordinated through tickets and manual intervention.

Agentic AI shifts hosting operations to closed-loop environment control. Agents can continuously assess system health, configuration drift, and policy compliance across on-premises, DevOps, and cloud environments. By standardizing run-time environments, exposing life cycle actions through APIs, and embedding policy as code, agents can autonomously handle routine activities such as rightsizing and maintaining capacity across environments. Executed well, this unlocks 20 to 40 percent in savings in initial deployments, with significantly higher automation potential over time.

In one example, a leading utilities provider is modernizing its hosting operations through a combination of standard and agentic automation. A bottoms-up assessment revealed that several areas, such as capacity management, were labor intensive and repetitive. The organization scaled infrastructure as code (IaC) for provisioning and is introducing AI agents to analyze multivariable trade-offs and orchestrate cross-functional workflows. As a result, approximately 15 percent of potential run-rate savings were identified through IaC and roughly an additional 20 percent of potential run-rate savings were identified through agentic AI.

⁴ "AI agents for mobile network," Deutsche Telekom press release, November 11, 2025.

Active cost and contract management

A significant share of infrastructure spend, which often accounts for 40 to 60 percent of total technology spend, is tied to external services such as cloud, software, and vendor contracts. This spend is typically managed through periodic reviews. As AI scales, costs become granular, multidimensional, and nonlinear, requiring stricter governance.

Agentic AI shifts cost optimization from periodic review to continuous management, unlocking 5 to 15 percent in savings. Agents can ingest real-time usage, licenses, contracts, and pricing benchmarks to automatically rightsize infrastructure, reclaim unused licenses, enforce budget guardrails, and flag uneconomic configurations. In parallel, procurement agents can monitor vendor performance, benchmark rates from historical data, validate invoices, and surface renegotiation triggers based on cost models with actual demand.

The CTO agenda: The first 90 days

Building an agent-ready infrastructure is not a quick fix, but the first 90 days are critical to setting direction, proving value, and building credibility. The most important actions include the following:

- *Redesign targeted processes.* Select one area with high volume, clear performance pain points, and strong potential for repeatable execution, such as service desk operations and incident management. Deconstruct the workflow into its component tasks and redesign the process so that routine tasks are executed automatically within defined boundaries, while engineers intervene when judgment or creativity is required. This redesign often simplifies the process itself.
- *Strengthen operational data.* Agents cannot compensate for inconsistent system records or unclear ownership. A practical starting point is to clarify the source of truth for assets, configurations, and dependencies. Standard naming conventions, consistent schemas, and explicit ownership reduce ambiguity. Determine if the underlying data and knowledge about the infrastructure are structured and consistent enough for machines to interpret and reuse.
- *Establish strong operating and governance practices.* Before agents are allowed to execute changes in production environments, CTOs need a clear framework that defines permissible actions, escalation thresholds, and accountability. Each agent should have a named owner, with clarity around which decisions can be made autonomously and which require review. Logging and audit capabilities must be comprehensive (see sidebar “The human and operating model shift behind agentic infrastructure”).

Find more content like this on the
McKinsey Insights App



Scan • Download • Personalize



The human and operating model shift behind agentic infrastructure

As agents assume responsibility for routine diagnostics, triage, and execution, human roles shift toward supervision, exception handling, architectural design, and systemic improvement. The primary impact will be a change in the nature of work required from human talent. Engineers spend less time resolving repetitive tickets and more time supervising autonomous execution, reducing recurrence, and strengthening resilience.

Realizing this shift requires redesigning operating processes so automation is embedded into workflows rather than layered onto legacy coordination models. Commercial models often need adjustment as well. Without aligning vendor contracts and labor structures to AI-enabled productivity, gains risk remaining embedded in delivery models rather than translating into financial outcomes.

The role of site reliability engineering (SRE) is also likely to evolve materially. As infrastructure becomes more distributed and dynamic, reliability can no longer be managed through runbooks and scripted automation alone. SRE work is shifting toward three priorities: designing and refining agents, building systems that validate and constrain agent behavior, and serving as a human in the loop for critical decisions.

SRE teams are uniquely positioned to build these systems. Their expertise in incident response, failure modes, and operational safety defines how agents operate in production. Increasingly, teams are codifying this knowledge into agents while also developing deterministic validation layers that define expected system behavior and establish guardrails for safe execution.

As agent-driven execution scales, validation and control become as critical as the agents themselves. Leading organizations are already restructuring IT operations around SRE-led models that bring together application and infrastructure expertise, with a mandate to systematically eliminate toil and continuously improve system resilience.

- *Put in place explicit agent management practices.* A formal registry that documents each agent's purpose, scope, and performance prevents fragmentation. Life cycle management ensures that outdated or redundant agents are retired. Visibility into performance and cost helps the organization understand where value is being created.

For the first time in decades, CTOs can redefine how infrastructure is built and how work is executed. Those who treat agentic AI as an incremental automation layer will see only localized gains. Those who build agent-ready foundations and reimagine infrastructure and operations will fundamentally change their organization's speed, resilience, and economics. Over time, infrastructure will shift from a supporting function to the platform that orchestrates and governs how work is executed across the enterprise, coordinating networks of agents at scale.

Arnaud Tournesac is a partner in McKinsey's Paris office, **Arun Gundurao** is a partner in the New York office, **Ling Lau** is a partner in the New Jersey office, and **Pankaj Sachdeva** is a senior partner in the Philadelphia office.

The authors wish to thank Doris Zhang, Ran Chen, Sandeep Nereyeth, Sebastian Flegr, and Vineetha Akkem for their contributions to this article.

This article was edited by Barr Seitz, an editorial director in the New York office.

Copyright © 2026 McKinsey & Company. All rights reserved.