AMD
together we advance_

PLANNING FOR *AGENTIC AI*

AN ENTERPRISE IT LEADER'S GUIDE TO INFRASTRUCTURE AND READINESS FOR THE AGENTIC ERA



WHAT'S INSIDE

THE NEXT SHIFT IN ENTERPRISE AI IS HERE 03

FROM CHATBOTS TO AGENTS: WHAT MAKES AI 'AGENTIC' 04

INSIDE AN AGENTIC AI WORKFLOW 05

AGENTIC AI PLACES NEW DEMANDS ON IT INFRASTRUCTURE 06

PREPARING INFRASTRUCTURE FOR A DIFFERENT KIND OF AI WORKLOAD 07

ENTERPRISE AGENTIC AI READINESS CHECKLIST 08

BUILD FOR AN AGENTIC FUTURE 09



THE NEXT SHIFT IN ENTERPRISE AI IS HERE

Generative AI changed how people interact with technology. AI assistants can answer questions and create content. But people still drive the process. They prompt the AI and decide what happens next.

Agentic AI changes that dynamic.

Instead of simply responding to a prompt, agentic AI can work toward defined goals with autonomy. It can determine the steps required, use tools and enterprise data, evaluate results, and work until the task is complete or requires human judgment.

This shift brings new planning considerations for IT teams. Agentic systems interact with the applications and data businesses already rely on, while creating new demands on the infrastructure that supports them. As adoption grows, IT leaders need to determine how these workloads will be governed, where they will run, and how the underlying infrastructure can support them at scale.

62% of organizations are experimenting with AI agents.[1]

23% are actively scaling agentic AI systems somewhere in their enterprise.[1]



FROM CHATBOTS TO AGENTS: WHAT MAKES AI 'AGENTIC'

The key difference between generative and agentic AI is who drives the next step.

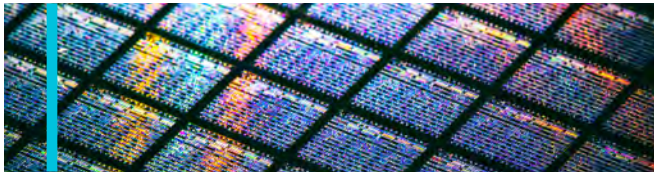
With traditional generative AI, manual intervention remains the center of each workflow. People initiate each interaction and decide how to use the response. Agentic systems shift more responsibility to the technology, allowing work to continue without constant human direction.

Once given a goal, an agent can break the work into steps and determine what happens next. Along the way, it may call APIs, query data, check permissions, or bring in other agents to complete specific tasks. These actions can span multiple tools and systems, creating a workflow that adapts as the agent gathers new information.

For organizations, this represents a new operating model for AI. People set the goals and boundaries, while agents take a more active role in carrying work forward.

GENERATIVE AI VS. AGENTIC AI

GENERATIVE AI	AGENTIC AI
Responds to a prompt	Interprets and works toward a goal
Person initiates each step	Agent determines next step
Produces content or information	Takes action using tools or systems
Relies on human direction to continue	Executes multiple steps autonomously
Typically involves a single AI interaction	Coordinates multiple agents, models, and applications



THE AMD PERSPECTIVE

A single goal can expand into many concurrent tasks across models, applications, and enterprise systems – so the unit of planning is no longer one application, it's a coordinated compute system. AMD builds across that whole system: AMD EPYC™ server CPUs for the orchestration, data access, and enterprise workloads agents depend on; AMD Instinct™ GPUs for model reasoning; and open, standards-based software so new models and tools can be adopted without re-architecting the stack.

INSIDE AN AGENTIC AI WORKFLOW

Agentic workflows are dynamic and iterative. Rather than moving through a fixed sequence, an agent evaluates context, selects the right resources, and adjusts its approach as new information emerges.

This activity occurs in eight key stages. Agents may revisit earlier steps, initiate additional tasks, or adapt based on what they learn. Each stage also places different demands on infrastructure, making this process an important starting point for IT planning.



THE AMD PERSPECTIVE

Agentic AI introduces a new capacity-planning challenge. Instead of sizing infrastructure for a single application, IT teams need to plan for a growing digital workforce of agents operating across systems and generating new activity as they work. Infrastructure demands can grow quickly as more agents run concurrently and take on new tasks.

AGENTIC AI PLACES NEW DEMANDS ON IT INFRASTRUCTURE

Infrastructure demands go beyond model inference. Agents generate more calls to applications, APIs, data sources, and other services as they work. And because they can initiate tasks and agents, the demands on infrastructure may not grow linearly as you scale.

Readiness also isn't just about capacity. It also requires flexibility in where workloads run. Most enterprises use a mix of cloud and on-premises infrastructure, making security and data protection more important than costs when it comes to choosing where workloads live. Infrastructure needs to support this hybrid reality as AI deployments grow.

For IT leaders, planning must start before agentic AI reaches scale. Decisions made today will shape how effectively your organization can support growing agent activity tomorrow.

60%

of AI workloads are inference in 2026, up from 40% in 2024.[2]

79%

of organizations require AI to be integrated into general-purpose workloads.[3]

75%

of organizations have already refactored, relocated, or redeployed applications or data to accommodate their AI initiatives, or will do so.[4]



THE AMD PERSPECTIVE

Hybrid isn't a transition state – it's the operating environment. AMD EPYC™ server CPUs are broadly available across major cloud providers and on-premises platforms alike, so enterprises can place agent workloads wherever sovereignty, latency, and cost requirements point, and move them later without re-platforming.

PREPARING INFRASTRUCTURE FOR A DIFFERENT KIND OF AI WORKLOAD

Supporting agentic AI requires more than adding capacity. IT teams need to consider the different roles infrastructure plays across the agentic environment, from supporting growing agent activity to powering the systems agents rely on.

PLAN FOR A GROWING CPU ROLE

In an agent-driven environment, the CPU does three distinct jobs – and each one has a different compute profile:

- Orchestrating – planning, tool calls, permission and policy checks, and agent-to-agent coordination reward high core density.
- Feeding the accelerator – data preparation, retrieval, and pre/post-processing keep GPUs utilized, rewarding high per-core frequency.
- Running the business – the databases, applications, and services agents act upon reward balanced general-purpose performance.

The AMD EPYC™ server CPU portfolio spans all three profiles on a single platform, so IT teams can match the processor to the role instead of compromising across them and consolidate onto fewer platforms as agent activity grows.

PREPARE EXISTING SYSTEMS FOR MORE ACTIVITY

Agents rely on databases, applications, APIs, and enterprise services that support the business. More agents working concurrently can increase demand on these systems rather than replace them.

PRESERVE FLEXIBILITY AS AI EVOLVES

Agentic workflows can span different models, tools, and environments. Open, standards-based infrastructure can make it easier to integrate new technologies and preserve flexibility as your requirements evolve.



THE AMD PERSPECTIVE

Agents don't replace the enterprise estate; they add load to it. The plans agents make, the tools they call, and the records they read can create additional demand on databases, applications, and services that were sized for human-paced demand. That makes readiness a question of the whole environment, not just the AI tier. AMD builds for both sides of it: AMD Instinct™ GPUs for model reasoning, and AMD EPYC™ server CPUs for the agent execution and the business systems agents act upon – on the same platform, under one capacity plan.

ENTERPRISE AGENTIC AI READINESS CHECKLIST

Agentic AI use cases will shape infrastructure requirements. Early IT involvement can help determine how agents should be used and where workloads should run.

Use this checklist to assess your current environment and identify areas that require preparation.

☐ DEFINE YOUR USAGE MODEL

Determine how many users and agents you need to support, which models you'll use, and what data must stay on premises.

In practice: Estimate how many agents may run concurrently for your first use case.

☐ MAP YOUR WORKFLOWS

Identify which workflow stages your first agentic use cases will involve. Then, map your existing infrastructure against those needs.

In practice: Identify where an agent will need to access data, call tools, or use a model.

☐ BUILD IN GOVERNANCE FROM THE START

Design permissions, policy enforcement, and audit trails into your agentic environment early rather than adding them later.

In practice: Define which systems an agent can access and which actions require human approval.

☐ PLAN FOR HYBRID ENVIRONMENTS

Prepare to run workloads across cloud, on-premises, and edge environments without making architecture decisions that lock you into a single venue.

In practice: Determine which workloads can run in the cloud and which need to stay on premises.

☐ PRIORITIZE OPENNESS AND FLEXIBILITY

As agentic workloads evolve, prioritize open software and ecosystems that will support new model and tool adoption.

In practice: Choose tools that can work across multiple models and infrastructure environments.

☐ PREPARE YOUR PEOPLE AND PROCESSES

Treat agentic AI as an operating-model change, not just a technology purchase. Plan for the skills and change management you'll need alongside your infrastructure.

In practice: Assign responsibility for monitoring agents and handling exceptions or escalations.

BUILDING FOR AN AGENTIC FUTURE

Agents change the shape of demand. A single goal can fan out into many concurrent tasks – planning steps, tool calls, data queries, policy checks, and handoffs to other agents – making it critical to set up infrastructure appropriately. Enterprises can implement a strong foundation for agentic AI with a compute strategy that treats CPUs, accelerators, and networking as one coordinated system rather than a set of separate purchases: AMD EPYC™ server CPUs for orchestration, data preparation, and agent execution; AMD Instinct™ GPUs for large-scale inference; and AMD Pensando™ DPUs to connect and scale the environment.

That coordination is the destination, not the starting point. AMD Solution Blueprints include ready-to-deploy starting points for agent workflows, and AMD Inference Microservices provide prebuilt, hardware-optimized inference containers – so early deployments can focus on the use case and the guardrails around it rather than on assembling the stack beneath it.

Where that foundation sits matters as much as what it's made of. Agents work where the data and the systems already are, and AMD EPYC™ server CPUs are broadly available across major cloud providers and on-premises platforms, so agent activity can run close to the applications it acts on – wherever sovereignty, latency, and cost requirements point – without re-platforming. The same pattern reaches the endpoint, where AMD Ryzen™ AI agent PCs support local agents that handle planning, tool use, and multi-step tasks on the device itself, under enterprise governance and alongside data center deployments. Infrastructure should follow the work, rather than forcing every task onto one class of processor.

Venue flexibility only holds if the software stack allows it. A single workflow may call several models and tools in one run and swap them as better options appear, which makes openness a practical requirement. AMD Inference Microservices and AMD Solution Blueprints build on open-source models, frameworks, Kubernetes, Helm charts, and OpenAI-compatible APIs, with the AMD ROCm™ software stack supporting widely adopted tools including PyTorch, Hugging Face, and vLLM. Components can be substituted and blueprints extended as the ecosystem moves, without redesigning the platform underneath.

Agentic AI is still evolving, but those that begin planning now to preserve flexibility as the technology matures may be better positioned to adapt their infrastructure as quickly as agent adoption grows.

PREPARE YOUR INFRASTRUCTURE FOR AGENTIC AI

LEARN MORE

[1]McKinsey & Company, The state of AI in 2026: On the road to ROI, 2026.
[2]AMD, AI Is Transforming Every Industry, 2026.
[3]IDC, White Paper, sponsored by AMD, The Future of AI Inferencing Lies in Open, Flexible Infrastructure Stacks, doc #US54516826, July 2026.
[4]451 Research S&P, AI demands shift enterprise IT priorities – Highlights from Digital Pulse, Cloud and Data Center, 2026.

© 2026 Advanced Micro Devices, Inc.

