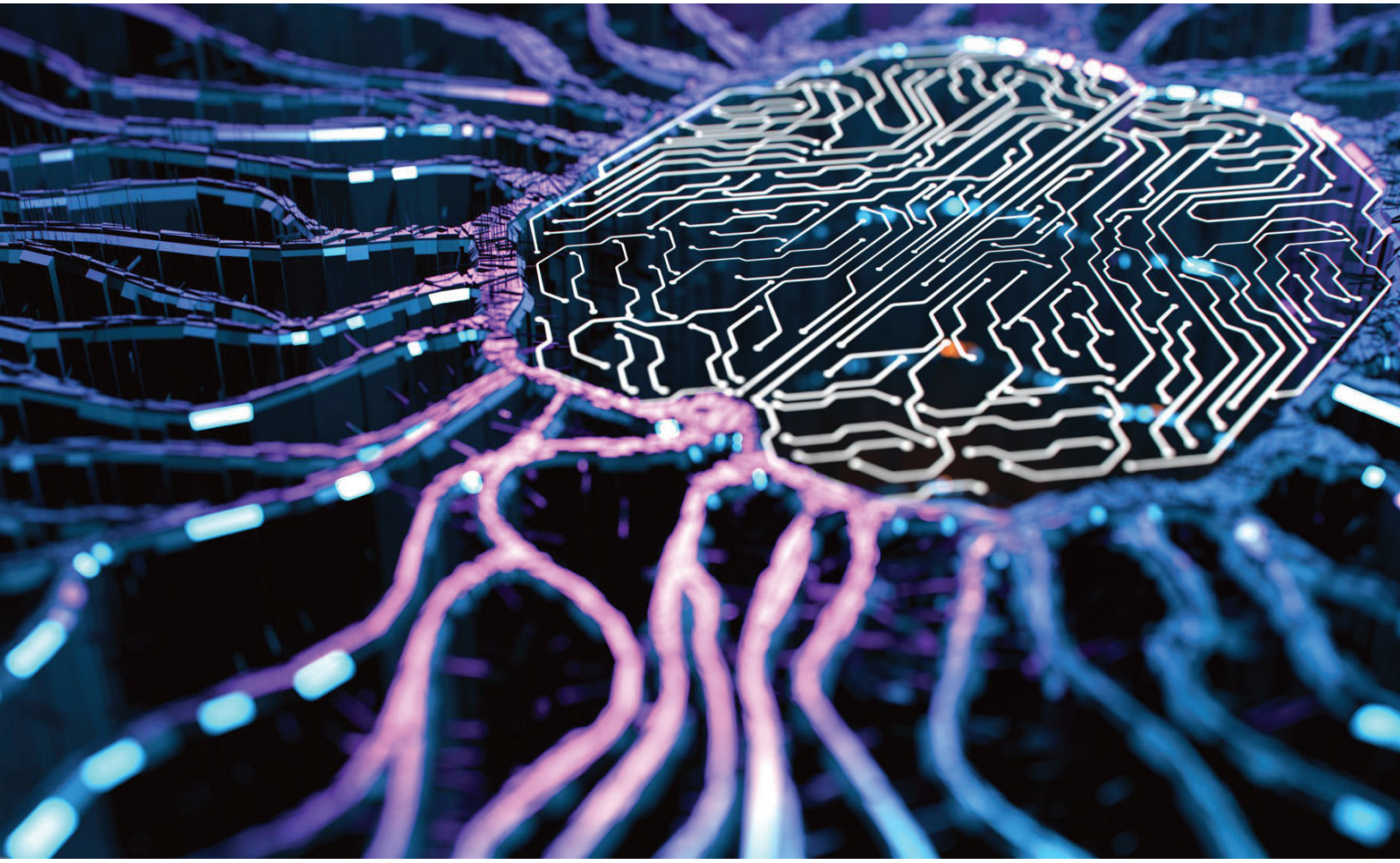
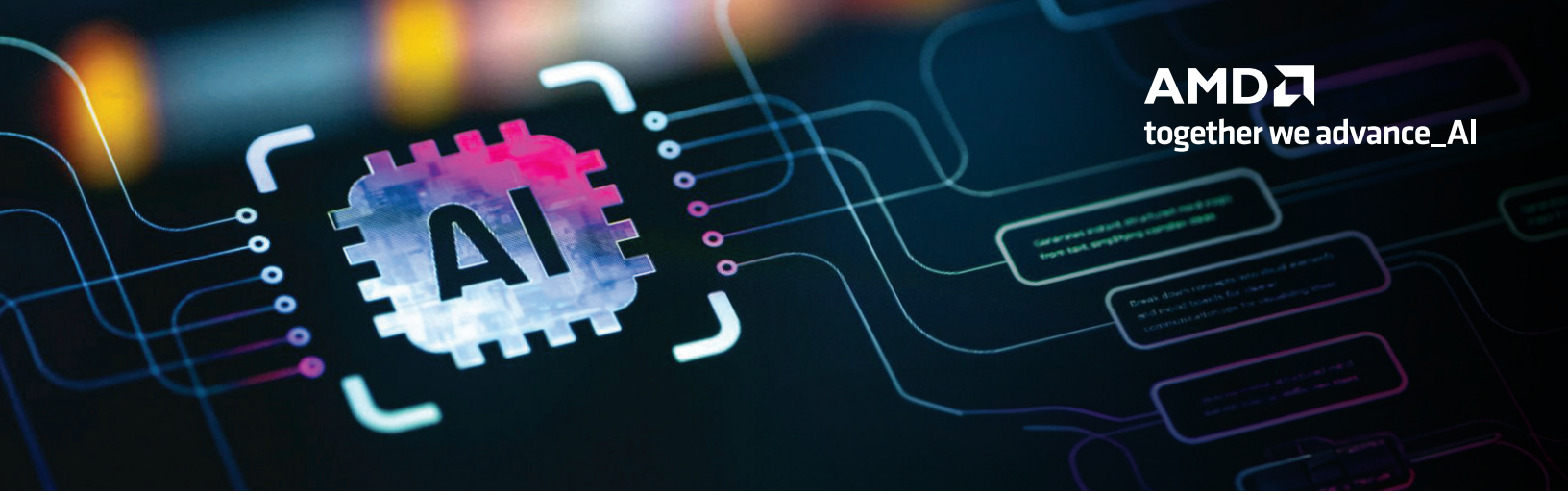


THE AGENTIC AI ERA AND THE CPU INFRASTRUCTURE THAT WILL POWER IT

WHITE PAPER



EXECUTIVE SUMMARY

Artificial intelligence has moved through a decisive inflection point. For the past several years, generative AI, through chatbots, has reshaped how people work by answering questions, drafting content, synthesizing information, and accelerating individual productivity. A more consequential period has begun: the Agentic AI Era. It can receive a goal, decompose it into tasks, dispatch other autonomous agents to execute those tasks across the digital world, evaluate their outputs, and act on the results, all without continuous human direction.

The infrastructure implications of enabling and hosting agentic AI are becoming profound. When every user can deploy dozens of autonomous agents simultaneously—and those agents further deploy hosts of *other* agents—it's conceivable that machine-to-machine traffic could rival, and in some cases exceed, human internet activity. Each of those agents can search the web, query databases, call APIs, authenticate to enterprise applications, read and write communications, and coordinate across systems. That work is inherently general-purpose and typically runs on CPUs.

The GPU remains indispensable as the AI brain, executing model training, reasoning, and content generation. But as AI systems evolve from generating answers to executing work, the CPU has re-emerged as a decisive infrastructure asset in the agentic era. Organizations scaling from single-agent pilots to multi-agent orchestration will quickly discover the limits of their infrastructure. Thread and core count, memory bandwidth and capacity, I/O performance, and per-core performance can determine how far and how fast agentic systems can scale. CPU demand is expected to grow not only with the quantity of agents but with the increasing volume of actions triggered by ever-more-capable AI models.

5th Generation AMD EPYC™ Server CPUs, based on "Zen 5" microarchitecture, scale to as many as 192 cores and deliver the throughput, versatility, and per-core performance that agentic workloads demand. With a comprehensive portfolio of CPUs, GPUs/accelerators, networking solutions, storage technologies, and software platforms, AMD products address the growing diversity and challenges of AI adoption and use, including agentic AI.

This paper explains what agentic AI is, how it works, what it requires from infrastructure, and how AMD and AMD EPYC™ Server CPUs provide a compelling foundation for organizations adopting and building it.

THE AI EVERYONE KNOWS—AND THE NEW AI THAT HAS ARRIVED

Imagine an AI agent built for software coding that is tasked by a human to fix an issue with software in production. The agent pulls logs from monitoring systems, searches recent commits, queries internal documentation, writes and tests multiple candidate fixes, runs validation checks, and escalates if human judgment is required. A few of those steps launch additional agents—one to analyze test failures, another to check security implications, another to prepare a deployment plan.

None of this work represents AI inference by a single model. It is a cascade of general-purpose tasks—retrieval, orchestration, tool execution, testing, and validation— typically executed concurrently across enterprise systems. This execution pattern—where many parallel actions coordinate toward a single outcome—applies to use cases such as customer support agents handling live cases or IT operations agents diagnosing incidents.

The generative AI chatbot systems that entered mainstream enterprise use beginning in 2022 represented a genuine technological breakthrough. For the first time, non-technical users could engage with AI systems through natural language, receiving substantive responses to complex questions without writing a single line of code. Generative AI accelerated research, compressed drafting cycles, improved software development velocity, and made powerful analytical capabilities broadly accessible.

Generative AI is a reactive system that responds only when prompted. A user composes the prompt, and the model generates a response. The user reads that response and might refine the prompt, add new context, or ask a follow-up question. Every step requires human initiation and action. One skilled professional using generative AI daily might be able to do the work of more than one or the equivalent of a highly experienced professional. That leverage is valuable, but it is bounded.

Agentic AI breaks that boundary. Rather than responding to prompts, agentic systems accept goals and pursue them autonomously. A user describes an objective—to research a market, monitor a competitor, find the lowest-cost shipping route that meets a set of criteria, compile a regulatory summary across three jurisdictions—and the agentic system goes to work. It does not ask for clarification at every step. It decomposes the goal into tasks, assigns those tasks to other agents, evaluates their outputs, and continues until the objective is met or it reaches a decision point that requires human judgment. And these agents are not contained within a single system or rack. They operate as a distributed fabric across the data center, interacting with multiple services, systems, and compute resources simultaneously.

This is a structural change in what AI does and how it operates, and therefore also what it demands from the infrastructure that supports it.

AGENTIC AI: MULTIPLYING AN EFFECTIVE WORKFORCE

The distinction between generative and agentic AI is operational: generative AI produces output—text, images, code, analysis—that a human can then act upon, while agentic AI produces action directly. It does not generate a report of the best available flight options; it searches for them, evaluates them against specified criteria, books the one that qualifies (quite possibly after first checking with the human for whom it is acting), and confirms the transaction. The human defines the goal, and the AI agent achieves it.

This distinction changes how organizations apply AI in their operations. Generative AI can multiply individuals' capabilities, while agentic AI can multiply the overall effectiveness of a workforce. An organization that deploys agentic AI enables an expanded digital workforce operating continuously across its systems, data sources, and external services, pursuing objectives that previously required human time and attention at every step.

While agents operate autonomously, they often only activate in response to a human request at first. They may also operate on a defined schedule, respond to a data event, or launch because another agent determined they should. The latter is what creates the compounding scale that defines the agentic era. Agents can trigger other agents, which trigger still others, creating complex, self-organizing workflows to achieve a single high-level goal.

As organizations move beyond initial deployments, the difference in scale becomes more pronounced. A single user can now initiate dozens of simultaneous autonomous processes, each operating across multiple systems to complete a portion of a larger objective. Agentic AI acts on information and intelligence in the world rather than simply producing output for humans to act on.

HOW AI AGENTS WORK

Agentic AI workloads operate through orchestration and parallel execution. A single user request can decompose into dozens of parallel agents, each handled by a different agent. One agent may query an internal database, while another searches public web sources, while a third accesses an application, and a fourth transforms data or validates results.

These tasks execute concurrently and return their outputs to an orchestrating agent that evaluates progress, determines next steps, or launches additional sub-agents to continue the work. As workflows expand, agents can trigger other agents across multiple layers of execution, creating highly dynamic and distributed compute activity across enterprise systems.

Not all agent tasks behave the same way. Some are latency-sensitive, making sequential decisions, handling authentication, or coordinating next steps. Others are throughput-heavy, processing documents, running validation checks, executing transformations, or testing outcomes at scale.

Agentic AI workflows can expand dynamically as work progresses. A single request may trigger multiple layers of autonomous tasks executing in parallel across enterprise systems, data sources, and applications. Agents execute wherever the required data, tools, or applications reside and can spread their activity across many servers and compute pools rather than occupying a single rack or cluster.

This differs from traditional enterprise applications, which typically follow structured, predictable workflows with bounded compute demand. Agentic AI workflows are more dynamic. As agents evaluate results and initiate additional tasks, infrastructure demand can grow in real time, turning a single request into many discrete computational tasks.

THE AI INFRASTRUCTURE EXPLOSION RESULTING FROM AGENTIC AI

The internet today supports billions of simultaneous users generating a continuous stream of requests—searches, logins, database queries, application interactions, and API calls. That activity defines the scale of modern enterprise computing. Agentic AI fundamentally changes that frame of reference.

In an agentic environment, a single user may direct dozens of autonomous processes operating simultaneously on their behalf. Each agent generates real computational activity: retrieving data, calling APIs, executing code, interacting with enterprise systems, and coordinating additional tasks. As organizations move from single-agent pilots to orchestrated multi-agent environments, infrastructure demand compounds quickly.

This creates a very different scaling dynamic from traditional generative AI workloads. Infrastructure leaders planning for agentic AI shouldn't assume linear growth in compute demand as deployments scale, increase concurrency, and gain autonomy.

THE AI "BRAIN" AND THE AI "WORKFORCE"

The architectural relationship between GPUs and CPUs is one of the most important—and most misunderstood—aspects of agentic AI infrastructure. GPUs earned their central role in AI because they excel at highly parallel, numerically intensive computation. Training and running large language models and generative AI require exactly the kind of matrix operations that GPUs were designed to accelerate.

The introduction of new, agentic workflows creates increasing demand for CPUs. As GPUs become more powerful and generate more tokens per second, they also drive more downstream activity: retrieval calls, tool invocations, orchestration steps, validation checks, and interactions with enterprise systems. Those tasks execute on general-purpose infrastructure, increasing demand for the CPU alongside the GPU. Faster accelerators do not replace CPU demand; they multiply it. This is because agentic workloads differ fundamentally from model inference itself. Searching the web, authenticating to an enterprise application, calling APIs, reading an email thread, querying databases, executing code, parsing results, and coordinating workflows are heterogeneous, often sequential tasks that depend on the flexibility of general-purpose computing.

Agentic AI also introduces a wider range of infrastructure requirements. Some stages benefit from strong per-core performance for fast orchestration and decision-making. Others require high core and thread density to support thousands of concurrent tasks. As autonomous workflows scale, infrastructure must support both simultaneously across diverse and constantly shifting workloads.

In traditional generative AI systems, CPUs primarily act as host processors for GPUs, handling scheduling, data movement, and system coordination, while the GPU performs the AI model's computational work. Agentic AI introduces a distinct execution layer for agent logic itself, supporting retrieval, orchestration, tool execution, integration, testing, validation, and data services. This layer operates alongside the generative AI layer and is naturally suited to CPU architecture.

As enterprises scale from single-agent pilots to orchestrated multi-agent production systems, CPU demand can accelerate sharply, in some cases faster than the GPU footprint itself. In these environments, CPU core count determines how many agents can execute simultaneously without resource contention. Throughput becomes the governing metric, driven by the number of cores available to support large volumes of concurrent, heterogeneous workloads.

AGENTIC AI IN THE ENTERPRISE: ADOPTION, PRIORITIES, AND CONCERNS

Enterprise adoption of agentic AI is moving from experimentation to production faster than most predicted. According to McKinsey's State of AI 2025 survey, 62% of organizations are experimenting with AI agents, with 23% actively scaling agentic AI systems somewhere within their enterprise. Adoption is accelerating across core business functions, including IT operations, software development, customer support, finance, procurement, and supply chain optimization. These are mission-critical domains where autonomous agents are augmenting and, in some cases, executing work that previously required continuous human involvement.

As deployments expand, organizations are prioritizing security, governance, consistency, and predictability of agent behavior, integration with existing enterprise systems, and the infrastructure required to operate agentic workloads at scale. These priorities are informing deployment strategies and driving architectures that embed observability, governance, resilience, and operational control alongside agent performance.

Scaling from pilot to production is also exposing infrastructure limitations. Orchestrating hundreds or thousands of simultaneous agents requires significantly higher levels of core density, memory bandwidth, I/O throughput, and power efficiency to meet the business case requirements.

AMD AND THE AGENTIC FUTURE

Because agentic AI is heterogeneous, no single processor characteristic defines the ideal infrastructure platform. Agentic AI requires a distributed architecture that combines high-frequency CPUs, dense-core CPUs, GPUs, storage, networking, and software.

AMD EPYC Server CPUs are purpose-built for the scale, throughput, and versatility that agentic AI workloads demand. The 5th Generation AMD EPYC 9005 Series scales to 192 cores and 384 threads per socket, offering the core density required to support large numbers of concurrent agent tasks. In multi-agent environments, core and thread counts directly influence how many autonomous processes can execute simultaneously without queuing or resource contention.

AMD EPYC Server CPUs also deliver significant per-core performance through the AMD "Zen 5" microarchitecture, with instruction-level and pipeline optimizations designed for dynamic, heterogeneous workloads such as web requests, database queries, code generation, API orchestration, and conditional branching. These workloads benefit from the broad instruction execution efficiency that "Zen 5" provides. They run natively across on-premises data centers, major public cloud environments, and hybrid environments. This consistency gives enterprises the flexibility to deploy agentic AI where governance, latency, cost, and data sovereignty requirements demand, without major architectural changes. Because agentic AI commonly operates within the existing x86 ecosystem, organizations can extend it across the same platforms, applications, operating systems, and middleware that already power the enterprise.

As agentic AI scales, CPU demand may grow beyond what many existing infrastructure plans were designed to support. AMD designed the EPYC roadmap to scale with the growing compute demands of agentic AI. The 6th Generation EPYC Server CPUs, codenamed "Venice," are expected in the second half of 2026 and will expand the platform across the dimensions that matter most for large-scale agentic workloads. "Venice" scales to 256 cores and 512 threads per socket on the "Zen 6" microarchitecture and is expected to deliver up to 70% greater compute performance, targeting double the memory bandwidth per socket, approaching 1.6 million terabytes per second (1.6 TB/s). AMD EPYC Server CPUs deliver the performance and density enterprises need today for agentic AI workloads, while "Venice" is designed to extend the platform for the larger-scale deployments still ahead.

EXPLORE AMD EPYC FOR AGENTIC AI

The agentic era demands infrastructure built for the general-purpose execution of real-world tasks as enterprises scale to multi-agent orchestration. AMD EPYC Server CPUs provide the thread and core count, per-core performance, memory bandwidth and capacity, I/O performance, and platform leadership the enterprise needs today, from pilot to production, across the world's leading cloud and data center environments.

LEARN MORE

Learn how AMD EPYC Server CPUs can accelerate your organization's agentic AI infrastructure strategy, visit amd.com/agenticai or contact your AMD representative to schedule a technical architecture review tailored to your deployment goals.

[1] McKinsey & Company, November 5, 2025, "The state of AI in 2025: Agents, Innovation, and Transformation", https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai