AMD

5 WAYS CPUs ARE DRIVING *FAST, EFFICIENT AGENTIC AI*

Here's how CPUs are enabling AI agents in the data center

As AI evolves into the agentic era with always-on workflows that plan, reason, and act, the CPU is stepping back into the spotlight as the system's orchestration engine. While GPUs handle the heavy AI math, the right CPU determines how quickly agents can juggle tools, memory, I/O, security, and real-time decisions at scale.

For five generations, AMD has been the industry standard in CPU leadership. AMD EPYC™ Server CPUs are purpose-built to run efficient agentic AI alongside all your other data center workloads. Read on to see the key CPU features and performance metrics that make a difference.

2.1X

Higher performance per core with a 5th Gen AMD EPYC™ Server CPU-based system compared to NVIDIA Grace™ CPU Superchip-based systems.[1]

1 HIGH CORE AND THREAD DENSITY

With more threads, your CPUs can manage more agents

Agentic platforms may run thousands of agent threads concurrently. CPUs must be ready to prepare data, orchestrate tasks, and handle microservices on a massive scale. The more cores and threads in a CPU, the more agents it can support.

In addition to high core counts, AMD EPYC Server CPUs feature Simultaneous Multithreading (SMT), which executes two threads per core.

- 5th Gen AMD EPYC™ Server CPUs currently have the highest available core count in x86 server processors with 192 cores (up to 384 threads).
- The upcoming 6th Gen AMD EPYC Server CPUs will offer up to 256 cores (up to 512 threads).

2 MULTIPLE CPU PERFORMANCE OPTIONS

Match a broad range of agentic workloads

You may need different CPU configurations to support different agentic workloads. High-frequency CPUs can maximize performance per core for latency-sensitive agents and fast orchestration logic. High throughput per core can make your agents highly efficient when handling multiple, simultaneous tasks.

With AMD EPYC Server CPUs, you can choose from a wide range of core counts and frequencies to optimize for specific goals.

- AMD EPYC™ 9005 Server CPUs deliver 2.3X the performance when compared to leading competitive offerings.[2]

AMD

3 LARGE MEMORY CAPACITY AND BANDWIDTH

Support long-context pipelines and rapid data retrieval

AI agents typically manage complex, multi-step tasks, often over an extended period. CPUs with large memory capacity and bandwidth can support agentic workflows involving vector databases, retrieval-augmented generation (RAG) pipelines, and other memory-intensive tasks.

- 5th Gen AMD EPYC Server CPUs support 12 DDR5-6400 memory channels (up to ~600 GB bandwidth) and up to 384 MB L3 cache.
- The upcoming 6th Gen AMD EPYC Server CPUs will support up to 2.6X the memory bandwidth (1.6 TB/s) over the previous generation.

4 FAST INPUT/OUTPUT (I/O)

Avoid bottlenecks with high-speed I/O

Agents require constant, high-speed interaction between processors, APIs, databases, and other network resources. Fast interconnect, including the latest generations of PCIe®, helps CPUs connect to data, tools, NVMe storage drives, and host nodes without delays.

- 5th Gen AMD EPYC Server CPUs offer up to 160 PCIe Gen 5 lanes.
- The upcoming 6th Gen AMD EPYC Server CPUs will offer ~2X more I/O bandwidth over the previous generation with PCIe Gen 6.

5 HIGH PERFORMANCE PER WATT

Maximize power efficiency with a high number of threads per watt

While agentic workloads are booming, data center architects are struggling to manage power consumption and cooling. AMD EPYC Server CPUs are optimized for performance per watt and threads per watt, which helps IT teams add capacity for agents within space and power constraints.

- A system with a 5th Gen AMD EPYC Server CPU delivers a 2.26X uplift on SPECpower®, measuring operations per watt against a comparable NVIDIA Grace Superchip-based system.[3]
- AMD EPYC 9005 Server CPUs offer energy-efficient server solutions with exceptional performance. For example, 2P servers using AMD EPYC™ 9965 CPUs deliver 1.6X more performance per system watt than Intel® Xeon® 6980P CPUs running SPECpower_ssj®2008.[4]

EXTENSIVE RANGE TO FIT YOUR *AGENTIC AI GOALS*

Agentic AI will require CPUs to do more than ever, and there's no one-size-fits-all solution. AMD offers a broad and flexible CPU portfolio, x86 compatibility, and an open-source ecosystem to minimize software friction.

By balancing core and thread count, frequency, memory, I/O, and performance per watt, you can add capacity for agents while managing costs.

Learn more about CPUs for agentic AI.

AMD

1. 9xx5-284: SPECrate®2017_int_base / Core comparison based on published and estimated results as of 06/01/2025.

   Configuration:
   2P AMD EPYC™ 9755 (2840 SPECrate®2017_int_base, 256 Total Cores / 128 Cores per CPU, https://www.spec.org/cpu2017/results/res2025q2/cpu2017-20250407-47519.html) versus 2P NVIDIA Grace™ CPU Superchip (estimated 740 SPECrate®2017_int_base, 144 Total Cores / 72 Cores per CPU, as per NVIDIA claim: https://developer.nvidia.com/blog/inside-nvidia-grace-cpu-nvidia-amps-up-superchip-engineering-for-hpc-and-ai/). SPECrate®2017_int_base / Core is calculated as (2840 / 256) / (740 / 144) = 2.16x. OEM published SPEC scores will vary based on system configuration. SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

2. 9xx5-002F: SPECrate®2017_int_base comparison based on published scores from www.spec.org as of 12/11/2025. 2P AMD EPYC 9965, 192C, 500W, $11988 USD, 3230, 6.460, 0.269, https://www.spec.org/cpu2017/results/res2025q2/cpu2017-20250324-47086.html 2P Intel Xeon 6780E, 144C, 330W, $8513 USD, 1420, 4.303, 0.167, https://www.spec.org/cpu2017/results/res2025q4/cpu2017-20251020-50067.html SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. AMD CPU prices as of 12/11/2025.

3. 9xx5-217: As of May 29, 2025, a 2P AMD EPYC™ 9755 system (128 cores) delivers a 2.26x SPECpower_ssj® 2008 overall ssj_ops/watt uplift versus a 2P NVIDIA Grace™ CPU Superchip system (144 cores), and a 2P AMD EPYC™ 9965 system (192 cores) delivers a 3.34x uplift versus the same Grace system.

   Configurations:
   2P EPYC 9755: 29,950 overall ssj_ops/watt: https://www.spec.org/power_ssj2008/results/res2024q4/power_ssj2008-20240924-01460.html. 2P NVIDIA Grace Superchip: 13,218 overall ssj_ops/watt: https://www.spec.org/power_ssj2008/results/res2024q3/power_ssj2008-20240515-01413.html. SPEC® and SPECpower_ssj® 2008 are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. Results based on SPECpower_ssj2008 weighted average (100%–10% load).

4. 9xx5-134: SPECpower_ssj® 2008 comparison based on published scores from www.spec.org as of 4/30/2025. 2P AMD EPYC 9965 (35920 ssj_ops/watt, 384 Total Cores, https://spec.org/power_ssj2008/results/res2024q4/power_ssj2008-20241007-01464.html) 2P Intel Xeon 6980P (21679 ssj_ops/watt, 256 Total Cores, https://spec.org/power_ssj2008/results/res2025q2/power_ssj2008-20250324-01511.html) SPEC®, SPEC CPU®, and SPECpower® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.