AMD

5 REASONS THE INDUSTRY IS CHOOSING *AMD EPYC™ 9006 SERIES SERVER CPUs*

World's best server CPUs for cloud, enterprise, and AI[1]

AMD EPYC™ 9006 Series Server CPUs unlock AMD leadership performance, thread density, performance per watt, and total cost of ownership (TCO). See why they're the best choice for demanding workloads across the spectrum.

1 GET THE RIGHT CPU FOR THE JOB

One size does not fit all. AMD EPYC™ 9006 Series Server CPUs come in four purpose-built processor families. Each family includes multiple SKUs optimized for different workloads and infrastructure constraints, so you can target unique workload demands with an optimum mix of thread density, frequency, memory bandwidth, and I/O for your power and budgetary requirements.

AMD EPYC™ 9006 SP7 SERVER CPUs
Leadership performance and density
Up to 256 cores and 512 threads deliver high-density compute for more VMs, vCPUs, and agent capacity per watt and per rack.

AMD EPYC™ 9006 SP8 SERVER CPUs
Enterprise and general-purpose workloads
Eight- to 128-core options with balanced compute, memory, and I/O support enterprise workloads alongside agentic AI while optimizing ROI.

AMD EPYC™ 9006 LP SERVER CPUs
Next-gen AI scale
Faster CPU-GPU communication and LPDDR memory deliver excellent host performance in GPU nodes.

AMD EPYC™ 9006X SP7 SERVER CPUs
HPC and technical computing
3X L3 cache[2] and up to 1.6 TB/s memory bandwidth speed time to insight for electronic design automation (EDA), financial services, HPC, and technical computing.

2 GET LEADERSHIP PERFORMANCE, THREAD DENSITY, AND PERFORMANCE PER WATT

AMD EPYC™ 9006 Series Server CPUs are the highest-performing server CPU (based on SPECrate2017®_int_base),[3] delivering leadership performance and power-efficiency gains, plus up to 1.6 TB/s memory bandwidth.

Up to **1.8X** **the performance of 5th Gen AMD EPYC™ Server CPUs**[4]

Up to **1.5X** **performance per CPU watt**[4]

Up to **1.4X** **performance per CPU dollar**[4]

Compares top-of-stack 6th Gen AMD EPYC™ Server CPUs with 5th Gen EPYC™ 9965 Server CPUs running SPECrate2017_int_base. Please see note 4 for details.

AMD

3 MEET SURGING AGENTIC DEMAND WITH THE RIGHT MIX OF CPUs

Agentic AI spans workloads, from large numbers of agents running in parallel to GPU-intense inference and general-purpose computing — all at a massive scale. AMD EPYC™ 9006 Series Server CPUs give you multiple options for the three essential CPU profiles required for agentic AI.

AGENT SANDBOX CPUs

Run more agents per watt, per dollar, and per rack[5,6,7]

With up to 512 threads per CPU, AMD EPYC™ 9006 Series Server CPUs support more agents.

AI HOST NODE CPUs

Maximize GPU utilization and token generation

At up to 5GHz, AMD EPYC™ 9006 Series Server CPUs deliver low-latency host performance for GPU platforms.

GENERAL-PURPOSE CPUs

Boost the enterprise systems agents use to complete tasks

In a 100 kW rack, AMD EPYC™ 9006 Series Server CPUs deliver up to 3.3X the compute of NVIDIA Vera CPUs.[8]

4 CONSOLIDATE SERVERS WITH LEADERSHIP TCO

Generational improvements in compute density and power efficiency make AMD EPYC™ 9006 Series Server CPUs ideal for consolidating servers and helping to deliver more compute capacity with every rack.

Use up to **77%** **less power**[9]

Run up to **92%** **fewer servers**[9]

Save with up to **40%** **lower TCO over five years**[9]

Estimated consolidation and savings when replacing servers equipped with 2P Intel® Xeon® Gold 6258R CPUs (Cascade Lake) with servers equipped with AMD EPYC™ 9996 Server CPUs. Please see note 9 for details.

5 EXTEND CONFIDENTIAL AI TO GPUs, NETWORKING, AND STORAGE

AMD EPYC™ 9006 Series Server CPUs feature AMD Infinity Guard[10] with AMD Secure Encrypted Virtualization (SEV), including TEE Device Interface Security protocol (TDISP), which can extend trusted domains over I/O to TDISP-enabled endpoints.

Build on a hardware root of trust

AMD Infinity Guard provides a robust set of security features that help complement industry ecosystem partners at the software and system levels.

Help keep data safe while in use

Spin up confidential VMs and containers for training, deployed models, and agentic workflows without workload partitioning or code changes.

The industry's most mature and broad confidential computing ecosystem[11]

AMD SEV is supported on AMD EPYC™ Server CPUs across major cloud service providers and data center operating sytems.

AMD

CHOOSE THE RIGHT CPU FOR THE JOB
WITH AMD EPYC™ 9006 SERIES SERVER CPUs

World's best server CPUs for cloud, enterprise, and AI[1]

With purpose-built designs for agentic AI, GPU racks, HPC, and general-purpose computing, AMD EPYC™ 9006 Series Server CPUs deliver the leadership performance, thread density, power efficiency, and scalability you need to meet the coming waves of new demand for data center computing.

Explore the portfolio

1. EPYC-029E: Claim of "best CPU for cloud, enterprise, and AI" is based on AMD's assessment of the best overall combination of projected performance, thread density, enterprise workload results, AI benchmark performance, energy efficiency, memory subsystem capabilities, I/O capabilities, process technology, and built-in security features among evaluated general-purpose enterprise server CPUs as of July 23, 2026. "Best" reflects the strongest overall combination across the evaluated cloud, enterprise, and AI criteria and does not imply leadership in every individual workload or metric. Products not available for comparable testing or not positioned as general-purpose enterprise server CPUs were not included.
2. AMD EPYC™ 9006X SP7 has 3X L3 cache compared to AMD EPYC™ 9686F SP7.
3. 9xx6-004A: SPECrate®2017_int_base comparison based on AMD estimates for top-of-stack 2P 6th Gen AMD EPYC™ 9996 estimated SPECrate2017_int_base score is 4900 (256C, 600W Default CPU Power, $14904 1ku).

   Highest overall SPECrate2017_int_base score as of July 22, 2026, is 3240 2P AMD EPYC™ 9965 (192C, 500W TDP, $11988 1ku) powered server, https://spec.org/cpu2017/results/res2026q2/cpu2017-20260518-51509.html.

   Highest published 2P Intel Xeon 6980P SPECrate2017_int_base score as of July 22, 2026, is 2510 (128C, 500W TDP, $13955 1ku), https://spec.org/cpu2017/results/res2025q4/cpu2017-20251126-50522.html.

   2P NVIDIA Vera performance based on AMD internal estimates as of July 22, 2026, estimated SPECrate2017_int_base score is 1459, (88C, 450W TDP, pricing not published), https://www.amd.com/content/dam/amd/en/documents/solutions/ai/methodology-description.pdf).

   ARM AGI performance based on AMD Internal estimates as of July 22, 2026, estimated SPECrate2017_int_base score is 1944 (136C, 300W TDP, pricing not published).

   Preliminary performance estimates based on AMD engineering projections or measurements as of July 22, 2026, and subject to change. Pricing not published for NVIDIA Vera, Intel pricing from ark.intel.com. Phoronix testing for NVIDIA Vera at https://www.phoronix.com/review/nvidia-vera-benchmarks/. SPEC® and SPECint® are registered trademarks of Standard Performance Evaluation Corporation. Learn more at spec.org. Starting with the 6th Gen AMD EPYC™ server processor family, AMD uses Default CPU Power to describe processor power consumption, succeeding AMD's historical TDP reference. Default CPU Power reflects total power consumed across the processor's compute and I/O dies for the stated performance target. Default CPU Power and TDP may both serve as processor power references for product comparison, platform planning, and performance-per-watt analysis.
4. 9xx6-001A: SPECrate®2017_int_base comparison based on AMD internal estimates for top-of-stack 2P 6th Generation EPYC™ CPU and 5th Gen EPYC™ measurements and pricing as of July 22, 2026. Preliminary performance estimates based on AMD engineering projections or measurements for respective product launch dates and subject to change.

   2P AMD EPYC™ 9996, 512C total, 600W Default CPU Power, $14904, 16x64GB 12800 MT/s MRDIMM, AOCC 6.0, performance determinism, SPECrate2017_int_base 4900, SPECrate2017_int_base/CPU W 4.083, SPECrate2017_int_base/CPU $ 0.329

   2P AMD EPYC™ 9965, 384C total, 500W TDP, $ 11988, 12x64GB DDR5-6400, AOCC 5.0, performance determinism, SPECrate2017_int_base 2740, SPECrate2017_int_base/CPU W 2.740, SPECrate2017_int_base/CPU $ 0.229

   SPEC® and SPECint® are registered trademarks of Standard Performance Evaluation Corporation. Learn more at spec.org. Starting with the 6th Gen AMD EPYC™ server processor family, AMD uses Default CPU Power to describe processor power consumption, succeeding AMD's historical TDP reference. Default CPU Power reflects total power consumed across the processor's compute and I/O dies for the stated performance target. Default CPU Power and TDP may both serve as processor power references for product comparison, platform planning, and performance-per-watt analysis.
5. 9xx6-014: Comparison based on published top-of-stack core counts and CPU W, Default CPU Power for 6th Gen EPYC™, and TDPs for 5th Gen EPYC™, Intel® Xeon®, and NVIDIA Vera for estimated highest agents/CPU W across 6th Gen AMD EPYC™ 256C at 400W, AMD EPYC™ 9965 (500W TDP), NVIDIA Vera (450W TDP), ARM AGI (300W TDP), Intel Xeon 6980P (500W TDP), and Intel Xeon 6990E+ (450W TDP) powered servers as of July 22, 2026. 2 threads per core (SMT). 1 thread per core for ARM AGI.

   Agent counts are estimates derived from available CPU thread resources used as a proxy under a consistent theoretical workload. Actual agent capacity and throughput will vary based on workload, model, memory, software, orchestration, and system configuration.

   Starting with the 6th Gen AMD EPYC™ server processor family, AMD uses Default CPU Power to describe processor power consumption, succeeding AMD's historical TDP reference. Default CPU Power reflects total power consumed across the processor's compute and I/O dies for the stated performance target. Default CPU Power and TDP may both serve as processor power references for product comparison, platform planning, and performance-per-watt analysis. Intel Xeon TDP from ark.intel.com. NVIDIA Vera TDP from https://developer.nvidia.com/blog/nvidia-vera-cpu-sets-a-new-standard-for-agentic-workloads-in-ai-factories/. ARM AGI Specifications from https://www.arm.com/products/cloud-datacenter/arm-agi-cpu#Specifications.

6. 9xx6-013: Comparison based on published top-of-stack core counts and 1ku pricing for estimated highest agents/CPU $ and threads/CPU $ AMD EPYC™ 9006 (512 threads), AMD EPYC™ 9005 (384 threads), and Intel® Xeon® 6 (256 threads) SKUs as of July 22, 2026. Threads derived as 2 threads per core (SMT). Intel and Xeon are trademarks of Intel Corporation or its subsidiaries.

   Source: https://www.amd.com/content/dam/amd/en/documents/solutions/ai/methodology-description.pdf

   Agent counts are estimates derived from available CPU thread resources used as a proxy under a consistent theoretical workload. Actual agent capacity and throughput will vary based on workload, model, memory, software, orchestration, and system configuration.

7. 9xx6-012: Based on estimated performance data for NVIDIA, Intel, and AMD EPYC™ server processors for agentic AI, the AMD EPYC™ 9996 provides the most agents per rack at a 100 kW power envelope per rack.

   Compared to the cores per rack of NVIDIA Vera (88c) powered server racks:
   - The AMD EPYC™ 9996 (256C) provides 2.08x the cores (and threads with SMT) per rack
   - The AMD EPYC™ 9965 (192C) provides 1.86x the cores (and threads with SMT) per rack
   - The Intel Xeon 6980P (128C) provides 1.24x the cores (and threads with SMT) per rack

   Source: https://www.amd.com/content/dam/amd/en/documents/solutions/ai/methodology-description.pdf

   Agent counts are estimates derived from available CPU thread resources used as a proxy under a consistent theoretical workload. Actual agent capacity and throughput will vary based on workload, model, memory, software, orchestration, and system configuration.

8. 9xx6-021: Rack-level general purpose performance comparison within a 100 kW rack power constraint based on AMD analysis as of July 2026 comparing 2P AMD EPYC™ "Venice" 9996 (512 total cores) vs 2P NVIDIA Vera (176 total cores, Olympus). Performance reflects geometric mean across 6 workloads: estimated SPECrate® 2017_int_base, Server-Side Java® Multi-JVM Max jOPS, Web serving (NGINX/WRK), Key-Value store (Redis), In-memory caching (Memcached), and Relational DB (TPROC-C). All platforms use 2-socket nodes.

   System manufacturers may vary configurations, yielding different results. See https://www.amd.com/content/dam/amd/en/documents/solutions/ai/methodology-description.pdf for full derivations and assumptions.

   TPROC-C workload is an open-source workload derived from TPC-Benchmark™ Standard, and as such is not comparable to published TPC-C™ results, as the results do not comply with the TPC-C Benchmark Standard. TPC, TPC Benchmark and TPC-C are trademarks of the Transaction Processing Performance Council.

   Preliminary performance estimates based on AMD engineering projections or measurements as of July 22, 2026 and subject to change. SPEC® and SPECint® are registered trademarks of Standard Performance Evaluation Corporation. Learn more at spec.org.

9. 9xx6TCO-001: This scenario contains many assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. The AMD Server & Greenhouse Gas Emissions TCO (total cost of ownership) Estimator Tool - version 1.57, compares the selected AMD EPYC™ and Intel® Xeon® CPU based server solutions required to deliver a TOTAL_PERFORMANCE of ~397,000 units of SPECrate®2017_int_base performance as of July 20, 2026. This analysis compares a 2P AMD 256 core (512 total cores) EPYC™ 9996 powered server with an estimated SPECrate2017_int_base score of 4900; versus legacy 2P Intel Xeon 28 core (56 total cores) Gold_6258R based server with a SPECrate2017_int_base score of 397, https://spec.org/cpu2017/results/res2020q3/cpu2017-20200915-23981.pdf

   Environmental impact estimates made leveraging data from the 2025 International Country Specific Electricity Factors and can be found at https://www.carbondi.com/#electricity-factors/ and the US EPA Greenhouse Gas Equivalencies Calculator used in this analysis was sourced on 09/04/2024 and can be found at https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator.

   For additional details, see https://www.amd.com/ en/legal/claims/epyc.html#q=9xx6TCO-001

10. GD-183A: AMD Infinity Guard features vary by EPYC™ Processor generations and/or series. Infinity Guard security features must be enabled by server OEMs and/or Cloud Service Providers to operate. Check with your OEM or provider to confirm support of these features. Learn more about Infinity Guard at http://www.amd.com/en/products/processors/server/epyc/infinity-guard.html.

11. EPYC-056: Confidential Computing on EPYC™ is enabled by the SEV security feature, which was introduced with 1st Generation EPYC™ in 2017. 2nd Gen EPYC™ powered the first confidential computing cloud instance in Google Cloud in 2020. EPYC™: powers the highest number of confidential VM options available on all major CSP; Supports both host and guest in the Linux® Kernel; Is available on all major Linux Distributions; Has support on VMware®; supports confidential containers.

    EPYC-069: Agent counts are estimates derived from available CPU thread resources used as a proxy under a consistent theoretical workload. Actual agent capacity and throughput will vary based on workload, model, memory, software, orchestration, and system configuration.

    EPYC-070: Starting with the 6th Gen AMD EPYC™ server processor family, AMD uses Default CPU Power to describe processor power consumption, succeeding AMD's historical TDP reference. Default CPU Power reflects total power consumed across the processor's compute and I/O dies for the stated performance target. Default CPU Power and TDP may both serve as processor power references for product comparison, platform planning, and performance-per-watt analysis.