AMD
together we advance_AI

AMD EPYC™ 9006 SERIES SERVER CPUs

WORLD'S BEST SERVER CPU PORTFOLIO FOR THE AGENTIC AI ERA

From agentic AI to cloud, enterprise, and HPC, the broad AMD EPYC™ 9006 Series Server CPU portfolio extends AMD leadership in performance, thread density, performance per watt, and total cost of ownership (TCO). With the best server CPUs for the agentic era,[1] you can scale more agents with dense compute,[2] maximize AI host-node throughput, and power the general-purpose tools agents use to complete tasks.

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

AMD EPYC™ 9006 SP7 SERVER CPUs

MAXIMIZE PERFORMANCE AND AGENT CAPACITY

- Highest-performing server CPU (based on SPECrate2017®_int_base)[3]
- Up to **256 CORES/512 THREADS** for the highest VM and agent density[2]
- Highest memory bandwidth[4] up to **1.6 TB/s**
- High frequency up to **5 GHz**

AMD EPYC™ 9006 SP8 SERVER CPUs

LEADERSHIP PERFORMANCE PER SYSTEM DOLLAR

- Broad selection with **8- TO 128-CORE** options
- Balanced compute, memory, and I/O design
- Optimized ROI for agentic AI and enterprise workloads

AMD EPYC™ 9006X SP7 SERVER CPUs

ACCELERATE HPC AND TECHNICAL COMPUTING

- **3X L3 CACHE**[5] for EDA, technical computing, financial services, and HPC workloads
- Highest frequency in the portfolio
- Highest memory bandwidth[4] up to **1.6 TB/s**

AMD EPYC™ 9006 LP SERVER CPUs

NEXT-GENERATION AI HOST NODE CPU

- High AI throughput with high frequency
- System power savings with **LPDDR** memory
- Faster CPU-GPU connectivity

AGENTIC AI NEEDS THREE CPU PROFILES

Agentic AI workflows require CPUs to run agents, host GPUs, and execute general-purpose tasks. AMD EPYC™ Server CPUs offer the broadest CPU portfolio for agentic AI.[6]

AGENT SANDBOX CPU

THE MOST AGENTS PER WATT, PER DOLLAR, AND PER RACK[7,8,9]

6th Gen AMD EPYC™ CPU-powered servers provide the most agent capacity compared to competing alternatives.

MOST AGENTS PER CPU WATT[7]

Leading thread density (up to 512 threads) and leadership threads per watt to run more agents and services concurrently

MOST AGENTS PER CPU DOLLAR[8]

Leading thread count and optimized TCO to maximize the number of agents per dollar

MOST AGENTS PER RACK[9]

Leading thread density that helps make over 40,000 agents available in a single rack from leading OEMs[10]

AI HOST NODE CPU

THE HOST CPU TO KEEP AI MOVING

GET MORE FROM GPU ACCELERATORS

High-frequency AI host node CPUs help maximize GPU utilization and token generation

MAXIMIZE CPU-TO-GPU COMMUNICATION

Industry-leading PCIe® Gen 6 I/O subsystem for 2X bandwidth per lane vs. the previous generation[11]

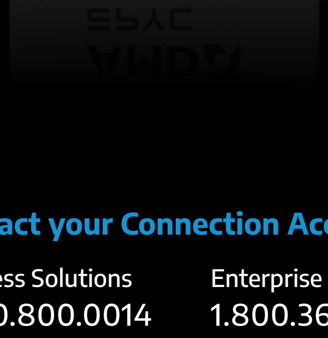
FAST MEMORY BANDWIDTH

Up to 1.6 TB/s per socket memory bandwidth with MRDIMM-12800, supporting large-context agentic AI reasoning

GENERAL-PURPOSE CPU

BEST PERFORMANCE FOR GENERAL-PURPOSE WORKLOADS[3]

- Leading performance per watt based on estimated SPECint® benchmark results[3]
- Leadership TCO through consolidation and enhanced power efficiency



- Leadership performance to orchestrate agentic planning, tool use, and subagents
- Confidential AI with a trusted domain anchored by the CPU and extended via AMD Secure Encrypted Virtualization (AMD SEV) with TDISP-enabled trusted I/O to compatible endpoints[12]

Connection
we solve IT®

Contact your Connection Account Team for more information.

Business Solutions
1.800.800.0014

Enterprise Solutions
1.800.369.1047

Public Sector Solutions
1.800.800.0019

www.connection.com/AMD

1. EPYC-067: Based on AMD internal analysis and modeling of representative infrastructure workloads underlying agentic AI systems as of June 2026. Performance comparisons reflect estimated rack-scale throughput within a modeled 100 kW deployment and include proxy workloads spanning web services, data infrastructure, and orchestration layers. Results are derived from a combination of AMD testing, third-party data, and projections, and may vary based on system configuration, software stack, and workload mix. See AMD Agentic AI rack-scale performance blog and AMD methodology description for details on assumptions, workloads, and measurement approach.
2. 9xx6-005A: vCPU support based on thread count, comparing top-of-stack 6th Gen EPYC™ 9996 CPU with 256 cores (512 threads) and 5th Gen EPYC™ 9965 with 192 cores (384 threads). In comparison, the highest thread count Intel® Xeon® processor is the Intel Xeon 6990E at 288 threads, and the highest thread count NVIDIA CPU is the NVIDIA Vera with 88 cores (176 threads).
   For reference, 6th Gen Intel Xeon CPUs support up to ~0.8X, ARM AGI is expected to support up to ~0.7X, and NVIDIA Vera is expected to support ~0.5X the thread count of 5th Gen AMD EPYC™ CPUs. Intel specifications from ark.intel.com. ARM specifications from https://www.arm.com/products/cloud-datacenter/arm-agi-cpu#Specifications. NVIDIA specifications from https://developer.nvidia.com/blog/nvidia-vera-cpu-delivers-high-performance-bandwidth-and-efficiency-for-ai-factories/.
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. EPYC-032D: AMD EPYC™ 9006 processors support up to 16 channels of 12,800 MT/s MRDIMM memory, providing up to 1,638.4 GB/s of theoretical memory bandwidth per socket. Intel® Xeon® 6 processors and ARM AGI processors support up to 12 channels of 8,800 MT/s MRDIMM memory, providing up to 844.8 GB/s of theoretical memory bandwidth per socket.
5. AMD EPYC™ 9006X SP7 has 3X L3 cache compared to AMD EPYC™ 9686F SP7.
6. EPYC-068: The AMD EPYC™ server CPU portfolio spans the industry's broadest ranges of data center deployments, from general-purpose enterprise, cloud, telecom, SMB, and HPC systems to emerging AI environments including sandboxed agentic AI deployments and GPU head node servers. AMD EPYC™ 6th Generation platforms extend this breadth by uniquely combining high core and thread density of up to 512 threads, advanced memory bandwidth of up to 16 channels of 12.8 GT/s MRDIMM support, next-generation PCIe® Gen 6 connectivity, and select SKUs with boost frequencies up to 5 GHz.
7. 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.
8. 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.
9. 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.
10. 9xx6-015: Comparison based on AMD internal estimates and NVIDIA published specifications as of July 22, 2026. 6th Gen AMD EPYC™ 256C at 400W powered server racks are estimated to deliver up to 1.91X the cores per rack at 26% less power and 2.57X the cores/rack W compared to a NVIDIA Vera (88C, 450W TDP) based server rack. AMD rack specifications based on Supermicro rack configurations and Gigabyte rack configurations (Model H295-D80-CS1). Calculations reflect total sled power including CPU, memory, other system W with 15% loss factor. NVIDIA specifications from: https://www.nvidia.com/en-us/data-center/products/vera-rack/ and https://developer.nvidia.com/blog/nvidia-vera-rubin-pod-seven-chips-five-rack-scale-systems-one-ai-supercomputer/. Supermicro rack specifications provided by Supermicro.
    MGX SMC Rack 1 SMC Rack 2 Gigabyte
    Vera 88C 450W 9965 192c 500W 6th Gen EPYC™ 256C 400W
    # 1U sleds 32 24 42 24
    # CPUs/sled 8 8 4 8
    # CPUs/rack 256 192 144 192
    # core/CPU 88 192 256 256
    Compute Rack Power Total 185104 144038.88 137461.8 269400
    Relative Rack Power 1 0.778 0.743 1.455
    Cores / Rack W 0.122 0.256 0.313 0.182
    Relative Cores / Rack W 1 2.103 2.571 1.499
    Cores/Rack 22528 36864 43008 49152
    Cores/Rack 1 1.636 1.909 2.182
    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.
    Results may vary based specific rack configuration used for comparison, including but not limited to wattage limits.
11. 9xx6-003: PCIe Gen comparison based on PCI-SIG published statements, https://pcisig.com/pci-express-6.0-specification. 2P 6th Gen EPYC™ CPU with 128 lanes of PCIe Gen 6 and 5th Gen EPYC™ with 128 lanes of PCIe Gen 5 as of June 3, 2025. PCIe is a registered trademark of PCI-SIG Corporation.
12. 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.

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.