AMD
EPYC

AMD EPYC™ SERVER CPU ARCHITECTURE AND PERFORMANCE OVERVIEW

WHITEPAPER | 2026

AMD EPYC™ 9006 SERIES SERVER CPUs: ARCHITECTURAL LEADERSHIP FOR THE NEXT-GENERATION DATA CENTER

Data center performance is no longer defined by CPU speed alone. As core counts rise and cloud, enterprise, high-performance computing (HPC) and AI workloads become more demanding, customers are increasingly constrained by rack power, memory bandwidth and capacity, data movement, software compatibility, and operational complexity.

AMD EPYC™ 9006 Series Server CPUs, are designed around those system-level constraints. Built on "Zen 6" and "Zen 6c" architectures, EPYC™ 9006 server CPUs combines up to 256 cores (512 threads), 16 channels of DDR5 memory, JEDEC-standard MRDIMM support for up to 12,800 MT/s, PCIe® Gen 6, CXL 3.1, advanced security and RAS (reliability, availability and serviceability) capabilities, and new power-management technologies. Just as important, it builds on five earlier generations of x86 platform and software enablement, as shown in Figure 1

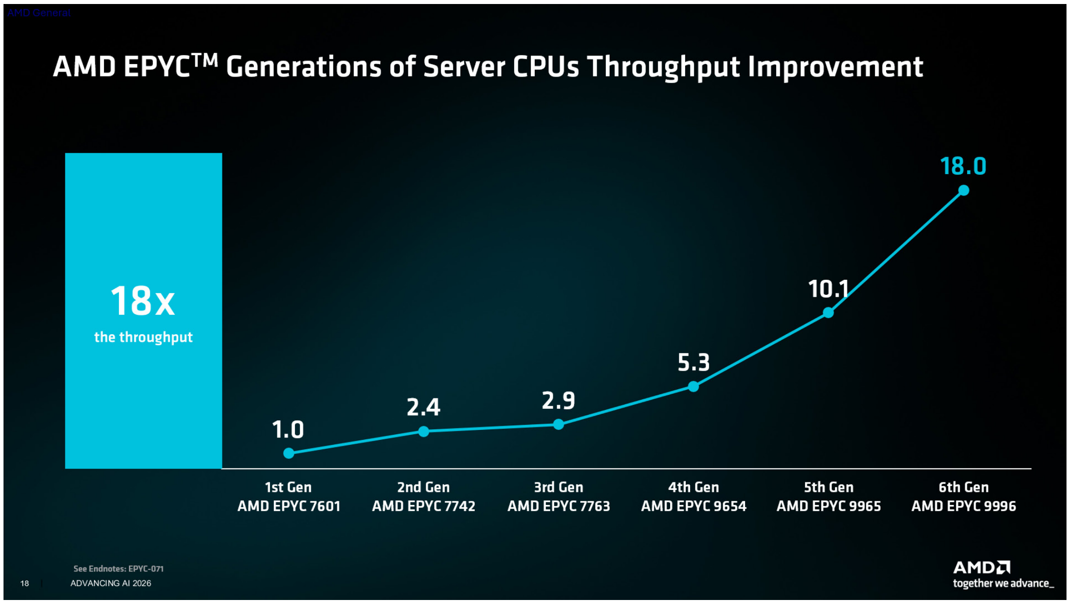


Figure 1: Generations of AMD EPYC™ Server CPUs throughput improvement.1

The result is not simply a faster processor. AMD EPYC™ 9006 server CPUs is a mature, balanced platform designed for more useful work per server and per rack across cloud, enterprise, HPC and AI infrastructures. Not only does each generation deliver more performance, but each new platform also typically gains new incremental capabilities. Figure 2 shows innovative features in 6th Gen AMD EPYC™ server CPUs.

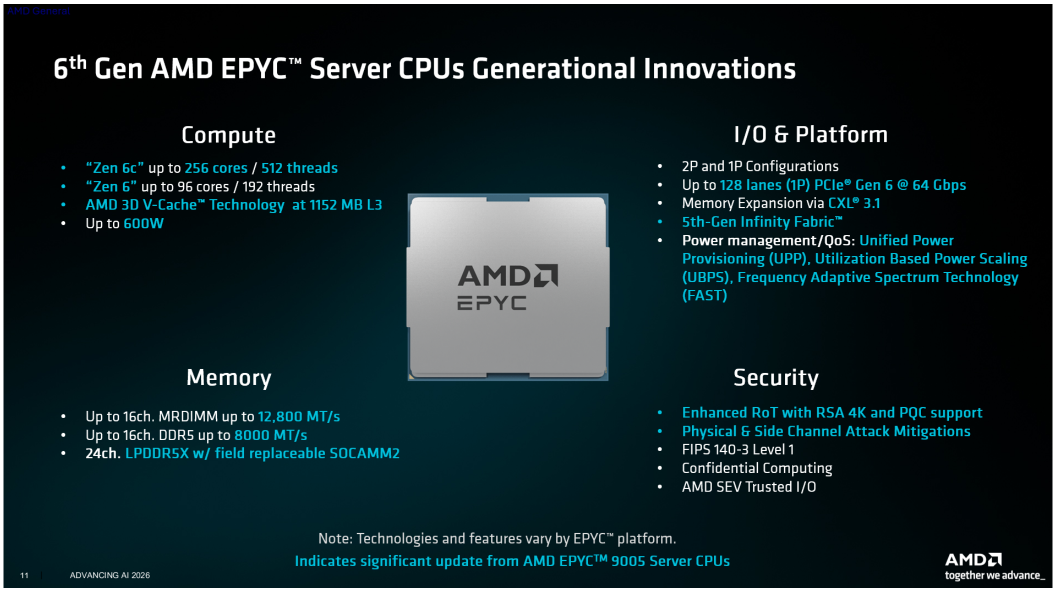


Figure 2: Select 6th Generation AMD EPYC generational innovations.

More than a new processor generation, the AMD EPYC 9006 server CPU extends a proven platform strategy built on innovation, execution discipline, software compatibility, industry standards and scalable system architecture.

This white paper examines the following system-level factors that make the AMD EPYC processor the most compelling data center CPU:

1. Maturity Built Through 6 Generations of EPYC Server CPU Innovation
2. x86 Ecosystem Advantage
3. Chiplet Architecture as the Foundation for Scale and Breadth of Solutions
4. Simultaneous Multithreading
5. AVX-512, Vector Compute and Future ACE Support
6. Memory Bandwidth, Capacity and MEXT Introduction
7. CXL as a Multigeneration Memory and I/O Fabric
8. Security and Confidential Computing
9. Reliability, Availability and Serviceability
10. Power-Aware Performance Management: UBPS, UPP and FAST
11. Operating-System Scheduler and Platform Enablement
12. Benchmark Leadership to Best Practices

1. MATURITY BUILT THROUGH SIX GENERATIONS OF EPYC SERVER CPU INNOVATION

Maturity of the AMD EPYC server CPU platform is the result of repeated architectural updates, platform validation, extensive software enablement and optimization, and production deployment in collaboration across a vast systems and software ecosystem and with customers, large and small. These collaborations inform the approach AMD takes when expanding server capability across core count, per-core performance, memory bandwidth, I/O bandwidth, security mechanisms, power-management functions and firmware interfaces. AMD places a high priority on compatibility with the proven x86-64 server software model.

Server platform maturity depends on more than processor microarchitecture. It requires validated BIOS and firmware flows, ACPI and CPPC interface support, operating-system and hypervisor enablement, compiler and library optimization, OEM system qualification, memory and I/O interoperability testing, management-controller integration, telemetry support, firmware-update procedures, and sustained operation under production workloads. The AMD EPYC server CPU business unit continues to invest heavily in this enablement work across cloud, enterprise, HPC and AI infrastructure environments.

Deployment breadth provides additional validation coverage. AMD EPYC server CPUs are offered across more than 450 OEM server platforms and more than 1,600 public cloud instances. These systems are deployed in hyperscale cloud infrastructure, enterprise data centers, HPC clusters, AI infrastructure and GPU-attached server configurations. The installed base is comprised of platforms across different chassis designs, memory populations, I/O topologies, accelerator configurations, operating systems, hypervisors, management stacks and workload classes.

AMD EPYC server CPUs support workloads used by billions of people each day through their usage in infrastructure that supports large-scale cloud services, enterprise applications, scientific computing systems and online services. This scale produces operational feedback across hardware behavior, firmware settings, OS scheduling, virtualization, memory reliability, I/O error handling, compiler output, library performance, telemetry and workload-specific tuning. AMD uses this feedback for continuous improvement.

AMD EPYC 9006 server CPUs are the latest iteration with the “Zen 6” architecture. They offer higher core density, expanded memory and I/O capabilities, updated security features, and additional power-management mechanisms over 5th Gen EPYC CPUs. While the 6th Gen platform is new, it enters an ecosystem that already includes optimized firmware, OS schedulers, virtualization stacks, performance tools, application-tuning practices, OEM platform designs, cloud instance definitions and system-management workflows. This existing enablement helps to reduce the number of new hardware and software variables introduced during platform qualification.

At large deployment scale, maturity also affects availability and serviceability. Production platforms require diagnostics, error containment, firmware-update workflows, memory and I/O qualifications, management integration, telemetry collection, and predictable behavior under sustained workloads. These requirements become more significant as systems increase core counts, memory capacity, accelerator attachment, I/O bandwidth and rack-level power density.

2. X86 ECOSYSTEM ADVANTAGE

The x86 ecosystem matters because the majority of data center software is written, compiled, validated and operated against well-understood x86 execution semantics. This includes operating systems, hypervisors, kernel modules, drivers, runtime libraries, databases, middleware, security agents, observability tools, deployment automation and infrastructure-control software. Across these layers, the deployed x86 software base represents billions of lines of production code.

This software base is not limited to application binaries. It includes build systems, packaging flows, test infrastructure, CI/CD pipelines, monitoring integrations, kernel interfaces, performance tools, failover procedures, support contracts and operational runbooks. These dependencies influence platform qualification because a workload must be validated across the full execution and operations environment, not just across the processor instruction set.

Platform transition costs can be significant for large deployments. Moving established workloads to a different instruction-set architecture can require source-code changes, dependency analysis, application recertification, build-system updates, compiler retuning, performance validation, ISV support verification, image rebuilds, automation changes, operational retraining and parallel infrastructure qualification. In large codebases, many migration issues occur outside the application logic, including build files, test infrastructure, packaging scripts, kernel modules, drivers, monitoring hooks and deployment configuration. For workloads already validated on x86 architecture, AMD EPYC server CPU-based systems can be introduced with less software retargeting than a transition to a different CPU architecture.

The relevance of the x86 platform continues in agentic AI systems. Although AI model execution may run on GPUs or other accelerators, the surrounding agent runtime frequently executes on general-purpose CPUs. Tool calling, sandbox orchestration, code interpreters, retrieval services, policy checks, API gateways, vector-database clients, logging agents, schedulers, authentication services, network proxies, storage services and control-plane components commonly run as existing Linux processes, containers, microservices or virtual machines. These components often reuse conventional x86 instruction set architecture, rather than requiring new AI-specific execution environments. While some will cite a single platform as the ideal agentic CPU, agentic AI has introduced new workflows and diverse workloads that require different points of optimization available in the AMD EPYC 9006 server CPU portfolio, as shown in Figure 3.

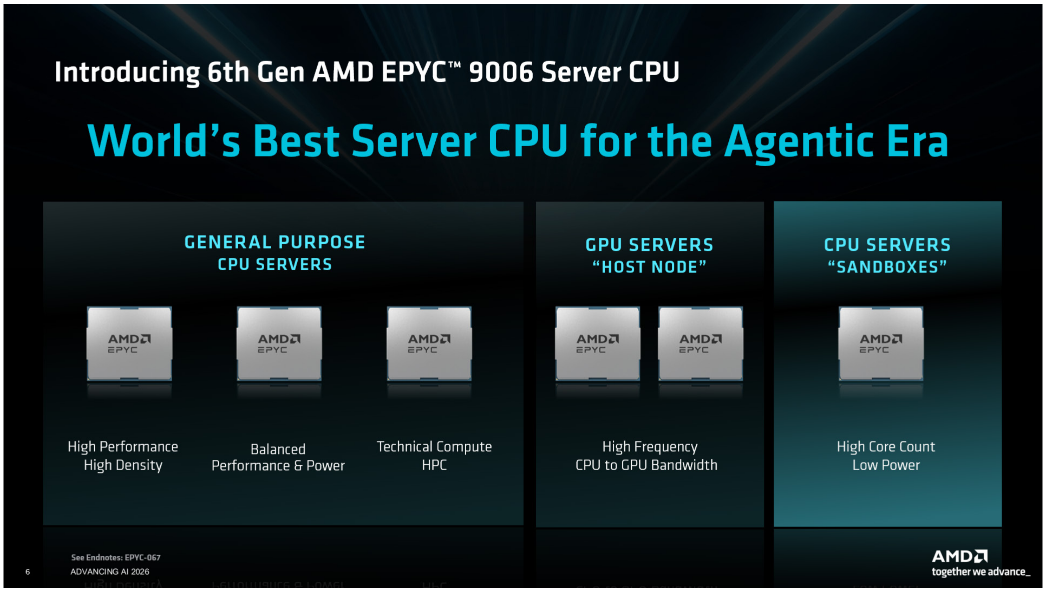


Figure 3: 6th Gen AMD EPYC™ Server CPU for agentic AI.2

Agentic AI workloads are not a single monolithic model invocation. A single user request can trigger multiple tool calls, retrieval operations, sandboxed code executions, database queries, policy evaluations, serialization steps, encryption operations and network transfers. Many of these tasks are latency-sensitive, branch-heavy, I/O-bound or dependent on existing enterprise services. Running them on an x86 platform allows AI infrastructure to reuse mature operating system services, container platforms, security controls, observability tools, networking stacks, storage clients and enterprise integration software.

With AMD EPYC 9006 server CPUs, the benefit is not only x86 architecture compatibility, but the software enablement accumulated across prior EPYC CPU generations. Existing x86 architecture operating-system images, hypervisor configurations, container platforms, orchestration frameworks, security controls, monitoring agents and enterprise applications can be reused when they are already certified for the target environment. In addition, prior EPYC server CPU generations have driven improvements in firmware interfaces, non-uniform memory access (NUMA) awareness, schedulers, vector-code generation, virtualization layers, libraries and application tuning. AMD EPYC 9006 server CPUs inherit this software path and extend it with new platform capabilities rather than starting from a new enablement baseline.

Founding members AMD and Intel, along with Microsoft, Google Cloud, Meta, Dell, HPE, Lenovo, Oracle, Red Hat and Broadcom, established the x86 Ecosystem Advisory Group to guide further innovation while unifying the x86 ecosystem. The group, shown in Figure 4, brings together major tech leaders to improve cross-platform compatibility, reduce instruction set fragmentation and simplify software development – with the objective of ultimately easing customer adoption of new functionality delivered by x86 systems.

Figure 4: The x86 Ecosystem Advisory Group.

This group's goals and purpose are to unite hardware and software leaders to guide future x86 ecosystem extensions, to prevent fragmentation issues seen with older instruction sets like AVX-512, and to standardize architectural interfaces and programming models across data centers and client devices. Some major technical milestones are FRED (Flexible Return Event Delivery); AVX10 standardized vector extension; the ACE unified approach for matrix multiplication capabilities; and the ChkTag collaborative hardware-based memory safety feature designed to mitigate buffer overflows.

3. CHIPLET ARCHITECTURE AS THE FOUNDATION FOR SCALE AND BREADTH OF SOLUTIONS

AMD developed chiplet architecture to scale modern server processors across core count, memory bandwidth, I/O capability, yield, power efficiency and product configuration. As server processors move toward hundreds of cores, monolithic die designs become increasingly difficult to manufacture economically, validate efficiently and adapt across product requirements. Chiplets address these constraints by decomposing the processor into modular building blocks that can be manufactured, validated and combined with greater flexibility.

For AMD EPYC 9006 server CPUs, compute chiplets (CCDs) include “Zen 6” or “Zen 6c” cores, while dedicated I/O logic chiplets (IODs) provide memory controllers, PCIe, CXL, coherency, security and platform connectivity, as shown in Figure 5. This separation improves performance and energy efficiency by allowing each domain to be optimized for its role. Compute chiplets can use process technology and physical design choices optimized for transistor density, frequency and core efficiency, while the I/O die can emphasize high-speed interfaces, memory PHYs, SerDes, reliability, security and longer-lived platform functions.

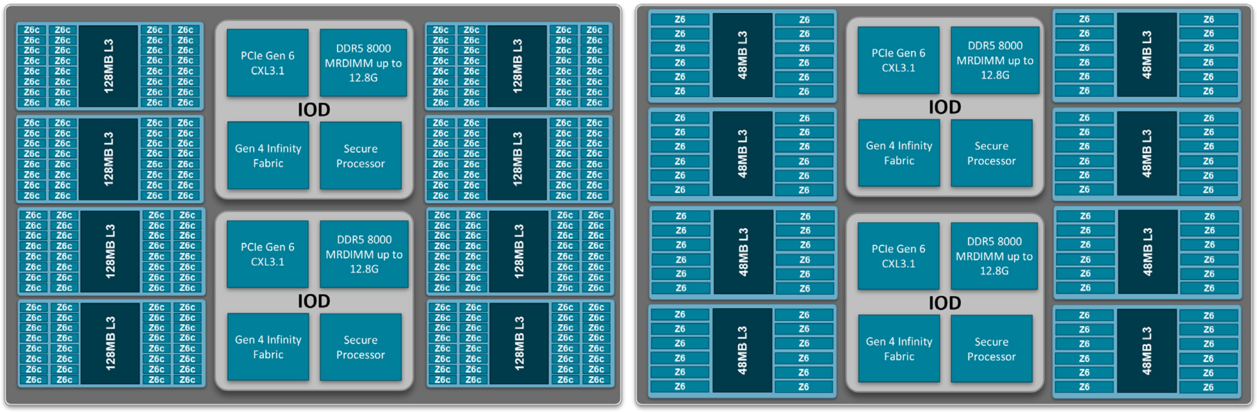


Figure 5: 6th Gen AMD EPYC™ Server CPU chiplet for “Zen6c” and “Zen6.”

The separation of compute and I/O also impacts engineering schedules and validation. CPU cores and I/O subsystems often have different design timelines, validation requirements, risk profiles and reuse patterns. Core microarchitecture advances with each generation to improve IPC, frequency, vector capability, cache effectiveness and power efficiency. I/O logic tracks platform standards such as DDR6, MRDIMM, PCIe Gen 6, CXL 3.1, security infrastructure, firmware interfaces and board-level electrical requirements. Chiplets allow these domains to evolve on partially independent schedules while supporting multiple SKUs from common building blocks.

The modular design of the processor also means interconnect technology is a crucial second pillar of this scalability model. AMD Infinity Fabric™ links chiplets, memory, I/O resources, CPUs and accelerators into a coherent system. xGMI extends this fabric model for high-bandwidth CPU-to-CPU and CPU-to-GPU communication, enabling tightly coupled systems to move data efficiently across devices. Looking forward, Infinity Fabric over Ethernet (IFoE) will extend fabric concepts beyond a single server into rack-scale and pod-scale deployments using Ethernet-based switching infrastructure.

Together, compute chiplets, Infinity Fabric Technology, xGMI and IFoE define a scalable architecture that extends from silicon modularity to system connectivity and rack-level infrastructure. This architecture supports performance scaling, energy-efficiency, product reuse and configuration flexibility across a range of data center requirements that would be difficult to address with a monolithic processor design.

4. SIMULTANEOUS MULTITHREADING

Simultaneous multithreading, or SMT, allows a physical CPU core to expose multiple hardware threads to software. When one hardware thread is stalled on memory access, branch resolution, synchronization or I/O, the core can issue work from another hardware thread if execution resources are available. This can improve aggregate throughput by increasing utilization of core resources that would otherwise be idle during stall cycles. This technique is well established in x86 architecture and is a key feature for most modern CPUs. Learning from real-world deployments, AMD has optimized SMT across EPYC™ Server CPU generations by transitionally dynamic scheduling, using watermarked resource boundaries and broadening thread-interleaving queues, which drastically reduce core resource contention stalls, thread synchronization latency and locking overhead during heavy concurrent workloads.

A single-threaded processor, as shown in Figure 6, will allocate all physical resources to processing a single thread before context switching to begin processing a second thread. An SMT-enabled Server CPU can be divided into two logical cores per physical core, allowing some level of work and resource parallelism in place of waiting for each thread to run to completion.

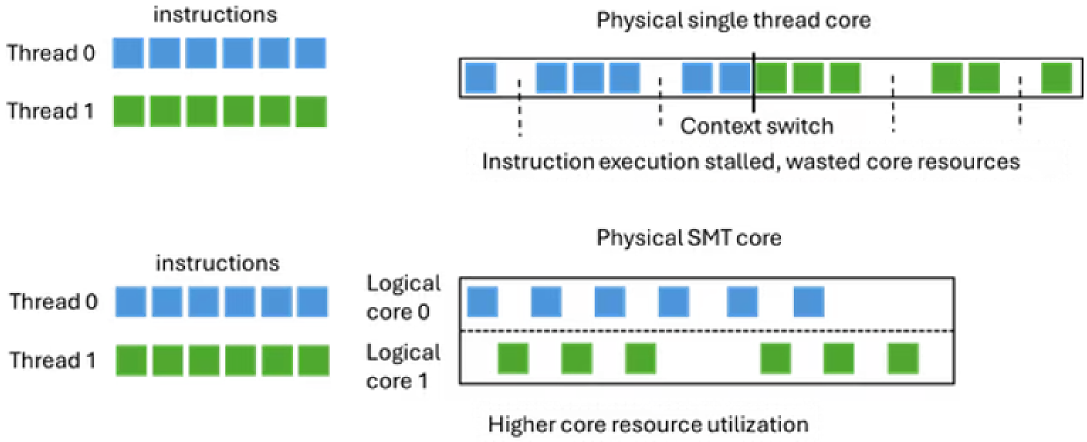


Figure 6: A simple process view of a single thread versus simultaneous multithreading.

SMT is most useful for workloads with sufficient thread-level parallelism and frequent latency stalls. Examples include virtualized infrastructure, containerized services, web and application servers, storage services, networking functions, control-plane software, build systems, test automation, and many host-side AI infrastructure tasks. These workloads often include many independent software threads or processes that can share a core without requiring each runnable thread to be mapped to a separate physical core.

SMT increases the maximum hardware-thread count available per socket. High-core-density configurations such as AMD EPYC 9006 server CPUs with 256-core processors can expose up to 512 hardware threads when SMT is enabled. This thread capacity is useful for environments that run many concurrent services, virtual machines, containers, orchestration components or short-lived tasks. The performance benefit depends on workload behavior, including cache locality, memory bandwidth demand, synchronization overhead, branch behavior and contention for shared core resources.

SMT is particularly relevant to agentic AI sandbox environments where numerous lightweight and partially independent tasks execute concurrently around the model inference path. Those tasks include tool invocations, code interpreters, retrieval agents, policy checks, observability services, API requests, simulation workloads and short-lived microservices. Because many of these activities spend significant time waiting on memory, storage, network or accelerator responses, SMT can improve host-side utilization by allowing stalled and ready threads to share core execution resources, increasing overall throughput and reducing
idle cycles.

On AMD EPYC server CPUs, SMT remains a configurable feature that can be enabled or disabled through platform firmware/BIOS settings at boot time, while operating systems can achieve similar scheduling behavior by limiting execution to a single software thread per core. In some deployments, however, SMT may be disabled to provide stronger isolation, more deterministic latency, simpler performance characterization, licensing compliance or addressing of specific security requirements.

5. AVX-512, VECTOR COMPUTE AND FUTURE ACE SUPPORT

Vector compute is used by many data center workloads that apply the same operation across large blocks of data. Support of the AVX-512 vector instruction set expands SIMD execution capability for software that is compiled or tuned to use wide vector instructions. Relevant workload classes include analytics, scientific computing, signal processing, compression, encryption, genomics, video processing, machine-learning preprocessing, embedding pipelines and vector-search workloads. In AMD EPYC 9006 server CPUs, AVX-512 instructions operate alongside scalar execution, SMT, cache hierarchy, memory bandwidth, compiler support, optimized libraries and workload-specific tuning.

The benefits of AVX-512 instructions depend on hardware support and software enablement. Compilers, math libraries, cryptographic libraries, media libraries, database engines, analytics frameworks and application runtimes must generate or select vectorized code paths for workloads. When these paths are available, AVX-512 can increase instruction-level data parallelism and reduce the number of instructions required for eligible operations. Delivered performance still depends on memory bandwidth, cache locality, instruction mix, frequency behavior and the fraction of the workload that can be vectorized.

In AI infrastructure, CPUs continue to execute many supporting functions even when GPUs or other accelerators perform model training or inference. Host-side work can include data preparation, tokenization, embedding pipeline support, retrieval-augmented generation, compression, encryption, networking, storage services, monitoring, scheduling and control-plane execution. Vectorized CPU paths can improve throughput for eligible portions of these functions and reduce host-side bottlenecks that limit accelerator utilization.

6. MEMORY BANDWIDTH, CAPACITY AND MEXT INTRODUCTION

Memory bandwidth, capacity and cost are primary design considerations for modern server platforms. As core counts increase, a processor's delivered performance depends on how effectively its memory subsystem can supply data to the cores, sustain concurrent access streams and keep large working sets resident in local memory. If memory bandwidth is insufficient, additional cores can remain underutilized. If memory capacity is insufficient, applications may spill to slower storage tiers or require more sockets and servers. If memory cost is too high, system configurations may be limited by economics rather than compute capability.

AMD EPYC 9006 server CPUs address these far-ranging requirements with a wide local-memory subsystem. The platform supports up to 16 channels of DDR5, JEDEC-standard second-generation MRDIMM support of up to 12,800 MT/s, providing approximately 1.6 TB/s peak memory bandwidth per socket, and large local memory configurations. With 16×256GB DIMMs, AMD EPYC Server CPUs can support up to 4TB of local memory capacity per socket. This combination of channel count, transfer rate and capacity is important for keeping high-core-count processors supplied with data while allowing larger datasets, caches and application state to remain in DRAM.[29]

The memory subsystem affects a broad range of workloads. Virtualization hosts require capacity to keep more virtual machines and containers resident. Databases, in-memory databases and analytics engines require capacity and bandwidth for tables, indexes, column scans, joins, caches and intermediate query state. HPC and engineering workloads require sustained memory throughput for large meshes, matrices, particles and state vectors. AI infrastructure uses CPU memory for preprocessing, tokenization, retrieval, embedding pipeline support, orchestration, networking, storage services and accelerator feeding. In each case, memory behavior can determine whether CPU cores and accelerators remain productive or stall while waiting for data.

Large local memory capacity is especially useful for workloads where performance depends on keeping working sets close to the processor. Enterprise databases, data warehouses, graph analytics, EDA, computational fluid dynamics, finite-element analysis, weather modeling, molecular dynamics, reservoir simulation and digital-twin workloads often operate on large data structures with irregular or bandwidth-intensive access patterns. Capacity of 4TB of local memory per socket increases addressable problem sizes, reduces dependence on storage tiers, improves cache and page locality, and reduces I/O bottlenecks for these memory-intensive applications.

Memory economics are also important because the mid-2026 market has proven that DRAM can represent a large portion of server bill-of-material cost. This has led many customers to closely scrutinize the amount of memory their servers need based on workload requirements. Customers are increasingly aware of options to evaluate the capacity and bandwidth requirements of the workload. The 16 memory channels available on select AMD EPYC 9006 CPUs provide compelling optionality in the definition of useful and affordable memory configurations to meet diverse needs.

More recently, new technology has emerged to provide even more alternatives. Some workloads require a large addressable memory footprint but do not access all pages with the same frequency. Memory-tiering approaches such as AMD MEXT memory optimization technology address this by using AI-based, software-managed memory management to support workloads split between DRAM and flash-based tiers. By identifying frequently accessed pages and moving colder pages to a lower-cost tier, memory tiering can increase effective capacity while controlling cost and masking the latency penalties of flash. For AMD EPYC 9006 server CPU-based systems, high-bandwidth local DRAM provides the performance tier, while tiering software can help manage larger working sets where access patterns do not require all data to remain in DRAM.

7. CXL AS A MULTIGENERATION MEMORY AND I/O FABRIC

CXL® coherent interconnect is the industry standard. It allows CPUs, memory expansion devices, accelerators and other platform components to participate in a flexible, memory-centric system architecture. For high-core-count servers, this capability is important because performance and efficiency increasingly depend on how well compute, memory and I/O resources can be scaled and shared across the platform.

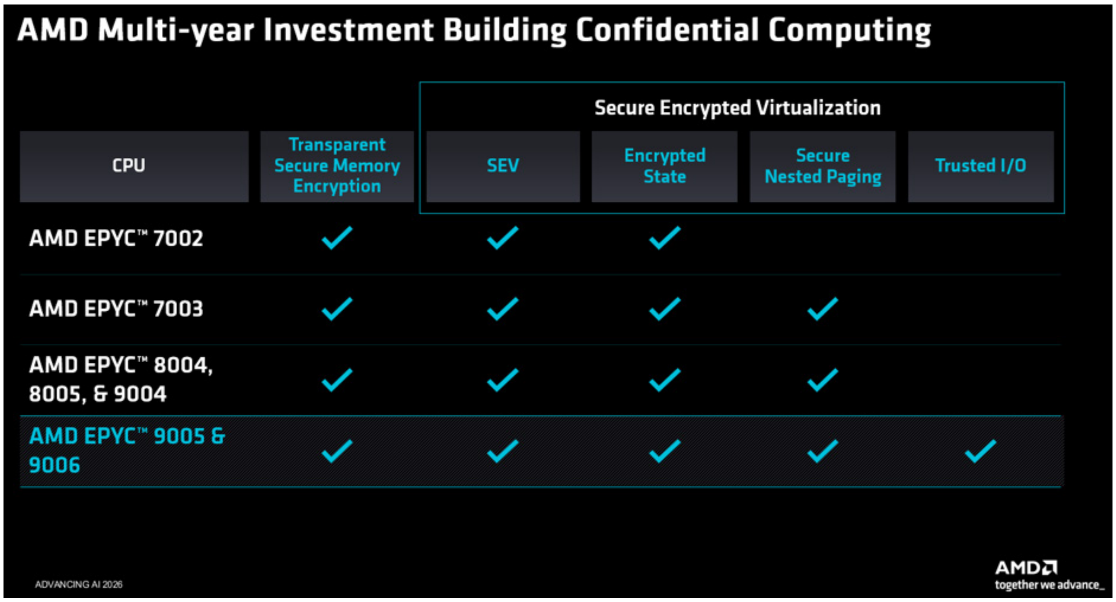
AMD EPYC 9006 server CPUs combine PCIe Gen 6 connectivity with CXL 3.1 support, extending the EPYC platform beyond traditional local DRAM and conventional PCIe-attached devices. This matters because CXL adoption depends on ecosystem maturity across processors, memory devices, switches, firmware, operating systems, hypervisors, orchestration tools and data center deployment models.

In AMD EPYC 9006 server CPU systems, CXL complements high core density, high local memory bandwidth and the mature chiplet-based EPYC architecture. Customers can use this platform to support diverse memory-expansion, accelerator-attachment, storage and networking strategies without being constrained to a single vertically integrated architecture.

In the near term, the most practical value of CXL is memory expansion and tiering with potential to scale to disaggregated or pooled memory, designed for cost advantages. CXL Type 3 memory devices can extend addressable capacity for workloads that exceed more economical local DRAM configurations, helping larger working sets remain closer to the CPU than storage-based tiers. In an era of constrained and expensive memory, CXL-based memory expansion provides another option for large system capacities utilizing more small DIMMs – or possibly allowing reuse of "legacy" DIMMs.

Over time, the CXL ecosystem has matured with production-ready devices, software-tiering enhancements and hyperscale data center deployments. AMD has been a strong advocate for ecosystem development, contributing not only to open-standard development but also to open-source software for tiered memory.

8. SECURITY AND CONFIDENTIAL COMPUTING



		Secure Encrypted Virtualization			
CPU	Transparent Secure Memory Encryption	SEV	Encrypted State	Secure Nested Paging	Trusted I/O
AMD EPYC™ 7002	✓	✓	✓		
AMD EPYC™ 7003	✓	✓	✓	✓	
AMD EPYC™ 8004, 8005, & 9004	✓	✓	✓	✓	
AMD EPYC™ 9005 & 9006	✓	✓	✓	✓	✓

Figure 7: AMD EPYC™ Server CPU generational security capabilities.

With AMD Infinity Guard technologies, AMD has continued to advance platform security capabilities across generations, as shown in Figure 7. AMD EPYC 9006 server CPUs provide security functions across processor boot, memory protection, virtualization, I/O, debug and cryptographic infrastructure. AMD EPYC 9006 server CPU security capabilities include enhanced Root of Trust, RSA 4K, post-quantum cryptography support, physical and side-channel mitigations, SEV, SEV-ES, SEV-SNP, SEV-TIO, trusted I/O concepts, secured debug, and FIPS 140-3 Level 1 targeting. These functions are relevant because modern workloads span CPU cores, memory, accelerators, CXL-attached devices, storage, networking, firmware, hypervisors and cloud control-plane software – any of which can be a potential attack point for malicious actors.

9. RELIABILITY, AVAILABILITY AND SERVICEABILITY

Reliability, availability and serviceability, or RAS, are required for large-scale server deployments because hardware faults, memory errors, I/O failures, firmware issues, link errors and service events can potentially affect workload continuity and system uptime. The data flowing through high-performance, high-capacity systems are under constant assault from heat, cosmic rays, materials decay and other threats that can cause data to change inadvertently. AMD EPYC 9006 server CPUs extend RAS capabilities developed across multiple EPYC generations and apply them to higher core counts, larger memory footprints, expanded I/O fabrics, accelerator-attached systems and denser fleet deployments.

EPYC 9006 server CPUs establish a comprehensive, server-class RAS architecture that spans the compute cores, the memory subsystem, on- and off-package interconnects, and the I/O and CXL domains. This ensures no major fault domain in the platform is left without detection, correction and reporting coverage. With DDR5 memory technology, the subsystem is capable of full single-device data correction on DDR5 RDIMM and MRDIMM, a level of resiliency that low-power (LP) memory technologies cannot structurally match, positioning EPYC 9006 CPUs ahead of competing architectures that trade memory correction strength for power or density. Additionally, RDIMM/MRDIMM form factors provide unmatched field-replaceable unit (FRU) capability for servicing. Beyond correction, EPYC 9006 server CPUs emphasize containment and continuity, isolating faults, recovering transient errors in hardware and permitting the platform to remain in service through conditions that might have forced a reset on older hardware, which can translate into higher measured availability at fleet scale.

The RAS framework for AMD EPYC 9006 server CPUs features mechanisms for error detection, error containment including data poisoning, recovery handling, enhanced diagnostics, secured debug, memory serviceability, PCIe® surprise hotplug and error containment, and hardened platform behavior under fault conditions. These functions are especially important as systems scale to more cores, more memory channels, larger DIMM populations, higher PCIe bandwidth, CXL-attached devices and accelerator-rich configurations.

RAS maturity depends on repeated validation across silicon revisions, firmware releases, memory configurations, PCIe-capable devices, CXL-capable devices, operating systems, hypervisors, OEM platforms and field deployments. Correctable memory errors, uncorrectable errors, link failures, device faults, firmware events and service actions must be reported with enough detail for operating systems, management controllers, fleet-management software and service teams to determine whether to recover, isolate, migrate, reset, replace or continue operation.

At fleet scale, RAS behavior affects fault isolation, failure analysis, firmware-update workflows, component replacement, workload migration and service-level management. The objective is to maintain predictable operations when individual components experience correctable errors, uncorrectable errors, link degradation, device faults or maintenance events. For AMD EPYC 9006 server CPUs, RAS is an essential part of the platform architecture needed to operate high-density cloud, enterprise, HPC and AI infrastructure with large socket counts, large memory footprints and complex I/O topologies.

10. POWER-AWARE PERFORMANCE MANAGEMENT: UBPS, UPP AND FAST

Modern data centers are increasingly constrained by power, cooling and rack-density limits, making efficient power management as important as raw compute performance. AMD is proud of the energy-efficiency leadership demonstrated in the prior generations of EPYC server CPUs but recognizes growing challenges. To address these, several complementary power-management capabilities available on 6th Gen AMD EPYC™ Server CPUprocessors help align system energy consumption with workload demands to minimize waste and address infrastructure constraints, shown in Figure 8.

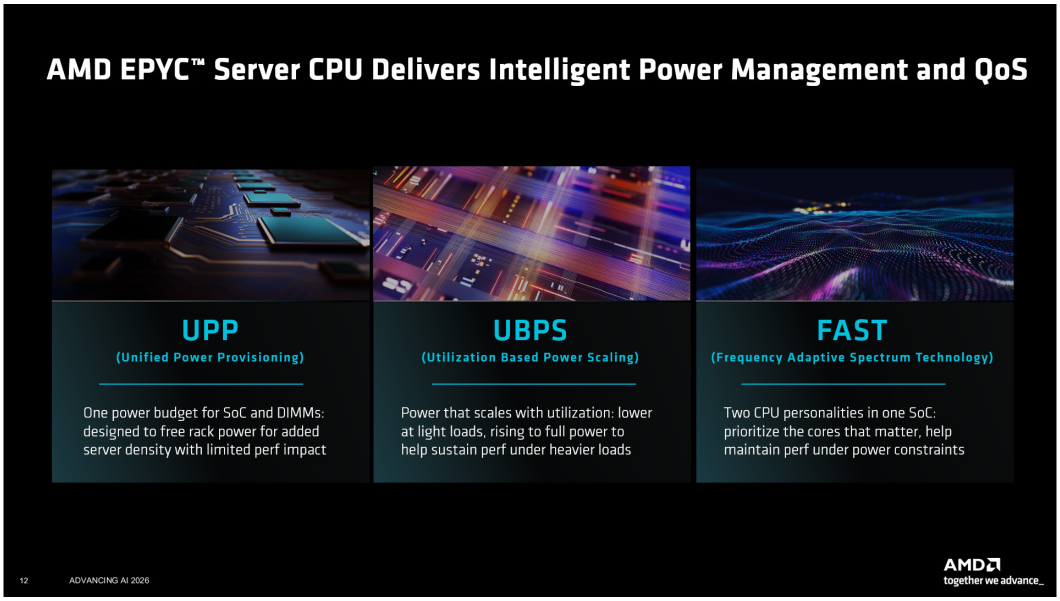


Figure 8: UPP, UBS and FAST Technologies.

Utilization Based Power Scaling (UBPS) dynamically adjusts platform power allocation according to measured workload utilization. Data center systems are often provisioned for worst-case power consumption even though many workloads operate well below peak utilization for extended periods. This can leave significant power capacity unused and reduce effective rack utilization. UBPS helps close the gap between provisioned and consumed power. It allows power to dynamically scale more closely to actual workload activity, improving overall power efficiency and reducing stranded power capacity.

Unified Power Provisioning (UPP) extends power management beyond the processor by treating platform power as a shared resource across the server. Modern servers include CPUs, memory, accelerators, networking, storage, cooling and management components that all operate within common power and thermal limits. UPP coordinates power allocation across these resources, helping align available power with the components and workloads that need it most at a given time. This platform-level approach can reduce inefficiencies caused by static component provisioning and improve utilization within fixed rack and facility power envelopes.

Frequency Adaptive Spectrum Technology (FAST) manages processor frequency behavior under changing workload, power and thermal conditions. Different workloads place different demands on the processor: Latency-sensitive applications may benefit from higher frequencies; highly parallel throughput-oriented workloads may require balancing frequency across many active cores. FAST uses platform telemetry and firmware policies to adjust frequency behavior dynamically within available power and thermal limits, helping the processor respond more effectively to varying workload characteristics.

Together, UBPS, UPP and FAST form a coordinated framework for power-aware performance management. UBPS aligns power consumption with actual workload utilization. UPP manages how power is allocated across the platform. And FAST optimizes processor frequency behavior within those power and thermal boundaries. The combined effect is a more efficient use of available power resources, improved infrastructure utilization and more predictable system behavior across diverse data center workloads.

11. OPERATING-SYSTEM SCHEDULER AND PLATFORM ENABLEMENT

Years of ecosystem collaboration across AMD EPYC server CPU generations have built a broad software-enablement base with operating-system vendors, hypervisor providers, cloud providers, OEMs, compiler and library developers, and system-management partners. This work carries forward into the AMD EPYC 9006 server CPUs. The platform exposes high core counts, chiplet topology, multiple memory domains, high-bandwidth I/O, CXL connectivity, accelerator attachment, RAS telemetry, security features and power-management controls for utilization. These capabilities are surfaced through the OS scheduler, memory manager, power-management stack, interrupt subsystem, virtualization layer, device drivers, firmware interfaces and system-management software.

Enablement of such technologies spans major enterprise, cloud, virtualization and regional operating-system environments. This includes Red Hat Enterprise Linux, SUSE Linux Enterprise Server, Ubuntu Server, Debian, Oracle Linux, Rocky Linux, AlmaLinux, Microsoft Windows Server, VMware ESXi, Nutanix AOS, hyperscaler-maintained cloud images, container-host operating systems and China-focused operating systems such as openEuler, OpenCloudOS and Anolis OS. This coverage reflects work across kernel support, boot flows, device enablement, virtualization, security, RAS reporting, management integration, cloud images and performance tooling.

AMD operating-system enablement is not limited to basic boot support. High-core-count systems require optimization across NUMA awareness, CCD topology, memory placement, scheduler behavior, SMT scheduling, CPU-frequency control, interrupt affinity, I/O locality, CXL device exposure, virtualization overhead, RAS event handling, secured virtualization and system telemetry. These optimizations have been developed and refined across prior generations of server processors, and AMD EPYC 9006 server CPUs benefit from that accumulated software work while adding enablement for new topology, memory, I/O, security, RAS and power-management capabilities.

NUMA behavior is not solved by hardware alone. A high-core-count chiplet processor exposes locality relationships among cores, caches, memory controllers, I/O devices, CXL devices and accelerators. If the OS, hypervisor or runtime treats the system as a flat pool of identical CPUs and memory, workloads can experience higher memory latency, reduced cache locality, unnecessary cross-domain traffic, interrupt misplacement and lower throughput. Correct topology exposure, scheduler policy, memory placement, interrupt affinity and virtualization integration allow software to place work closer to the resources it uses. This is why broad OS enablement and tuning across Linux, Windows Server, VMware, cloud images and regional operating systems is important to production behavior.

AMD EPYC™ 9006 Server CPU features such as UBPS, UPP, FAST, CXL, SEV-SNP, SEV-TIO, RAS telemetry and secured debug require OS and firmware coordination. Their effectiveness depends on telemetry collection, platform policy, power and thermal limits, CPU-frequency control, scheduler decisions, workload placement, device exposure, security attestation, error reporting, and management software integration. The cumulative AMD software-enablement base reduces the amount of new integration work required for each EPYC server CPU while still allowing new CPU-specific features to be exposed and tuned for production deployments.

12. BENCHMARK LEADERSHIP TO BEST PRACTICES

A defining characteristic of the success realized by AMD EPYC server CPUs has been its commitment to industry-standard benchmarks, open evaluation methodologies and full-stack software optimization. Benchmark leadership is most meaningful when it reflects reproducible, comparable, standards-based evidence across real workloads rather than isolated peak-performance demonstrations.

Instead of relying solely on proprietary demonstrations, AMD has invested heavily in supporting recognized benchmarks such as SPEC, TPC, MLPerf, STAC and other application-level workloads that reflect real-world deployments across cloud, enterprise, AI and HPC environments.

Over six generations of EPYC server CPU innovation, AMD has established leadership across hundreds of benchmark results spanning virtualization, databases, technical computing, analytics, cloud infrastructure and AI workloads. These results are not simply the consequence of hardware innovation; they reflect a broader philosophy of co-optimizing processors, firmware, compilers, runtime libraries, operating systems, performance tools and application software.

Modern performance depends far more than microarchitecture or frequency. Delivered performance increasingly depends on compiler quality, NUMA awareness, scheduling efficiency, memory placement, vectorization, virtualization performance and application-specific optimization. AMD sustained investment in Linux kernel optimization and tool chain and ecosystem enablement helps ensure that “Venice” hardware innovations translate into meaningful real-world application performance.

General Purpose Computing

The SPEC CPU® 2026 benchmark announced in May 2026 is the newest update and successor to the widely used SPEC CPU® 2017 test suite. Its tests are highly regarded, regulated and audited as the industry standard for assessing computing intensive workloads. The SPEC CPU 2026 benchmark rigorously tests processors, memory subsystems and compilers across a variety of systems, providing valuable insights into performance and efficiency. Its 52 benchmarks are divided into four suites. Among these four suites, the SPECrate® 2026 Integer suite consists of 14 benchmarks and is among the most foundational points of comparison that customers request for general-purpose computing and CPU performance evaluation. (As this document is published, results from both AMD and Nvidia were not yet suitable for publishing based on SPEC rules regarding public system availability. The results discussed below are measured results that must be considered estimates based on publishing rules.)

Results of normalized platform-level performance as measured by overall throughput show the 6th Gen AMD EPYC™ 9996 server CPU delivers 2.37x that of Intel Xeon 6980P and 2.24x (2.37/1.06=2.24) that of Nvidia Vera for the SPECrate 2026_int_base workload. Figure 9 shows the results.

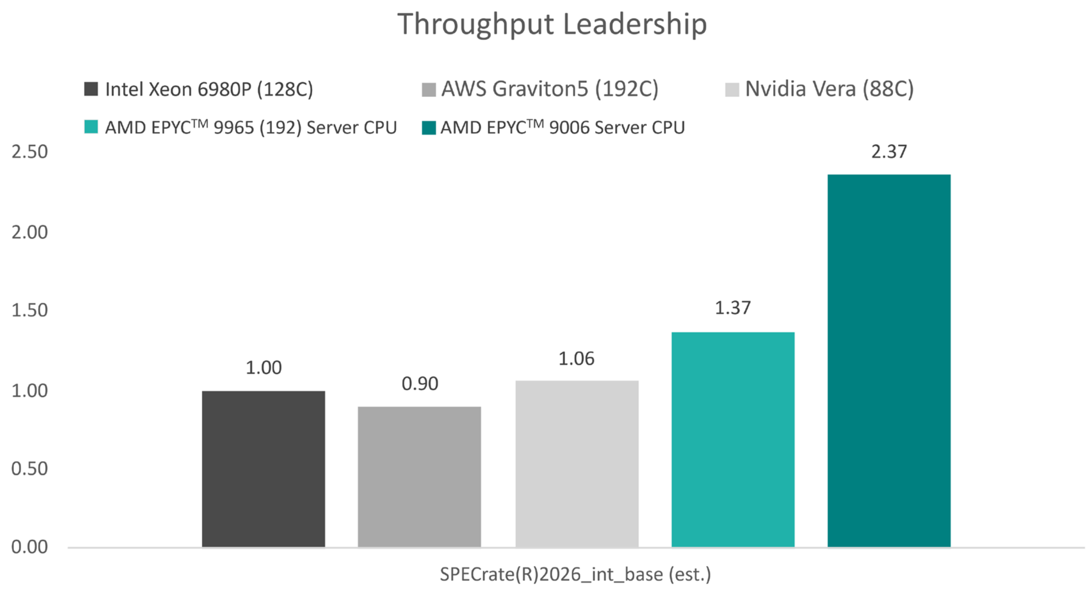


Figure 9: Top-of-stack general-purpose system computing performance (GCC Compiler).[3]

High per-core performance is critical for latency-sensitive workloads such as databases, user interactive services, live translations and specific phases of agentic AI. (To accurately reflect production environments, per-core performance is benchmarked under full system load.) Figure 10 illustrates the normalized per-core performance in which 6th Gen AMD EPYC™ 9006 (96C) Server CPU delivers 1.20x of Nvidia Vera for the SPECrate 2026_int_base workload.

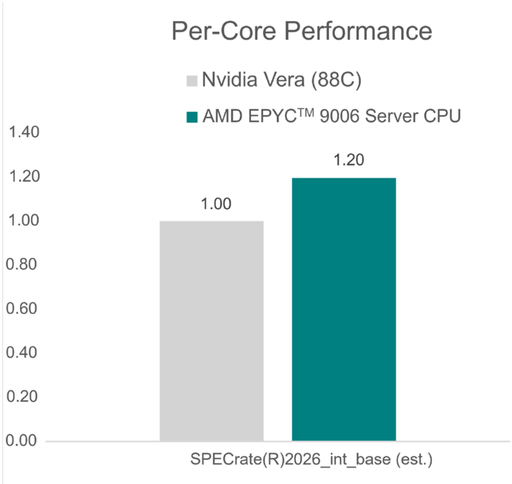


Figure 10: Mid-stack per-core system performance comparison (GCC Compiler).[4, 5]

On Agentic AI-Focused Testing, 6th Gen AMD EPYC™ Server CPU Defeats Nvidia

Four of the SPEC CPU 2026 subtests that demonstrate agentic AI performance were recently highlighted by Nvidia. In its published comparison, the company used a 5th Gen AMD EPYC server CPU system. When a 6th Gen AMD EPYC™ 9006 (96C) Server CPU is tested versus an Nvidia Vera 88C on a per-core basis in the same subtests, AMD EPYC™ 9006 Server CPU outperforms Vera. Figure 11 shows the results.

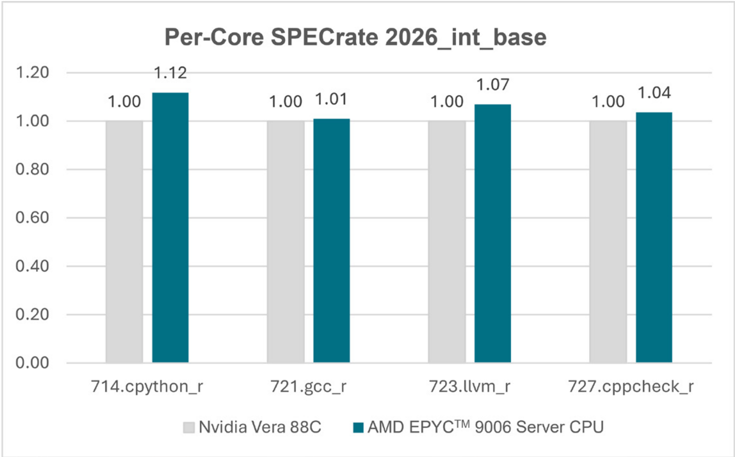


Figure 11: Mid-stack per-core performance (GCC Compiler) results for subtests published by Nvidia.6

Additionally, 6th Gen EPYC CPUs lead Vera in per-core performance across all 14 subtests of the SPECrate 2026 Integer suite, as shown in Figure 12. This wider testing comparison offers a more detailed view into the results in Figure 10, which highlighted the 1.2x geomean per-core performance advantage for the 96-core EPYC 9006 CPU across the suite.

Results of the testing display a processor's capabilities across a wide range of production deployments that include agentic AI. Agentic AI inference-focused deployments along with harnesses are complex and diverse, and they are much better represented by all subtests than a selected few cherry-picked by a competitor.

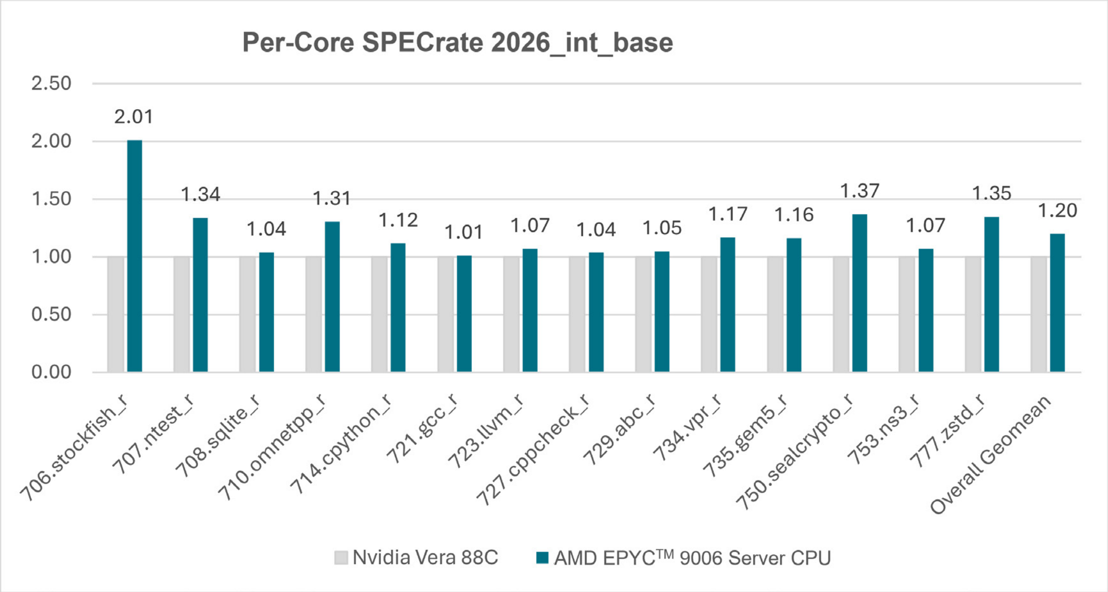


Figure 12: Mid-stack per-core performance (GCC Compiler) across all SPECrate 2026 Integer subtests[7]

Sustainable Memory Bandwidth

A high memory bandwidth platform is essential for memory-intensive workloads like generative AI model inference, HPC simulations, large-scale graph analytics and enterprise database systems that constantly ingest and process massive real-time data streams. **STREAM Triad** is the most representative and widely cited kernel of the STREAM Benchmark, an industry-standard synthetic program used to measure sustainable main memory bandwidth (in MB/s or GB/s).

AMD EPYC 9006 server CPUs with 16 memory channels and support for 12,800 MT/s MRDIMM memory offer higher peak bandwidth of 1.6 TB/s than the LPDDR5x-based Nvidia Vera platform with 1.2TB peak memory bandwidth.

Customers want measured results, not "theoretical peaks." Results comparing AMD-measured data for 1P AMD EPYC™ 9006 Server CPU platform using DDR5 12,800 MRDIMMs and Phoronix-published data for 1P Nvidia Vera 88C platform using LPDDR5X memory are displayed in Figures 13a and 13b.

Figure 13a shows overall system-level sustainable memory bandwidth where 6th Gen AMD EPYC™ 9006 Server CPU (96C) delivers 1.18x of Nvidia Vera (88C) for STREAM Triad.

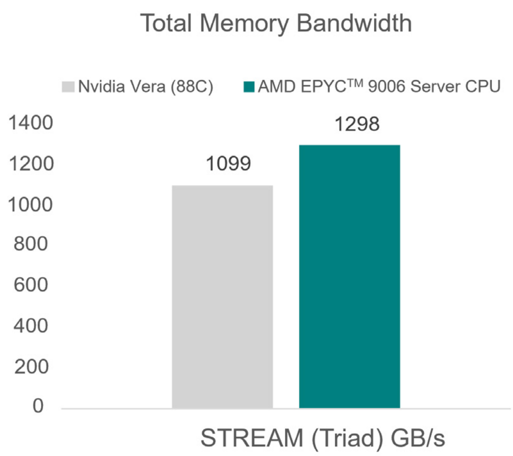


Figure 13a: Overall system STREAM Triad.[8]

In addition to overall high memory bandwidth, high per-core memory bandwidth is critical to prevent individual execution units from being starved for data during data-intensive operations, ensuring that performance scales linearly as more cores are assigned to a workload. Figure 13b shows the normalized per-core sustainable memory bandwidth where 6th Gen AMD EPYC™ 9600 (96C) Server CPU delivers 1.08x of Nvidia Vera (88C) for STREAM Triad.

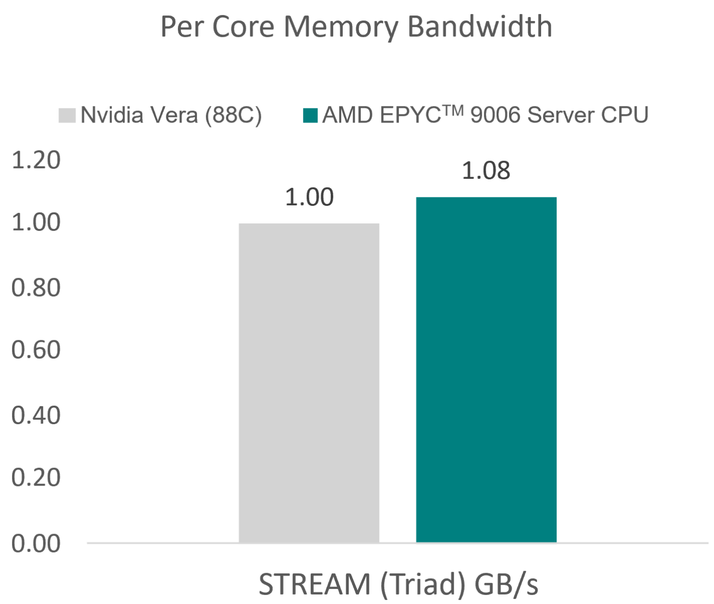


Figure 13b: Per-core STREAM Triad.[9]

Enterprise and Cloud Native

As demonstrated, AMD EPYC architecture provides outstanding performance across a wide range of applications, including critical enterprise and cloud-native workloads. Next, AMD considered other testing standards that represent a broad spectrum of enterprise applications and cloud-native workloads. They also offer insights into how processors perform under real-world conditions. While top-tier SKU performance and efficiency are critical for cloud service providers aiming to maximize VM density, core-to-core comparisons are especially relevant for enterprise environments, where software stacks are often licensed per core.

6th Gen AMD EPYC™ 9996 Server CPU delivers platform-level performance increases of 2.4x for server-side Java, 2.5x for OpenSSL, 3.5x for MongoDB with YCSB, 2.9x for Redis Benchmark, 3.7x for NGINX with WRK, and 2.6x for transaction processing systems based on TPC-C kit when normalized to an Intel Xeon 6980P (128C)-powered system. Figure 14 shows the results.

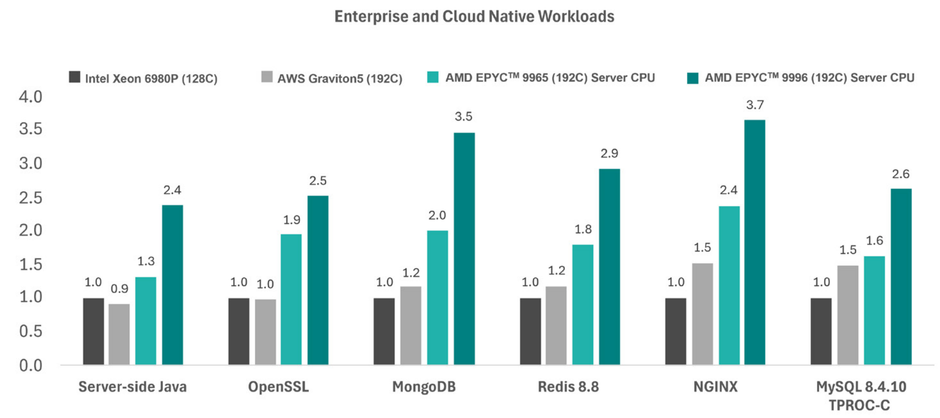


Figure 14: Enterprise and cloud native workload performance.[10, 11, 12, 13, 14, 15]

Java®, a platform-independent programming language, enables seamless application deployment across diverse environments. To evaluate server-side Java performance, AMD used a benchmark that simulates corporate IT workloads, such as point-of-sale transactions, online purchases and data mining. This benchmark provides the key metric m-jOPS, measuring "maximum sustained throughput" for Java e-commerce transactions. AMD EPYC Server CPUs deliver industry-leading Java performance because their core architectural strengths align with the fundamental resource demands of the Java Virtual Machine.

Security and cryptographic performance are critical for workloads ranging from secure web services and APIs to identity management, encrypted storage and messaging. **OpenSSL**, a widely trusted open-source toolkit, offers robust libraries and utilities for TLS/SSL, encryption, hashing and key management. Its versatility makes it essential for tasks such as establishing secure connections, managing certificates, performing encryption/decryption, creating CSRs and ensuring compliance with industry standards. For this evaluation, AMD tested Scrypt KDF (memory-hard key derivation), AES-256-GCM (wrap/unwrap the signing key) and ECDSA-P256 (sign the attestation + verify). AMD EPYC™ Server CPU excels at these workloads due to dedicated hardware accelerators, ISA (instruction set architecture) updates and a capacity for parallel execution.

MongoDB is an open-source, document-based NoSQL database designed to handle large amounts of unstructured or semi-structured data efficiently. It is widely used in modern applications such as social media platforms, content management systems, mobile apps and e-commerce sites where flexibility, scalability and performance are essential. For competitive analysis, AMD utilized the popular Yahoo! Cloud Serving Benchmark (YCSB), an open-source framework that evaluates and compares the performance of various cloud databases and storage systems under different workloads. MongoDB relies heavily on large concurrent client connections, aggressive in-memory caching, rapid serialization/deserialization of JSON/BSON data and rapid storage I/O. AMD EPYC™ Server CPU excels at the measured tasks across these four major architectural pillars.

Redis™, a popular in-memory data store, works as a key-value database, cache and message broker with optional durability features. It's popular for cloud environments and supports functions that help user productivity, including streaming, microservices and data analytics. Redis can handle various data types, including strings, lists, maps and others. Its in-memory design provides excellent performance. AMD used a tool called Redis-benchmark to measure how well Redis performs. This tool simulates multiple clients connecting to the server and measures how long it takes to complete requests. The results show the average number of requests a Redis server can handle per second. AMD EPYC™ 9006 Server CPU dominates Redis-benchmark tests through targeted architectural features that directly overcome traditional bottlenecks in memory access latency, single-core instructions-per-clock (IPC) and CPU cache sizes.

NGINX™ is a popular web server known for its flexibility to act as a reverse proxy, load balancer, mail proxy and HTTP cache. It's designed to efficiently handle client requests and deliver web content. NGINX can operate as a standalone web server or boost performance and security by serving as a reverse proxy for other servers. AMD used the popular WRK tool to evaluate performance by generating significant HTTP loads during benchmark testing. AMD EPYC™ Server CPUs deliver leading performance for NGINX by combining high core density for large concurrent connection handling with specialized AVX-512 cryptographic extensions that drastically accelerate TLS/SSL handshakes.

MySQL™, among the most widely adopted open-source database management systems, is commonly used in enterprise and cloud-native environments for decision support and transaction processing. To evaluate its performance, AMD leveraged TPROC-C based on the TPC-C benchmark® for transaction processing systems that provide targeted insights into how MySQL handles diverse enterprise demands. AMD EPYC™ Server CPU delivers leading MySQL performance by combining high core density and large DDR5 memory bandwidth to process high-concurrency transaction streams simultaneously while slashing latency through its large on-die L3 cache.

High-Performance Computing (HPC)

HPC applications often require the absolute best performance available. And it's been a major focus for AMD since 1st Gen AMD EPYC Server CPU launched in 2017, with customers consistently valuing the performance and efficiency these processors provide. With each successive CPU generation, AMD delivers steady advancements in HPC performance, bringing exceptional computing power to sectors ranging from commercial enterprises to national laboratories.
A comparison of the generational and competitive performance across a wide range of HPC workloads for 6th Gen Intel® Xeon® 6980P with 5th Gen AMD EPYC™ 9965 Server CPUs and latest 6th Gen AMD EPYC 9996 Server CPUs displays platform-level performance. It shows 6th Gen AMD EPYC™ 9996 Server CPU delivering 3.08x for NAMD, 1.80x for Quantum Espresso, 3.13x for GROMACS, and 2.90 for WRF when normalized to 6th Gen Intel Xeon 6980P. Figure 15 shows the details.

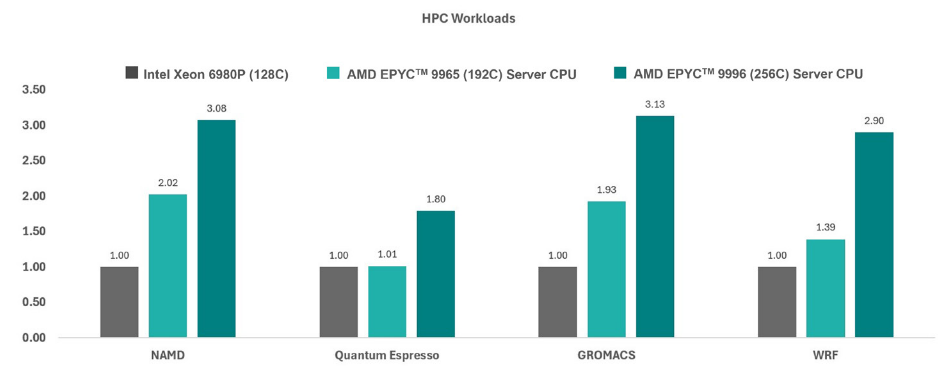


Figure 15: High-performance computing workload performance.[16, 17, 18, 19]

The **NAMD** benchmark, a standardized performance suite for evaluating high-performance computing hardware using molecular dynamics simulations, features key systems that include ApoA1 (92,224 atoms), F1-ATPase (327,506 atoms) and STMV (1,066,628 atoms). AMD EPYC Server CPUs excel at NAMD by utilizing full 512-bit AVX-512 execution pipelines to rapidly compute complex atomic forces while leveraging high core densities to maximize local parallel scaling.

Quantum ESPRESSO is an open-source suite for nanoscale electronic-structure calculations and materials modeling using density-functional theory, plane waves and pseudopotentials. The Quantum ESPRESSO 7.0 ausurf benchmark was used to compare system performance. AMD EPYC Server CPUs deliver leading performance for Quantum ESPRESSO by leveraging AVX-512 vector pipelines to accelerate dense quantum-chemical math, while utilizing memory subsystem to satisfy the high data-bandwidth demands of plane-wave simulations.

Molecular dynamics application **GROMACS** simulates Newtonian motion for systems ranging from hundreds to millions of particles. GROMACS runs exceptionally well on AMD EPYC Server CPUs because its highly parallel molecular dynamics algorithms map perfectly onto EPYC's Server CPU native AVX-512 execution pipelines, high core counts supported with large memory bandwidth.

The Weather Research and Forecasting (**WRF**) model is an advanced mesoscale numerical weather prediction system tailored for both atmospheric research and operational forecasting. It includes two dynamic cores, a data assimilation system and software that enables parallel computation and easy extensibility. WRF is versatile, supporting various meteorological applications across scales ranging from tens of meters to thousands of kilometers. AMD EPYC Server CPU dominate WRF weather modeling by using high core counts and high memory bandwidth to scale parallel calculations seamlessly.

Agentic AI Workloads

Unlike traditional single-shot inference workloads, agentic AI systems execute complex end-to-end workflows: characterizing user requests, assembling context, planning and routing actions, retrieving enterprise knowledge, querying databases, invoking APIs, parsing documents, executing transient tools, verifying results and streaming grounded responses. Overall performance depends on the combined efficiency of computing, memory, storage, networking, orchestration software and runtime isolation. In this environment, the server CPU plays a pivotal role, not merely as a supporting component for accelerators, but as the engine that drives orchestration, retrieval, tool execution and response generation across the production agentic AI pipeline.

To evaluate this emerging class of workloads, AMD worked closely with customers and industry partners to identify representative applications and develop benchmarks that reflect real-world agentic AI deployments. As the first step, distinct agentic AI workflow stages were identified: gateway, assemble context, plan and route, retrieve context and similarity search, reasoning, enterprise tools, ephemeral tools, verification, and stream Response. As a next step, representative workloads were mapped to agentic AI pipeline execution phases.

5th Gen AMD EPYC server CPUs already deliver leadership performance across the diverse workloads that define modern agentic AI deployments, as shown in Figure 16. AMD EPYC 9006 Series server CPUs build on that leadership with additional performance, scale and platform capability, enabling enterprises to support increasingly complex AI pipelines. As agentic AI continues to evolve, infrastructure demands are expected to grow rapidly with increased demands for concurrency, orchestration complexity, retrieval volume and tool execution. Designed for agentic AI, industry leading memory capacity and platform throughput, AMD EPYC 9006 Series server CPUs help future-ready AI infrastructure investments by providing the performance headroom and scalability needed for the next generation of enterprise agentic AI applications.

6th Gen AMD EPYC™ 9996 Server CPU delivers platform-level performance increases of 2.4x for server-side Java, 2.5x for OpenSSL, 3.5x for MongoDB with YCSB, 2.9x for Redis Benchmark, 3.7x for NGINX with WRK, and 2.6x for transaction processing systems based on TPC-C kit when normalized to an Intel Xeon 6980P (128C)-powered system. Figure 14 shows the results.

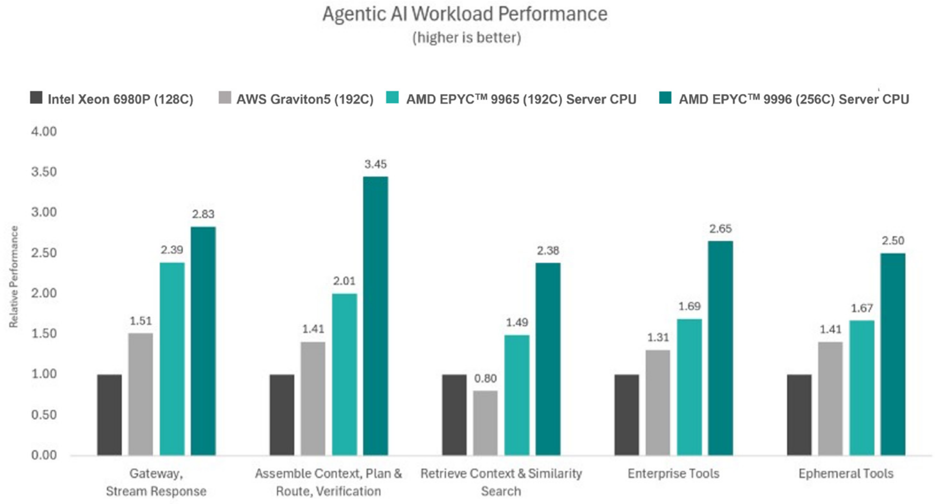


Figure 16: Agentic AI workload performance (CPU Centric) by stage.[20, 21, 22, 23, 24, 25]

Rack level

Agentic AI is changing the shape of infrastructure. As enterprises move from isolated AI experiments to production agentic systems, supporting CPU infrastructure becomes critical: Orchestration services, databases, web front ends, caches, middleware, APIs and control-plane services all need to scale efficiently within real rack power and thermal limits. Customers do not deploy benchmark headlines; they deploy racks constrained by power, cooling, floor space, software compatibility and operational readiness. Evaluated through this lens, AMD EPYC™ Server CPUs demonstrate clear rack-scale leadership.

Modelling a 100 kW rack capacity scenario, AMD has previously published estimated performance results reproduced here. 5th Gen AMD EPYC™ 9965 Server CPU delivers an estimated 2.37x the rack-level throughput of the Nvidia Vera baseline and roughly 1.6x that of Intel Xeon 6980P, as shown in Figure 17. 6th Gen AMD EPYC™ 9996 Server CPUs extend this comparison to 3.30x of Nvidia Vera-powered platform performance. (These estimates are calculated using overall geomean of six workloads including SPECrate® 2017_int_base workload.)

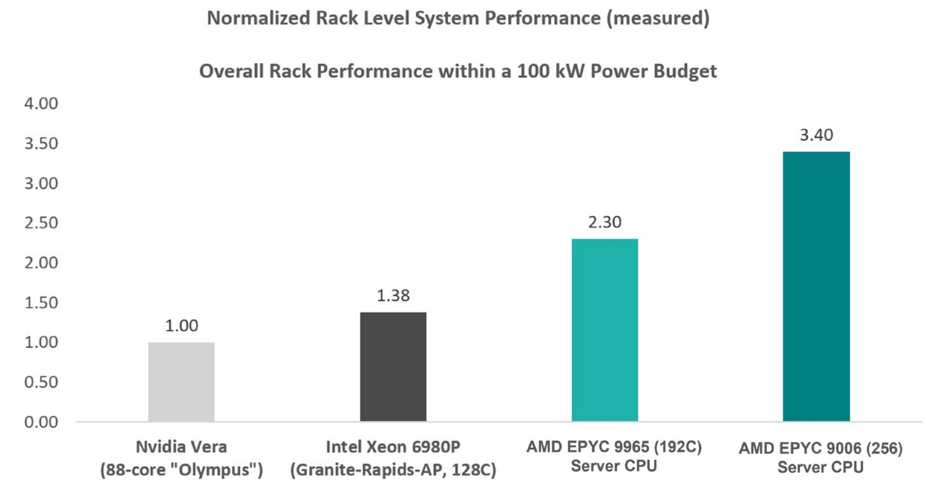


Figure 18: AMD EPYC rack-level performance leadership (measured, estimated).27

Conclusion

6th Gen AMD EPYC™ Server CPUs are best understood as platform advancements, rather than as traditional CPU refresh. New EPYC™ Server CPUs represent the culmination of six generations of EPYC innovation, combining "Zen 6" and "Zen 6c" compute architectures with industry-leading thread density, expanded memory bandwidth, next-generation I/O, advanced security, enterprise-class reliability and intelligent power management. By integrating up to 256 cores and 512 threads, 16 channels of DDR5 memory, MRDIMM support of up to 12,800 MT/s, PCIe Gen 6 connectivity and CXL 3.1, the AMD EPYC 9006 server CPU portfolio of platforms outlined in Figure 19 demonstrates the long-standing AMD philosophy of balanced system design – where optimized compute, memory, I/O, security and efficiency evolve together rather than in isolation.

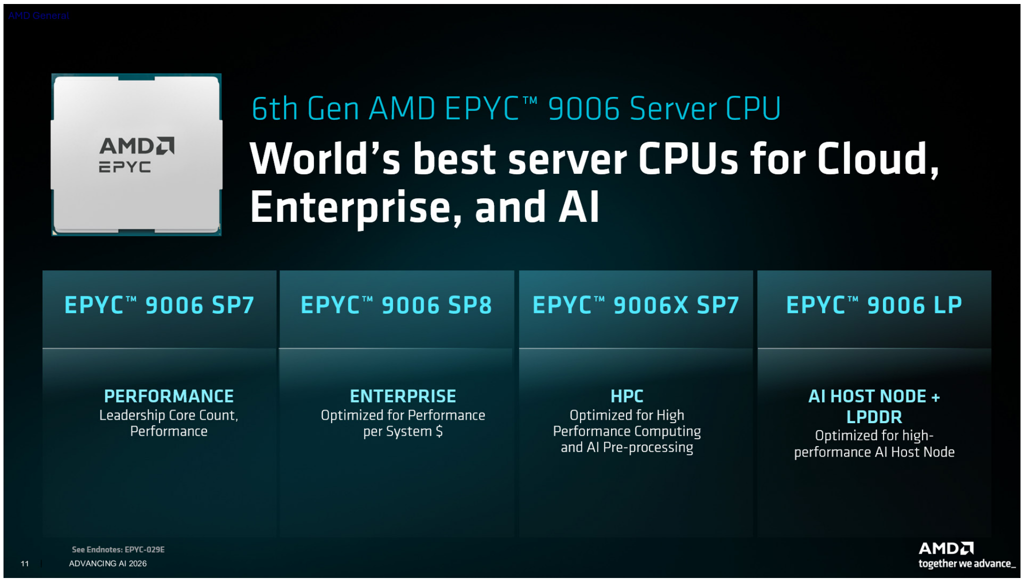


Figure 19: The AMD EPYC 9006 server CPU portfolio.2, 28

The significance of 6th Gen AMD EPYC™ Server CPUs extends beyond raw processor performance. Modern data centers are increasingly constrained by data movement, memory capacity, power availability, infrastructure utilization and software scalability. "Venice" addresses these challenges through a combination of chiplet-based scalability, AMD Infinity Fabric technology connectivity, expanded memory bandwidth, confidential-computing capabilities, enhanced RAS features, advanced vector processing and adaptive power-management technologies such as UBPS, UPP and FAST. Together, these innovations help maximize infrastructure efficiency while enabling customers to consolidate workloads, improve utilization and support increasingly complex application environments.

For cloud providers, AMD EPYC™ 9006 Server CPU delivers the density, bandwidth and efficiency required to increase service capacity and improve rack-level economics. For enterprise organizations, it provides a secure, highly scalable platform for virtualization, databases, analytics, storage and business-critical workloads. For HPC environments, AMD EPYC™ 9006 Server CPU offers the memory bandwidth, core density and vector-compute capability required for demanding technical applications. For AI infrastructure, it strengthens the CPU's role as the host processor, orchestration engine, control plane and data-movement hub that feeds accelerators and coordinates distributed AI services.

Perhaps most important, AMD EPYC™ 9006 Server CPU builds upon the maturity, ecosystem breadth and software compatibility established across six generations of EPYC CPUs leadership. Customers benefit not only from the latest architectural innovations but also from a proven platform that has been extensively validated across hyperscale cloud deployments, enterprise data centers, HPC systems and emerging AI environments. In this way, AMD EPYC™ 9006 Server CPU delivers a unique combination of innovation and continuity – enabling organizations to adopt next-generation capabilities while preserving existing software investments, operational practices and infrastructure expertise.

As the industry moves toward increasingly heterogeneous, AI-driven and data-intensive computing environments, the role of the CPU will expand rather than diminish. AMD EPYC server CPUs are designed for the future. 6th Gen EPYC CPUs provide the compute performance, memory architecture, connectivity, security and platform intelligence required to support the next generation of cloud, enterprise, HPC and AI infrastructure, reinforcing the AMD vision of a scalable, efficient and open data center platform for the decade ahead.

AMD EPYC

Footnotes and Disclaimers

1. EPYC-071: Generational SPECrate®2017_int_base throughput comparison across the AMD EPYC server processor family, 1st Gen (2017) through 6th Gen (2026), based on 2P (dual-socket) scores. 1st Gen AMD EPYC 7601 (272) is a published spec.org result at launch (Nov 2017): https://spec.org/cpu2017/results/res2017q4/cpu2017-20171114-00833.html. All other generational scores are AMD estimates based on AMD reference measurements at product launch.. Normalized to 2P AMD EPYC 7601 (272) = 1.00: 2nd Gen EPYC 7742 (est. 657) 2.4x; 3rd Gen EPYC 7763 (est. 779) 2.9x; 4th Gen EPYC 9654 (est. 1450) 5.3x; 5th Gen EPYC 9965 (est. 2740) 10.1x; 6th Gen EPYC 9996 (est. 4885) 18.0x. 6th Gen 9006 Series is a pre-production AMD internal estimate 6th Gen AMD EPYC uses Default CPU Power to describe processor power consumption (see EPYC-070). SPEC® and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. Results may vary due to factors including system configurations, software versions and BIOS settings.

2. 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 7/23/2026. "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.

3. 9xx6-050: SPECrate®2026_int_base throughput comparison based on AMD internal estimates for the 2P 6th Gen AMD EPYC 9996 and 2P NVIDIA Vera as of 7/22/2026. Preliminary performance estimates based on AMD engineering projections and subject to change. 2P 6th Gen AMD EPYC 9996: GCC 15.2, estimated SPECrate®2026_int_base score 2070. NVIDIA Vera: 88 cores, GCC 15.2, estimated SPECrate®2026_int_base score 925. For ~2.2x the performance https://nvdam.widen.net/s/nmw5vblpqd/nvidia_vera_cpu_architecture_whitepaper?nvid=nv-tblg-543584. SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. Results may vary.

4. 9xx6-049: SPECrate®2026_int_base per-core comparison based on AMD internal estimates for the 6th Gen AMD EPYC 96c HF processor and NVIDIA Vera as of 7/22/2026. Preliminary performance estimates based on AMD engineering projections and subject to change. 2P 6th Gen AMD EPYC 96c HF processor, GCC 15.2, estimated SPECrate®2026_int_base score 1210 (6.3 per core). 2P NVIDIA Vera: 88 cores, GCC 15.2, estimated SPECrate®2026_int_base score 925 (5.3 per core). For: ~1.2x the performance / core Nvidia Vera performance from: https://nvdam.widen.net/s/nmw5vblpqd/nvidia_vera_cpu_architecture_whitepaper?nvid=nv-tblg-543584. SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. Results may vary.

5. 9xx6-047: SPECrate®2017_int_base per-core comparison based on AMD internal projections and measurements as of 07/20/2026. 6th Gen AMD EPYC™ 9006 Server CPU delivers leadership 1P SPECrate®2017_int_base per-core performance versus three competing server processors, each at the AMD core configuration noted. Versus NVIDIA Vera: 1P 96C 6th Gen AMD EPYC 9996 estimated score 1350, per-core 1350 / 96 = 14.06, versus 1P 88C NVIDIA Vera score 729.5, per-core 729.5 / 88 = 8.29.. Versus Intel Xeon 6980P: 1P 128C 6th Gen AMD EPYC estimated score 1565, per-core 1565 / 128 = 12.23, versus 1P 128C Intel Xeon 6980P score 1255, per-core 1255 / 128 = 9.80; . Versus Arm AGI: 1P 96C 6th Gen AMD EPYC 9996 estimated score 1350, per-core 14.06, versus 1P 136C Arm AGI score 972, per-core 972 / 136 = 7.15. Configurations: 6th Gen AMD EPYC (96C and 128C) using AOCC 5.1 with 16x64GB 4R MRDIMM 12800 MT/s; Intel Xeon 6980P using OneAPI with 12x64GB MRDIMM 8800 MT/s; Arm AGI using GCC 13 with 12x64GB MRDIMM 8800 MT/s; NVIDIA Vera using GCC 13 with 8x96GB SOCAMM2. SPEC® and SPECint® are registered trademarks of Standard Performance Evaluation Corporation. Learn more at spec.org. Results may vary.

6. 9xx6-052: Estimated SPECrate®2026_int_base per-core comparison on the four integer subtests highlighted by NVIDIA, based on AMD internal estimates for the 6th Gen AMD EPYC 96-core CPU (downcored from 256C) and NVIDIA Vera (88-core), as of 7/22/2026. Preliminary performance estimates based on AMD internal testing and subject to change. 6th Gen AMD EPYC 96-core CPU (downcored from 256C): GCC 16.1, MRDIMM-8000, Performance Determinism. NVIDIA Vera: 88 cores, scores from NVIDIA published architecture whitepaper. Results (per-core basis): Subtest EPYC_96C_score Vera_88C_score Per-core_ratio 714.cpython_r 1513 1240 1.12x 721.gcc_r 900 817 1.01x 723.llvm_r 1061 909 1.07x 727.cppcheck_r 1006 890 1.04x Reference: https://nvdam.widen.net/s/nmw5vblpqd/nvidia_vera_cpu_architecture_whitepaper. SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. Results may vary.

7. 9xx6-053: Estimated SPECrate®2026_int_base per-core comparison across all 14 integer subtests, based on AMD internal estimates for the 6th Gen AMD EPYC 96-core CPU (downcored from 256C) and NVIDIA Vera (88-core), as of 7/22/2026. Preliminary performance estimates based on AMD engineering projections and subject to change. 6th Gen AMD EPYC 96-core CPU (downcored from 256C): GCC 16.1, MRDIMM-8000, Performance Determinism. NVIDIA Vera: 88 cores, scores from NVIDIA published architecture whitepaper. Results (per-core basis): Subtest EPYC_96C_score Vera_88C_score Per-core_ratio 706.stockfish_r 3010 1370 2.01x 707.ntest_r 1210 830 1.34x 708.sqlite_r 841 744 1.04x 710.omnetpp_r 1200 842 1.31x 714.cpython_r 1513 1240 1.12x 721.gcc_r 900 817 1.01x 723.llvm_r 1061 909 1.07x 727.cppcheck_r 1006 890 1.04x 729.abc_r 942 823 1.05x 734.vpr_r 1039 815 1.17x 735.gem5_r 1640 1300 1.16x 750.sealcrypto_r 1219 816 1.37x 753.ns3_r 1950 1670 1.07x 777.zstd_r 710 483 1.35x SPECrate2026_int_base(geomean) 1210.31 924.87 1.20x Reference: https://nvdam.widen.net/s/nmw5vblpqd/nvidia_vera_cpu_architecture_whitepaper. SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. Results may vary.

8. 9xx6-054: STREAM Triad sustainable memory bandwidth comparison based on AMD internal measurement of a 1P 6th Gen AMD EPYC 96-core (downcored from 256C) system and NVIDIA Vera (88-core), as of 7/22/2026. Preliminary performance estimates based on AMD engineering projections and subject to change. 1P 6th Gen AMD EPYC 9996 (96-core): 16x 4R MRDIMM-12800 64GB, NPS2, Power Determinism, TDP=PPT=600W, Ubuntu 24.04 kernel 6.14.0-37-generic, OS boot: amd_iommu=on iommu=pt tsc=nowatchdog nmi_watchdog=0 nokaslr amd_pstate=active. NVIDIA Vera (88-core): STREAM Triad score per Phoronix published review (https://www.phoronix.com/review/nvidia-vera-benchmarks/4). Results (overall system): Processor STREAM_Triad_GBps Normalized NVIDIA Vera 88C 1099 1.00 6th Gen AMD EPYC 96C 1298 1.18 Results may vary.

9. 9xx6-055: STREAM Triad per-core sustainable memory bandwidth comparison based on AMD internal measurement of a 1P 6th Gen AMD EPYC 96-core (downcored from 256C) system and NVIDIA Vera (88-core), as of 7/22/2026. Preliminary performance estimates based on AMD engineering projections and subject to change. 1P 6th Gen AMD EPYC 96-core (downcored from 256C): 16x 4R MRDIMM-12800 64GB, NPS2, Power Determinism, TDP=PPT=600W, Ubuntu 24.04 kernel 6.14.0-37-generic, OS boot: amd_iommu=on iommu=pt tsc=nowatchdog nmi_watchdog=0 nokaslr amd_pstate=active. NVIDIA Vera (88-core): STREAM Triad score per Phoronix published review (https://www.phoronix.com/review/nvidia-vera-benchmarks/4). Results (per-core basis): Processor Triad_GBps Cores Per-core_GBps Normalized NVIDIA Vera 88C 1099 88 12.49 1.00 6th Gen AMD EPYC 96C 1298 96 13.52 1.08 Results may vary.

10. 9xx6-037: Redis 8.8 Open Source Stable request-throughput comparison based on AMD internal testing as of 7/20/2026. 1P AMD EPYC 9996 (256C) delivers up to 2.9x the Redis 8.8 request throughput of 1P Intel Xeon 6980P (128C) Software: Redis 8.8 Open Source Stable, redis-benchmark SET and GET, one Redis server per 8 vCPUs aggregated, KVM, Ubuntu 24.04. Configurations: 1P Intel Xeon 6980P: 1P Intel Xeon 6980P, 128 cores (256 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, Supermicro SYS-222HA-TN, BIOS v1.5, Max performance BIOS profile, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P Graviton5: 1P Graviton5 (AWS m9gd.metal-48xl), 192 cores (192 vCPU), SMT OFF, 768GB memory, Ubuntu 24.04, kernel Linux 6.17.0-1009-aws. 1P AMD EPYC 9965: 1P AMD EPYC 9965, 192 cores (384 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, AMD reference platform, BIOS RVOT1006C, TDP=500W, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P AMD EPYC 9996: 1P AMD EPYC 9996, 256 cores (512 vCPU), SMT ON, 64GB DDR5-8000 RDIMM fully populated, AMD reference platform, BIOS BKC10.3 90RC4, 600W Default CPU Power, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. Data presented: 1P Intel Xeon 6980P 42636320, 1P Graviton5 50057444, 1P AMD EPYC 9965 76703115, 1P AMD EPYC 9996 124130835.50. Ratio 9996 vs 6980P: 124130835.50 / 42636320 = 2.91 (rounds to 2.9x). Results may vary. Results for AWS Graviton5 were obtained from a publicly available cloud instance. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. 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 not be directly comparable to physical server configurations due to differences in platform implementation, firmware, operating environment, memory configuration, and system tuning.

11. 9xx6-038: NGINX request-throughput comparison based on AMD internal testing as of 7/20/2026. 1P AMD EPYC 9996 (256C) delivers up to 3.7x the NGINX request throughput of 1P Intel Xeon 6980P (128C) Software: NGINX served with wrk load generator (1KB file), one NGINX server per 8 vCPUs aggregated, KVM, Ubuntu 24.04. Configurations: 1P Intel Xeon 6980P: 1P Intel Xeon 6980P, 128 cores (256 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, Supermicro SYS-222HA-TN, BIOS v1.5, Max performance BIOS profile, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P Graviton5: 1P Graviton5 (AWS m9gd.metal-48xl), 192 cores (192 vCPU), SMT OFF, 768GB memory, Ubuntu 24.04, kernel Linux 6.17.0-1009-aws. 1P AMD EPYC 9965: 1P AMD EPYC 9965, 192 cores (384 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, AMD reference platform, BIOS RVOT1006C, TDP=500W, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P AMD EPYC 9996: 1P AMD EPYC 9996, 256 cores (512 vCPU), SMT ON, 64GB DDR5-8000 RDIMM fully populated, AMD reference platform, BIOS BKC10.3 90RC4, 600W Default CPU Power, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. Data presented: 1P Intel Xeon 6980P 10526326, 1P Graviton5 15929450, 1P AMD EPYC 9965 24907610, 1P AMD EPYC 9996 38451232. Results for AWS Graviton5 were obtained from a publicly available cloud instance. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. 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 not be directly comparable to physical server configurations due to differences in platform implementation, firmware, operating environment, memory configuration, and system tuning.

12. 9xx6-039: MySQL 8.4.10 online transaction processing throughput comparison (HammerDB TPROC-C, a workload derived from the TPC-C Benchmark Standard. Results are based on TPC-C derived workloads and are not comparable to published TPC benchmark results.) based on AMD internal testing as of 7/20/2026. 1P AMD EPYC 9996 (256C) delivers up to 2.6x the MySQL 8.4.10 online transaction processing throughput (HammerDB TPROC-C) of 1P Intel Xeon 6980P (128C) Software: HammerDB TPROC-C on MySQL 8.4.10, 500 warehouse dataset, users scaled to vCPU count, KVM, Ubuntu 24.04. Configurations: 1P Intel Xeon 6980P: 1P Intel Xeon 6980P, 128 cores (256 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, Supermicro SYS-222HA-TN, BIOS v1.5, Max performance BIOS profile, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P Graviton5: 1P Graviton5 (AWS m9gd.metal-48xl), 192 cores (192 vCPU), SMT OFF, 768GB memory, Ubuntu 24.04, kernel Linux 6.17.0-1009-aws. 1P AMD EPYC 9965: 1P AMD EPYC 9965, 192 cores (384 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, AMD reference platform, BIOS RVOT1006C, TDP=500W, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P AMD EPYC 9996: 1P AMD EPYC 9996, 256 cores (512 vCPU), SMT ON, 64GB DDR5-8000 RDIMM fully populated, AMD reference platform, BIOS BKC10.3 90RC4, 600W Default CPU Power, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. Data presented: 1P Intel Xeon 6980P 15499304, 1P Graviton5 23075914, 1P AMD EPYC 9965 25168617, 1P AMD EPYC 9996 40804653. + Results for AWS Graviton5 were obtained from a publicly available cloud instance. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. 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 not be directly comparable to physical server configurations due to differences in platform implementation, firmware, operating environment, memory configuration, and system tuning.

13. 9xx6-040: OpenSSL cryptographic throughput comparison based on AMD internal testing as of 7/20/2026. 1P AMD EPYC 9996 (256C) delivers up to 2.5x the OpenSSL cryptographic throughput of 1P Intel Xeon 6980P (128C) Software: OpenSSL cryptographic throughput workload, KVM, Ubuntu 24.04. Configurations: 1P Intel Xeon 6980P: 1P Intel Xeon 6980P, 128 cores (256 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, Supermicro SYS-222HA-TN, BIOS v1.5, Max performance BIOS profile, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P Graviton5: 1P Graviton5 (AWS m9gd.metal-48xl), 192 cores (192 vCPU), SMT OFF, 768GB memory, Ubuntu 24.04, kernel Linux 6.17.0-1009-aws. 1P AMD EPYC 9965: 1P AMD EPYC 9965, 192 cores (384 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, AMD reference platform, BIOS RVOT1006C, TDP=500W, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P AMD EPYC 9996: 1P AMD EPYC 9996, 256 cores (512 vCPU), SMT ON, 64GB DDR5-8000 RDIMM fully populated, AMD reference platform, BIOS BKC10.3 90RC4, 600W Default CPU Power, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. Data presented: 1P Intel Xeon 6980P 552433496.01, 1P Graviton5 545579628, 1P AMD EPYC 9965 1073669574.93, 1P AMD EPYC 9996 1389064585.24. Results for AWS Graviton5 were obtained from a publicly available cloud instance. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. 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 not be directly comparable to physical server configurations due to differences in platform implementation, firmware, operating environment, memory configuration, and system tuning.

14. 9xx6-041: MongoDB throughput comparison based on AMD internal testing as of 7/20/2026. 1P AMD EPYC 9996 (256C) delivers up to 3.5x the MongoDB throughput of 1P Intel Xeon 6980P (128C) Software: MongoDB throughput workload, KVM, Ubuntu 24.04. Configurations: 1P Intel Xeon 6980P: 1P Intel Xeon 6980P, 128 cores (256 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, Supermicro SYS-222HA-TN, BIOS v1.5, Max performance BIOS profile, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P Graviton5: 1P Graviton5 (AWS m9gd.metal-48xl), 192 cores (192 vCPU), SMT OFF, 768GB memory, Ubuntu 24.04, kernel Linux 6.17.0-1009-aws. 1P AMD EPYC 9965: 1P AMD EPYC 9965, 192 cores (384 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, AMD reference platform, BIOS RVOT1006C, TDP=500W, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P AMD EPYC 9996: 1P AMD EPYC 9996, 256 cores (512 vCPU), SMT ON, 64GB DDR5-8000 RDIMM fully populated, AMD reference platform, BIOS BKC10.3 90RC4, DEFAULT CPU POWER 600W, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. Data presented: 1P Intel Xeon 6980P 1794003, 1P Graviton5 2088232, 1P AMD EPYC 9965 3590497, 1P AMD EPYC 9996 6195941. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. Results for AWS Graviton5 were obtained from a publicly available cloud instance. 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 not be directly comparable to physical server configurations due to differences in platform implementation, firmware, operating environment, memory configuration, and system tuning.

15. 9xx6-057: Estimated SPECjbb®2015 MultiJVM max-jOPS (server-side Java) comparison based on AMD internal testing as of 01 September 2026, normalized to 1P Intel Xeon 6980P = 1.00 (higher is better). 1P AMD EPYC 9996 (256C) delivers up to 2.4x, and 1P AMD EPYC 9965 (192C) 1.3x, the SPECjbb2015 MultiJVM max-jOPS of 1P Intel Xeon 6980P (128C); 1P AWS Graviton5 (192C) delivers 0.9x. Configuration: SPECjbb2015 MultiJVM. AMD EPYC parts: 32x 64GB 4R MRDIMM-12800, NPS2, SMT enabled, Power Determinism, CPPC enabled, Ubuntu 24.04.3 kernel 6.14, firmware BKC 10.3 (BIOS 90RC5); 6th Gen AMD EPYC 9996 at 600W Default CPU Power. 1P Intel Xeon 6980P (128C). 1P AWS Graviton5 (192C): AWS m9gd.metal-48xl, 768GB, Ubuntu 24.04. Data presented (max-jOPS): 1P Intel Xeon 6980P (128C) = 536905 (1.00x); 1P AWS Graviton5 (192C) = 494180 (0.9x); 1P AMD EPYC 9965 (192C) = 706963 (1.3x); 1P AMD EPYC 9996 (256C) = 1281876 (2.4x). Results for AWS Graviton5 were obtained from a publicly available cloud instance; cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, and other provider variables. SPEC® and SPECjbb® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. Results may vary.

16. 9xx6-031: GROMACS (benchPEP) performance comparison based on AMD internal testing as of 07/15/2026. Software: AOCC 5.1 with AOCL 5.1 (lp64) and OpenMPI 4.1.5 on the AMD EPYC parts; Intel compiler 2023.0.0 with MKL 2023.0.0 and OpenMPI 4.1.5 on Intel Xeon 6980P. All results are the mean of 3 runs. Intel Xeon 6980P: 2P (128C per socket, 256C total, SMT off, 1 thread per core), 24x 64 GiB DDR5 at 8800 MT/s (1536 GiB total), BIOS 1.4, Supermicro SYS-222HA-TN, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.55.1). AMD EPYC 9755 (5th Gen): 2P (128C per socket, 256C total, SMT off, 1 thread per core), 24x 64 GiB RDIMM DDR5 at 6400 MT/s (1536 GiB total), BIOS TVOT1009C, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.52.1). AMD EPYC 9965 (5th Gen): 2P (192C per socket, 384C total, SMT off, 1 thread per core), 24x 64 GiB RDIMM DDR5 at 6400 MT/s (1536 GiB total), NPS4, BIOS TVOT100BB, High Performance Mode, power determinism, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.56.1). 6th Gen AMD EPYC 9996 (RDIMM): 2P (256C per socket, 512C total, SMT off, 1 thread per core), 32x 64 GiB RDIMM DDR5 at 8000 MT/s (2048 GiB total), NPS2, BIOS TNIV0090D, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.56.1). 6th Gen AMD EPYC 9996 (MRDIMM): 2P (256C per socket, 512C total, SMT off, 1 thread per core), 32x 64 GiB MRDIMM DDR5 at 12800 MT/s (2048 GiB total), NPS2, BIOS TNIV0090D, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.56.1). Data presented (ns/day): 6980P 2.937; 9755 4.425; 9965 5.663; 9996 RDIMM 8.249; 9996 MRDIMM 9.195. Results may vary due to factors including system configurations, software versions and BIOS settings.

17. 9xx6-032: NAMD (20stmv) performance comparison based on AMD internal testing as of 07/15/2026, results in ns/day. Software: AOCC 5.1 with AOCL 5.1 (lp64) and OpenMPI 4.1.5 on the AMD EPYC parts; Intel compiler 2023.0.0 with MKL 2023.0.0 and OpenMPI 4.1.5 on Intel Xeon 6980P. All results are the mean of 3 runs. Intel Xeon 6980P: 2P (128C per socket, 256C total, SMT off, 1 thread per core), 24x 64 GiB DDR5 at 8800 MT/s (1536 GiB total), BIOS 1.4, Supermicro SYS-222HA-TN, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.55.1). AMD EPYC 9755 (5th Gen): 2P (128C per socket, 256C total, SMT off, 1 thread per core), 24x 64 GiB RDIMM DDR5 at 6400 MT/s (1536 GiB total), BIOS TVOT1009C, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.52.1). AMD EPYC 9965 (5th Gen): 2P (192C per socket, 384C total, SMT off, 1 thread per core), 24x 64 GiB RDIMM DDR5 at 6400 MT/s (1536 GiB total), NPS4, BIOS TVOT100BB, High Performance Mode, power determinism, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.56.1). 6th Gen AMD EPYC 9996 (RDIMM): 2P (256C per socket, 512C total, SMT off, 1 thread per core), 32x 64 GiB RDIMM DDR5 at 8000 MT/s (2048 GiB total), NPS2, BIOS TNIV0090D, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.56.1). 6th Gen AMD EPYC 9996 (MRDIMM): 2P (256C per socket, 512C total, SMT off, 1 thread per core), 32x 64 GiB MRDIMM DDR5 at 12800 MT/s (2048 GiB total), NPS2, BIOS TNIV0090D, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.56.1). Data presented (ns/day): 6980P 0.556; 9755 0.858; 9965 1.125; 9996 RDIMM 1.72; 9996 MRDIMM 1.71. Results may vary due to factors including system configurations, software versions and BIOS settings.

18. 9xx6-034: Quantum ESPRESSO (Ta205) performance comparison based on AMD internal testing as of 07/15/2026, results in wallclock time (s). Software: AOCC 5.1 with AOCL 5.1 (lp64) and OpenMPI 4.1.5 on the AMD EPYC parts; Intel compiler 2023.0.0 with MKL 2023.0.0 and OpenMPI 4.1.5 on Intel Xeon 6980P. All results are the mean of 3 runs. Intel Xeon 6980P: 2P (128C per socket, 256C total, SMT off, 1 thread per core), 24x 64 GiB DDR5 at 8800 MT/s (1536 GiB total), BIOS 1.4, Supermicro SYS-222HA-TN, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.55.1). AMD EPYC 9755 (5th Gen): 2P (128C per socket, 256C total, SMT off, 1 thread per core), 24x 64 GiB RDIMM DDR5 at 6400 MT/s (1536 GiB total), BIOS TVOT1009C, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.52.1). AMD EPYC 9965 (5th Gen): 2P (192C per socket, 384C total, SMT off, 1 thread per core), 24x 64 GiB RDIMM DDR5 at 6400 MT/s (1536 GiB total), NPS4, BIOS TVOT100BB, High Performance Mode, power determinism, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.56.1). 6th Gen AMD EPYC 9996 (RDIMM): 2P (256C per socket, 512C total, SMT off, 1 thread per core), 32x 64 GiB RDIMM DDR5 at 8000 MT/s (2048 GiB total), NPS2, BIOS TNIV0090D, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.56.1). 6th Gen AMD EPYC 9996 (MRDIMM): 2P (256C per socket, 512C total, SMT off, 1 thread per core), 32x 64 GiB MRDIMM DDR5 at 12800 MT/s (2048 GiB total), NPS2, BIOS TNIV0090D, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.56.1). Data presented (wallclock time (s)): 6980P 2864.897; 9755 2287.51; 9965 2831.003; 9996 RDIMM 1798.487; 9996 MRDIMM 1594.417. Results may vary due to factors including system configurations, software versions and BIOS settings.

19. 9xx6-035: WRF (conus2.5km) performance comparison based on AMD internal testing as of 07/15/2026, results in mean time (s). Software: AOCC 5.1 with AOCL 5.1 (lp64) and OpenMPI 4.1.5 on the AMD EPYC parts; Intel compiler 2023.0.0 with MKL 2023.0.0 and OpenMPI 4.1.5 on Intel Xeon 6980P. All results are the mean of 3 runs. Intel Xeon 6980P: 2P (128C per socket, 256C total, SMT off, 1 thread per core), 24x 64 GiB DDR5 at 8800 MT/s (1536 GiB total), BIOS 1.4, Supermicro SYS-222HA-TN, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.55.1). AMD EPYC 9755 (5th Gen): 2P (128C per socket, 256C total, SMT off, 1 thread per core), 24x 64 GiB RDIMM DDR5 at 6400 MT/s (1536 GiB total), BIOS TVOT1009C, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.52.1). AMD EPYC 9965 (5th Gen): 2P (192C per socket, 384C total, SMT off, 1 thread per core), 24x 64 GiB RDIMM DDR5 at 6400 MT/s (1536 GiB total), NPS4, BIOS TVOT100BB, High Performance Mode, power determinism, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.56.1). 6th Gen AMD EPYC 9996 (RDIMM): 2P (256C per socket, 512C total, SMT off, 1 thread per core), 32x 64 GiB RDIMM DDR5 at 8000 MT/s (2048 GiB total), NPS2, BIOS TNIV0090D, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.56.1). 6th Gen AMD EPYC 9996 (MRDIMM): 2P (256C per socket, 512C total, SMT off, 1 thread per core), 32x 64 GiB MRDIMM DDR5 at 12800 MT/s (2048 GiB total), NPS2, BIOS TNIV0090D, Red Hat Enterprise Linux 10.1 (Linux 6.12.0-124.56.1). Data presented (mean time (s)): 6980P 1.51; 9755 1.066; 9965 1.084; 9996 RDIMM 0.671; 9996 MRDIMM 0.52. . Results may vary due to factors including system configurations, software versions and BIOS settings

20. 9xx6-046: Gateway and Stream Response tier comparison based on AMD internal testing as of 07/11/2026. NGINX 1.24.0 with the WRK 4.2.0 load generator, aggregate max requests/sec across pinned instances. System configurations: Intel Xeon 6980P (128C/256T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, SNC3, SuperMicro SYS-222HA-TN, BIOS v1.5, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AWS Graviton5 (192C/192T, 1P), DDR5-8800, AWS m9gd.metal-48xl, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. Since cloud, all default BIOS and system config. AMD EPYC 9755 (128C/256T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, AMD reference platform, BIOS RVOT1006C, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AMD EPYC 9965 (192C/384T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, AMD reference platform, BIOS RVOT1006C, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AMD EPYC 9996 (256C/512T, 1P), SMT ON, 64GB DDR5-8000 RDIMM, AMD reference platform, BKC10.3 90RC4, Default CPU Power 600W cppc on power determinism, SMT ON, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. Summary results (max requests/sec): Intel Xeon 6980P = 10162179; AWS Graviton5 = 15331108; AMD EPYC 9755 = 17906196; AMD EPYC 9965 = 24320476; AMD EPYC 9996 = 28789170. Results for AWS Graviton5 were obtained from a publicly available cloud instance. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. 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 not be directly comparable to physical server configurations due to differences in platform implementation, firmware, operating environment, memory configuration, and system tuning.

21. 9xx6-042: Assemble Context, Plan and Route, and Verification tier comparison based on AMD internal testing as of 07/11/2026. TPCx-AI kit 2.0.0 end-to-end AI pipeline at Scale Factor 30, multi-instance, SMT/HT enabled on x86, median AIUCpm. Configurations Intel Xeon 6980P (128C/256T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, SNC3, SuperMicro SYS-222HA-TN, BIOS v1.5, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AWS Graviton5 (192C/192T, 1P), DDR5-8800, AWS m9gd.metal-48xl, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. Since cloud, all default BIOS and system config. AMD EPYC 9755 (128C/256T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, AMD reference platform, BIOS RVOT1006C, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AMD EPYC 9965 (192C/384T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, AMD reference platform, BIOS RVOT1006C, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AMD EPYC 9996 (256C/512T, 1P), SMT ON, 64GB DDR5-8000 RDIMM, AMD reference platform, BKC10.3 90RC4, Default CPU Power 600W cppc on power determinism, SMT ON, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. Summary results (AIUCpm (median)): Intel Xeon 6980P = 1750.36; AWS Graviton5 = 2444.80; AMD EPYC 9755 = 2704.19; AMD EPYC 9965 = 3458.79; AMD EPYC 9996 = 5982.91. Results for AWS Graviton5 were obtained from a publicly available cloud instance. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. 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 not be directly comparable to physical server configurations due to differences in platform implementation, firmware, operating environment, memory configuration, and system tuning.

22. 9xx6-043: Retrieve Context and Similarity Search tier comparison based on AMD internal testing as of 07/11/2026. FAISS IVF4096,PQ128x4fs index, sift1m dataset, k=10, aggregate QPS across pinned instances. Configurations Intel Xeon 6980P (128C/256T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, SNC3, SuperMicro SYS-222HA-TN, BIOS v1.5, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AWS Graviton5 (192C/192T, 1P), DDR5-8800, AWS m9gd.metal-48xl, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. Since cloud, all default BIOS and system config. AMD EPYC 9755 (128C/256T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, AMD reference platform, BIOS RVOT1006C, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AMD EPYC 9965 (192C/384T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, AMD reference platform, BIOS RVOT1006C, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AMD EPYC 9996 (256C/512T, 1P), SMT ON, 64GB DDR5-8000 RDIMM, AMD reference platform, BKC10.3 90RC4, Default CPU Power 600W cppc on power determinism, SMT ON, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. Summary results (aggregate QPS (mean)): Intel Xeon 6980P = 316069; AWS Graviton5 = 119179; AMD EPYC 9755 = 369252; AMD EPYC 9965 = 472079; AMD EPYC 9996 = 751453. Results for AWS Graviton5 were obtained from a publicly available cloud instance. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. 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 not be directly comparable to physical server configurations due to differences in platform implementation, firmware, operating environment, memory configuration, and system tuning.

23. 9xx6-044: Enterprise Tools tier comparison based on AMD internal testing as of 07/11/2026. TPC-H and TPC-C derivatives on MySQL, Redis benchmark, MongoDB-YCSB in 32-vCPU VMs; per-VM geometric mean of 5x subsets. Results are based on TPC-H and TPC-C derived workloads and are not comparable to published TPC benchmark results. System Configuration: Intel Xeon 6980P (128C/256T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, SNC3, SuperMicro SYS-222HA-TN, BIOS v1.5, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AWS Graviton5 (192C/192T, 1P), DDR5-8800, AWS m9gd.metal-48xl, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. Since cloud, all default BIOS and system config. AMD EPYC 9755 (128C/256T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, AMD reference platform, BIOS RVOT1006C, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AMD EPYC 9965 (192C/384T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, AMD reference platform, BIOS RVOT1006C, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AMD EPYC 9996 (256C/512T, 1P), SMT ON, 64GB DDR5-8000 RDIMM, AMD reference platform, BKC10.3 90RC4, Default CPU Power=600w cppc on power determinism, SMT ON, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. Summary results (geomean throughput): Intel Xeon 6980P = 2284701; AWS Graviton5 = 2982203; AMD EPYC 9755 = 2546290; AMD EPYC 9965 = 3867149; AMD EPYC 9996 = 6054748. Results for AWS Graviton5 were obtained from a publicly available cloud instance. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. 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 not be directly comparable to physical server configurations due to differences in platform implementation, firmware, operating environment, memory configuration, and system tuning.

24. 9xx6-045: Ephemeral Tools tier comparison based on AMD internal testing as of 07/11/2026. Multi-persona literal replay (Development, Context Compaction, Security, Media, RAG Ingestion); raw geometric mean of five per-persona throughput rates. System configurations: Intel Xeon 6980P (128C/256T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, SNC3, SuperMicro SYS-222HA-TN, BIOS v1.5, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AWS Graviton5 (192C/192T, 1P), DDR5-8800, AWS m9gd.metal-48xl, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. Since cloud, all default BIOS and system config. AMD EPYC 9755 (128C/256T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, AMD reference platform, BIOS RVOT1006C, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AMD EPYC 9965 (192C/384T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, AMD reference platform, BIOS RVOT1006C, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AMD EPYC 9996 (256C/512T, 1P), SMT ON, 64GB DDR5-8000 RDIMM, AMD reference platform, BKC10.3 90RC4, Default CPU Power 600W cppc on power determinism, SMT ON, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. Summary results (geomean throughput): Intel Xeon 6980P = 1.779; AWS Graviton5 = 2.505; AMD EPYC 9755 = 2.317; AMD EPYC 9965 = 2.970; AMD EPYC 9996 = 4.451. Results for AWS Graviton5 were obtained from a publicly available cloud instance. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. 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 not be directly comparable to physical server configurations due to differences in platform implementation, firmware, operating environment, memory configuration, and system tuning.

25. 9xx6-048: Geometric mean across the five CPU-centric Agentic AI pipeline tiers (Gateway and Stream Response; Assemble Context, Plan and Route, and Verification; Retrieve Context and Similarity Search; Enterprise Tools; Ephemeral Tools) based on AMD internal testing as of 07/11/2026. System configurations: Intel Xeon 6980P (128C/256T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, SNC3, SuperMicro SYS-222HA-TN, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AWS Graviton5 (192C/192T, 1P), DDR5-8800, AWS m9gd.metal-48xl, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AMD EPYC 9755 (128C/256T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, AMD reference platform, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AMD EPYC 9965 (192C/384T, 1P), SMT ON, 64GB DDR5-6400 RDIMM, AMD reference platform, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. AMD EPYC 9996 (256C/512T, 1P), SMT ON, 64GB DDR5-8000 RDIMM, AMD reference platform, Ubuntu 24.04.4 LTS kernel 6.17.0-29-generic, performance governor. Per-tier ratios vs Intel Xeon 6980P = 1.00: Gateway and Stream Response (AMD EPYC 9755 = 1.76x, AMD EPYC 9965 = 2.39x, AMD EPYC 9996 = 2.83x, AWS Graviton5 = 1.51x); Assemble Context, Plan and Route, and Verification (AMD EPYC 9755 = 1.54x, AMD EPYC 9965 = 2.01x, AMD EPYC 9996 = 3.45x, AWS Graviton5 = 1.41x); Retrieve Context and Similarity Search (AMD EPYC 9755 = 1.17x, AMD EPYC 9965 = 1.49x, AMD EPYC 9996 = 2.38x, AWS Graviton5 = 0.38x); Enterprise Tools (AMD EPYC 9755 = 1.11x, AMD EPYC 9965 = 1.69x, AMD EPYC 9996 = 2.65x, AWS Graviton5 = 1.31x); Ephemeral Tools (AMD EPYC 9755 = 1.30x, AMD EPYC 9965 = 1.67x, AMD EPYC 9996 = 2.50x, AWS Graviton5 = 1.41x). Consolidated geometric mean normalized to Intel Xeon 6980P = 1.00: AMD EPYC 9755 = 1.36x, AMD EPYC 9965 = 1.82x, AMD EPYC 9996 = 2.74x, AWS Graviton5 = 1.08x. See 9xx6-042, 9xx6-043, 9xx6-044, 9xx6-045, 9xx6-046 for more info. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. Results may vary due to factors including system configurations, software versions and BIOS settings.

26. 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™ 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.

27. 9xx6-056: Estimated rack-level throughput within a 100 kW rack (Dell IR7000), based on AMD internal modeling using AMD-measured and NVIDIA-disclosed performance across six workloads (SPECrate®2026_int_base, SPECjbb®2015 MultiJVM max-jOPS, NGINX, Redis, memcached, TPROC-C MySQL), as of 9/2026. Preliminary performance estimates based on AMD engineering projections and subject to change. Rack model: NVIDIA Vera 88C 450W (176 CPUs/rack, 2S), 5th Gen AMD EPYC 9965 192C 500W (150 CPUs/rack, 2S), 6th Gen AMD EPYC 9996 256C 600W (126 CPUs/rack, 2S). Results (rack throughput, normalized to Vera): Platform Rack-throughput_ratio NVIDIA Vera 88C baseline 1.00 Intel Xeon 6980P 128C 1.38 5th Gen AMD EPYC 9965 192C 2.30 6th Gen AMD EPYC 9996 256C 3.40 Results may vary. SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. Results may vary.

28. EPYC-025D: As of July 2026, 6th Gen EPYC 9996 has 256 cores and 512 threads with SMT enabled which is higher than any other publicly disclosed 1P CPU

29. 9xx6-006A: 6th Gen EPYC CPUs can support up to 2.6X memory bandwidth (up to 16 channels of 12,800MT/s DDR5 for 1.6TB/s of memory bandwidth/socket compared to 614GB/s memory bandwidth/socket of 5th Gen EPYC CPUs (9005) with 12 channels of 6,400MT/s DDR5). 6th Gen EPYC CPUs can support up to 2X the IO bandwidth (PCIe Gen 6 supported at 64GT/s per lane, EPYC 9005 platforms support PCIe Gen 5 at 32GT/s per lane) Competitor memory bandwidth/socket: Vera: ~1.2 TB/s ARM AGI = 12 * 8 * 8800 = 844800 = ~0.8 TB/s 6980P = 12 * 8 * 8800 = 844800 = ~0.8 TB/s 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/. ARM AGI specifications from: https://www.arm.com/static/az/pdf/product-brief/arm-agi-cpu-product-brief.pdf