AGENTIC AI NEEDS RACK-SCALE CPU PERFORMANCE – AMD EPYC™ DELIVERS IT TODAY



Agentic AI is changing the shape of infrastructure. As enterprises move from isolated AI experiments to production agentic systems, the 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 that lens, AMD EPYC™ processors demonstrate clear rack-scale leadership. Under the modeled 100 kW rack scenario, AMD EPYC™ 9965 delivers an estimated 2.37x the rack-level throughput of the NVIDIA Vera baseline and roughly 1.6x that of Intel Xeon 6980P. Next-generation AMD EPYC "Venice" is projected to extend the Vera comparison to 3.30x.[1] Just as important, this is infrastructure customers can build today on standard x86 platforms, not a future architecture they have to wait for.

AGENTIC AI NEEDS CPU-RICH INFRASTRUCTURE

It is easy to frame the AI buildout as a GPU story. But production agentic systems are not just model inference, they are sprawling, continuously running service environments. Every agent depends on orchestration logic, transactional databases, web and API endpoints, key-value stores, in-memory caches, and middleware that coordinate work, hold state and brokers requests across the system. These services are overwhelmingly CPU-bound, and they scale with the number of concurrent agents rather than the size of any single model.

As agentic deployments move into production, the volume of this supporting infrastructure grows with them. The processor platform that hosts these services becomes a primary determinant of how many agents an enterprise can actually run, and at what cost. This is the layer where general-purpose CPU capacity, not accelerator peak performance, sets the ceiling.

WHY RACK-LEVEL PERFORMANCE IS THE RIGHT METRIC

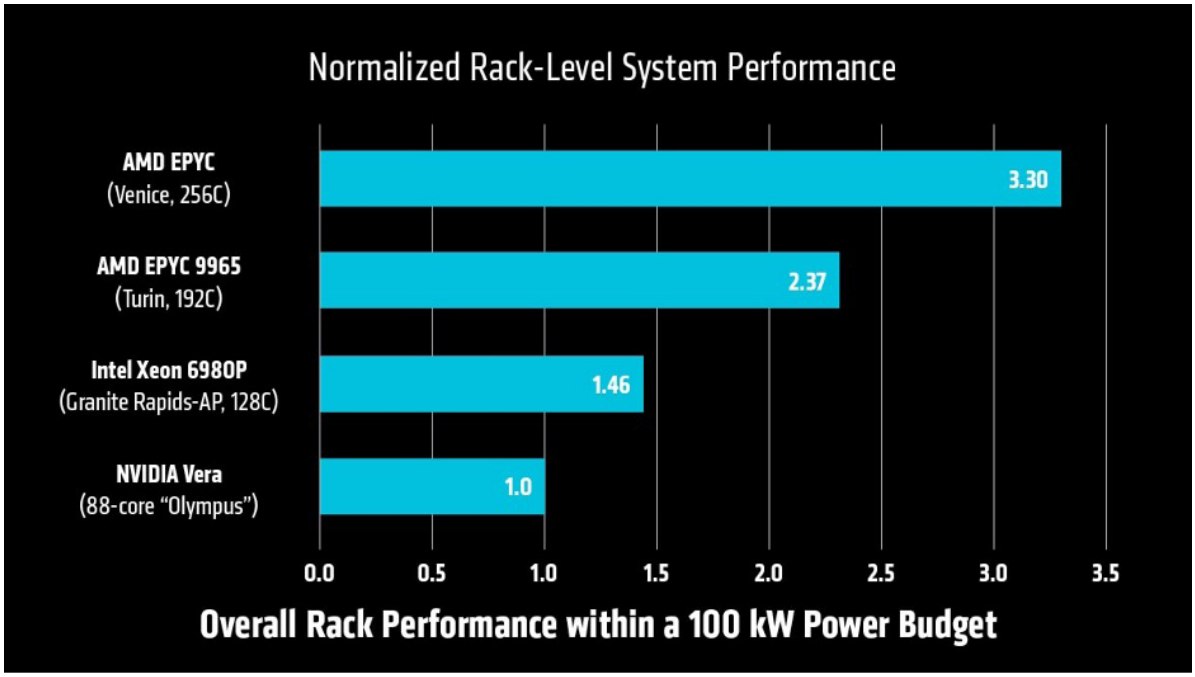
Component benchmarks describe a chip. They do not describe what a customer can deploy. Data centers are provisioned in racks, and racks are bounded by a fixed power and thermal budget, finite floor space, software-compatibility requirements, and operational readiness. The question that determines real capacity is not "how fast is one socket" but "how much useful work fits inside a 100 kW rack."

That's the lens this analysis uses. All configurations are normalized to a modeled 100 kW rack built on 2P (two-processor) platforms, so the comparison reflects deployable service capacity rather than isolated peak processor behavior. Higher-density configurations translate directly into more service capacity per rack. This is what drives capital efficiency, floor-space utilization and operational simplicity.

AMD EPYC RACK-LEVEL PERFORMANCE LEADERSHIP

Across the evaluated workloads – general-purpose compute, server-side Java, web serving, key-value, in-memory caching and relational databases – AMD EPYC leads the modeled rack-level results decisively. AMD EPYC 9965 ("Turin," 192C) delivers a 2.37x normalized geometric mean advantage over NVIDIA Vera (88-core "Olympus"), with Intel Xeon 6980P ("Granite Rapids-AP," 128C) turning in 1.46x over NVIDIA Vera. When AMD EPYC "Venice" (256C) arrives, it extends AMD's advantage to 3.30x. The gains hold across the entire workload set rather than depending on a single favorable benchmark.

The pattern is consistent: As core density rises within the fixed power envelope, aggregate service throughput rises with it. For the transactional, web-serving and middleware tiers that surround agentic systems, that means materially more concurrency and responsiveness per rack, the qualities that ultimately govern how many agents an environment can sustain.

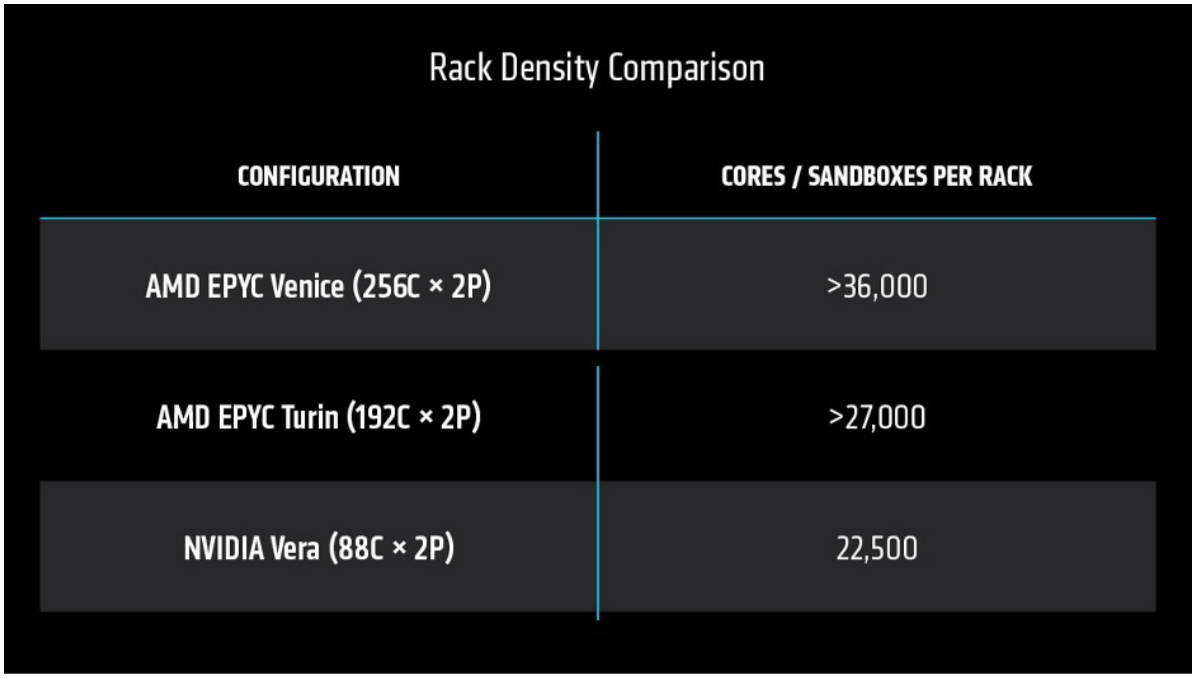


Server CPU Rack Configuration Summary (Normalized)

PARAMETER	NVIDIA VERA	INTEL XEON GNR-AP 6980P	AMD EPYC (TURIN) 9965	AMD EPYC (VENICE)
Cores Per Socket	88	128	192	256
Core Per Socket (Normalized)	1.0	1.45	2.18	2.90
2P System Power (Normalized)	1.0	1.18	1.18	1.41
Nodes Per Rack (Normalized)	1.0	0.85	0.85	0.71
Rack Power Budget	100 kW	100 kW	100 kW	100 kW

SHIPPING DENSITY TODAY, NOT PROPRIETARY PROMISES

Rack density has become a headline metric, and rightly so; it's a direct proxy for deployable value, and it's where AMD's currently available solutions stand out. An AMD EPYC "Turin" deployment in a Dell PowerEdge IR7000, or any comparable liquid-cooled rack, supports more than **27,000 CPU cores** per rack today; next-generation AMD EPYC "Venice" is architected to scale beyond **36,000 cores in the same rack class**. Sandboxes and CPU cores aren't directly equivalent, but as a directional measure of rack-scale compute density the picture is clear: The density positioned as future-looking is already being exceeded with standard infrastructure available now.



These AMD deployments run on standard liquid-cooled data center equipment and the x86 software ecosystem enterprises already operate, with no new rack architecture required – preserving software continuity, reducing migration friction and shortening time-to-production.

METHODOLOGY AND WORKLOAD DETAILS

The workload suite spans the infrastructure dimensions most relevant to agentic AI service environments, using established benchmarks as proxies:

- **General-purpose computing:** SPEC CPU 2017 Integer Rate
- **Server-side Java**: a SPECjbb2015-derived workload measuring throughput and latency-sensitive business-logic execution
- **Web serving:** NGINX with the WRK tool, under sustained concurrent request load
- **Key-value store:** redis-benchmark, for high-speed in-memory operations
- **In-memory caching/analytics:** Memcached with memtier_benchmark
- **Relational databases:** TPROC-C, a TPC-C-derived OLTP proxy, on MySQL

The set doesn't model full end-to-end agent pipelines; it isolates the infrastructure layers those pipelines depend on. Comparisons are performed at rack level using a reference 100 kW envelope with 2P platforms, with system power and nodes-per-rack normalized to NVIDIA Vera. Because the "Venice" and Vera figures reflect modeled and projected configurations, results are presented as estimates within the stated rack-power constraint.

THREADED PERFORMANCE

In addition to rack-level performance and energy efficiency, per-core performance remains a critical consideration for some workloads. AMD has consistently led on this metric for demanding workloads such as databases, analytics, simulations and host processing in multi-GPU server environments. Our "Venice" 64-core CPU is estimated to deliver a 27% performance-per-core advantage compared to the Vera 88-core processor. Even at a higher core count, the 96-core "Venice" CPU is projected to still deliver 11% higher performance-per-core than the Vera 88-core processor.

CONCLUSION: DEPLOYABLE PERFORMANCE WINS

Agentic AI infrastructure should be planned at the rack level, not around isolated component claims. On that basis the conclusion is straightforward: AMD EPYC delivers higher deployable CPU throughput, x86 software continuity and a standards-based path to dense, AI-supporting infrastructure. And it's available today on shipping platforms. For enterprises scaling toward production agentic AI, that combination of density, compatibility and deployability is what turns performance into capacity.

ENDNOTES

1. Results are based on modeled rack-level configurations using publicly available and internal benchmark data. Results may not reflect actual deployed system performance.

DISCLAIMERS

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18u.

COPYRIGHT NOTICE

© 2026 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, Versal, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective owners. Certain AMD technologies may require third-party enablement or activation. Supported features may vary by operating system. Please confirm with the system manufacturer for specific features. No technology or product can be completely secure. PID5091750