Performance Projections Methodology:

Computing in the Agentic AI Era

June 2026

Introduction

This document defines the performance methodology used for estimating performance data for Nvidia, Intel and AMD server CPUs for agentic AI.

This document demonstrates rack-level performance for platforms powered by the following processors within a 100 kW rack power constraint:

Nvidia Vera (88C)
Intel® Xeon® 6980P (128C)
AMD EPYC™ 9965 (192C)
AMD EPYC™ "Venice" (256C)

Assumptions and Considerations

All performance projections are based on existing data from published results, AMD internal testing and third-party testing. For the purposes of fair comparisons, the following considerations were used:

- Performance estimations assume all nodes are 2-socket systems.
- A single rack consists of dual-socket (2P) nodes with a 100kW power constraint.
- Performance at rack-level is within the 100kW power constraint.

This document discusses the six workload projections below:

- Estimated SPECrate®2017_int_base
- Server-Side Java® Multi-JVM Max jOPS
- Web serving (NGINX with WRK)
- Key-Value store (Redis)
- In-memory caching (Memcached)
- Relational databases (TPROC-C)*

*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.

Performance Estimation for Node-Level (2-Socket) Platforms:

AMD performance references is based on published AMD testing of the NVIDIA Grace Superchip and comparison systems. This blog has data for the following platforms:

- Nvidia Grace 2x72C Superchip
- Intel® Xeon® 6980P (128C)
- AMD EPYC 9965 (192C)

Nvidia Vera 88C Node Level (2P) Platform Performance Estimation

The Nvidia Vera performance of the six workloads are derived by scaling above Nvidia Grace Superchip platform performance by applying a 1.63x factor, based on overall geometric mean of select test results from the Phoronix blog published May 26, 2026 (Link).

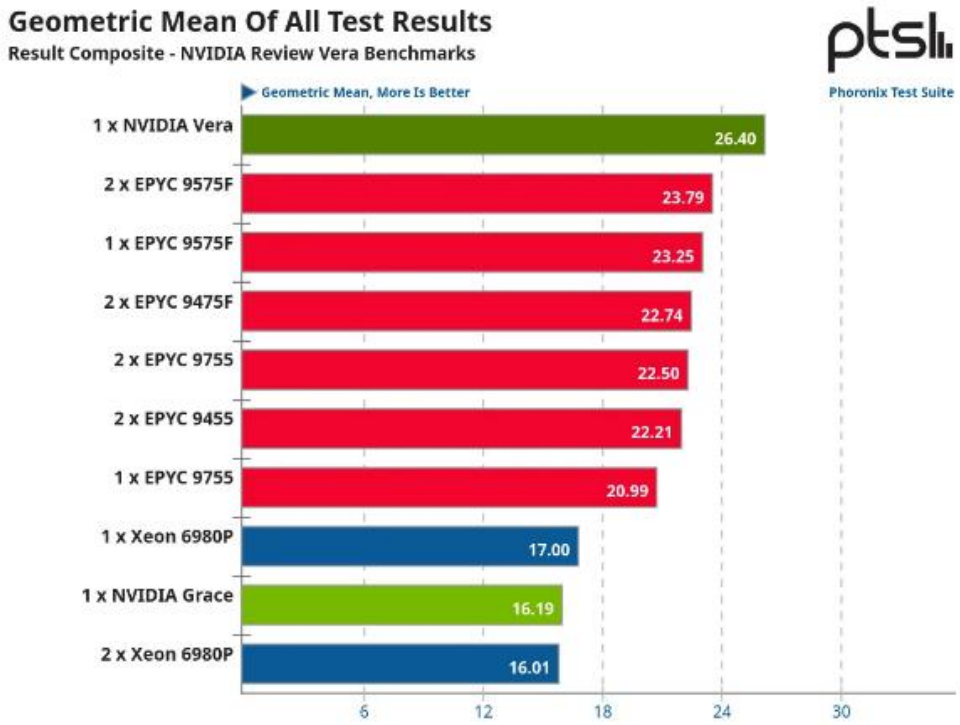

Fig. 1 – Geometric Mean of Test Results, Phoronix Media

Rack Level Power and Performance Derivations Method

Estimation for the power used by a single node and nodes per rack within 100 kW power constraint shown below are normalized for Nvidia Vera:

	Nvidia Vera (88C)	Intel Xeon 6980P (128C)	AMD EPYC 9965 (192C)	AMD EPYC "Venice" (256C)
Cores Per CPU	88	128	192	256
Cores Per CPU (Normalized)	1.0	1.45	2.18	2.90
2P Node 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

*2P server power based on processor TDPs and component estimates

Using the estimated power of a dual-socket (2P) node as above, number of nodes for each platform are calculated applying a 100 kW power constraint. Then rack level performance is derived using formula:

Rack level performance = *number of nodes supported @100 kW* x *single-node performance*

	Nvidia Vera (88C)	Intel Xeon 6980P (128C)	AMD EPYC 9965 (192C)	AMD EPYC "Venice" (256C)
Cores Per Node (Normalized)	1.0	1.45[i]	2.18[ii]	2.90
Cores Per Rack (Normalized)	1.0	1.24	1.86	2.08
Nodes Per Rack (Normalized)	1.0	0.85	0.85	0.71
Single Node Performance Derivation	Estimated as ~1.63x of the Nvidia Grace[iii]	N/A	N/A	Estimated as ~1.7x of the AMD EPYC 9965 SPECrate®2017_int_base and internal testing
Rack-Level Performance Derivation[iv]	Number of Nodes x Node level performance of Nvidia Vera	Number of Nodes x Node level performance of Intel Xeon 6980P	Number of Nodes x Node level performance or AMD EPYC 9965	Number of Nodes x Node level performance of AMD EPYC "Venice" 256C

(i) Based on published AMD testing
(ii) Based on AMD Internal Testing, October 2024
(iii)Based on Phoronix Data, Nvidia Vera shows ~1.63× improvement in mixed real-world workloads.
(iv) Performance derivations are not directly comparable but suggest similar generational gains, subject to workload characteristics

Performance Projections

By following the above derivation method, we arrive at the relative performance-per-rack estimates below:

	Nvidia Vera (88C)	Intel Xeon 6980P "GNR-AP" (128C)	AMD EPYC 9965 "Turin" (192C)	AMD EPYC "Venice" (256C)
SPECrate®2017_int_base	1 (est.)	1.47	1.60	2.40 (est.)
Server-side Java® multi-JVM max	1	2.34	2.93	3.76
Web Serving (NGINX)	1	1.18	2.37	3.30
Key-Value Store (Redis)	1	1.31	2.23	3.10
In-Memory Caching (Memcached)	1	0.93	2.49	3.47
Relational Databases (TPROC-C)	1	1.99	2.91	4.05
Geometric Mean	**1**	**1.46**	**2.37**	**3.30**

Threaded Performance Derivations Method

In addition to rack-level performance and energy efficiency, per-core performance is also very important metric. AMD has consistently led on this metric for demanding workloads such as databases, analytics, simulations, and host processing in multi-GPU server environments.
To estimate per-core performance, performance is estimated for Nvidia Vera 88C as described above in this document. For AMD EPYC "Venice" 64C and 96C per-core performance, a similar two-socket platform performance is estimated and then it is normalized to Nvidia Vera 88C two socket platform per-core performance using industry standard established practice of using SPECrate®2017_int_base.

	Nvidia Vera (88C)	AMD EPYC "Venice" (64C)	AMD EPYC "Venice" (96C)
Cores Per CPU	88	64	96
Node Level SPECrate®2017_int_base*	1.0	0.92	1.21
Per-core SPECrate®2017_int_base*	1.0	1.27	1.11

*2P Server estimates

AMD EPYC "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 Vera 88-core processor for the power used by a single core.

Conclusion

Under the 100 kW rack constraint used in this analysis, rack-level throughput is determined by the product of estimated dual-socket (2P) node performance and supported node count per rack. Applying this method across the six modeled workloads yields normalized geometric mean normalized rack-level performance of 1.00 for Nvidia Vera, ~1.46 for Intel Xeon 6980P “GNR-AP”, ~2.37 for AMD EPYC 9965 “Turin” and ~3.30 for AMD EPYC “Venice.” These results indicate that, within the assumptions of this model, higher rack compute density and stronger rack-level performance can materially increase deployable throughput for general-purpose agentic AI infrastructure.

In addition, greater cores per rack may be relevant for agentic AI deployments, where infrastructure often supports a mix of concurrent orchestration, retrieval, data services and application-layer processing around model execution. Within that context, higher core density could suggest the ability to support more concurrent agent-driven workflows or related services within a given power and space envelope, which may have implications for infrastructure utilization, service capacity per rack and the economics of scaling.

Because these estimates rely on published results, internal measurements and projection-based scaling factors, they are intended to provide directional comparison rather than direct measured rack benchmarks.