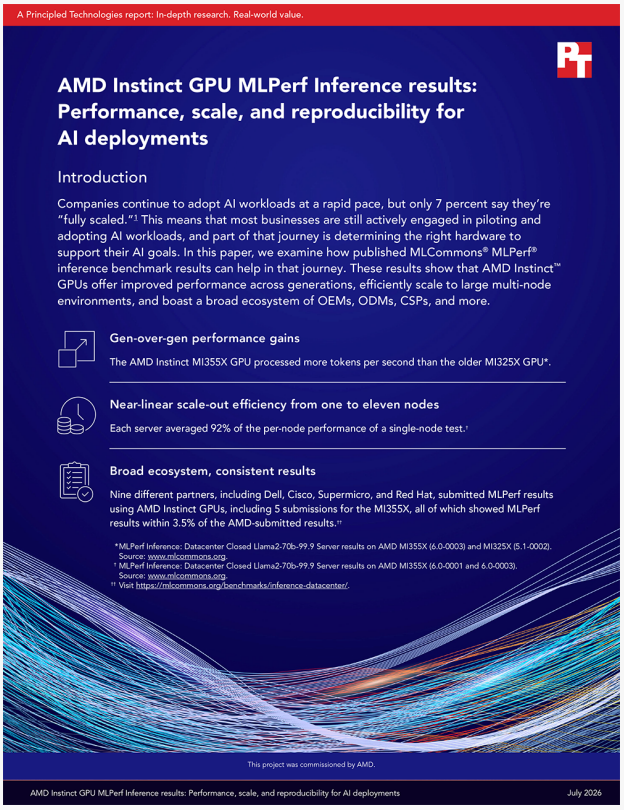
PT releases research study showcasing MLPerf Inference results for AMD Instinct GPUs

New report examines performance, scalability, and consistency of AMD Instinct GPU deployments using published MLPerf inference benchmark results

SAN JOSE, CA, UNITED STATES, July 8, 2026 /EINPresswire.com/ -- Principled Technologies (PT) has published a new research report analyzing publicly available MLPerf® Inference benchmark results for AMD Instinct™ GPUs. The study explores generational performance improvements, large-scale deployment efficiency, and result consistency across multiple hardware vendors. Based on published MLPerf data, the report found that the AMD Instinct MI355X GPU delivered substantially higher inference throughput than the previous-generation MI325X and scaled efficiently across a multi-node deployment.



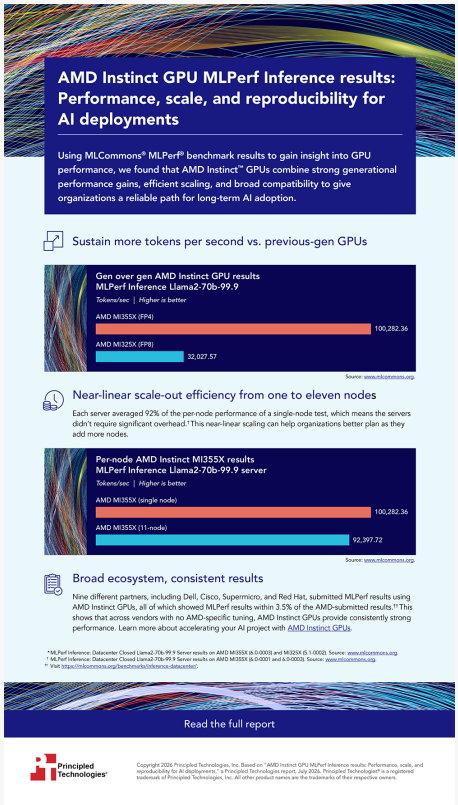
Report: AMD Instinct GPU MLPerf Inference results: Performance, scale, and reproducibility for AI deployments

The report, AMD Instinct GPU MLPerf Inference results: Performance, scale, and reproducibility for AI deployments, reviews MLPerf Datacenter Inference submissions with a focus on Llama 2 70B workloads. PT compared results from AMD- and partner-submitted benchmarks to evaluate how AMD Instinct GPUs perform across generations and across a broad ecosystem of OEM platforms.

According to the report, AMD-submitted MLPerf results showed the AMD Instinct MI355X achieving 100,282.36 tokens per second on the Llama2-70B-99.9 Server benchmark, compared to 32,027.57 tokens per second for the AMD Instinct MI325X, representing approximately 3.1 times the throughput. Partner-submitted results demonstrated similar gains, reinforcing the reproducibility of the results across vendors.

The report also examined multi-node scaling performance. An 11-node cluster containing 87 AMD Instinct MI355X GPUs achieved more than 1 million tokens per second, with each node maintaining roughly 92 percent of the throughput observed in the single-node benchmark. According to PT, this level of efficiency indicates near-linear scaling for large AI inference deployments.

“These results tell us two things: first, that the new generation of this AMD Instinct GPU delivered improved performance over the older generation, and second, that non-AMD submitters achieved performance within 3.5 percent of the AMD-submitted results for both generations,” the report states. “The fact that AMD and other OEMs achieved such similar results shows that customers might reasonably expect AMD performance to remain strong and consistent regardless of the underlying OEM hardware.”



Infographic: AMD Instinct GPU MLPerf Inference results: Performance, scale, and reproducibility for AI deployments

PT also highlighted the value of benchmark transparency when evaluating AI infrastructure investments.

“MLPerf is a targeted benchmark suite that can provide insight into GPU performance. It helps us gauge gen-over-gen performance, scaling performance, and consistency across partner implementations,” the report notes. “When we look at these results for the AMD Instinct GPU family, notably the MI355X, we see that AMD customers can enjoy generational performance boosts, near-linear scaling in multi-node environments, and broad OEM support for their GPU-enabled AI inference workloads.”

The study further observed that nine organizations—including Cisco, Dell, HPE, Oracle, Supermicro, and Red Hat—submitted MLPerf v6.0 inference results using AMD Instinct GPUs, demonstrating support across a wide ecosystem of hardware and software providers.

To learn more, read the full Principled Technologies report or view the infographic.

About Principled Technologies, Inc.

Principled Technologies, Inc. is the leading provider of technology marketing and learning & development services.

Principled Technologies, Inc. is located in Durham, North Carolina, USA. For more information, please visit www.principledtechnologies.com.

Sharon Horton
Principled Technologies, Inc.
press@principledtechnologies.com
Visit us on social media:
LinkedIn
Facebook
YouTube
X
