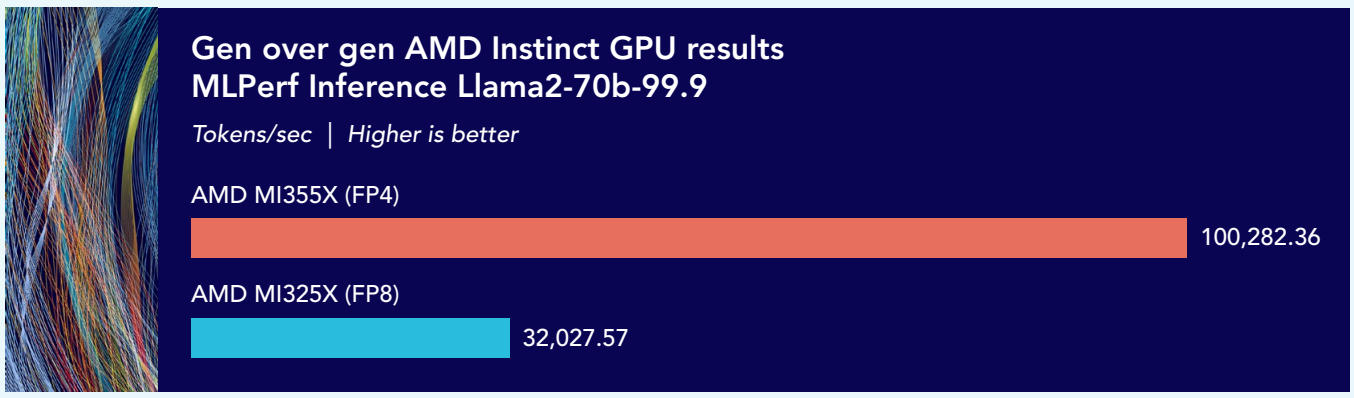
AMD Instinct GPU MLPerf Inference results: Performance, scale, and reproducibility for AI deployments

Using MLCommons® MLPerf® benchmark results to gain insight into GPU performance, we found that AMD Instinct™ GPUs combine strong generational performance gains, efficient scaling, and broad compatibility to give organizations a reliable path for long-term AI adoption.

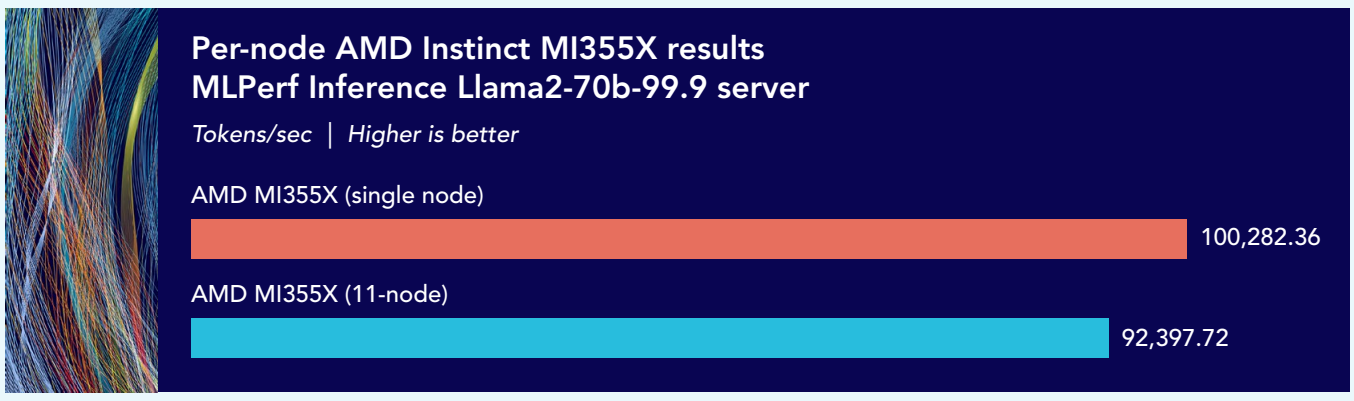
Sustain more tokens per second vs. previous-gen GPUs



Source: www.mlcommons.org.

Near-linear scale-out efficiency from one to eleven nodes

Each server averaged 92% of the per-node performance of a single-node test, which means the servers didn't require significant overhead.† This near-linear scaling can help organizations better plan as they add more nodes.



Source: www.mlcommons.org.

Broad ecosystem, consistent results

Nine different partners, including Dell, Cisco, Supermicro, and Red Hat, submitted MLPerf results using AMD Instinct GPUs, all of which showed MLPerf results within 3.5% of the AMD-submitted results.†† This shows that across vendors with no AMD-specific tuning, AMD Instinct GPUs provide consistently strong performance. Learn more about accelerating your AI project with AMD Instinct GPUs.

*MLPerf Inference: Datacenter Closed Llama2-70b-99.9 Server results on AMD MI355X (6.0-0003) and MI325X (5.1-0002). Source: www.mlcommons.org.
† MLPerf Inference: Datacenter Closed Llama2-70b-99.9 Server results on AMD MI355X (6.0-0001 and 6.0-0003). Source: www.mlcommons.org.
†† Visit https://mlcommons.org/benchmarks/inference-datacenter/.

Read the full report

Principled Technologies®