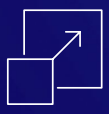
AMD Instinct GPU MLPerf Inference results: Performance, scale, and reproducibility for AI deployments

Introduction

Companies continue to adopt AI workloads at a rapid pace, but only 7 percent say they're "fully scaled."[1] This means that most businesses are still actively engaged in piloting and adopting AI workloads, and part of that journey is determining the right hardware to support their AI goals. In this paper, we examine how published MLCommons® MLPerf® inference benchmark results can help in that journey. These results show that AMD Instinct™ GPUs offer improved performance across generations, efficiently scale to large multi-node environments, and boast a broad ecosystem of OEMs, ODMs, CSPs, and more.



Gen-over-gen performance gains

The AMD Instinct MI355X GPU processed more tokens per second than the older MI325X GPU*.

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

Each server averaged 92% of the per-node performance of a single-node test.†

Broad ecosystem, consistent results

Nine different partners, including Dell, Cisco, Supermicro, and Red Hat, submitted MLPerf results using AMD Instinct GPUs, including 5 submissions for the MI355X, all of which showed MLPerf results within 3.5% of the AMD-submitted results.††

*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/.

This project was commissioned by AMD.

What is MLPerf?

Run by the MLCommons consortium, the MLPerf benchmark suite measures AI and machine learning performance. In this paper, we'll look at public submissions for the MLPerf Inference: Datacenter benchmark, which measures system performance of pre-trained models under specific latency constraints. Submissions marked as "Closed" must use fixed-model and pre-processing parameters to keep submissions comparable. "Open" submissions are useful proof points, but they should not be presented as direct like-for-like comparisons with Closed results. Here, we look mainly at Closed submissions, with some references to Open submissions. We did not perform our own testing for this report.

The MLPerf Inference: Datacenter benchmark shows how quickly a system can execute each model in three scenarios: offline (measuring maximum throughput), server (representing real-time workloads such as chatbots), and interactive (stricter latency requirements that are more realistic for live workloads).

The benchmark outputs a throughput metric in terms of tokens per second. Models can run in the 99th percentile mode or 99.9th percentile mode, meaning that 99% or 99.9% of requests must meet the accuracy and latency constraints. Read more.

Generational momentum

The GPU market is exploding alongside AI growth, with projections for a market share of $325.96 billion by 2031.[2] GPU manufacturers must continue evolving to keep up with increasing AI workload demands. Customers want to know that their investment in a certain GPU ecosystem will continue to support their needs, so they can avoid an expensive rip-and-replace in the future.

Comparing MLPerf results from older and newer AMD GPU models show a roadmap that continues to deliver more usable performance for inference workloads over time, allowing continued optimizations in your investment. In this paper, we compare MLPerf submissions for the older MI325X GPU based on the CDNA 3 architecture and the newer MI355X based on the CDNA 4 architecture to show the performance improvement of AMD Instinct GPUs from 2024 to 2025.

Why precision matters

In the comparisons we assess in this report, the precision for the older MI325X GPU tests was 8-bit floating point (FP8) while the precision for the newer MI355X GPUs was 4-bit floating point (FP4). This contributes to the superior performance of the MI355X, as precision affects both throughput and accuracy. In MLPerf Llama 2 70B results, FP8 represents each model weight in 8 bits, while FP4 halves that to 4 bits, allowing larger batch sizes within the same GPU memory capacity.

While FP4 increases throughput performance, lower precision can reduce accuracy. The MLPerf 99.9 accuracy threshold requires that a system achieve scored throughput only when its output remains within 0.1 percent of the reference model's quality target. However, when we compare the 99.9th percentile test results in this report with the llama2-70b-99 tests in the same MLPerf submissions, the throughput for each GPU remains the same.[3] This confirms that neither submission sacrificed output quality to achieve its throughput numbers, which can assure customers deploying these configurations in production environments where model accuracy is non-negotiable.

We compare two different sets of results: AMD-submitted results, and OEM-submitted results. AMD-submitted results represent highly tuned reference submissions. (Note: This set of results uses different MLPerf versions [5.1 vs. 6.0], which means the underlying test harnesses may differ in ways that impact performance.) The OEM-submitted results may or may not reflect the best GPU tunings. Together, these results paint a good picture of performance gains of AMD GPUs across generations.

AMD-submitted results: AMD Instinct MI355X GPU vs. AMD Instinct MI325X GPU

When comparing the Llama2-70b-99.9 Server MLPerf results from submission 6.0-0003 (AMD Instinct MI355X GPU on v6.0) and submission 5.1-0002 (AMD Instinct MI325X GPU on v5.1), we see that the MI355X (at FP4 precision) with 100,282.36 tokens/sec[4] delivered 3.1x the performance of the MI325X (with FP8 precision), with 32,027.57 tokens/sec[5] (see Figure 1). For customers, this 3.1x increase means more throughput per server, more inference capacity, and potentially better infrastructure efficiency.

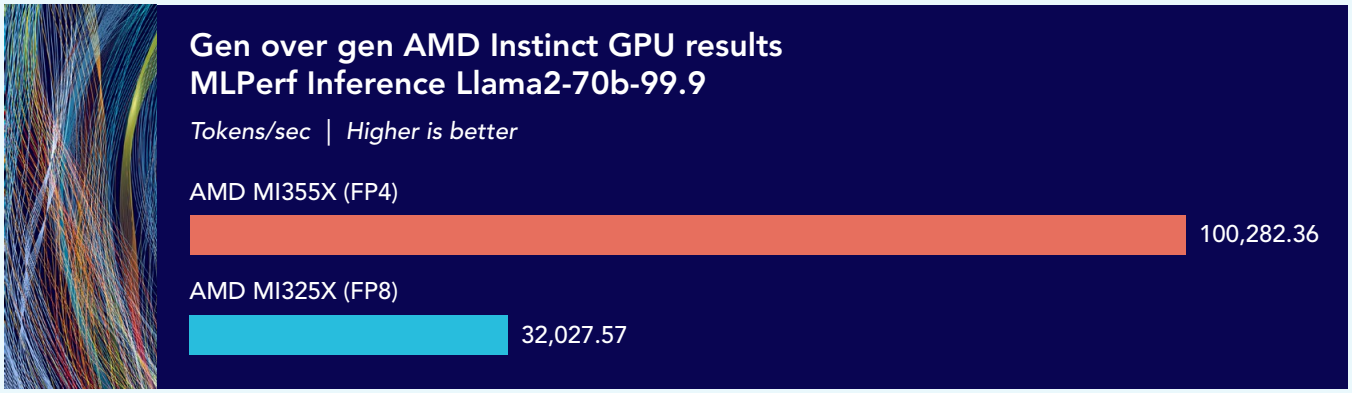


Figure 1: MLPerf Inference: Datacenter Closed Llama2-70b-99.9 Server results on AMD MI355X (6.0-0003) and MI325X (5.1-0002). Source: www.mlcommons.org.

OEM-submitted results: AMD Instinct MI355X GPU vs. AMD Instinct MI325X GPU

When comparing the llama2-70b-99.9 Server MLPerf submission 6.0-0027 (MI355X on Dell) and submission 6.0-0070 (MI325X on MangoBoost), we see that the MI355X GPU, with 96,734.37 tokens/sec[6] at FP4 precision, outperformed the MI325X GPU, with 31,888.9 tokens/sec[7] at FP8 precision, by 3x on the Llama2-70b-99.9 model test (see Figure 2). These partner-submitted results shows similar results to the AMD-submitted gen-over-gen result, reinforcing the fact that these types of gains are reproducible across vendors.

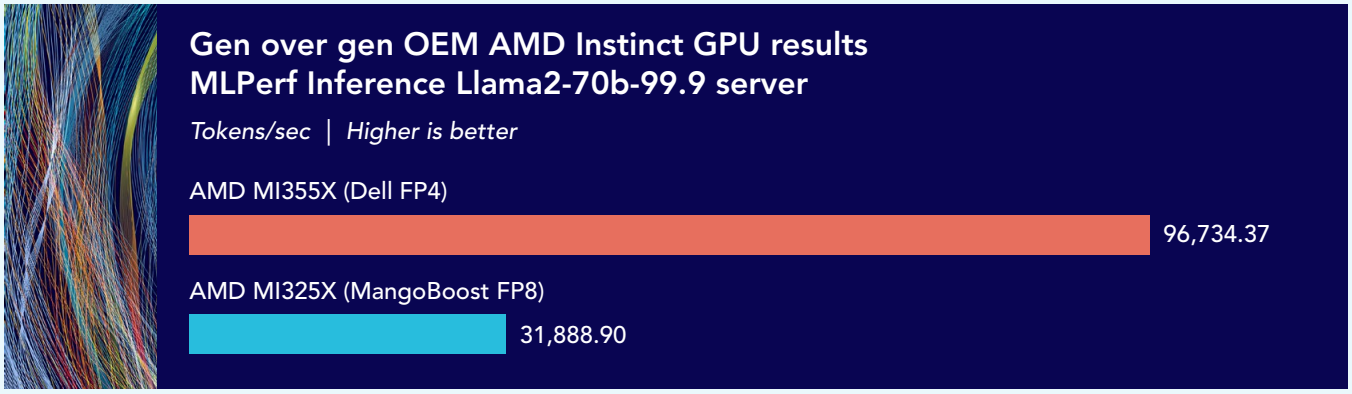


Figure 2: MLPerf Inference: Datacenter Closed Llama2-70b-99.9 Server results on AMD MI355X (6.0-0027) and MI325X (6.0-0070). Source: www.mlcommons.org.

What this means for you

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.

As noted in a previous section, precision is important, and according to the AMD spec sheet for the MI355X GPU, the theoretical peta floating point operations per second (PFLOPs) at FP4 is 10.1 and 5 at FP8. This shows that the MI355X could handle up to twice the throughput at FP4 versus at FP8. If we assume that the 2:1 ratio holds true in the MLPerf results, the new Instinct GPU model would still deliver 1.5x the performance of the older model.

Buyers may fear that manufacturers are marketing products with optimized performance data that they will find difficult to replicate. 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.

For organizations looking to improve their text summarization workloads, chatbots, Q&A systems, reasoning, agentic workflows, and other LLM inference use cases, these MLPerf results show that investing in AMD GPUs can provide performance that is consistent across tech vendors and will likely continue to improve with new generations.

Scaling and workload support

Over 1 million tokens/second across 11 nodes

For customers looking to deploy large, multi-node AI workloads, determining GPU performance is not as simple as multiplying results from a single-node test by the number of nodes. The more nodes you add to a single workload, the more overhead is required to keep all the parts communicating effectively.

To show how GPU and server frameworks may scale, we present AMD-submitted MLPerf Llama2-70b results on an 11-node cluster containing a total of 87 MI355X GPUs. Because this multi-node test (Public ID 6.0-0001) used the same precision (FP4), MLPerf version (6.0) and GPU, we can easily calculate how efficiently the AMD Instinct MI355X can scale. The 11-node cluster broke the one million tokens/sec barrier, which, according to AMD, is their first time reaching this milestone.[8] With a Llama2-70b-99.9 Server score of 1,016,375 tokens/sec, each of the 11 nodes achieved a throughput of 92,397 tokens/sec. In the single-node test, the system achieved 100,282.36 tokens/sec. This means that when scaled to an 11-node cluster, each node operated at roughly 92 percent efficiency, which is very close to linear scaling (see Figure 3).

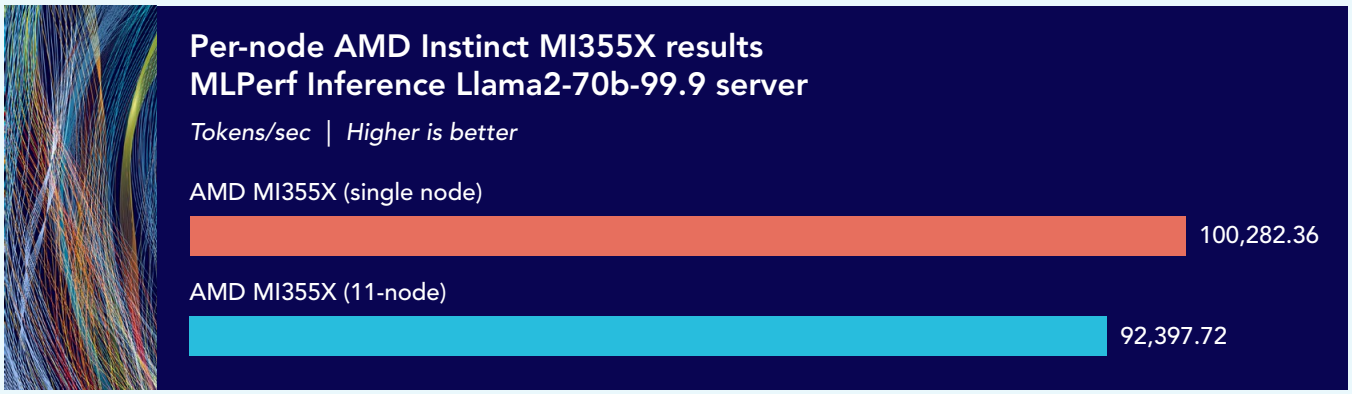


Figure 3: MLPerf Inference: Datacenter Closed Llama2-70b-99.9 Server results on AMD MI355X (6.0-0001 and 6.0-0003).
Source: www.mlcommons.org.

More than just Llama2

Though this report focuses on MLPerf Datacenter: Inference Llama2-70b-99.9 server results, there are also public MLPerf inference results on AMD Instinct GPUs for other types of AI workloads. If your AI workloads aren't represented by Llama 2 testing, visit the MLCommons website for other AMD Instinct results. For example, MLPerf provides Closed results for the gpt-oss-120b benchmark, which gives you an idea of the tokens/sec the GPUs can achieve on a model for multi-step problem solving such as an agentic AI pipeline.[9] There are also mixtral-8x7b and stable-diffusion-xl results, representing workloads such as multilingual customer support pipelines and text-to-image.[10] Additionally, the Open category shares AMD Instinct results for Wan-2.2-t2v text-to-video.[11]

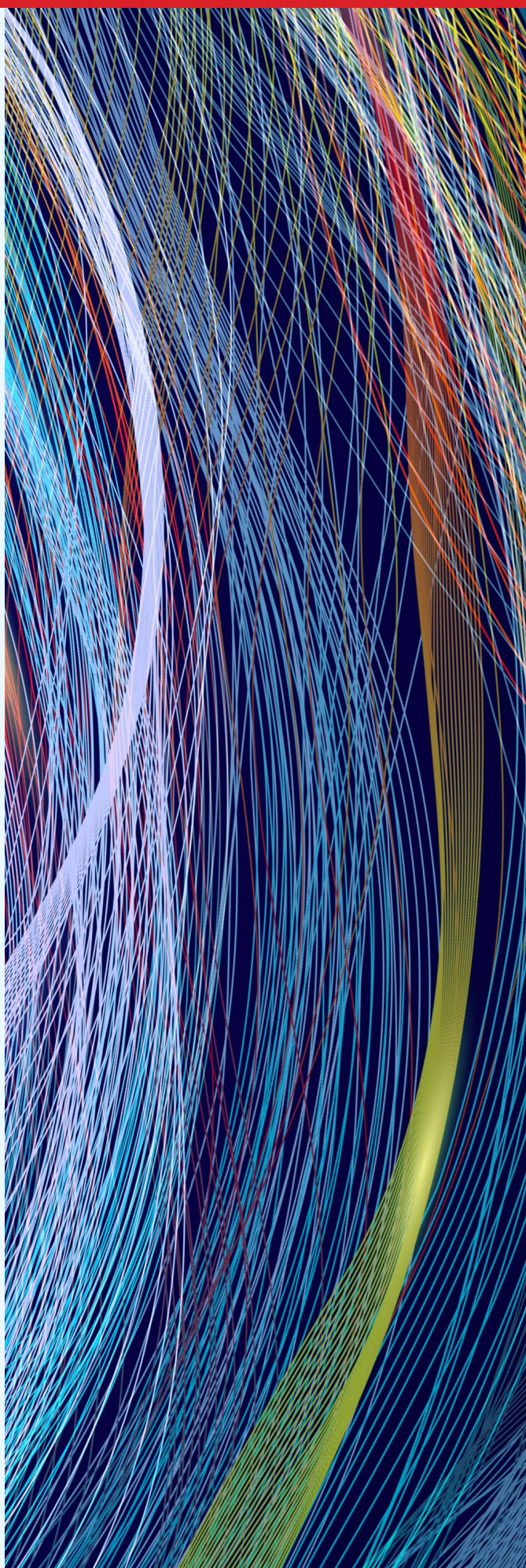
Regardless of which type of inference workload you're planning to deploy, you can browse the MLCommons website for submitted AMD Instinct results to gauge how your workload can benefit from using AMD Instinct GPUs.

AMD ecosystem

AMD partners have shown how AMD Instinct GPUs perform in their environments, thus proving that the GPUs can perform across a broad ecosystem. In the MLPerf v6.0 Closed inference results, nine different organizations, including Cisco, Dell, GigaComputing, HPE, MangoBoost, Mitac, Oracle, Supermicro, and Red Hat, submitted results on one or more AMD Instinct GPU models.[12] Customers looking to purchase AMD Instinct GPUs for their AI workloads can be confident that they can run them successfully on a broad spectrum of OEM systems. Part of AMD's effort in ensuring generational performance increases is their investment in their ROCm software, as shown by their near monthly release update cadence over the last year.[13] And results show that the AMD Instinct MI355X model can provide consistent results regardless of OEM hardware. These MLPerf results show repeatability, consistent performance, and broad OEM support for AMD Instinct GPUs.

Conclusion

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. 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. Customers can evaluate AMD Instinct GPUs for inference deployments knowing the platform is improving across generations, scaling efficiently, and offers similar performance across partner platforms. Learn more about AMD Instinct GPUs.



1. Textero.io, "AI Adoption Statistics: What Percentage of Companies Use AI Worldwide in 2026," accessed June 10, 2026, https://textero.io/research/what-percentage-of-companies-use-ai.
2. Mordor Intelligence, "GRAPHICS PROCESSING UNIT (GPU) MARKET SIZE & SHARE ANALYSIS - GROWTH TRENDS AND FORECAST (2026 - 2031)," accessed June 10, 2026, https://www.mordorintelligence.com/industry-reports/graphics-processing-unit-market.
3. MLPerf® v6.0 Inference Closed Llama2-70b-99.9 server and closed Llama2-70b-99 server. Retrieved from https://mlcommons.org/benchmarks/inference-datacenter/ 21 May 2026, entries 6.0-0027 and 6.0-0070. Results verified by MLCommons Association. The MLPerf name and logo are registered and unregistered trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.
4. MLPerf® v6.0 Inference Closed Llama2-70b-99.9 server. Retrieved from https://mlcommons.org/benchmarks/inference-datacenter/ 21 May 2026, entry 6.0-0003. Result verified by MLCommons Association. The MLPerf name and logo are registered and unregistered trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.
5. MLPerf® v5.1 Inference Closed Llama2-70b-99.9 server. Retrieved from https://mlcommons.org/benchmarks/inference-datacenter/ 21 May 2026, entry 5.1-0002. Result verified by MLCommons Association. The MLPerf name and logo are registered and unregistered trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.
6. MLPerf® v6.0 Inference Closed Llama2-70b-99.9 server. Retrieved from https://mlcommons.org/benchmarks/inference-datacenter/ 21 May 2026, entry 6.0-0027. Result verified by MLCommons Association. The MLPerf name and logo are registered and unregistered trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.
7. MLPerf® v6.0 Inference Closed Llama2-70b-99.9 server. Retrieved from https://mlcommons.org/benchmarks/inference-datacenter/ 21 May 2026, entry 6.0-0070. Result verified by MLCommons Association. The MLPerf name and logo are registered and unregistered trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.
8. AMD, "MLPerf 6.0: AMD Instinct™ MI355X GPUs Surpass 1M Tokens/Sec, Power New Workloads and Demonstrate Distributed Inference," accessed June 10, 2026, https://www.amd.com/en/blogs/2026/amd-delivers-breakthrough-mlperf-inference-6-0-results.html.
9. MLCommons, "MLPerf Inference: Datacenter," accessed June 10, 2026, https://mlcommons.org/benchmarks/inference-datacenter/.
10. MLCommons, "MLPerf Inference: Datacenter."
11. MLCommons, "mlcommons/inference_results_v6.0," accessed June 10, 2026, https://github.com/mlcommons/inference_results_v6.0/tree/main/open/AMD/results/8xMI355X_2xEPYC_9597F.
12. MLCommons, "MLPerf Inference: Datacenter."
13. AMD, "ROCm release history," accessed June 10, 2026, https://rocm.docs.amd.com/en/latest/release/versions.html.

This project was commissioned by AMD.

Primary contributors

Tech: Sarah C.

Writing: Jennifer V.

Design: Laura K.

PM: Scott Luchene

How we created this report

A PT team, which includes the contributors we've listed and others, created this report and performed the technical work behind it. We used AI to complete some research and for some initial drafting of text.

Principled Technologies®

Facts matter.®