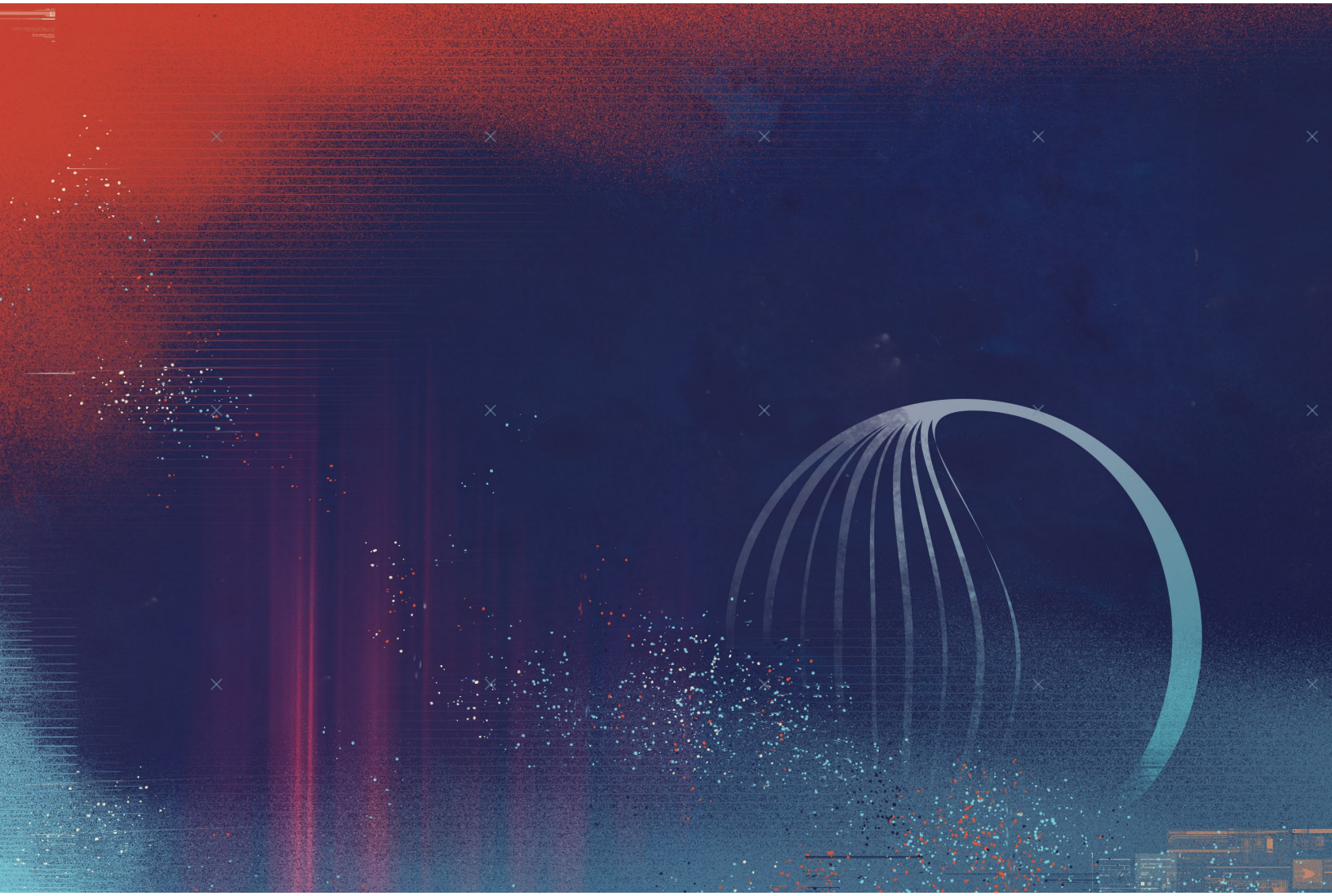
signal65



AMD Instinct MI355X Examining Next-Generation Enterprise AI Performance

AUTHOR

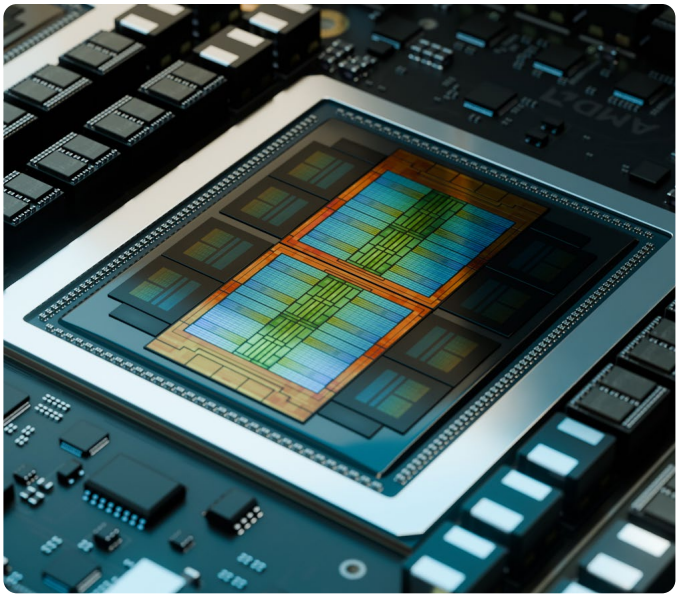
Russ Fellows
VP, Labs | Signal65

IN PARTNERSHIP WITH

AMD
together we advance_

JUNE 2025

Executive Summary



Few technologies have experienced a faster adoption rate than AI, which has emerged as one of the most important enterprise workloads over the past decade. Companies that excel at implementing these new AI workloads cost effectively will gain a significant competitive advantage. While the leading GPU supplier has continued to deliver enhancements, the AI accelerator landscape has changed markedly over the past few generations, with AMD emerging as the leading challenger to the existing hegemony.

Enterprise users have come to understand the trade-offs between the instant availability of cloud deployments vs. the challenges presented by regulations and data governance. Moreover, many enterprises often utilize a mix of both cloud hosted environments to rapidly trial and prototype solutions, while considering on premises AI deployments to maximize their spending while ensuring data privacy.

Having shipped dedicated AI accelerators for nearly a decade, AMD continues to be a key player in the AI market. Through concentrated efforts on their Instinct hardware portfolio and AI software library, AMD is now strategically positioned to advance its market share and challenge the top competitor.

Signal65 was asked to analyze and evaluate the performance of the new AMD Instinct MI355X compared to a leading competitor, the NVIDIA B200 GPU for common enterprise AI workloads. Working with AMD and utilizing AMD's labs, Signal65 tested several LLM workloads and compared them to published results from NVIDIA for the B200.

Looking specifically at several benchmarks, the AMD MI355X:

- Running DeepSeek-R1, the MI355X produced from **1x to 1.5x higher throughput**
  - For low latency, with concurrency of 16, AMD produced **up to 1.25x higher throughput**
- Outpaces B200 on Llama3.1-405B inferencing, delivering from near parity (1.02x) to **more than 2x of NVIDIA's** published results
  - The Geometric Mean across 11 configurations showed **AMD was 1.35x of B200**
- When running MLPerf LoRA fine-tuning of the Llama2 70B model, AMD showed, a **10% advantage**
- Achieves a **2.93x generation-over-generation uplift** versus the AMD MI300X
  - A single MI355X - 8 GPU system completed LoRA MLPerf runs **9.6% faster** than a four node - 32 GPU MI300X solution

The remainder of this paper covers the AMD Instinct MI355X performance in more detail, along with an overview of why the MI355X outperforms the B200. This includes the chip's large, high-bandwidth memory advantages and AMD's software library ROCm compared to NVIDIA CUDA library.

Enterprise AI Requirements

Many AI benchmarks showcase massive clusters, often containing more than 1,000 GPUs delivering LLM training results in a matter of minutes. These tests are focused on hyper-scalers creating foundational models. While these results may be impressive, they are often not applicable to the majority of enterprises who understand that data gathering, data preparation, toolsets choices and model optimization can all take far longer to perfect than the few minutes required to train an LLM on a large cluster.

Enterprise AI deployments are rightly concerned with utilizing AI for productivity gains, rather than research. Moreover, the focus for firms is on production inferencing, including RAG, agentic workloads and fine-tuning. With the current state of the art GPUs such as the AMD MI355X, a single cluster of 64 GPUs in one or two racks can support hundreds of simultaneous inferencing sessions or rapidly perform fine-tuning on the largest publicly available LLMs, such as Llama-405B and DeepSeek-R1 models, in a matter of minutes.

Additionally, we analyzed recent submissions on the MLPerf Training, version 5.0 results by AMD partner MangoBoost and the ability to effectively scale out AMD Instinct systems to multi-node clusters. Based upon these submitted results, and detailed later in this paper, users can effectively scale AMD Instinct systems to multi-node clusters.

Within this context, the following aspects are important considerations:

- Hardware Performance – ability to run the latest LLMs at industry leading speed
- Software Libraries – tool set support for the latest AI models and optimizations
- TCO Considerations – desire leading performance, with long life at lower cost

Hardware Performance

AMD GPU hardware has rapidly evolved, enabling AMD to match or exceed the performance of other leading AI accelerator providers. For inferencing, one area of advancement has been to quantize, or reduce the size of the data types containing the LLM model weights. AMD's performance in terms of floating-point operations, memory bandwidth and capacity can match or exceed NVIDIA B200 GPUs across a range of measurements.

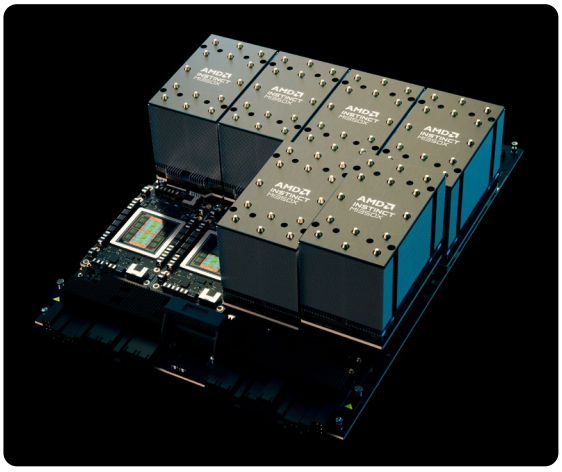
Software Libraries

It is critical that any software advances be leveraged by the hardware for platforms to remain relevant. To date, a significant amount of emphasis has been placed on GPU software libraries, such as NVIDIA CUDA and AMD ROCm as important considerations. As such, AMD is investing heavily to make ROCm a leading AI library. The foundational support for AMD GPUs is in place for many AI frameworks, with open-source tools such as vLLM, Hugging Face ONNX, SGLang, and Modular Max all providing support for AMD ROCm alongside CUDA.

TCO Considerations

Finally, investment protection is a consideration for enterprise users. The question of whether a resource will still be valuable in 3 years or even longer are important financial questions, which directly impact TCO calculations. The longer a resource remains relevant, the lower the effective cost is per year. With LLM models growing increasingly large, one of the primary factors dictating a GPU's usefulness is the size of the memory. As we will discuss, one competitive edge the AMD Instinct line of GPUs has is a memory capacity advantage compared to an equivalent NVIDIA GPU. The increased memory enables AMD GPUs to maintain their usefulness and relevancy for a longer duration, which ultimately provides a TCO advantage.

AMD Instinct GPU Architecture



The AMD Instinct MI355X represents a significant step forward in GPU architecture and memory technology. It pairs a next-generation compute engine with industry leading 288 GB of HBM3e memory per card, a combination that delivers high performance across a range of critical enterprise AI tasks, including LLM training, fine-tuning, and retrieval-augmented generation (RAG), along with new agentic workloads. These trends, including reasoning and agentic models are driving an increase in inferencing context windows.

Memory Capacity and Low-Precision Optimizations

Perhaps the most significant architectural advantage of the MI355X is its massive 288 GB of HBM3e memory. This is a full 96 GB more than the Blackwell B200's 192 GB, and the additional capacity has profound implications for real-world AI workloads. Additionally, the MI355X peak bandwidth is 8 TB/s compared to 5 TB/s for the NVIDIA B200.

With 288 GB of HBM3e, a single MI355X can accommodate:

- A full Llama3 70B checkpoint in FP8, along with the optimizer states for Low-Rank Adaption (LoRA) fine tuning
- 128k-token KV caches for a 70B decoder without needing to spill over to host RAM
- The entire Llama-405B-parameter FP4 model on every GPU, eliminating the need for tensor-parallel communications and improving performance

The MI355X implements matrix cores, which similar to the B200 tensor cores support low-precision data formats like FP8 and FP4. Using smaller, quantized data types helps accelerate AI inferencing workloads by reducing the memory and computational requirements of models. The AMD AI Tensor Engine for ROCm (AITER) inference library is bundled with recent ROCm library releases to optimize operations for all supported data sizes, including FP4. This also helps increase processing speed and efficiency without intervention from users implementing inferencing or training using the ROCm library.

Benchmark Results: AMD MI355X vs. NVIDIA B200

The environment for the observed testing was similar for every test and workload run. A single Supermicro server with 8 – AMD MI355X GPUs was used for all testing.

Signal65 Comments: *With the exception of the DeepSeek-R1 model runs, all NVIDIA B200 data points reference published data by NVIDIA or MLCommons. This was done for several reasons, with the primary rationale being that it enables each vendor to promote their own, validated performance under optimal conditions.*

The data reported for NVIDIA B200 comes from the GitHub repository performance metric page, with results pulled on June 4, 2025. More configuration details are provided in the Appendix, including firmware versions, OS versions, drivers, software stacks and other important components of the AI stacks utilized. Over time, it is expected that both AMD and NVIDIA will work to enhance their libraries, and thus their performance results.

MI355X Performance Overview

Overall, the MI355X performed better than the NVIDIA B200, across multiple workloads when compared to NVIDIA's own published benchmarks. To date, there have not been a significant number of published results outside of NVIDIA's GitHub repository and MLPerf Training and Inferencing results. Of the values published, AMD has favorable comparisons to many of these results, only falling behind in a few corner cases.

Signal65 Comments: *Across a wide range of inferencing, pre-training and fine-tuning workloads, we observed the AMD MI355X deliver results on real-world scenarios that matched or surpassed NVIDIA's latest B200 GPU. Importantly, the results reported are for common AI workload scenarios, not theoretical Tera-flop or other synthetic scenarios. This includes running MLPerf training workloads along with inferencing the very largest available LLMs, including Llama3.1-405B and DeepSeek-R1. These results show that AMD can deliver industry leading AI acceleration, and price / performance results.*

Scalability is also an important consideration, particularly for fine-tuning workloads. Due to the rapid increase in computational efficiency, along with the increases in GPU memory by AMD, the need to scale inferencing workloads beyond 8 GPUs is decreasing. This is evidenced by the capability of the MI355X to inference Llama3.1-405B or models up to 570B parameters on a single GPU when using FP4 quantization. Even when running 16-bit, full weights the MI355X can perform large context inferencing using as few as 4 GPUs, thereby enabling multiple inferencing instances to run on a single node.

In pre-training tasks, the MI355X matches or even exceeds the performance of NVIDIA's formidable Blackwell B200, demonstrating up to 3% greater speed for Llama3-70B at FP8. For large-model inferencing using Llama3.1-405B, the MI355X demonstrates a notable advantage over the B200. Our test results indicate it can deliver an average of 1.3x more offline tokens per second, with specific workloads showing an impressive gain of up to 1.6x. Furthermore, in online serving scenarios with 16 concurrent users, the MI355X achieves up to 1.25x higher throughput.

The use of Low Rank Adaptation (LoRA) has become a popular method to enhance the speed of fine-tuning LLMs. The MI355X achieves a 2.93x generational performance increase compared to its predecessor, the MI300X. As a result, a single 8 GPU MI355X system can complete a LoRA MLPerf run 9.6% faster than four MI300X nodes. This result is described in the section titled "MLPerf Llama2 70B LoRA Fine-Tuning", and demonstrates the ability to scale AMD GPU performance with using MangoBoost software and networking accelerators.

The architectural advantages and software advancements of the MI355X translate into tangible performance gains in a variety of real-world AI workloads. The following section will summarize the results of six independent tests, highlighting the MI355X performance in pre-training, fine-tuning, and inference tasks. Unless otherwise noted, the results are presented in tokens per second per GPU and represent the median of three to five runs.

Pre-Training Performance

In the demanding realm of LLM pre-training, the MI355X demonstrates a performance advantage over the B200, as shown in the following two test cases.

Llama3-8B Pre-Training (FP8)

In an FP8 pre-training task with the Llama3 8B model, an 8-GPU MI355X platform running MegatronLM achieved a throughput of 31,190 tokens/second/GPU, making it 3% faster than an 8-GPU B200 platform running NeMo 25.04, which reached 30,411 tokens/second/GPU.

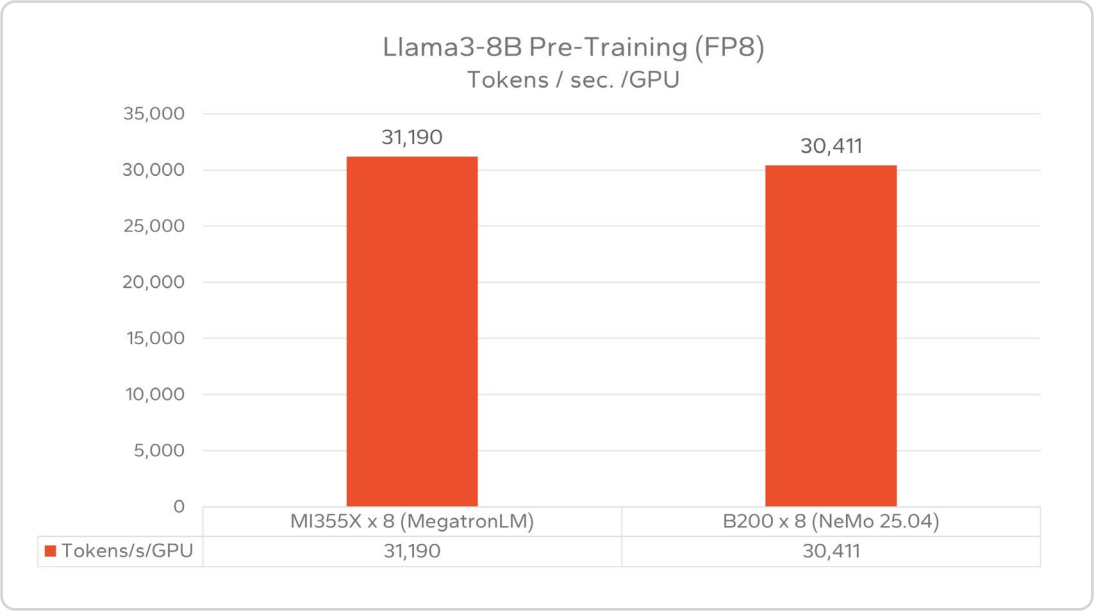


Figure 1: *(Source: Signal65 / AMD)*

Note: Published NVIDIA B200 Reference: nvidia.com | nemo-framework | Performance

Llama3-70B Pre-Training (BF16)

When training the larger Llama3 70B model with BF16 precision, the MI355X lead widens to **12% advantage**. An 8-GPU MI355X system reached a throughput of 2,154 tokens/second/GPU, compared to 1,918 tokens/second/GPU for an 8-GPU B200 system.

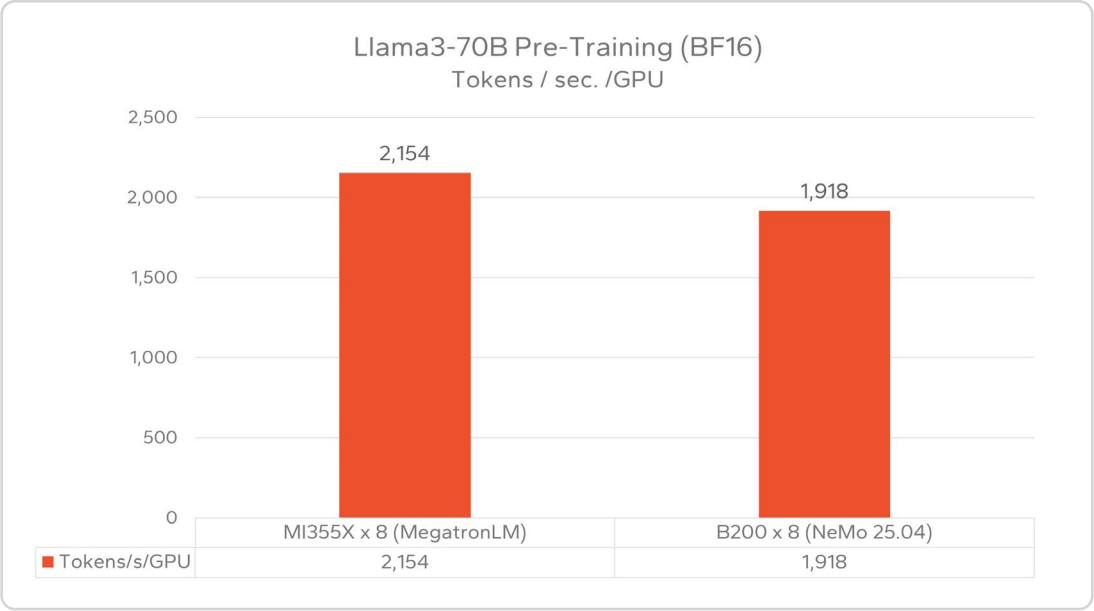


Figure 2: *(Source: Signal65 / AMD)*

Note: No Published NVIDIA Reference. See testing results in Appendix.

Signal65 Comments: *The Llama3-70B BF16 results for the B200 have not been published by NVIDIA. The data presented here was gathered during testing of the MI355X and B200, in part to help show comparative performance for commonly used data types, such as BF16.*

Llama3-70B Pre-Training (FP8)

In evaluating the Llama3-70B pre-training workload using FP8 precisions, an 8 GPU MI355X system achieved similar performance as an 8 GPU NVIDIA B200. Specifically, the AMD system achieved a **3% higher token rate**, as seen below in the chart.

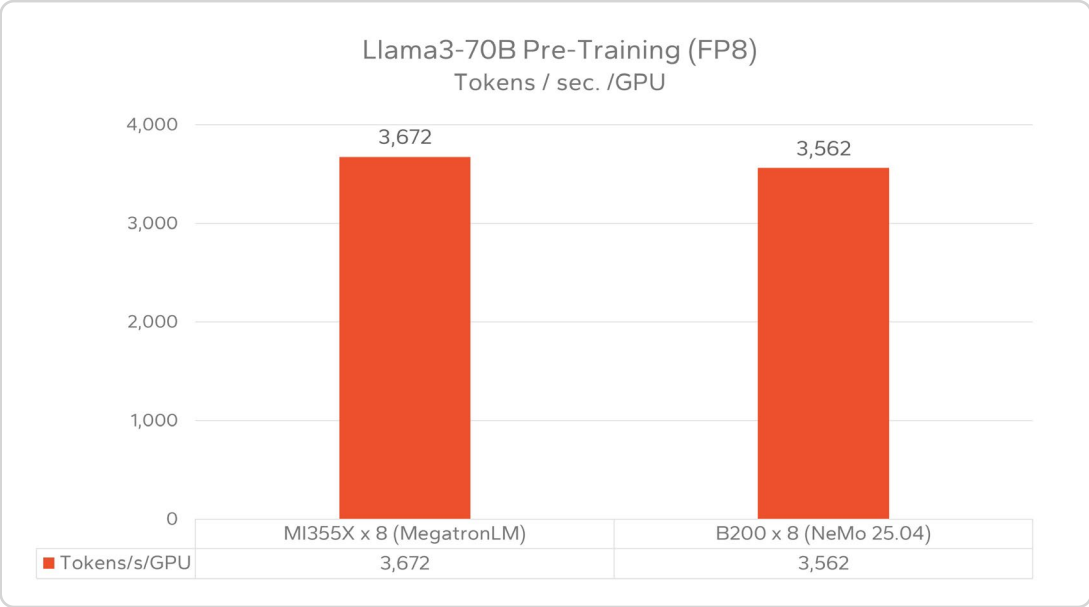


Figure 3

Note: Published NVIDIA B200 Reference: nvidia.com | nemo-framework | Performance

Fine-Tuning and Inferencing – Key Enterprise Workloads

The MI355X architectural advantages, particularly its large HBM capacity and high memory bandwidth, become even more apparent in fine-tuning and inference workloads, where latency and throughput are critical. The following three test cases highlight these advantages.

MLPerf Llama2-70B LoRA Fine-Tuning

MLPerf is an industry standard set of benchmarks covering a range of AI related workloads. There are a variety of training workloads, one of which is the MLPerf Llama2-70B fine-tuning using LoRA technique. This particular workload has become one of the most popular in the MLPerf Training category, with a high number of submissions from a variety of vendors. There are several factors for this, but one of the most important is that it represents a realistic workload scenario, where an enterprise may want to fine-tune an LLM using a parameter efficient technique such as LoRA to enhance – or fine tune – the model using specific corporate data.

For Llama2-70B fine tuning, NVIDIA utilized their Nemo framework version 25.04, which is designed as a proof-of-concept PyTorch library for scaling LLM training. NVIDIA partners show a range of results for a single node, 8 GPU system using B200 accelerators, with the best result being 11.209 minutes (MLPerf Training ID: 5.0-0089).

Note: Link to MLPerf Training results v5.0 is: MLCommons Training: Version 5.0 Results

Signal65 observed this same workload, (MLPerf LoRA fine-tuning of the Llama2 70B model) on a single, 8-GPU MI355X system, with it completing the task in under 10 minutes. Across multiple runs, using the MLPerf scoring methodology, the AMD MI355X completed this workload in **9.96 minutes, a 10% advantage**. There are three interesting comparisons available for this workload:

1. AMD has made generational improvements, comparing the results for a 4 node (32 GPU) AMD MI300X with MangoBoost, to a single node (8 GPU) AMD MI355X system, results shown in Figure 4.
2. In a matching 8 GPU setup, MI355X shows a 2.93x improvement compared to the MI300X. (29.25 vs 9.96 mins)
3. The AMD MI355X produced better performance (lower time) than the best published NVIDIA B200 result, showing **10% better performance**, as shown below in Figure 5.

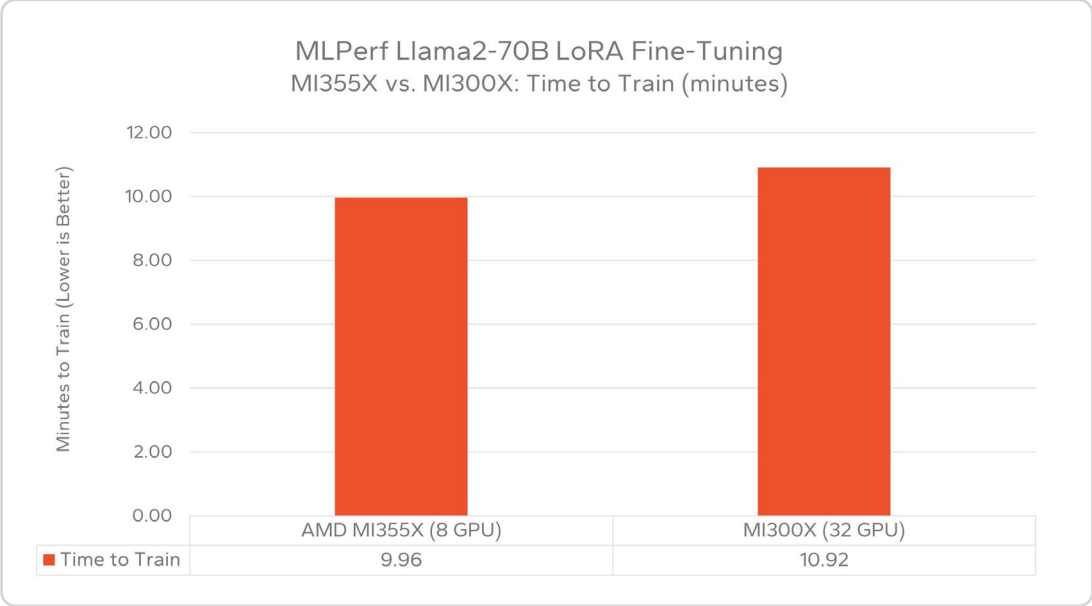


Figure 4: *(Source: Signal65 / MLPerf)*

Note: MLPerf submission IDs for Figure 4 are MI355X: N/A and MI300X 4-node: 5.0-0058.

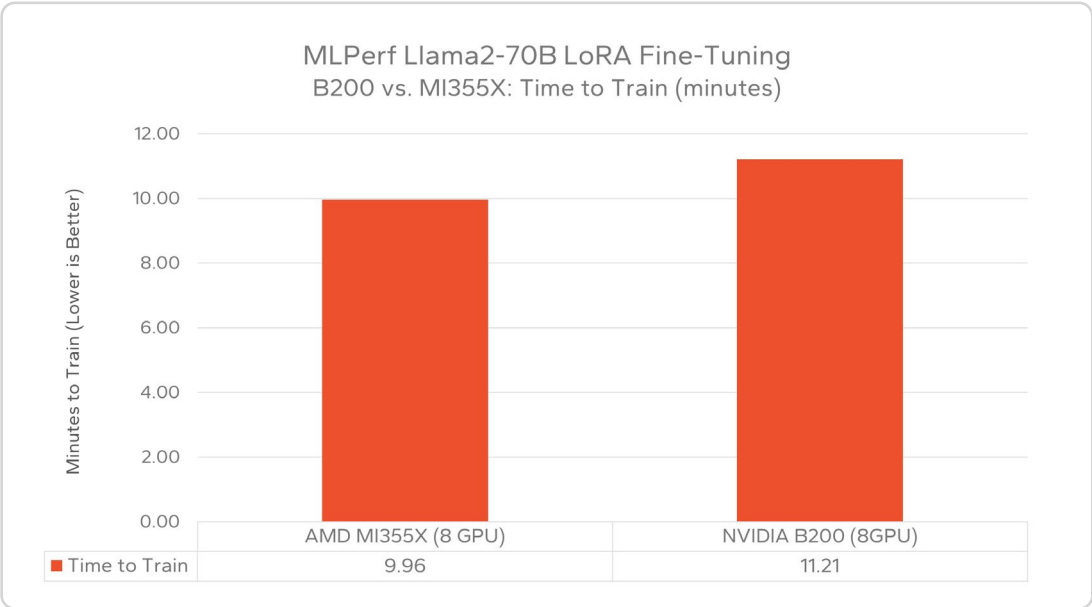


Figure 5: *(Source: Signal65 / MLPerf)*

Note: MLPerf submission IDs for Figure 5 are MI355X: N/A and B200 is: 5.0-0089.

DeepSeekR1 Online Serving (FP4)

When running DeepSeek-R1 at FP4, we compared the MI355X to published NVIDIA B200 results. This showed advantages across two areas:

1. As the number of concurrent requests increased, the AMD MI355X performance increasingly outpaced NVIDIA B200 performance. DeepSeek-R1 at FP4 precision on a single node (TP = 8)
2. The MI355X system produced up to 1.25x higher throughput at a concurrency of 16, in a low latency environment

Signal65 Comments: *Note that there is not a specific "Online Serving Scenario" for inferencing workloads. However, it has become standard practice to create thresholds, such as 30 ms. ITL as a way to measure and compare inferencing performance.*

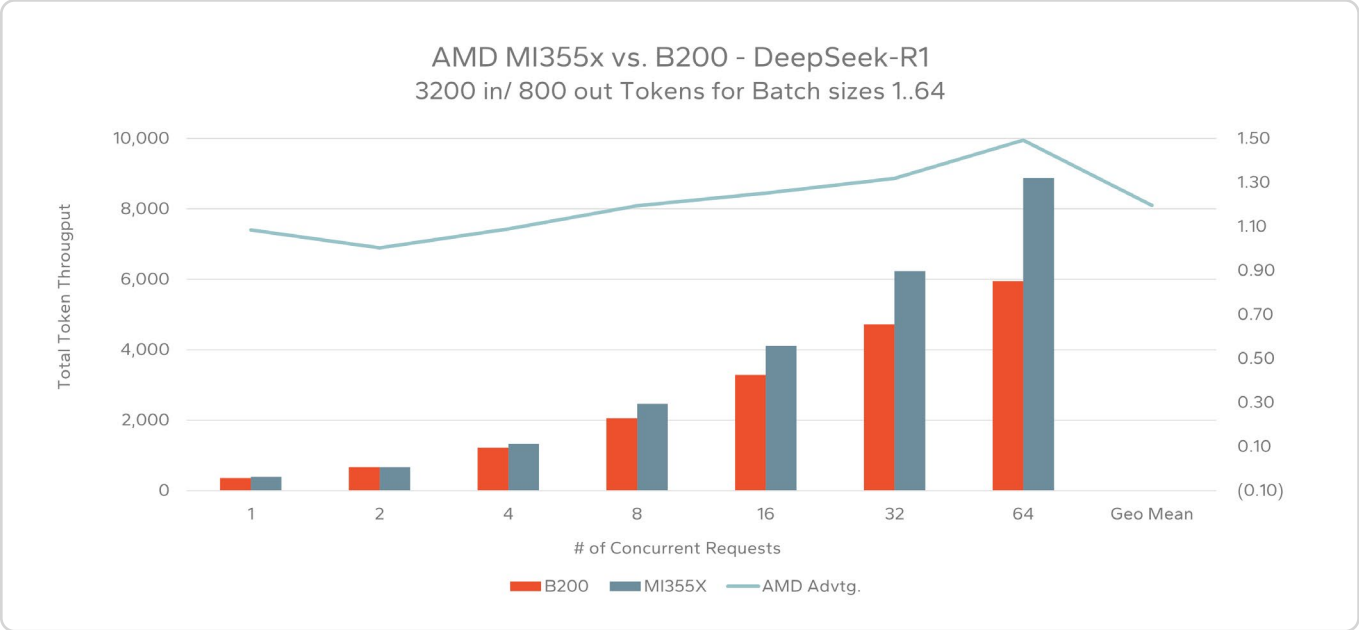


Figure 6

Note: Link to DeepSeek-R1 NVIDIA Results: Huggingface | nvidia/DeepSeek-R1-FP4

Signal65 Comments: *The choice of running DeepSeek-R1 with eight instances on a system (TP = 1), vs. running with the model spread across all GPUs (TP = 8) is both a question of capability and the solution that provides the best results. Given that NVIDIA chose to report their results with TP = 8 we believe that was chosen as their best result. Similarly, AMD chose to run with TP = 1, as this produced their highest result, producing a fair comparison.*

Llama3.1-405B Offline Inference (FP4)

A system with 8 - MI355X GPUs running vLLM delivers a speedup ranging from near parity (1.02%) to **more than 2X of NVIDIA's** published throughput results. The Geometric Mean across all 11 configurations showed a 35% advantage (**AMD was 1.35X of B200**). This test reproduced each configuration that Nvidia published for an 8 GPU B200 system running TensorRT-LLM, performing text generation using Llama-3.1-405B with FP4 quantization.

Signal65 observed text generation in an offline throughput scenario with the Llama-3.1-405B-Instruct model using a variety of input and output lengths. Testing was performed with a synthetic dataset, using a quantized FP4 model, exactly as NVIDIA did for their testing. With FP4, the entire model can fit and run effectively on a single MI355X GPU. For offline scenarios, running a separate copy of the model on each GPU can drive higher throughput than using tensor parallelism to spread a single copy of the model across multiple GPUs.

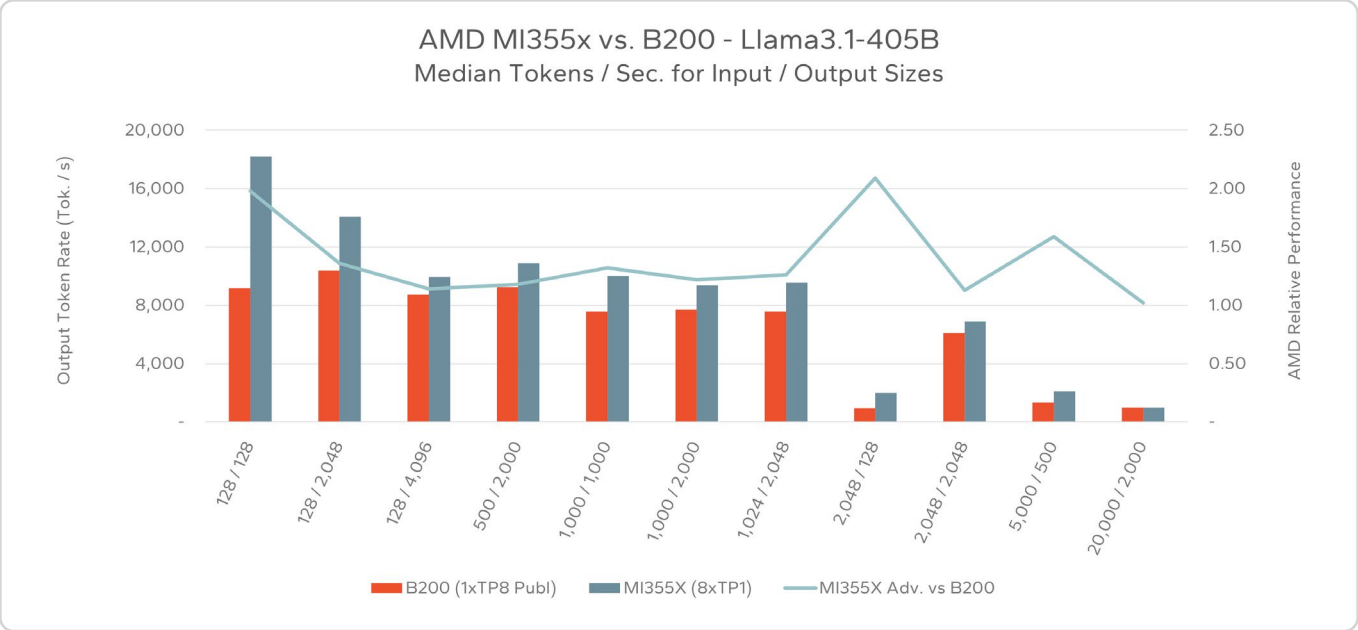


Figure 7

Note: General Link for Llama3.1-405B NVIDIA Results: GitHub NVIDIA/TensorRT-LLM Perf-Overview

For offline inferencing with the very large Llama3.1 405B model, the MI355X's ability to fit the entire model on a single GPU delivers significant performance advantages. This is due in part to the elimination of tensor-parallel communications, which are required on the B200 due to its smaller memory capacity.

Why the AMD MI355X Leads in Tested Workloads

The benchmark results are not isolated data points; they are the direct result of deliberate architectural decisions that give the AMD Instinct MI355X a fundamental advantage in these memory-centric AI workloads. The demonstrated performance advantages lie in the greater memory capacity and higher bandwidth, combined with improving intelligent software that leverages these hardware capabilities.

Memory Capacity and Bandwidth for Demanding Workloads

The memory specification of 288 GB of HBM3e per MI355X GPU represents a critical threshold for modern large language models. For many demanding scenarios, this capacity enables single-card execution of the largest models, as shown during the testing. For other workloads such as running a 70-billion-parameter decoder model at FP8 precision, along with its key/value (KV) cache for a very long 128,000-token sequence, fits entirely within the memory of a single MI355X GPU. The same workload would either not be possible to run within the confines of the NVIDIA B200's - 192 GB memory or leave little room for large input / output context windows.

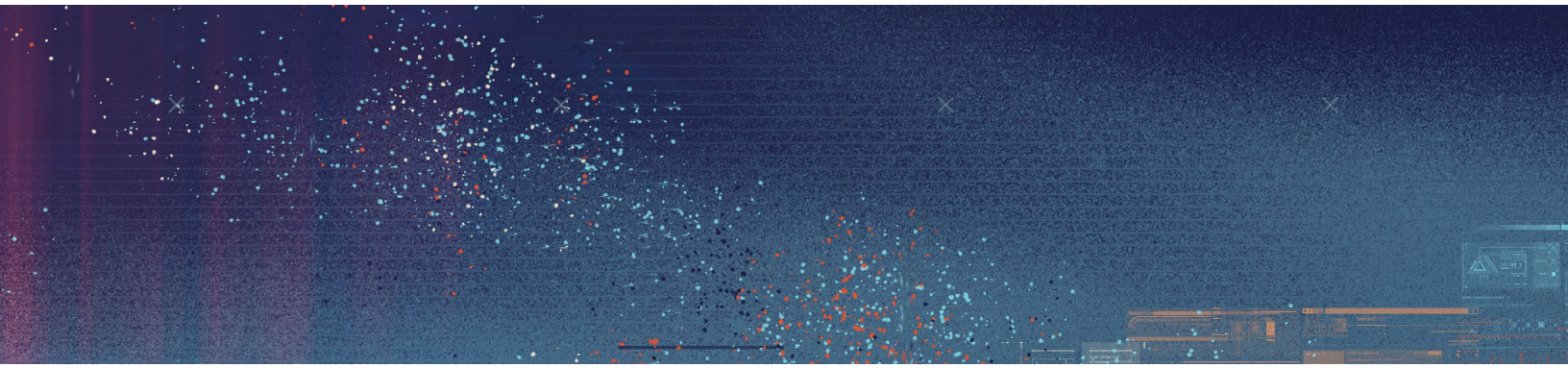
The MI355X memory advantage is not limited to capacity. Along with increased capacity comes significantly higher bandwidth, as the MI355X delivers an aggregate bandwidth of approximately 8 TB/s, which is more than 50% greater than the 5.0 TB/s offered by the B200.

This bandwidth advantage is most evident in workloads where memory access is the primary limiting factor. A prime example is LLM training using higher-precision formats, where the size of attention activations consumes a large portion of the memory bus. For tasks involving long sequences and large batches, higher memory bandwidth is not just beneficial, it is essential for achieving state-of-the-art performance.

The Power of Single-GPU Execution vs. Tensor Parallelism

To run a 405-billion-parameter model, the NVIDIA B200 solution splits the model across eight GPUs (TP=8), requiring communication between them. The AMD MI355X leveraged the 288 GB of HBM memory, allowing it to fit the entire model on a single GPU (TP=1). This removes the need for inter-GPU communication for tensor parallelism, resulting in a 1.35x to 1.59x lead in offline inference performance on that model.

- **Accelerated Fine-Tuning:** For LoRA (Low-Rank Adaptation) fine-tuning, the MI355X is the leading performer, outpacing the best publicly available NVIDIA B200 results. This allows enterprises to iterate on custom models more rapidly.
- **Denser Long-Context Chat:** With its 288 GB of HBM, a single MI355X GPU can maintain the full KV cache for two simultaneous 128,000-token chat sessions with a 70B model. This doubles the user density per GPU compared to accelerators with less memory, allowing for more efficient serving of long-context RAG and summarization applications without resorting to performance-sapping memory paging.
- **Faster VectorDB Ingestion:** In RAG pipelines, the speed of embedding and indexing is crucial. The MI355X's larger HBM allows a GPU embedding instance to pin a larger index directly in GPU memory alongside the encoder model. AMD measures a **2.1x speedup in index-build time** compared to GPUs with smaller memory that are forced to stream the embeddings over a slower PCIe bus.
- **Superior High-Concurrency Serving:** As demonstrated in the DeepSeekR1 test, the MI355X can handle **64 concurrent sessions** at a strict 30ms latency, compared to just 32 sessions for the B200. The resulting higher token throughput enables enterprises to serve more users from a system and reduce infrastructure costs.



Economic Considerations

Ultimately, infrastructure decisions come down to economics. While the list price of a single GPU is a factor, a true Total Cost of Ownership (TCO) analysis must consider the entire picture: node count, power consumption, cooling, networking, and software licensing. This paper does not examine TCO but instead focuses on the MI355X high performance along with what has typically been shown to be a lower acquisition or cloud consumption pricing.

Azure Cloud Pricing

Signal65 Comments: *All pricing data below was verified as accurate on June 10th, 2025 by Signal65. However, by definition, cloud pricing is highly dynamic and subject to change.*

The on-demand pricing for ND96isr AMD MI300X v5 instances in the West US region is about 50% less than the H100 v5 instances at $62.853/hour vs $127.816/hour. Comparing the ND96isr MI300X v5 to the H200 v5 instances, the MI300X still maintains a significant price advantage at $62.853/hour vs $110.24/hour for the H200.

In the East US 2 region, the AMD MI300X again costs about 50% less than the H100 at $48 vs $98.32. Similarly, the on-demand price of the H200 is $84.80, still much more than the AMD MI300x. Based upon this data, the AMD MI300X instances provide a lower cost than the H100 and H200 instances in both regions.

Signal65 Comments: *As cloud instances with the MI355X become available, we believe the past AMD pricing advantage seen for MI300X and MI325X will continue, providing customers with the ability to utilize the MI355X to achieve performance that is roughly equivalent to the NVIDIA B200, at a more cost effective price point.*

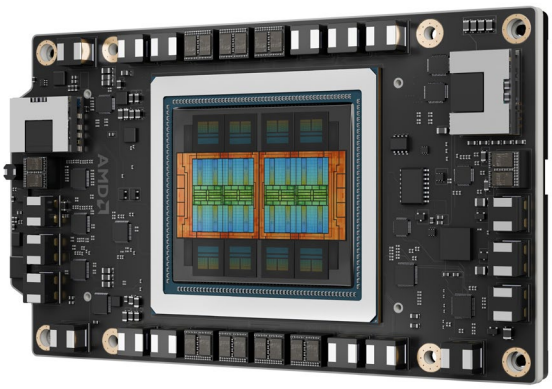
From a licensing perspective, ROCm uses the Apache 2.0 license with no per-GPU fee. NVIDIA's AI Enterprise software is available via a subscription model listed at $4,000 per GPU annually. (*Pricing based on NVIDIA public materials, 2025)*

Final Thoughts and Outlook

For decades, NVIDIA CUDA-based accelerators have been the dominant force in the world of high-performance computing and artificial intelligence. However, the rapid evolution of large language models (LLMs) and the increasing demand for GPU infrastructure have created an opening for competitive alternatives. Markets where multiple vendors are actively competing with leading products provide enterprises with choices, helping to improve the overall ecosystem.

Signal65 Comments: *During testing across multiple workloads, we found the AMD Instinct MI355X can provide leading performance and price / performance for enterprise inferencing, agentic, RAG and fine-tuning workloads, both on premises and in cloud deployments.*

Across each benchmark analyzed, the MI355X demonstrates compelling performance advantages:



- **Equal Training Performance:** The MI355X delivers similar pre-training performance as B200 across several different measurements.
- **Better Inferencing Performance:** The MI355X delivers **up to 2.0x higher throughput** on large-scale models like Llama3.1 405B and DeepSeekR1.
- **Memory Headroom:** The 288 GB HBM memory enables longer contexts, more users per GPU, and larger batch sizes without performance-killing offloads.
- **Compelling Economics:** Based upon historical price advantages, combined with zero software license fees, the MI355X could provide better dollar-per-token metrics.

While NVIDIA B200 does have advantages for large deployments when training foundational models, AMD's momentum in key enterprise workloads is undeniable. When performance is gated by memory throughput or context length, as has become common, the AMD Instinct MI355X shows it can provide greater performance for agentic AI, fine-tuning, RAG, and inferencing workloads.

Looking ahead, industry trends suggest this advantage may grow. AMD's roadmap indicates an annual refresh, with successor accelerators expected to provide higher computational speed, along with more memory and higher bandwidth. As AI models continue to demand longer contexts and more complex, multi-model agentic pipelines, the MI355X memory headroom advantage is poised to become even more critical. With AMD's ROCm library rapidly being integrated into the broader AI toolset, the barriers to switching to AMD Instinct line are insignificant compared to the economic benefits for enterprises running common AI workloads in cloud-hosted or on-premises deployments.

Important Information About this Report

CONTRIBUTORS

Russ Fellows
VP, Labs | Signal65

PUBLISHER

Ryan Shrout
President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

IN PARTNERSHIP WITH

AMD
together we advance_

ABOUT SIGNAL65

Signal65 is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

signal65

signal65

CONTACT INFORMATION
Signal65 I signal65.com

Appendix

Configurations

The following hardware configurations are the same for each test scenario except as otherwise stated.

MI355X: Supermicro AS-4126GS-NMR-LCC

Summary Description	Supermicro AS-4126GS-NMR-LCC with 2x AMD EPYC 9575F Processors, 8x AMD Instinct MI355X (288GiB, 1400W) GPUs, Ubuntu 22.04, and a pre-release build of ROCm 7.0.0
System Model	Supermicro AS-4126GS-NMR-LCC
System BIOS	1.4a
CPU	2x AMD EPYC 9575F Processors (2 sockets, 64 cores per socket, 2 threads per core)
NUMA Config	1 NUMA node per socket
Memory	3072 GiB (24 DIMMS, 6400 mts, 128 GiB/DIMM)
Disk	Root drive: 1x 3840GB Micron 7450 (MTFDKCC3T8TFR) Data drive: 1x 3840GB Micron 7450 (MTFDKCC3T8TFR)
GPU	8x MI355X, 288GiB, 1400W
Host OS	Ubuntu 22.04.5 LTS with Linux kernel 6.8.0-59-generic
Host GPU Driver	ROCm 7.0.0 (pre-release build 16047) + amdgpu 6.14.5 (build 2168543)
Firmware	BKC 25.08

MI355X vs B200: Llama3-8B – FP8 – 8GPU

Same Config.

MI355X vs B200: Llama3-70B – FP8 – 8GPU

Same Config.

MI355X vs B200: Llama3-70B – BF16 – 8GPU

Host GPU Driver Bare metal host run ROCm 7.0.0 (pre-release build 16047) + amdgpu 6.14.5 (build 2168543); container runs ROCm 6.5.0-56

MI355X vs B200: Llama-3.1-405B FP4 – 8 GPU

Firmware BKC 25.08 w/ brp700b_sfo_vrupdate

Nvidia publishes B200 Offline Throughput results for Llama-3.1-405B FP4 with tensor parallelism of 8 (TP=8).

MI355X vs MI300X: MLPerf Training Llama-2-70B Lora

Hardware Config

Machine	smci355-ccs-aus-e06-13	MI300X Public Result
CPU Processor	AMD EPYC 9575F 64-Core Processor	AMD EPYC 9575F 64-Core Processor
Number of CPU Cores	2 X 64c	2 X 64c
Number of nodes	1	1
Number of GPU	8	8
Memory Capacity/GPU (GB)	256	192
Host Memory Capacity	2TB	2TB
Interconnect	Infinity Fabric	Infinity Fabric
BIOS	RP 700D, SFO, PMFW 86.42.150	Not disclosed
Firmware	RP 700D, SFO, PMFW 86.42.150	Not disclosed

Software Stack

	1 node MI355	**1 node MI300X**	**4 node MI300X**
OS version	Ubuntu 22.04.5	Ubuntu 22.04.5	Ubuntu 22.04.5
ROCm/CUDA version	ROCm 6.5	ROCm 6.5	ROCm 6.5
PyTorch	2.6.0+gitc37337f	2.6.0+gitd70a91b	2.6.0+gitd70a91b
Python	3.10.17	3.10.6	3.10.6
Transformer Engine	1.14.0.dev0+1e19799	commit: b9c3f6a8	commit: b9c3f6a8
Flash Attention	-	2.7.3	2.7.3
Apex	1.6.0	1.6.0	1.6.0
hipblasLt / cuBLAS	0.15.0-14a20666d06	b5870a2	b5870a2
NeMo	2.0.0b0	v2.0.0.rc0.beta	v2.0.0.rc0.beta
Megatron	0.10.0rc0	commit: 190213a	commit: 190213a

MI355X vs B200: DeepSeek-R1 Online Serving

Firmware	BKC 25.08 w/ brp700b_sfo_vrupdate

B200: Supermicro Super Server

Summary Description	Supermicro Super Server with 2x Intel Xeon 6960P processors, 8x Nvidia B200 (NVLink 192G, 1000W) GPUs, Ubuntu 22.04, and Nvidia driver 570.124.06
System Model	Supermicro Super Server
System BIOS	1.0
CPU	2x Intel Xeon 6960P (2 sockets, 72 cores per socket, 2 threads per core)
NUMA Config	3 NUMA nodes per socket
Memory	2304 GiB (24 DIMMS, 6400 mts, 96 GiB/DIMM)
Disk	Root drive: 1x 960GB Samsung PM9A3 (MZ1L2960HCJR-00A07) Data drive: 4x 3.84TB Micron 7450 (MTFDKCC3T8TFR)
GPU	8x Nvidia B200, 192GiB, 1000W
Host OS	Ubuntu 22.04.5 LTS with Linux kernel 5.15.0-136-generic
Host GPU Driver	570.124.06
Firmware	97.00.6E.00.07

Test Procedures

MI355X vs B200: Llama3-8B – FP8 – 8GPU

MI355X

Testing on MI355X was executed with a Docker container that was built with run scripts and dependencies from ROCm/Megatron-LM and packaged with Llama3 8B model elements.

Pull MI355X Docker image from internal registry:
docker pull docker.gpuperf:5000/mad_workload_port/rocm/megatron-lm-llama-training:20250602_next_gen
Run MI355X container:
docker run -it --device /dev/dri --device /dev/kfd --network host --ipc \
host --group-add video --cap-add SYS_PTRACE --security-opt \
seccomp=unconfined \
--shm-size 64G --privileged \
-v ~/dockerx:/dockerx \
-e NVTE_CK_IS_V3_ATOMIC_FP32=0 \
-e NVTE_CK_USES_FWD_V3=1 \
-e NVTE_CK_USES_BWD_V3=1 \
-e NVTE_CK_HOW_V3_BF16_CVT=2 \
docker.gpuperf:5000/mad_workload_port/rocm/megatron-lm-llama-training:20250602_next_gen
Execute MI355X run:
bash examples/llama/train_llama3.sh MBS=4 BS=512 TP=1 TE_FP8=1 SEQ_LENGTH=8192 OPTIMIZER=adam MODEL_SIZE=8 TOKENIZER_MODEL=meta-llama/Meta-Llama-3-8B GEMM_TUNING=0 RECOMPUTE=0 FSDP=0

Metric Calculation

Training throughput was reported in units of tokens per second per GPU (tokens/sec/GPU).

MI355X vs B200: Llama3-70B – FP8 – 8GPU

MI355X

Testing on MI355X was executed with a Docker container that was built with run scripts and dependencies from ROCm/torchtitan and packaged with Llama3 70B model elements.

Pull MI355X Docker image from internal registry:
docker pull docker.gpuperf:5000/rocm/pytorch-training-private:torchtitan
Run MI355X container:
*docker run -it --device /dev/dri --device /dev/kfd --network host *
*--ipc host --group-add video --cap-add SYS_PTRACE *
*--security-opt seccomp=unconfined --privileged *
*-v $HOME:$HOME *
*-v $HOME/.ssh:/root/.ssh *
*-w /workspace/torchtitan *
*--shm-size 64G *
*--name torchtitan_bench *
docker.gpuperf:5000/rocm/pytorch-training-private:torchtitan
Execute MI355X run:
bash run_llama_70B_torchtitan_fp8.sh 6 8192

Metric Calculation

Training throughput was reported in units of tokens per second per GPU (tokens/sec/GPU).

MI355X vs B200: Llama3-70B – BF16 – 8GPU

A system running 8 - MI355X GPUs delivered a training throughput that was 2154 tokens per second per GPU using a per-GPU batch size of 8 and a sequence length of 8192.

A system running 8 - B200 GPUs delivered a training throughput that was 1918 tokens per second per GPU using a per-GPU batch size of 16 and a sequence length of 8192.

These values are the median of 5 independent executions, per system (listed below):

- MI355X: 2156, 2152, 2154, 2157, 2152 **(Median: 2154 tokens/sec/GPU)**
- B200: 1918, 1919, 1916, 1918, 1919 **(Median: 1918 tokens/sec/GPU)**

MI355X

Testing on MI355X was executed with a Docker container that was built with run scripts and dependencies from ROCm/torchtitan and packaged with Llama3 70B model elements.

Pull MI355X Docker image from internal registry:
docker pull docker.gpuperf:5000/rocm/pytorch-training-private:torchtitan
Run MI355X container:
*docker run -it --device /dev/dri --device /dev/kfd --network host *
*--ipc host --group-add video --cap-add SYS_PTRACE *
*--security-opt seccomp=unconfined --privileged *
*-v $HOME:$HOME *
*-v $HOME/.ssh:/root/.ssh *

*-w /workspace/torchtitan *
*--shm-size 64G *
*--name torchtitan_bench *
docker.gpuperf:5000/rocm/pytorch-training-private:torchtitan
Execute MI355X run:
bash run_llama_70B_torchtitan_bf16.sh 8 8192

B200

Testing on B200 was executed with a NeMo base container pulled from the Nvidia registry and packaged internally with the run scripts and Llama3 70B model elements.

Pull NeMo B200 container:
docker pull docker.gpuperf:5000/nvidia/nemo_25.04:nemo
Run NeMo B200 container:
*docker run --gpus all -it --network host --ipc host *
*-v $HOME:$HOME *
*-v $HOME/.ssh:/root/.ssh *
*--shm-size=64G *
*-w /workspace/llama3 *
docker.gpuperf:5000/nvidia/nemo_25.04:nemo
Execute B200 run:
bash run_70B_bf16.sh

Metric Calculation

Training throughput was reported in units of tokens per second per GPU (tokens/sec/GPU).

MI355X vs B200: Llama-3.1-405B FP4 – 8 GPU

Testing was conducted using the performance team's Lucid automation framework. Lucid executes the workload in a manner that is equivalent to the manual process described here.

If you do not already have a HuggingFace token, open your user profile (https://huggingface.co/settings/profile), select "Access Tokens", press "+ Create New Token", and create a new Read token.

For tests on AMD GPUs we used a pre-quantized copy of the model available to AMD users at https://huggingface.co/amd/Llama-3.1-405B-Instruct-wmxfp4-amxfp4-kvfp8-scale-uint8.

We used the following vLLM Docker image (Note: this is an internal development container that includes optimizations for MI355X):

rocm/pytorch-private:vllm-fp4-405b-250601-asm-rc2 (sha256:5850fcc1)

Launch/run the container.

Adjust input_tokens, output_tokens, num_prompts, and max_num_seqs for each configuration.

For the 20000/2000 configuration additional configuration changes are needed.

Metric Calculation

We are reporting the output tokens per second (Output Token Throughput) extrapolated to an 8-GPU server. In vLLM benchmark_throughput.py runs the throughput is reported as:

Throughput: xxx.xx requests/s, xxx.xx total tokens/s, xxx.xx output tokens/s

Using the "output tokens/s" value from this output, we extrapolated to an 8-GPU server by scaling the output token throughput by the number of model replicas that would fit on 8 GPUs. For MI355X using tensor parallelism of 1, the scaling factor was 8.

MI355X vs MI300X: MLPerf Training Llama-2-70B Lora

Test Procedures

Test procedure is the same as described in the "MI355X: MLPerf Training Llama-2-70B Lora 8-GPU Training Score" document.

Test procedure for 4 node MI300X submission by MangoBoost is published here.

Metric Calculation

Described in the "MI355X: MLPerf Training Llama-2-70B Lora 8-GPU Training Score" document.

MI355X vs B200: DeepSeek-R1 Online Serving

Testing was conducted using the performance team's Lucid automation framework. Lucid executes the workload in a manner that is equivalent to the manual process described here.

If you do not already have a HuggingFace token, open your user profile (https://huggingface.co/settings/profile), select "Access Tokens", press "+ Create New Token", and create a new Read token.

Tests on MI355X GPUs were run with the SGLang framework using a pre-quantized version of the DeepSeek-R1 model available to AMD users at https://huggingface.co/amd/DeepSeek-R1-WMXFP4-AMXFP4-Scale-UINT8-Attn-MoE-Quant.

On B200 testing was performed with TensorRT-LLM v0.20.0rc3 using Nvidias FP4 quantized model for DeepSeek: https://huggingface.co/nvidia/DeepSeek-R1-FP4.

Users with access to these models can download them.

The following Docker images were used:

- B200: tensorrt_llm/release:0.20.0rc3
- MI355X: rocm/aigmodels-private:experiment_950_5_30_jp (Note: this is an internal development container that includes optimizations for MI355X.)

MI355X

Start the server / run the benchmark.

Run the test with max_concurrency values 1, 2, 4, 8, 16, 32, 64 – _note that the server can be started once and reused for multiple executions of the client.

B200

Download the model.

Build a docker image for selected TensorRT-LLM release using instructions available from NVIDIA. (retrieved from https://nvidia.GitHub.io/TensorRT-LLM/installation/build-from-source-linux.html on June 3, 2025; a copy of this page is archived with the other Test artifacts). Tag the docker image with the GitHub tab to make it easily identifiable like tensorrt_llm/release:0.20.0rc3.

Start the server.

Start a docker container to run the client.

Execute the benchmark inside the client docker container.

Run the test with max_concurrency values 1, 2, 4, 8, 16, 32, 64.

Metric Calculation

Each framework reports metrics from the test including the median Inter Token Latency (ITL) and Total Token Throughput (tokens/second). From these values, we can determine the maximum throughput across any of the tests which had an ITL less than n ms.

In SGLang, the Median ITL is reported in the summary near the end of the output in the line “Median ITL (ms)”, while the total token throughput is in the line “Total token throughput (tok / s)

In genai-perf the output figures of merit are written to a JSON file named ‘profile_export_genai_perf.json’ in the current working directory. The median ITL latency is reported in field `inter_token_latency.p50`with units ‘milliseconds’. The Total Token Throughput is calculated using multiple fields in the output JSON file. An example calculation is shown below.

*TotalOutputTokens = RequestCount * OutputSequenceLength.avg*
*TotalInputTokens = RequestCount * InputSequenceLength.avg*
TestDurationSeconds = TotalOutputTokens/OutputTokenThroughput.avg
TotalTokenThroughput = (TotalOutputTokens+TotalInputTokens)/TestDurationSeconds