AMD FINANCIAL ANALYST DAY 2025

Vamsi Boppana
SVP, Artificial Intelligence Group

AMD

together we advance_

Cautionary Statement

This presentation contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) such as the features, functionality, performance, availability, timing and expected benefits of AMD products including AMD Instinct™ MI400 series accelerators and AMD Instinct MI500 series accelerators; AMD's ability to deliver leadership roadmap on an annual cadence; the strategic partnership with OpenAI and the deployment of six gigawatts of AMD Instinct GPUs and the timing thereof; the strategic partnership with Oracle and the deployment of 50,000 GPUs and the timing thereof; the expected benefits of AMD AI compute strategic partnerships; and AMD's ability to advance AI leadership, which are made pursuant to the Safe Harbor provisions of the Private Securities Litigation Reform Act of 1995. Forward-looking statements are commonly identified by words such as "would," "may," "expects," "believes," "plans," "intends," "projects" and other terms with similar meaning. Investors are cautioned that the forward-looking statements in this presentation are based on current beliefs, assumptions and expectations, speak only as of the date of this presentation and involve risks and uncertainties that could cause actual results to differ materially from current expectations. Such statements are subject to certain known and unknown risks and uncertainties, many of which are difficult to predict and generally beyond AMD's control, that could cause actual results and other future events to differ materially from those expressed in, or implied or projected by, the forward-looking information and statements. Investors are urged to review in detail the risks and uncertainties in AMD's Securities and Exchange Commission filings, including but not limited to AMD's most recent reports on Forms 10-K and 10-Q.

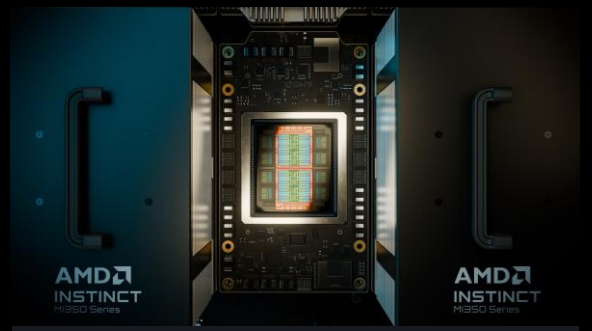
AMD does not assume, and hereby disclaims, any obligation to update forward-looking statements made in this presentation, except as may be required by law.

AMD INSTINCT

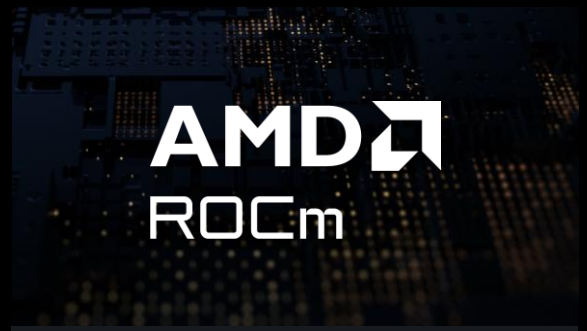
Entering Our Next Phase of Growth

AMD INSTINCT

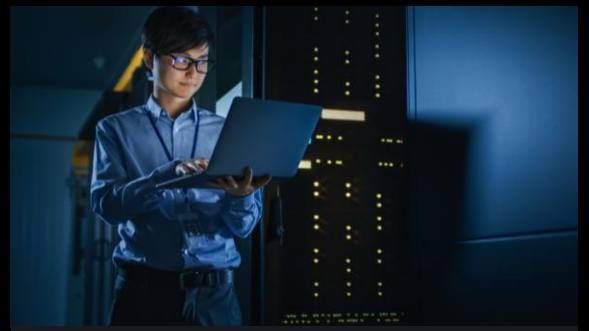
The Journey is Accelerating



AI Performance on Annual Cadence



Open-Source Software Ecosystem Adoption



Expanded Routes to Market

Rapidly Growing Partners & Customers

Rapid Expansion of Partners & Customers

7 of Top 10 AI Companies Have Deployed AMD Instinct™ at Scale

GTM Partners

Microsoft, ORACLE, HUMAIN, IBM, DELL Technologies, Lenovo, SUPERMICRO, Hewlett Packard Enterprise, CISCO, EVIDEN an atos business, core42, VULTR, Crusoe, DigitalOcean, TENSORWAVE, IREN, Cirrascale, EVERGRID, ASUS, ASRock Rack, GIGABYTE, Celestica, COMPAL, MiTAC, PEGATRON, QCT, wiwynn

Software Ecosystem

PyTorch, Hugging Face, vLLM, SGL, TensorFlow, llm-d, ONNX, deepspeed, UbiOps, MLIR, Red Hat OpenShift AI, Open Platform for Enterprise AI, OpenXLA, LLAMA STACK, CLEAR|ML, clarifai, unsloth, ComfyUI, LlamaIndex

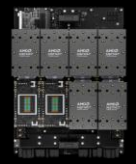
Customers

Meta, Microsoft, OpenAI, xAI, TESLA, cohere, G42, Luma AI, SAMSUNG, Jio, absci, character.ai, CHAI, Sciforium, Fireworks AI, Liquid, seekr, 310.AI, ZYPHRA, Lawrence Livermore National Laboratory, OAK RIDGE National Laboratory, LUMI, GENCI, HLRS, CSCS, eni

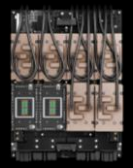
AMD Instinct™ MI350 Series

- Next Gen AI Compute • FP4 | FP6 | FP8
- Leadership Memory • HBM3E
- Industry Standard Design • OCP

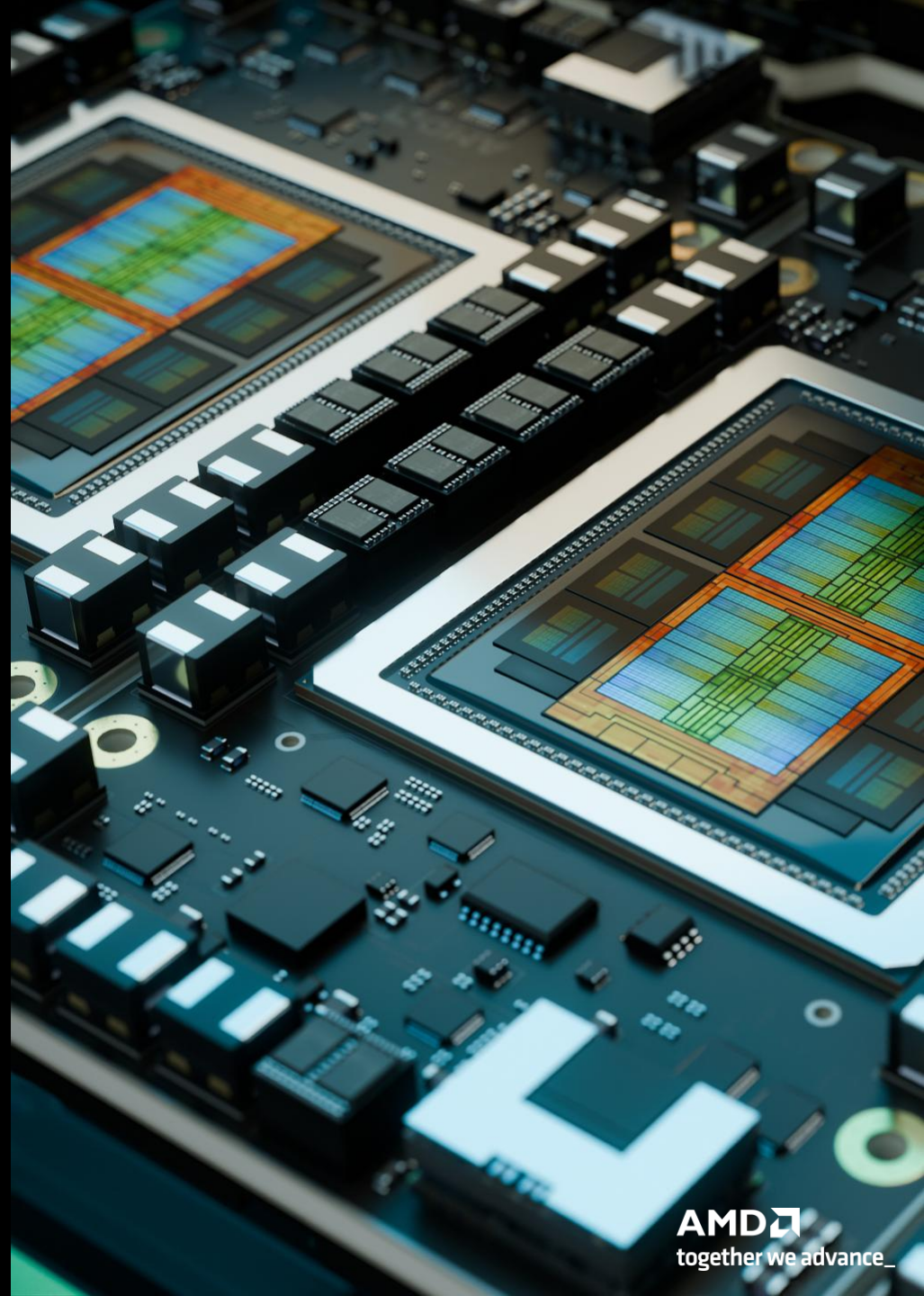
AMD CDNA 4



AMD Instinct™
MI350X



AMD Instinct™
MI355X

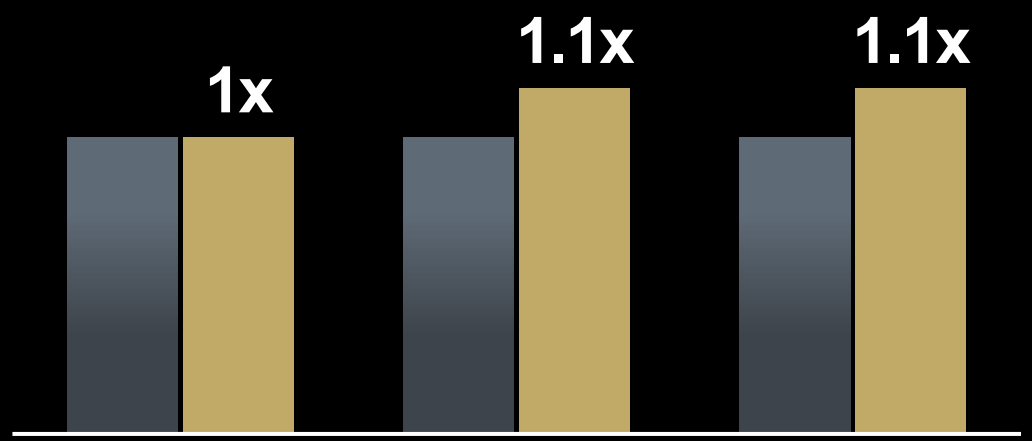


*See endnote MI350-006

AMD Instinct™ MI350 with ROCm™ Delivers Performance Leadership

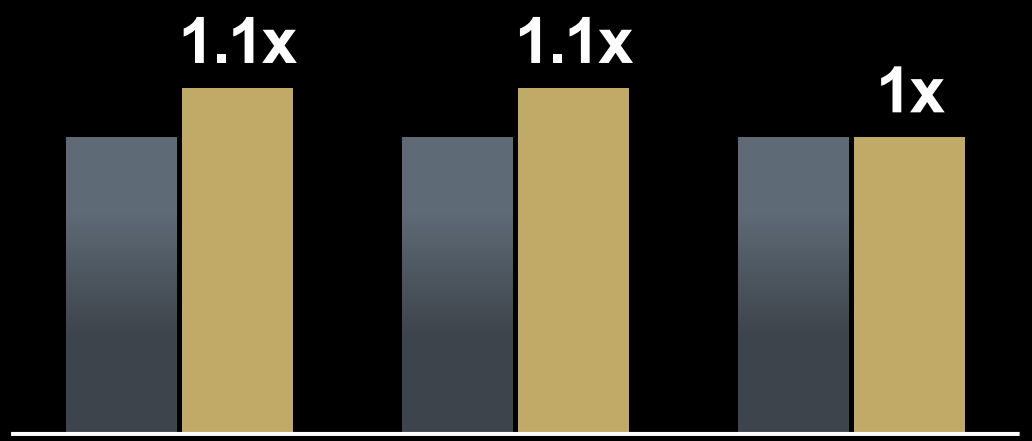
Inference Performance

B200 MI355X



Training Performance

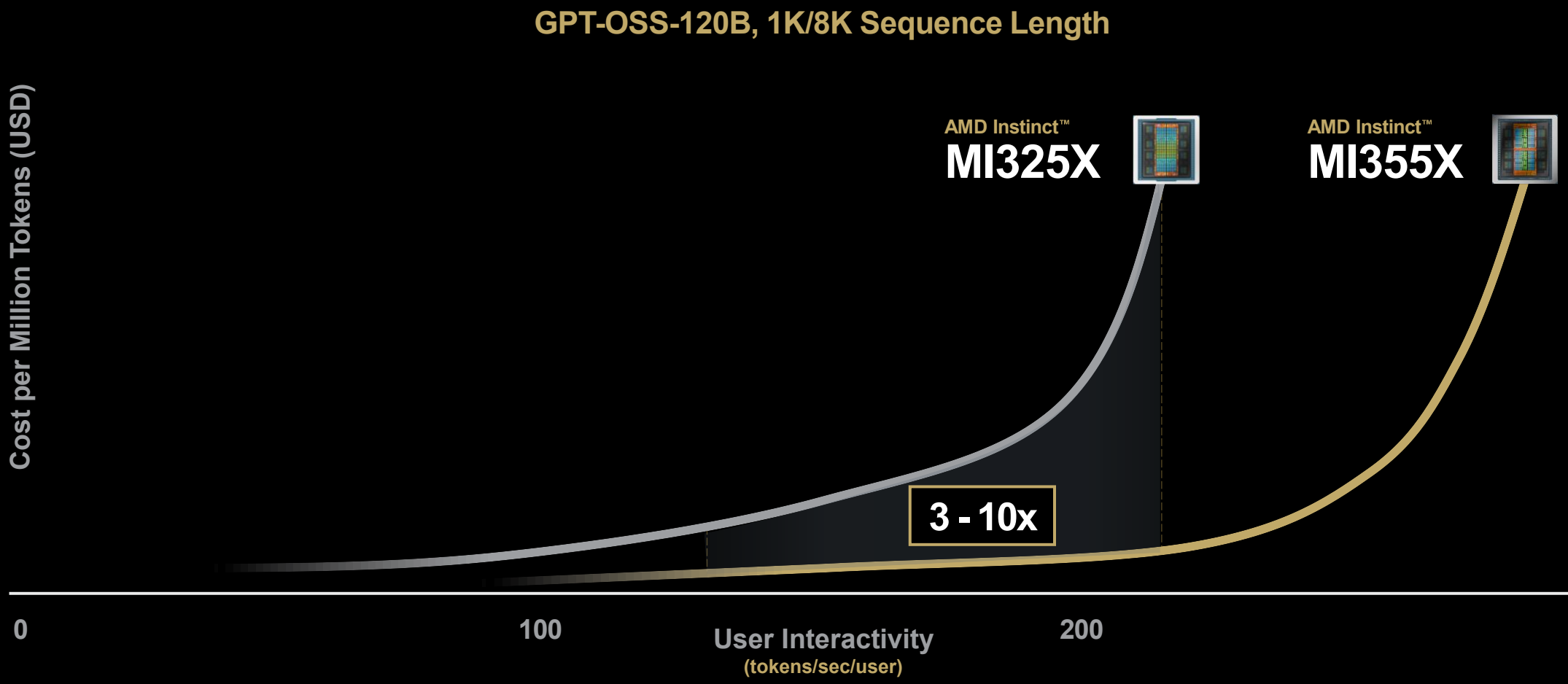
B200 MI355X



See endnotes MI350-062, 063, 064, 065, 066, & 067.

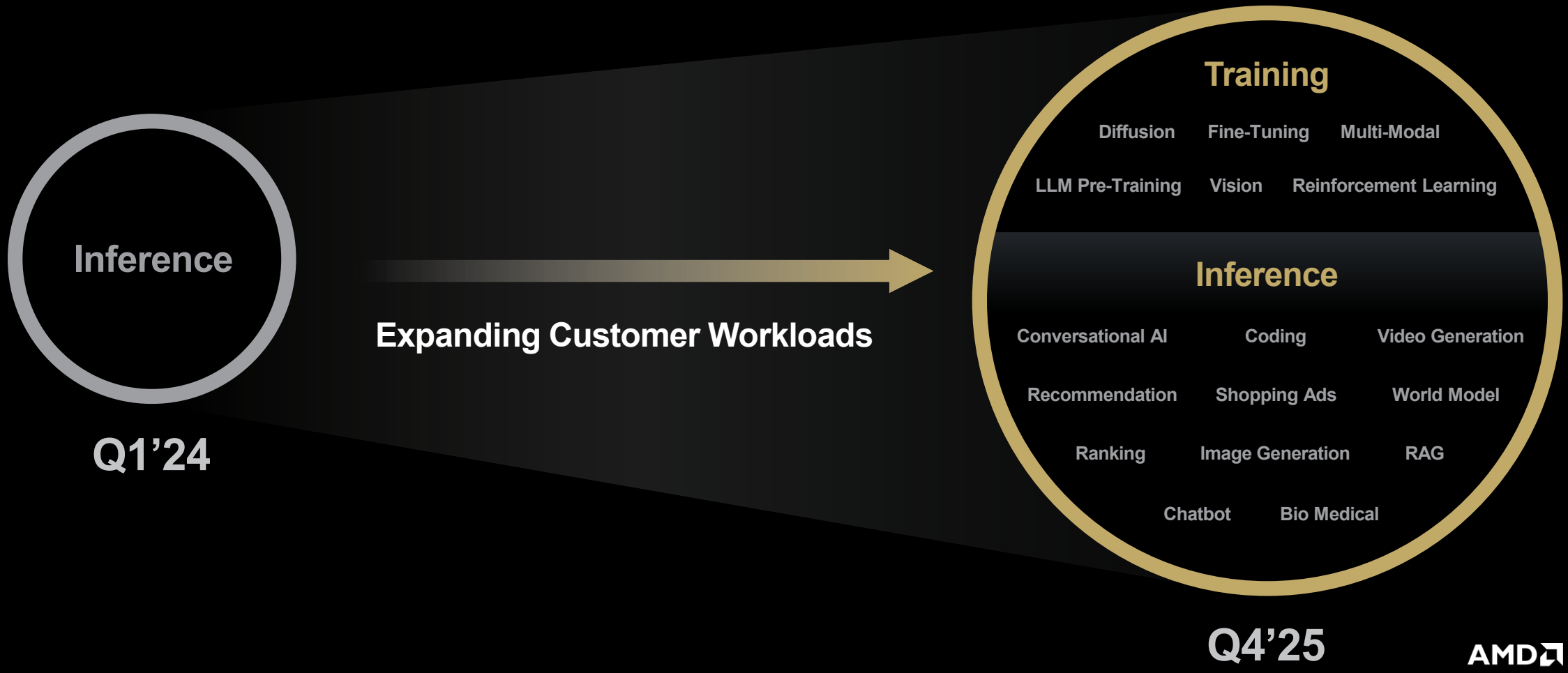
AMD Instinct™ MI350: Generational Efficiency Gains

Up to 10x Inference Cost-per-Token Reduction Gen-on-Gen



Based on SemiAnalysis' InferenceMax, AMD data (10/27)

Growing Workload Deployments on AMD Instinct™ GPUs



Enabling Open Innovation at Scale

AMD ROCm

Relentless Focus on Developers

Accelerated Inference Capabilities	Expanded Support for Training	Richer Out-of-the-Box Experience	Developer-First Approach to Enablement
Accelerated Release Cadence	Day-0 Support for Leading Models	Deepening Ecosystem Partnerships	Industry Benchmarks

AMD ROCm™ Momentum

10x	Top 10	2M
Increase in ROCm Software Downloads Y/Y	AI Open-Source GPU Projects* Natively Supported	Hugging Face Models Supported

*Top OSS projects = Pytorch, TensorFlow, JAX, Triton, Hugging Face Transformers, Unsloth, Ollama, ComfyUI, vLLM, SGLang

Expanding the Developer Community

Growing Learning & Compute Access for AI Communities

Early Support for Research Labs / Start-Ups

HARVARD UNIVERSITY · Stanford University · MIT · UC Berkeley

100,000+ Learners Through AMD AI Education Program

AMD AI Academy · coursera · DeepLearning.AI

Active Engagement Through Developer Events

AMD AI DevDay · AMD Hackathon · GPU MODE

Scalable Developer Cloud for Instant GPU Access

AMD Developer Cloud

Advancing AMD ROCm™ Strategy

Open Source

Driving Rapid Innovation

Abstraction

Enabling Higher Productivity

PyTorch · Hugging Face · vLLM · SGL · TensorFlow · llm-d

ONNX · deepspeed · UbiOps · Red Hat OpenShift AI · Open Platform for Enterprise AI · OpenXLA · MLIR · unsloth

Advancing AMD ROCm™ Strategy

AI Assist

Automating GPU Programming

Open Source

Driving Rapid Innovation

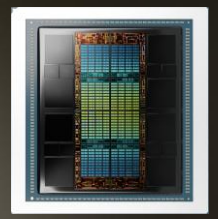
Abstraction

Enabling Higher Productivity

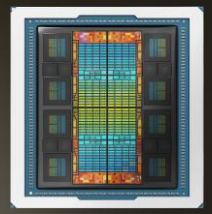
PyTorch · Hugging Face · vLLM · SGL · TensorFlow · llm-d · ONNX · deepspeed · UbiOps · Red Hat OpenShift AI · Open Platform for Enterprise AI · OpenXLA · MLIR · unsloth

Leadership Roadmap on Annual Cadence

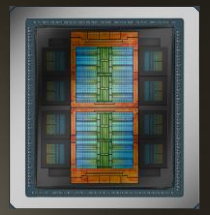
Accelerating Customer Adoption



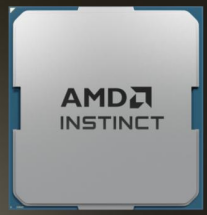
MI300X
2023



MI325X
2024



MI350 Series
2025



MI400 Series
2026

Roadmap subject to change.

AMD Instinct™ MI450 Series

Most Advanced AI Accelerator

40 PF 20 PF	432 GB 19.6 TB/s	3.6 TB/s	300 GB/s
FP4 • FP8 Flops	HBM4 Memory	Scale Up Bandwidth	Scale Out Bandwidth

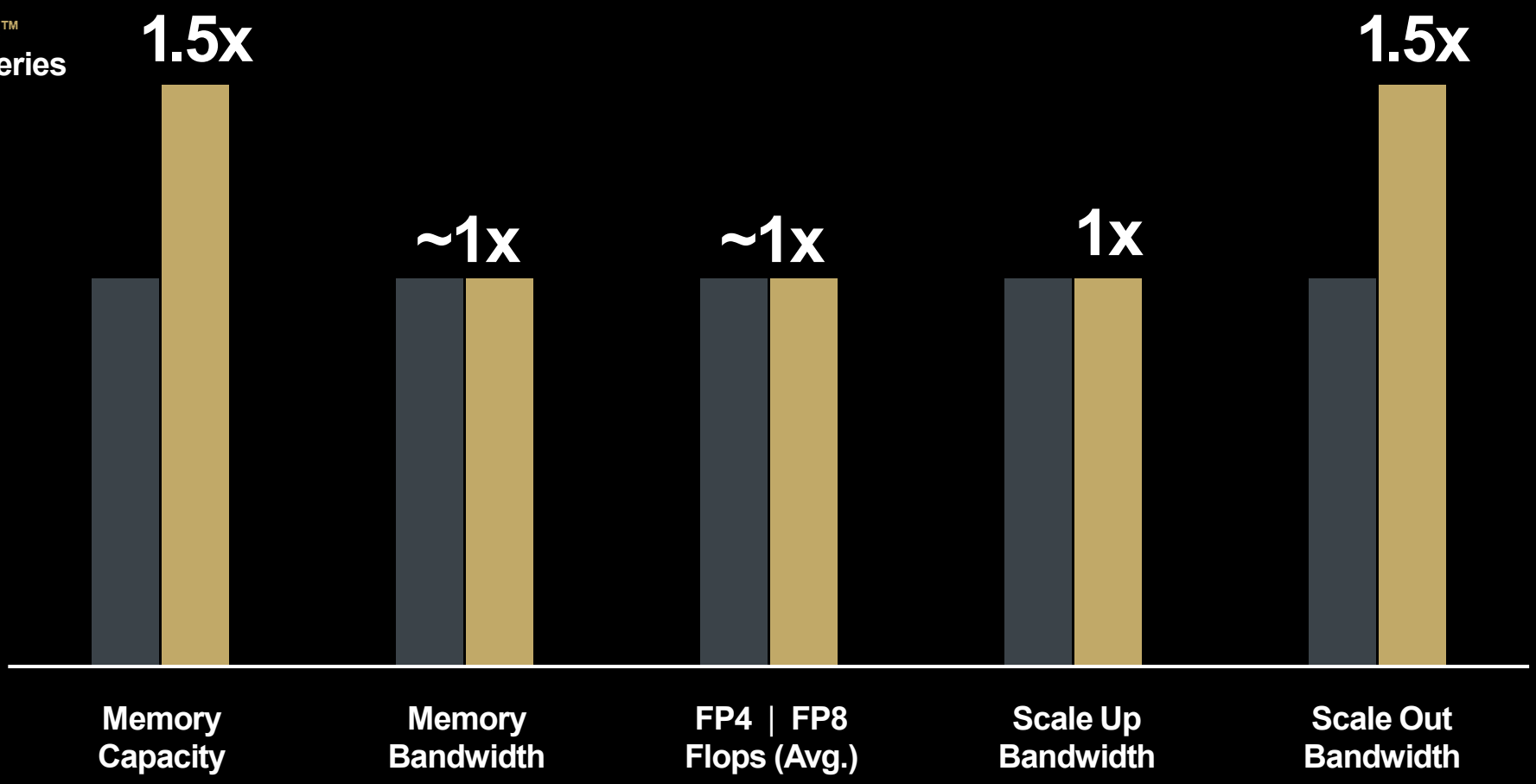
Advanced Process Technology	Leadership IP	Chiplet Architecture	3.5D Packaging	Co-Designed Hardware & Software

Source: AMD specs based on engineering projections, results subject to change.

Rack-Scale Performance Leadership With MI450 Series

Nvidia
Vera Rubin

AMD Instinct™
MI450 Series



Source: AMD specs based on engineering projections, results subject to change MI450-002.

AMD
together we advance_

AMD Instinct™ MI400 Series Portfolio

Leadership Across AI & Scientific Computing

AMD Instinct™

MI455X

At Scale AI Training & Inference

AI Compute | Scale Out Performance | HBM4 Memory

AMD Instinct™

MI430X

Sovereign AI & HPC

Hybrid Compute | Hardware-Based FP64 | HBM4 Memory

Strategic Partnerships Driving The Next Generation of AI Compute

AMD × OpenAI

Frontier AI Development

6GW

AMD Instinct™ Accelerators
Starting in 2H'2026

AMD × ORACLE

Zetta-Scale Compute

50,000

AMD Instinct™
MI450 Series GPUs
Starting in 2H'2026

AMD × Meta

Co-Designed Open Infrastructure

AMD "Helios" Rack
Co-defined with Meta
and Previewed at OCP

AMD × U.S. DEPARTMENT of ENERGY

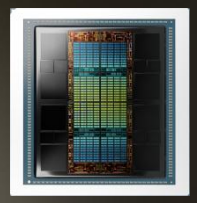
Extending US HPC Leadership

Lux: First US AI Factory
AMD Instinct MI355X Series

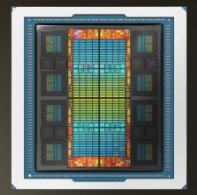
Discovery: Flagship
AI Supercomputer
AMD Instinct MI430X Series

Leadership Roadmap on Annual Cadence

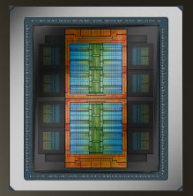
Accelerating Customer Adoption



MI300X
2023



MI325X
2024



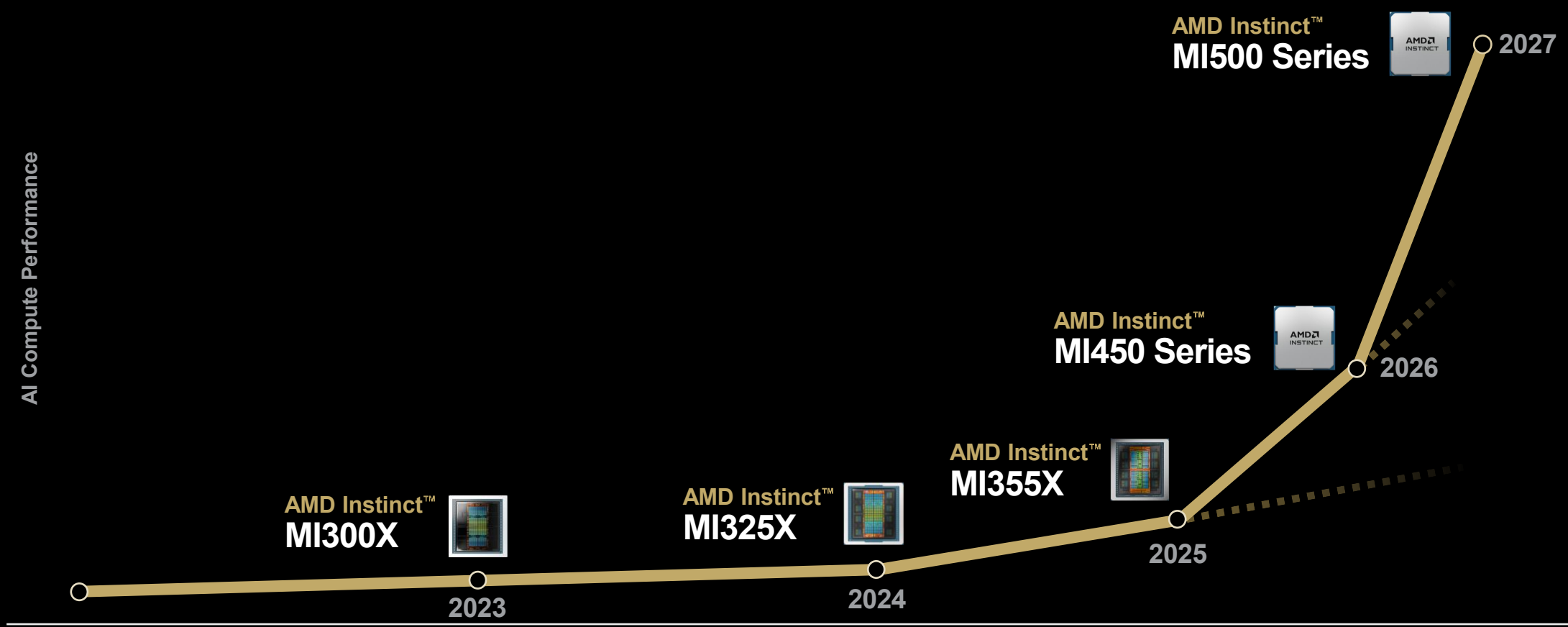
MI350 Series
2025

MI400 Series
2026

MI500 Series
2027

Roadmap subject to change.

Next Big Leap in AI Performance With MI500 Series



Based on AMD internal projections as of Nov. 2025.

AMD INSTINCT

Advancing AI Leadership

Leadership Roadmap	Deep Partnerships at Scale	Open Innovation Ecosystem
Delivering Transformative AI Performance	Building the Future of Compute Together with AI Leaders	Driving Adoption Through Open Platforms

Disclaimer & Attribution

DISCLAIMER: The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18u.

Endnotes

- MI350-006: Calculations by AMD Performance Labs in May 2025, based on the published memory capacity specifications of AMD Instinct™ MI350X / MI355X OAM GPU vs. an NVIDIA Hopper H200 GPU and Server manufacturers may vary configurations, yielding different results.

- MI350-062: Based on testing by AMD Performance Labs on Oct 24, 2025, using a Grok 2.5 xAI model to measure the average end-to-end latency improvement on an AMD Instinct 8x GPU MI355X platform vs. an Nvidia 8x GPU B200 platform. End-to-end latency measured during an inference performance cycle, with an 8K input length and 1k output length. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. AMD system configuration: Dual Core AMD EPYC 9575F 64-core processor , 8x AMD Instinct MI355X GPU platform, AMD ROCm 7.0.0 software, SGLang 0.5.4, PyTorch 2.9.0, Ubuntu® 22.04. NVIDIA system configuration: 2x Intel Xeon 6960P processors, 8x Nvidia B200 (NVLink 192G, 1000W) GPU platform, Nvidia driver 570.133.20, and Ubuntu 22.04.

- MI350-063: Based on testing by AMD Performance Labs on Oct 23, 2025, using an OpenAI GPT-OSS-120b model to measure the average throughput per second (TPS) of an 8x GPU AMD Instinct MI355X platform vs. an 8x GPU Nvidia B200 platform TPS measured during an inference performance cycle with an 8K input length and 1k output length. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. AMD system configuration: One (1) dual core AMD EPYC 9575F 64-core processor, 8x GPU AMD Instinct MI355X platform, System Bios: 1.4a, ROCm 7.0.2, amdgpu version: 6.14.14 (2218231), Ubuntu® 22.04.5 LTS. NVIDIA system configuration: 2x Intel Xeon 6960P processors, 8x Nvidia B200 (NVLink 192G, 1000W) GPU platform, Nvidia driver 570.133.20, and Ubuntu 22.04.

- MI350-064: Based on testing by AMD Performance Labs on Oct 27, 2025, measuring the Llama3 70B model's pretraining throughput (tokens/second/GPU) on an AMD Instinct 8x GPU MI355X platform running Primus Torchtitan vs. an NVIDIA 8x GPU B200 platform, running NVIDIA Nemo. Each config tested with BF16 data precision and a maximum sequence length of 8192 tokens and a per-GPU batch size of 8. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. AMD system configuration: Dual Core AMD EPYC 9575F 64-core processor, 8x AMD Instinct MI355X GPU platform, System BIOS 1.4a, 4 NUMA nodes per socket, Host OS Ubuntu 22.04.5 LTS with Linux kernel 5.15.160-generic, Host GPU driver AMD ROCm 7.0.1 + amdgpu 6.14.12, PyTorch deep learning framework 2.9.0, and AMD ROCm 7.0.1 software. NVIDIA system configuration: 2x Intel Xeon 6960P processors, 8x Nvidia B200 (NVLink 179GB, 1000W) GPU platform, System Bios 1.0, 3 NUMAnodes per socket, Host OS Ubuntu 22.04.5 LTS with Linux kernel 5.15.0-156-generic, Host GPU driver 580.82.07, PyTorch deep learning framework 2.8.0, and CUDA 13.0 software (MI350-064)

Endnotes

- MI350-065: Based on testing by AMD Performance Labs on Oct 27, 2025, measuring the Mixtral 8x7B model's pretraining throughput (tokens/second/GPU) on an AMD Instinct 8x GPU MI355X platform running Primus Megatron-LM, vs. an NVIDIA 8x GPU B200 platform, running NVIDIA NeMo. Both configs tested with BF16 precision datatype, a maximum sequence length of 8192 tokens, and a per-GPU batch size of 8. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. AMD system configuration: Dual Core AMD EPYC 9575F 64-core processor, AMD Instinct 8x GPU MI355X platform, System BIOS 1.4a, 4 NUMA nodes per socket, Host OS Ubuntu 22.04.5 LTS with Linux kernel 5.15.160-generic, Host GPU driver ROCm 7.0.1 + amdgpu 6.14.12, PyTorch deep learning framework 2.9.0, and AMD ROCm 7.0.1 software. NVIDIA system configuration: 2x Intel Xeon 6960P processors, Nvidia 8x GPU B200 (NVLink 179GB, 1000W) GPU platform, System Bios 1.0, NUMA 3 nodes per socket, Host OS Ubuntu 22.04.5 LTS with Linux kernel 5.15.0-156-generic, Host GPU driver 580.82.07, PyTorch deep learning framework 2.8.0, and CUDA 13.0 software.

- MI350-066: Based on testing by AMD Performance Labs on Oct 27, 2025, measuring the Mixtral 8x7B model's pretraining throughput (tokens/second/GPU) on an AMD Instinct 8x GPU MI355X platform running Primus Megatron-LM, vs. an NVIDIA 8x GPU B200 platform, running NVIDIA NeMo. Both configs tested with BF16 precision datatype, a maximum sequence length of 8192 tokens, and a per-GPU batch size of 8. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. AMD system configuration: Dual Core AMD EPYC 9575F 64-core processor, AMD Instinct 8x GPU MI355X platform, System BIOS 1.4a, 4 NUMA nodes per socket, Host OS Ubuntu 22.04.5 LTS with Linux kernel 5.15.160-generic, Host GPU driver ROCm 7.0.1 + amdgpu 6.14.12, PyTorch deep learning framework 2.9.0, and AMD ROCm 7.0.1 software. NVIDIA system configuration: 2x Intel Xeon 6960P processors, Nvidia 8x GPU B200 (NVLink 179GB, 1000W) GPU platform, System Bios 1.0, NUMA 3 nodes per socket, Host OS Ubuntu 22.04.5 LTS with Linux kernel 5.15.0-156-generic, Host GPU driver 580.82.07, PyTorch deep learning framework 2.8.0, and CUDA 13.0 software.

- MI350-067: Based on time-to-complete testing by AMD as of 10/17/2025 on an 8x GPU AMD Instinct™ MI355X platform for overall GPU-normalized training throughput for fine-tuning, using the Llama2-70B model with LoRA (FP8 precision). AMD test results were compared to the respective published performance results of an 8x GPU Nvidia B200 Platform (FP8). Server manufacturers may vary configurations, yielding different results. Performance may vary based on the use of the latest drivers and optimizations. AMD system configuration: Supermicro smci355-ccs-aus-n05-17: 2x AMD EPYC 9575F Processors (2 sockets, 64 cores per socket, 2 threads per core), 8x AMD Instinct™ MI355X (256GB, 2TB host memory) GPUs, BIOS RP 700D, SFO, PMFW 86.42.150, ROCm 7.0.0, Ubuntu 22.04.5 LTS, PyTorch 2.8.0++git64359f59. Docker image: rocm/mlperf:llama2_70b_training_5.1_2025-10-09-01-36-33 vs. Nvidia B200 8xGPU platform Results submitted here - https://mlcommons.org/benchmarks/training

Endnotes

- MI450-002: Preliminary performance comparison for AMD Instinct™ MI450 AI Rack Versus Nvidia Vera Rubin Rack. AMD Instinct specifications based on AMD engineering projections as of 11.10.2025 and are subject to change. Nvidia Vera Rubin specifications are from published data by Nvidia and may also be subject to change. AMD Rack vs. Nvidia Rack, Memory Capacity (HBM4) - 1.5x, Memory Bandwidth - 1x, FP4|FP8 FLOPS (avg) - 1x, Scale up bandwidth - 1x, Scale out bandwidth - 1.5x.