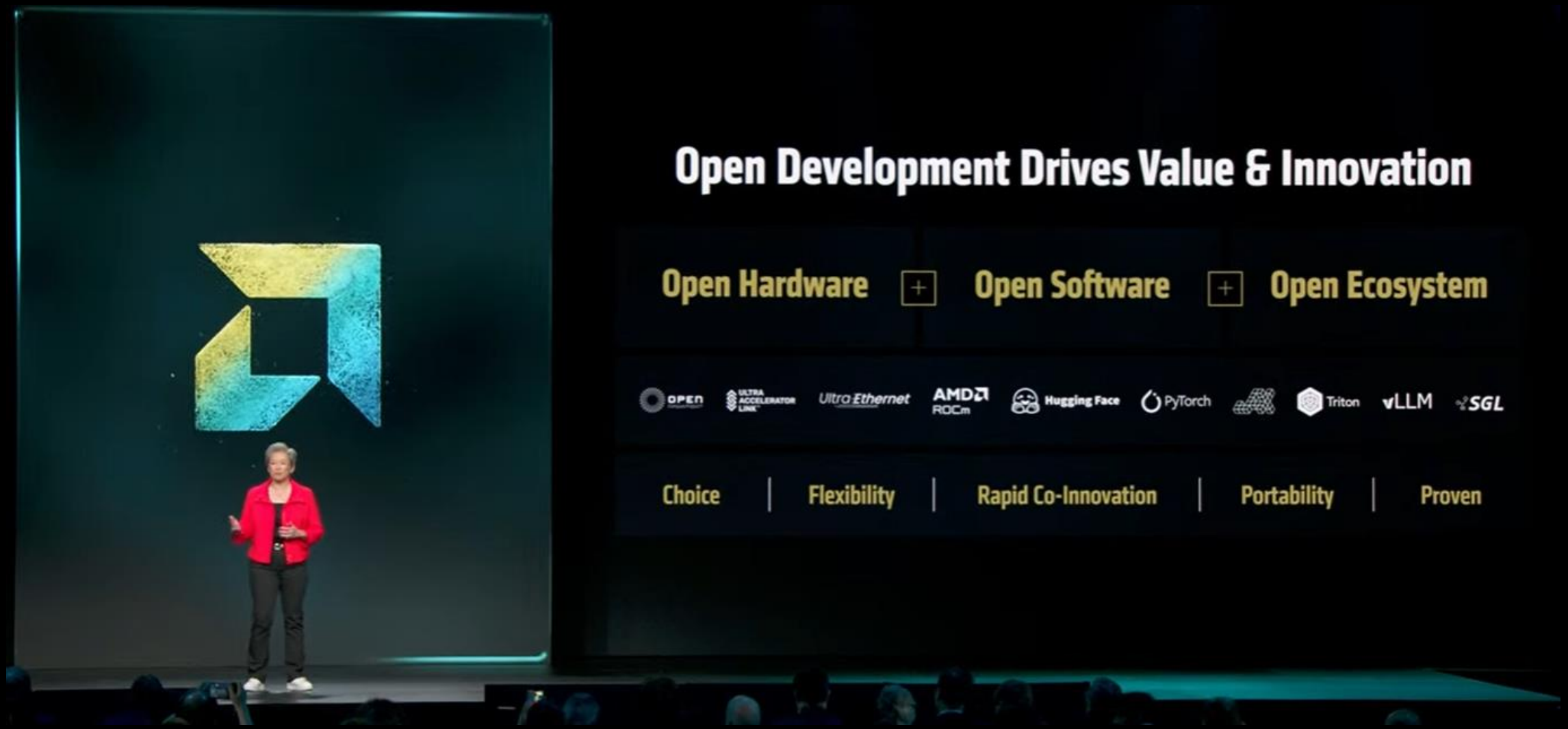

ADVANCING
AI 2025
AMDのAI最新動向
2025/6/27 AMD FAE 石橋史康
AMD

オープンなAI技術への取り組み



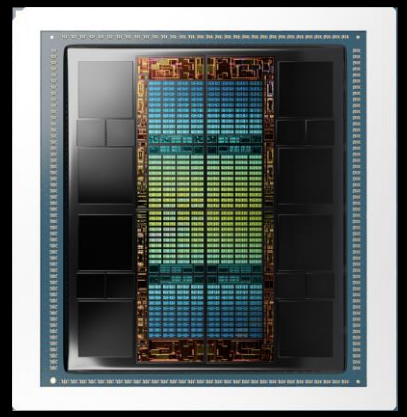
AMD | AIデータセンターを進化させる

See endnotes EPYC-029C. ZT Systems announced acquisition expected to close 1H 2025

AMD INSTINCT

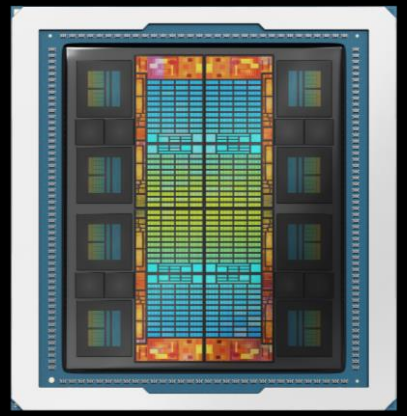
年次ロードマップのコミットメントの履行

AMD Instinct™
MI300A/X



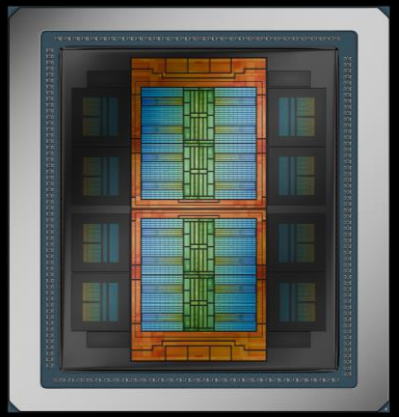
2023

AMD Instinct™
MI325X



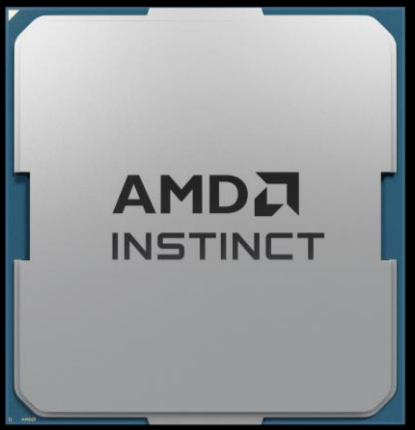
2024

AMD Instinct™
MI350 SERIES



2025

AMD Instinct™
MI400 SERIES

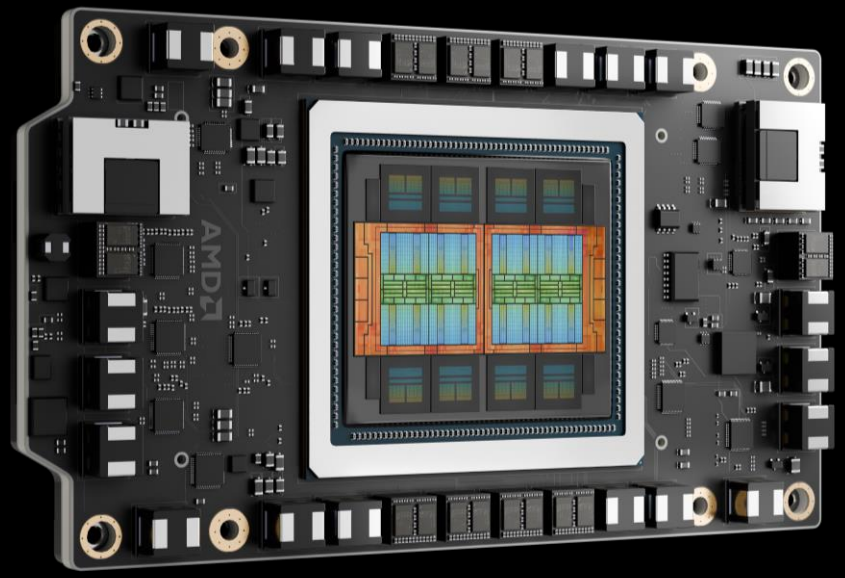


2026

Roadmap subject to change

AMD Instinct™

MI350 Series

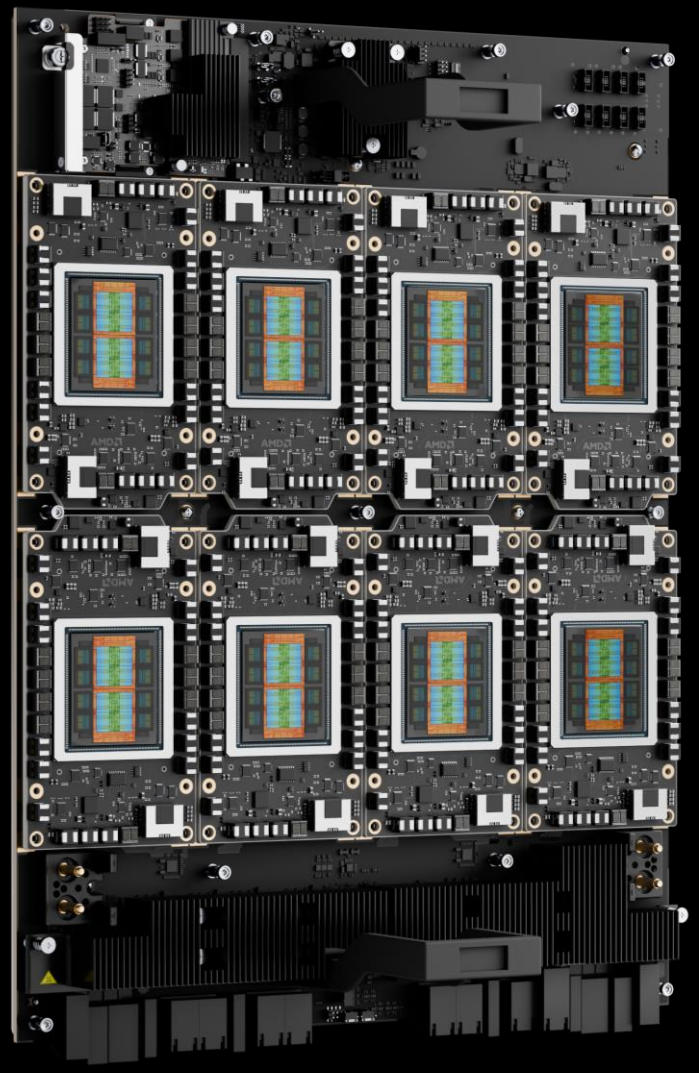


	Instinct™ MI355X
MEMORY	288 GB HBM3E
MEMORY BANDWIDTH	8 TB/s
FP64	79 TF
FP16	5 PF
FP8	10 PF
FP6	20 PF
FP4	20 PF
TBP	1400W

Instinct™ MI350 Series 競合比較

	vs. GB200	vs. B200
MEMORY	1.6x	1.6x
MEMORY BANDWIDTH	1.0x	1.0x
FP64	2.0x	2.1x
FP16	1.0x	1.1x
FP8	1.0x	1.1x
FP6	2.0x	2.2x
FP4	1.0x	1.1x

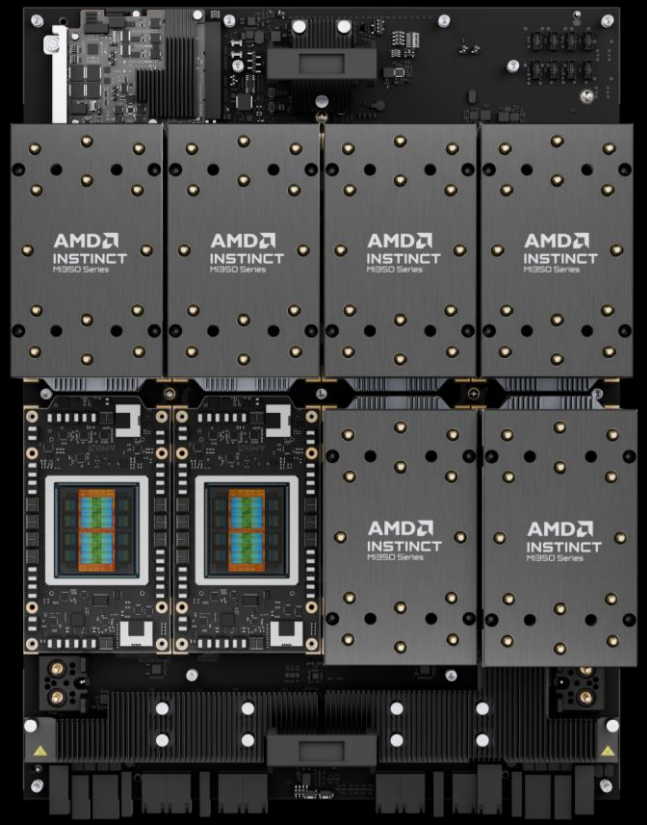
See endnote: MI350-008, 009



AMD Instinct™

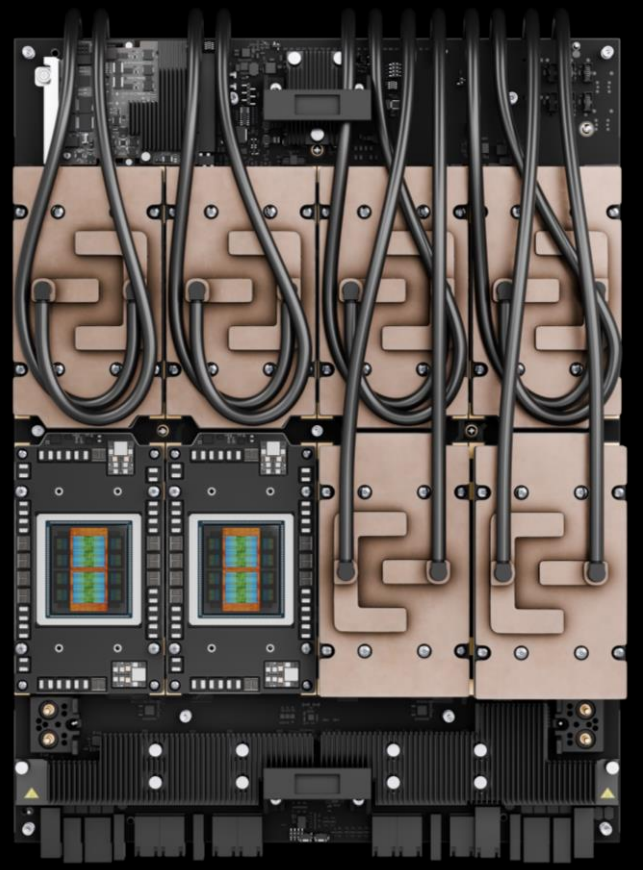
MI350 Series プラットフォーム

	Instinct™ MI355X • 8x
MEMORY	**2.3** TB HBM3E
MEMORY BANDWIDTH	**64** TB/s
FP64	**0.63** PF
FP8	**81** PF
FP6	**161** PF
FP4	**161** PF



空冷

AMD Instinct™ **MI350 Series**



液冷

AMD Instinct™ **MI350 Series**

MI355Xは世代を超えた性能の飛躍を実現

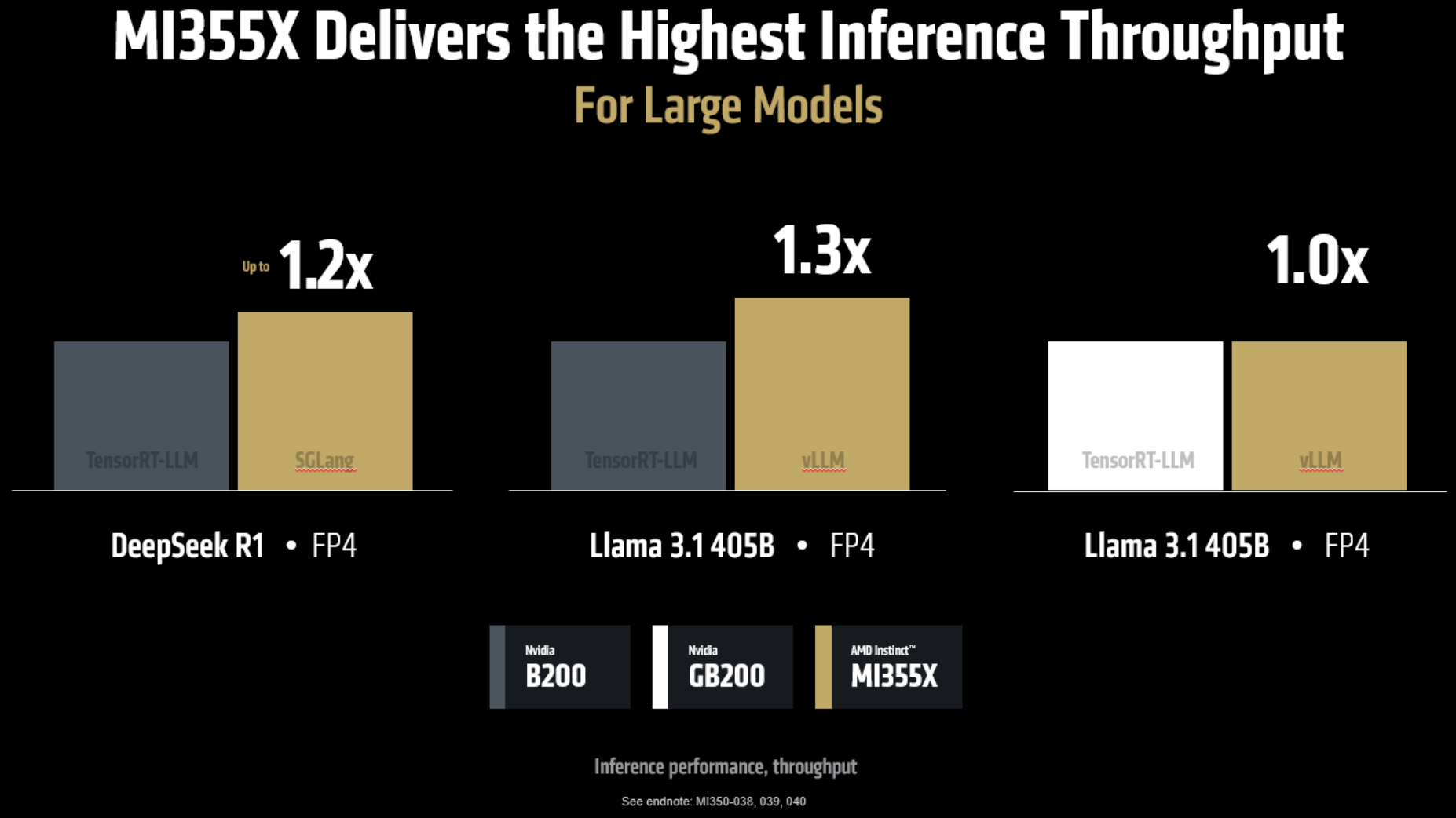
超低遅延推論

Llama 3.1 405B

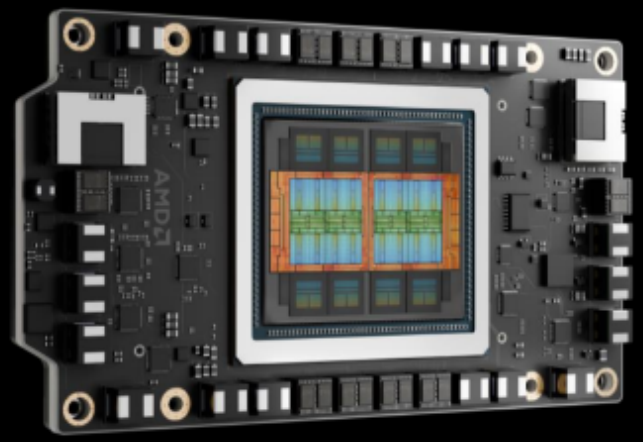
Inference performance, throughput • MI355X (FP4) and MI300X (FP8)

See endnote: MI350--044

MI355X推論性能（競合比較）



MI355X コスト効率のよいGPU

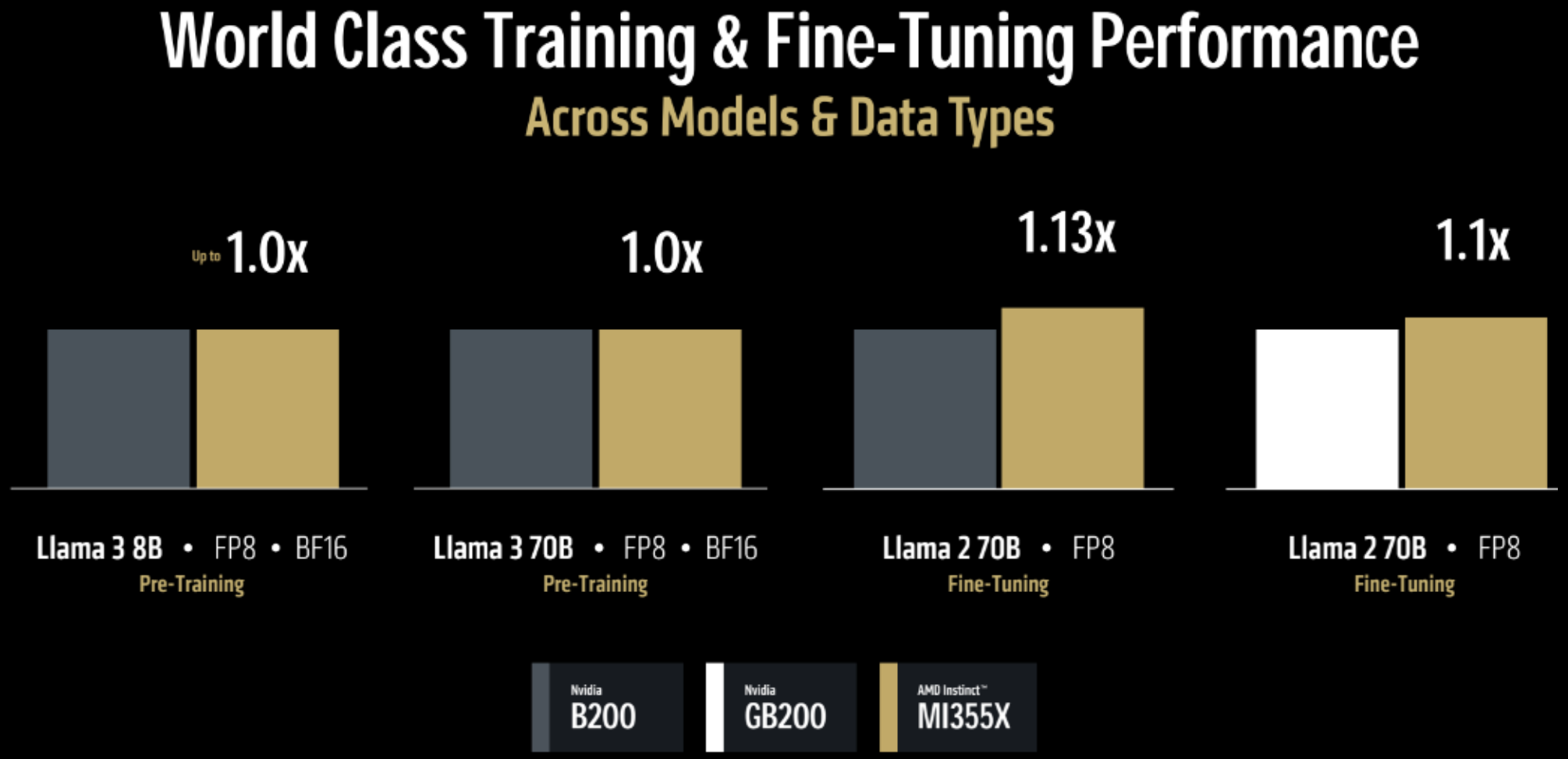


Up to 40% More Tokens / $

Using AMD Instinct™ MI355X vs. B200

See endnote: MI350-049

MI355X学習性能（競合比較）



See endnote: MI350-031, 032, 030, 033

AMD ROCm™ 7

AIイノベーションの加速と開発者生産性の向上

Latest Algorithms & Models	Advanced Features for Scaling AI	MI350 Series Support	Cluster Management	Enterprise Capabilities

ROCm 7新機能

Growing Inference Capabilities

New AMD ROCm Features

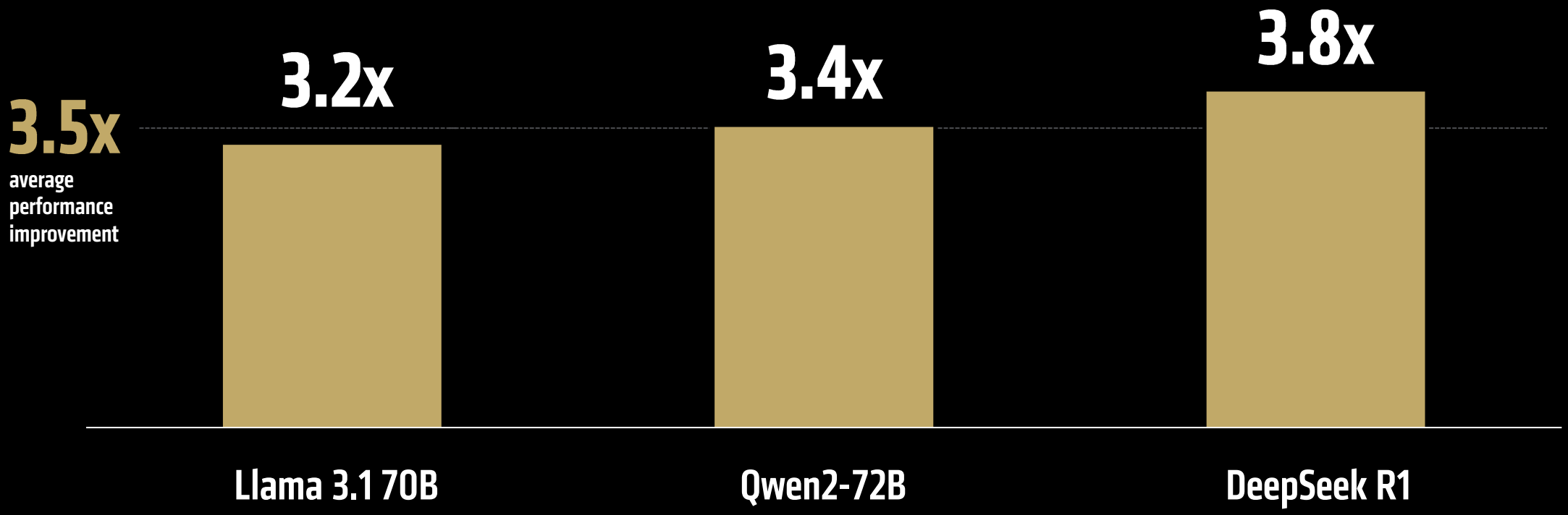
Enhanced Frameworks	vLLM v1	llm-d	SG Lang	
Serving Optimization	Distributed Inference	Prefill	Disaggregation	
Kernels & Algorithms	GEMM Autotuning	MoE	Attention	Python-Based Kernel Authoring
Communication	rocSHMEM	GPU Direct Access	RCCL	
Advanced Data Types	FP8	FP6	FP4	Mixed

Growing Training Capabilities

New AMD ROCm Features

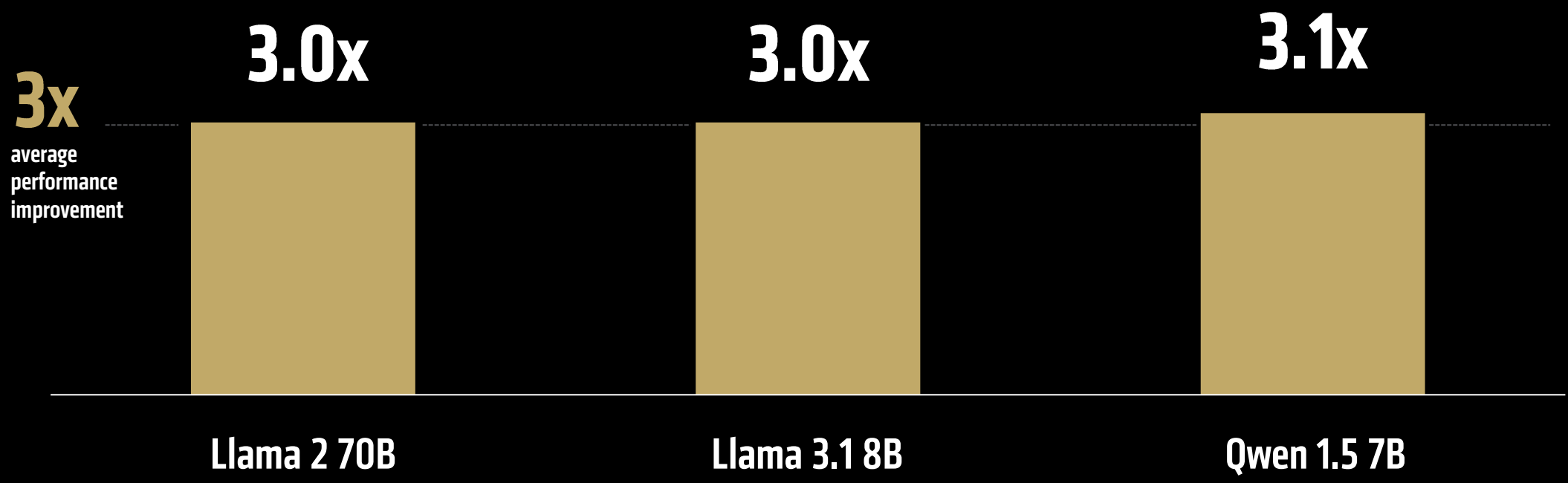
AMD Open-Source Models	Text-to-Text	Text-to-Image	European Models	Multimodal	Game Agent	
Enhanced Frameworks	PyTorch	JAX Maxtext	Torchtune	Torch-titan		
Parallelization	DP	PP	TP	FSDP	CP	EP
Kernels and Algorithms	GEMM		Attention			
Advanced Data Types	BF16		FP8			

ROCm7による推論性能の向上



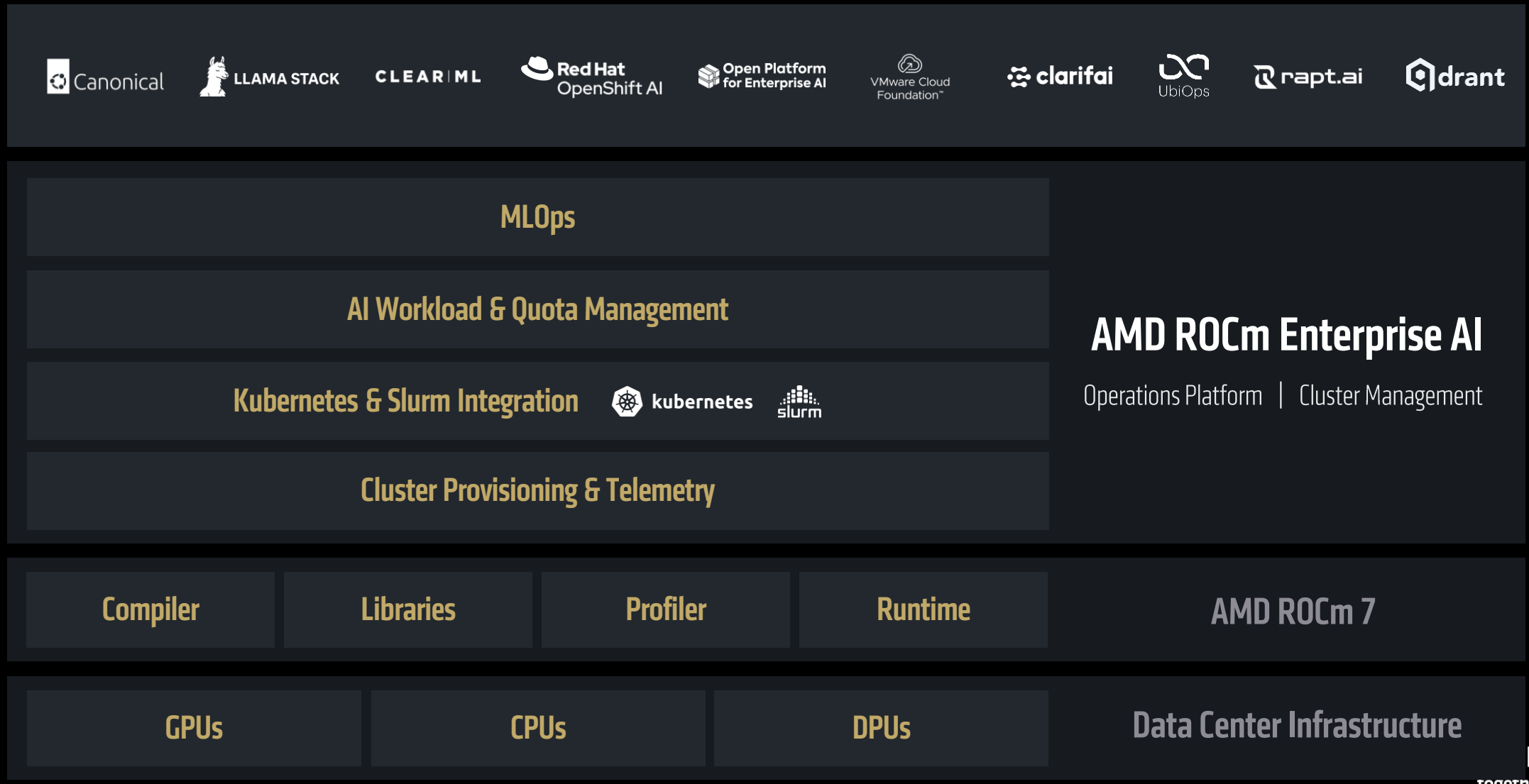
See endnote: MI300-080

ROCm7学習性能の向上

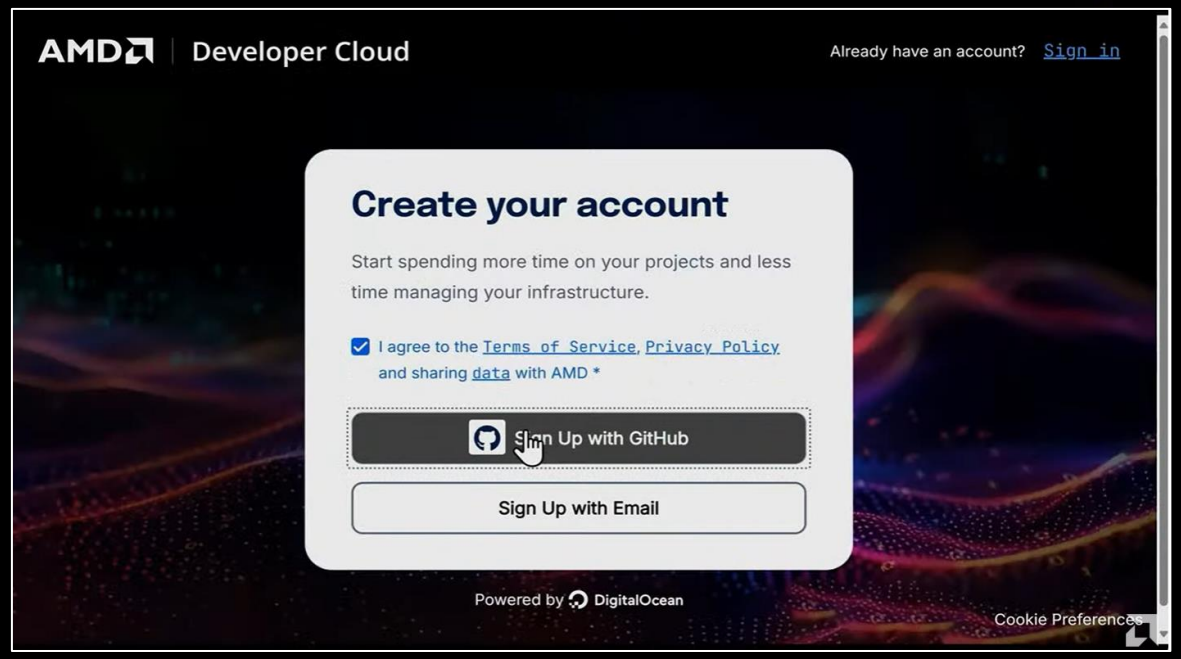


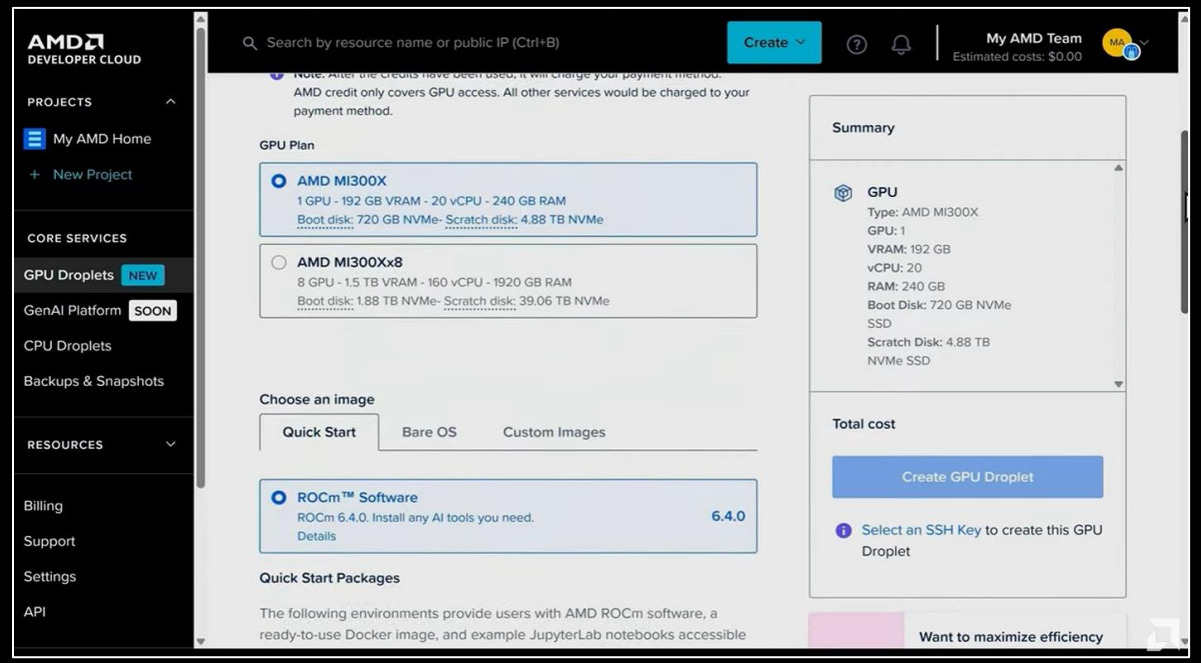
See endnote: MI350-081

AMD ROCm for Enterprise AI



AMD Developer Cloud





25時間の無料GPU利用時間（MI300X GPUインスタンス1台分相当の約$50 US分のクレジット）を10日間ご利用いただけるキャンペーンを展開中

https://www.amd.com/en/blogs/2025/introducing-the-amd-developer-cloud.html

ROCmクライアント対応

Expanding AMD ROCm on Client

AI-Assisted Coding | Customization | Automation | Advanced Reasoning | Model Fine-Tuning

AMD Ryzen AI 300

Up to 24B parameters

AMD Ryzen AI Max

Up to 70B parameters

In-Box Linux Support

2H 2025 — Red Hat EPEL

2H 2025 — Ubuntu

NEW — OpenSUSE

Fedora

Full Windows Support

NEW — PyTorch
Preview Q3 2025

NEW — ONNX-EP
Preview July 2025

HIP SDK

Linux in Windows WSL

AMD Threadripper™
+ Radeon™ AI

Up to 128B parameters

Ultra Ethernet Consortium

AIとHPCの増大する大規模なネットワーク需要を満たすために、イーサネットをオープンで相互運用可能な高性能なフルコミュニケーションスタックアーキテクチャへと進化させる

UEC 1.0 スペック - 昨日公開！

性能 | スケーラビリティ | コスト効率

ステアリングメンバー

AMD, ARISTA, BROADCOM, CISCO, EVIDEN an atos business, Hewlett Packard Enterprise, intel, Meta, Microsoft, ORACLE

一般メンバー

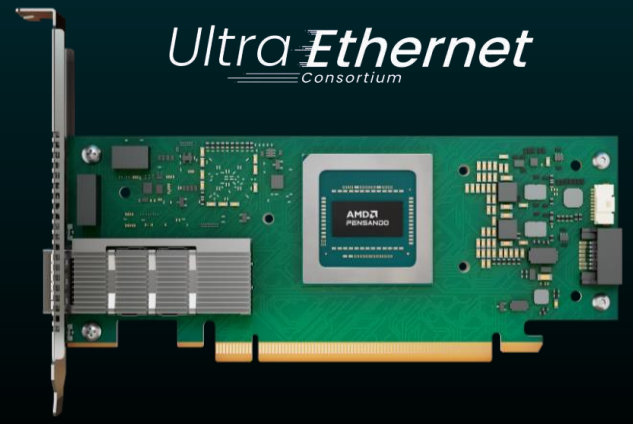
Alibaba Cloud, ARRCUS, Baidu 百度, 世纪互联 VNET, ByteDance, cadence, CORNELIS NETWORKS, DELL Technologies, enfabrica, HUAWEI, IBM, JUNIPER NETWORKS, KEYSIGHT TECHNOLOGIES, Lawrence Livermore National Laboratory, Lenovo, MARVELL, H3C, NOKIA, NVIDIA, Preferred Networks, PURESTORAGE, Qualcomm, Spirent Promise. Assured., SYNOPSYS, ZTE

計97メンバー

業界初のUltra Ethernet Consortium対応AI NIC

AMD Pensando™ Pollara 400 AI NIC

完全に プログラマブル ハードウェア パイプライン	最大1.4倍 パフォーマンスの向上	400 Gbps	オープン エコシステム スイッチの選択



インテリジェントな ロード バランシング

- アダプティブなパケット分散
- 完全にプログラマブルなトランスポート
- ロスのある/ロスレス スケールアウト イーサネット ネットワークの両方をサポート

輻輳の管理

- 輻輳の回避
- パスを考慮した輻輳制御
- 対処可能なテレメトリー

高速フェイルオーバーと 損失回復

- 選択的確認応答（SACK）
- RCCLプロキシ
- ジョブ完了の高速化

AMD Pensando™ “Vulcano”

クラスター向け次世代NIC

3nm	800G	Up to 8x	UAL | PCIe®
Process Node	Network Throughput	Scale Out Bandwidth per GPU	Host Interface

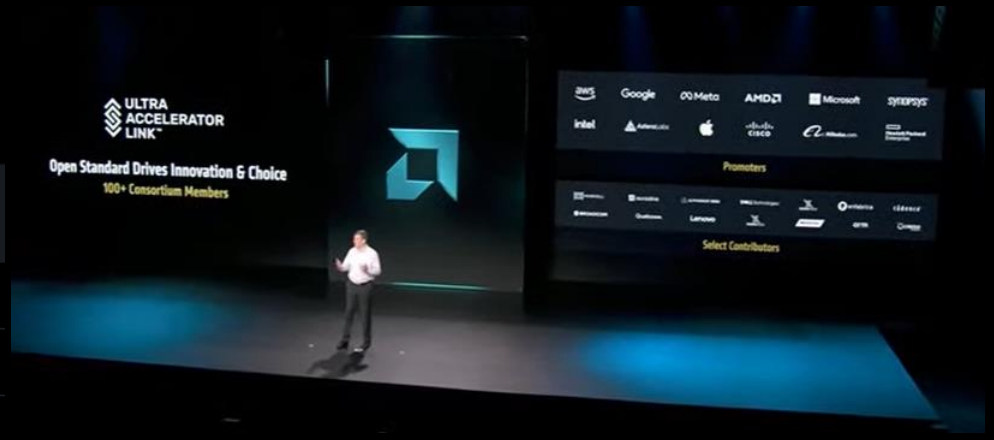
Ultra Ethernet Consortium

2026年登場予定

UAリンク(Ultra Accelerator Link)の実装

Ultra Accelerator Link, the Truly Open Standard

	UALINK	NVLINK FUSION
LOW ROUND TRIP LATENCY	Yes	Yes
HIGH SPEED I/O	224 Gbps	224 Gbps
MAXIMUM SCALABILITY	1024 GPUs	*576* GPUs
CPU USE	Any	Only to Nvidia GPU
GPU USE	Any	Only Nvidia CPU
MANAGEMENT SOFTWARE	Open	Nvidia Proprietary
SPECIFICATION	Fully Open	Closed



ULTRA ACCELERATOR LINK™

AMD BROADCOM cisco Google Hewlett Packard Enterprise intel Meta Microsoft

AMD EPYC™ “Venice”

最高性能のサーバー用CPU

Up to **256 cores**	**2.0x**	**1.7x**	**1.6** TB/s
2nm • Zen 6	CPU to GPU Bandwidth	Gen vs. Gen Performance	Memory Bandwidth

2026年登場予定

See endnote: VEN-003

AMD Instinct™ MI400

リーダーシップ生成AIアクセラレーター

40 PF | 20 PF	432 GB	19.6 TB/s	300 GB/s
FP4 • FP8 Flops	HBM4 Memory Capacity	Memory Bandwidth	Scale Out Bandwidth / GPU

2026年登場予定

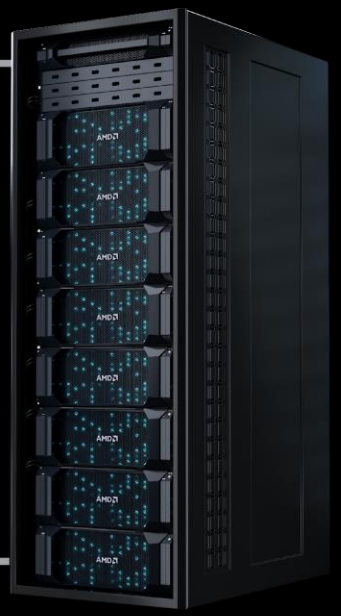
Specs based on engineering projections

AIインフラソリューションの進化

2024

AMD EPYC™
"GENOA"

AMD Instinct™
MI300 SERIES



2025

AMD EPYC™
"TURIN"

AMD Instinct™
MI350 SERIES

AMD Pensando™
POLLARA 400

2026

AMD EPYC™
"VENICE"

AMD Instinct™
MI400 SERIES

AMD Pensando™
"VULCANO"

Roadmap subject to change

Advancing AI 2025プレビュー

AMD “Helios”

最適化されたAIラックソリューション

AMD EPYC | AMD INSTINCT | AMD PENSANDO | AMD ROCm

2026登場予定

OPEN Compute Project | ULTRA ACCELERATOR LINK | Ultra Ethernet Consortium

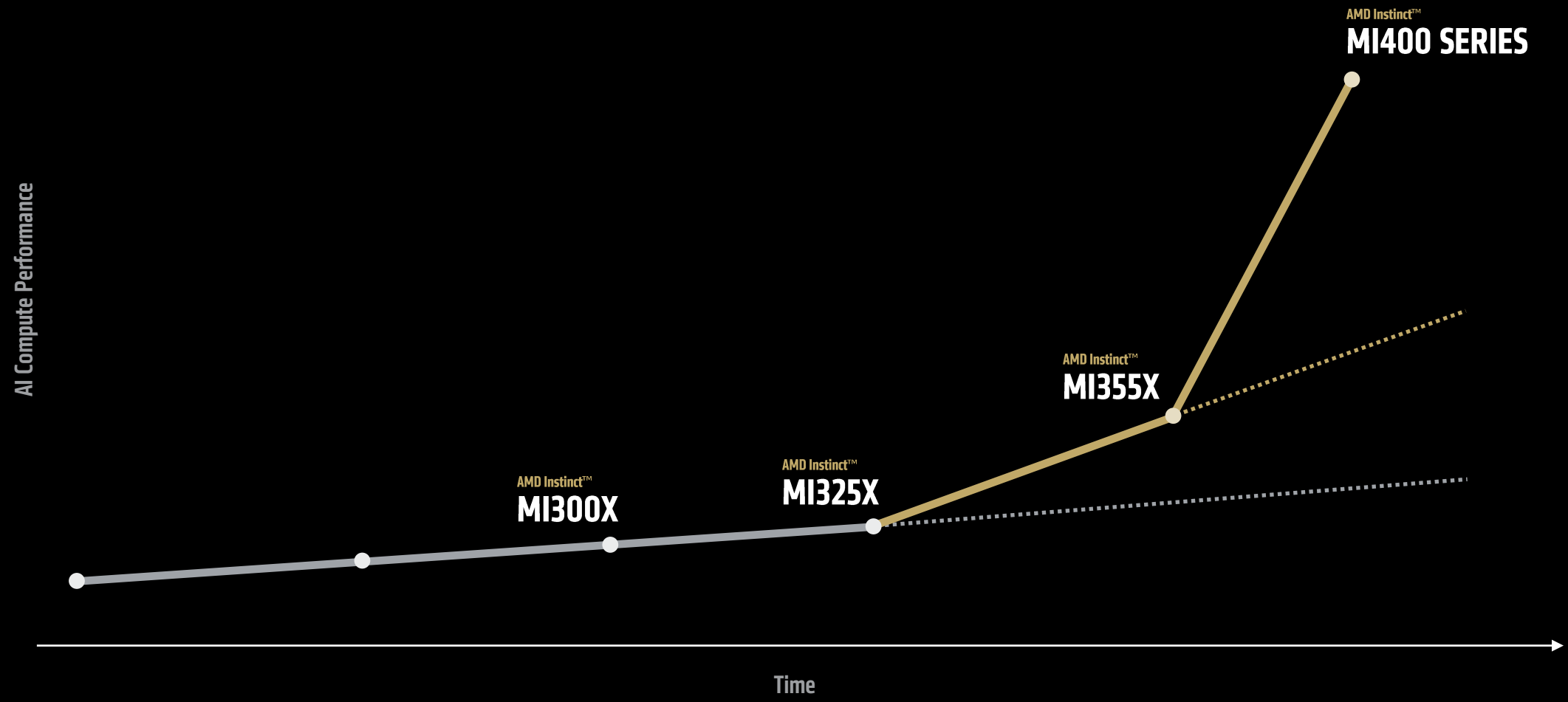
AMD “Helios” AI Rack

Rack Scale AI Performance Leadership

	AMD Instinct™ MI400 Series “Helios”	vs. Vera Rubin Oberon
GPU DOMAIN	72	1.0x
SCALE UP BANDWIDTH	260 TB/s	1.0x
FP4 • FP8 FLOPS	2.9 EF • 1.4 EF	1.0x
HBM4 MEMORY CAPACITY	31 TB	1.5x
MEMORY BANDWIDTH	1.4 PB/s	1.5x
SCALE OUT BANDWIDTH	43 TB/s	1.5x

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

AI Compute Performance



See endnote: MI350-047


AMD

Endnotes

MI350-008: Based on measurements taken by AMD Performance Labs in May 2025, of the peak theoretical precision performance of an AMD Instinct™ MI355X GPU with FP64 datatype with Matrix vs. Nvidia Grace Blackwell GB200 accelerator with FP64 datatype with Tensor; MI355X: FP32 with Matrix vs. GB200: FP32 datatype with Vector; and MI355X: FP6 datatype with Sparsity vs. GB200: FP6 datatype with Sparsity. Results may vary based on configuration, datatype. **MI350-008**

MI350-009: Based on calculations by AMD Performance Labs in May 2025, to determine the peak theoretical precision performance for the AMD Instinct™ MI350X / MI355X GPUs, when comparing FP64, FP32, TF32, FP16, FP8, FP6 and FP4, INT8, and bfloat16 datatypes with Vector, Matrix, Sparsity or Tensor with Sparsity as applicable, vs. NVIDIA Blackwell B200 accelerator. Server manufacturers may vary configurations, yielding different results.

MI350-030: Based on calculations by AMD internal testing as of 6/4/2025. Using 8 GPU AMD Instinct™ MI355X Platform for overall GPU-normalized Training Throughput (processed tokens per second) for text generation using the Llama3-70B chat model running Torchtitan (FP8) when using a maximum sequence length of 8192 tokens compared to published 64 GPU Nvidia B200 Platform performance running NeMo (FP8) when using a maximum sequence length of 8192 tokens. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-031: Based on calculations by AMD internal testing as of 6/4/2025. Using 8 GPU AMD Instinct™ MI355X Platform for overall GPU-normalized Training Throughput (processed tokens per second) for text generation using both LLaMA3-70B and LLaMA3-8B chat models running Torchtitan (BF16) or Megatron-LM (BF16) where applicable when using a maximum sequence length of 8192 tokens compared to 8 GPU Nvidia B200 Platform performance running NeMo (BF16) when using a maximum sequence length of 8192 tokens. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-032: Based on calculations by AMD internal testing as of 6/4/2025. Using 8 GPU AMD Instinct™ MI355X Platform for overall GPU-normalized Training Throughput (processed tokens per second) for text generation using both LLaMA3-70B and LLaMA3-8B chat models running Torchtitan (BF16) or Megatron-LM (BF16) where applicable when using a maximum sequence length of 8192 tokens compared to 8 GPU Nvidia B200 Platform performance running NeMo (BF16) when using a maximum sequence length of 8192 tokens. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-033: Based on calculations by AMD internal testing as of 6/5/2025. Using 8 GPU AMD Instinct™ MI355X Platform for overall GPU-normalized Training Throughput (time to complete) for fine-tuning using the Llama2-70B LoRA chat model (FP8) compared to published 8 GPU Nvidia B200 and 8 GPU Nvidia GB200 Platform performance (FP8). Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-038: Based on testing by AMD internal labs as of 6/6/2025 measuring text generated throughput for LLaMA 3.1-405B model using FP4 datatype. Test was performed using input length of 128 tokens and an output length of 2048 tokens for AMD Instinct™ MI355X 8xGPU platform compared to NVIDIA B200 HGX 8xGPU platform published results. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-039: Based on Lucid automation framework testing by AMD internal labs as of 6/6/2025 measuring text generated throughput for LLaMA 3.1-405B model using FP4 datatype. Test was performed using 4 different combinations (128/2048) of input/output lengths to achieve a mean score of tokens per second for AMD Instinct™ MI355X 4xGPU platform compared to NVIDIA DGX GB200 4xGPU platform. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI350-040: Based on testing (tokens per second) by AMD internal labs as of 6/6/2025 measuring text generated online serving throughput for DeepSeek-R1 chat model using FP4 datatype. Test was performed using input length of 3200 tokens and an output length of 800 tokens with concurrency up to 64 looks, serviceable with 30ms ITL threshold for AMD Instinct™ MI355X 8xGPU platform median total tokens compared to NVIDIA B200 HGX 8xGPU platform results. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-044:

Based on AMD internal testing as of 6/9/2025. Using 8 GPU AMD Instinct™ MI355X Platform measuring text generated online serving inference throughput for Llama 3.1-405B chat model (FP4) compared 8 GPU AMD Instinct™ MI300X Platform performance with (FP8). Test was performed using input length of 32768 tokens and an output length of 1024 tokens with concurrency set to best available throughput to achieve 60ms on each platform, 1 for MI300X (35.3ms) and 64ms for MI355X platforms (50.6ms). Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-049: Based on performance testing by AMD Labs as of 6/6/2025, measuring the text generated inference throughput on the LLaMA 3.1-405B model using the FP4 datatype with input length of 128 tokens and an output length of 2048 tokens on the AMD Instinct™ MI355X 8x GPU, and published results for the NVIDIA B200 HGX 8xGPU. Performance per dollar calculated with current pricing for NVIDIA B200 and Instinct MI355X based cloud instances. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. Current customer pricing as of June 10, 2025, and subject to change

MI300-080: Testing by AMD Performance Labs as of May 15, 2025, measuring the inference performance in tokens per second (TPS) of AMD ROCm 6.x software, vLLM 0.3.3 vs. AMD ROCm 7.0 preview version SW, vLLM 0.8.5 on a system with (8) AMD Instinct MI300X GPUs running Llama 3.1-70B (TP2), Qwen 72B (TP2), and Deepseek-R1 (FP16) models with batch sizes of 1-256 and sequence lengths of 128-204. Stated performance uplift is expressed as the average TPS over the (3) LLMs tested. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of the latest drivers and optimizations.

MI300-081: AMD Instinct MI300X platform (8x GPUs) and AMD ROCm 7.0 preview version software running Llama2-70B, Qwen1.5-14B, Llama3.1-8B, Megatron-LM using the FP16 and FP8 datatypes,shows a combined average of 3.04x or average of 304%) better training performance (TFLOPS) vs. AMD Instinct MI300X platform (8x GPUs) with ROCm 6.0 SW.