AMD AI Datacenter Solutions 2025

2025.11

AMD

together we advance_

AMD

AI Strategy



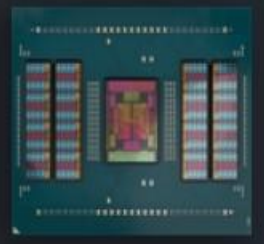
Leadership
Compute Engines

Open
Ecosystem

Full Stack
Solutions

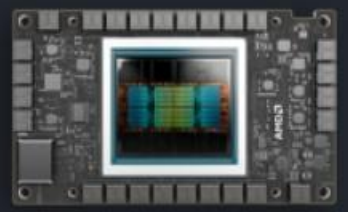
Best End-to-End AI Compute Portfolio in the Industry

AMD EPYC™ Processors



Leading server CPU

AMD Instinct™ Accelerators



World's best GPU accelerator

AMD Pensando™ Networking



Premier programmable DPUs & AI NICs

AMD Ryzen™ AI AMD Radeon™ AI Processors



Most powerful client AI processors

AMD Versal™ Adaptive SOCs



Leadership AI Processing at the edge

See endnote: SHO-000

Open Development Drives Value & Innovation

Open Hardware + **Open Software** + **Open Ecosystem**

OPEN Compute Project | ULTRA ACCELERATOR LINK™ | Ultra Ethernet | AMD ROCm | Hugging Face | PyTorch | Triton | vLLM | SGL

Choice | Flexibility | Rapid Co-Innovation | Portability | Proven

Investing in Full-Stack Solutions

Acquisitions Span Entire AI Value Chain



















Over 25 AI Acquisitions & Investments in the Last Year Alone

EPYC Momentum Accelerates...

>18x Server CPU Market Share Growth



Industry Leaders Run on EPYC™

Cloud	aws, Microsoft, Google, ORACLE
Digital	NETFLIX, Uber, Meta, zoom
Enterprise	BEST BUY, IBM, Emirates NBD, NISSAN
OEM	DELL Technologies, Hewlett Packard Enterprise, Lenovo, SUPERMICRO, CISCO

Source: Mercury

AMD ROCm

Rapidly Advancing Open Software Capabilities

Day-0 Support
for Leading Models

Accelerated
Pace of Innovation

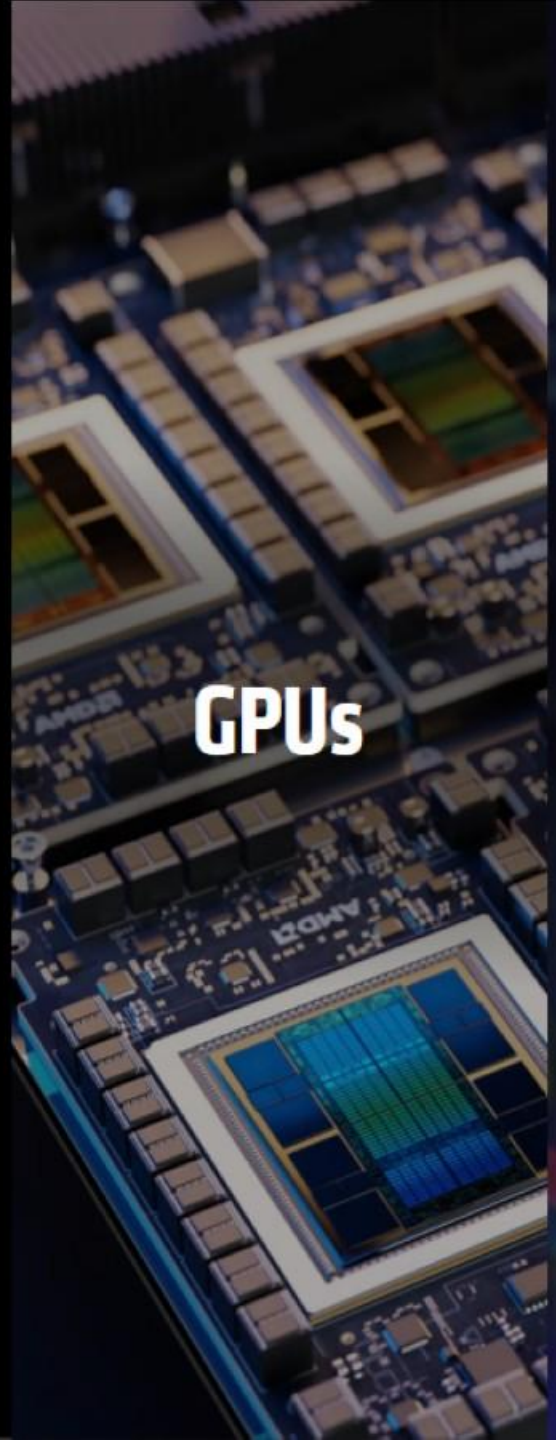
Broadening
Ecosystem Partnerships

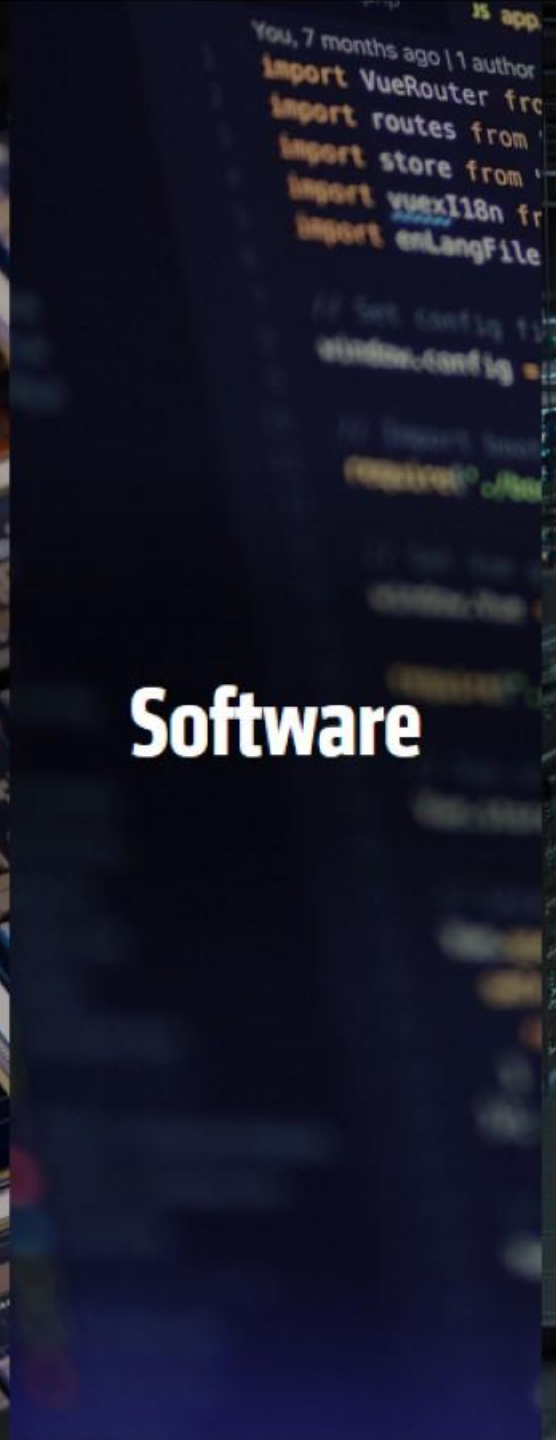
Developer First
Approach to Enablement

PyTorch · Triton · Hugging Face · TensorFlow · ONNX · vLLM · SGL

GPT · LLaMA 4 (Meta) · deepseek · Grok · Gemma 3 · Qwen · Command A · MISTRAL AI_

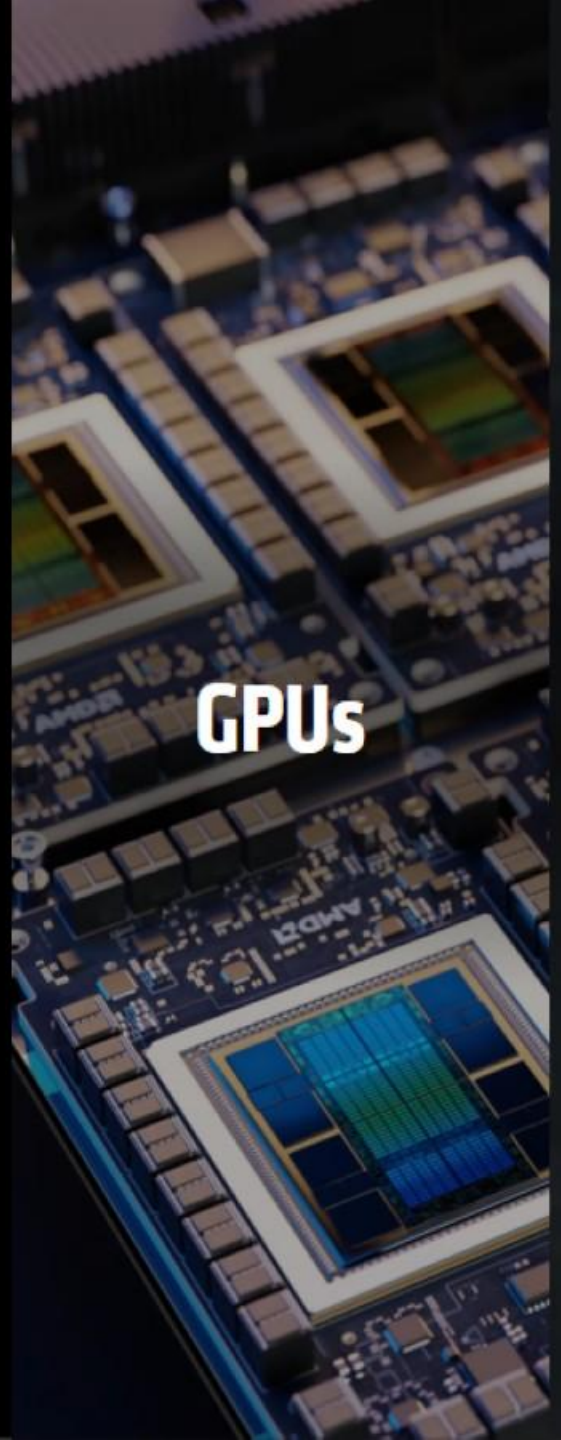
AMD AI Solutions







AMD AI Solutions



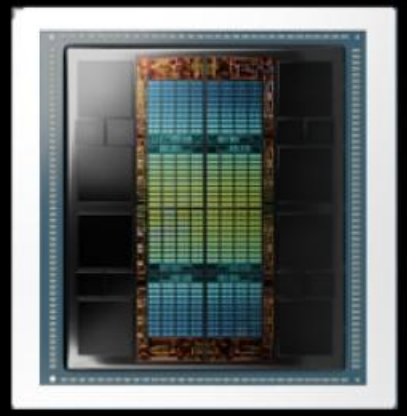
Software

Solutions

AMD INSTINCT

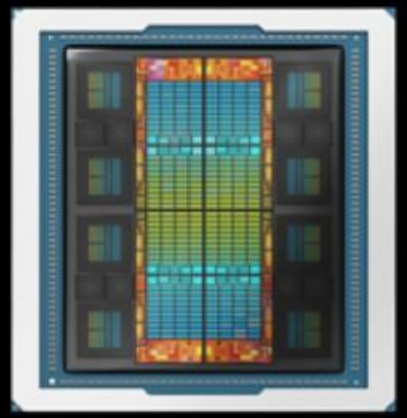
Delivering on Annual Roadmap Commitment

AMD Instinct™
MI300A/X



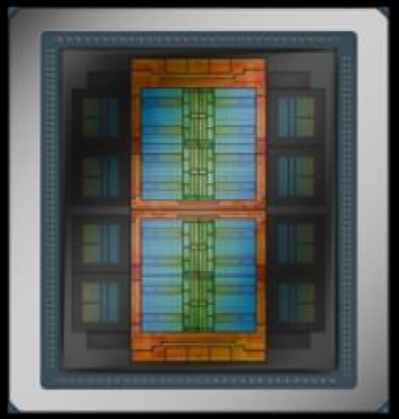
2023

AMD Instinct™
MI325X



2024

AMD Instinct™
MI350 SERIES



2025

AMD Instinct™
MI400 SERIES

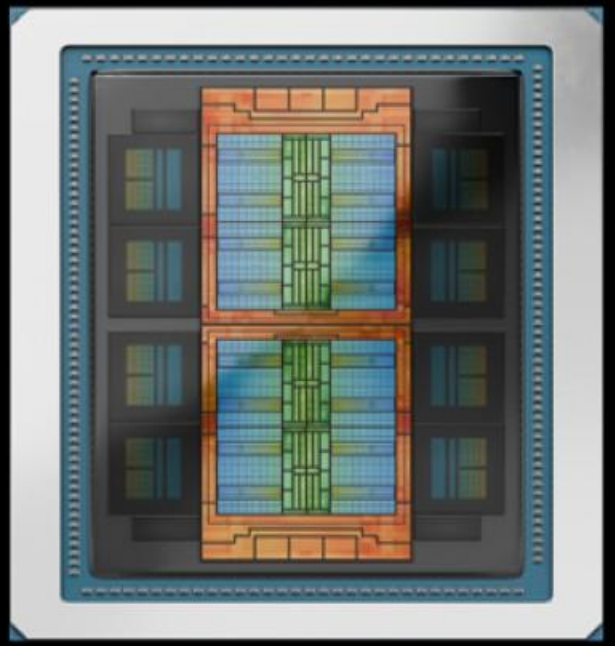


2026

Roadmap subject to change

AMD Instinct™ MI350 Series

Continued Generative AI Leadership



AMD Instinct™ **MI350 Series**

4th Gen Instinct™ architecture

3nm process node

185 billion transistors

FP4 & FP6 new gen AI datatypes

HBM3E leadership capacity

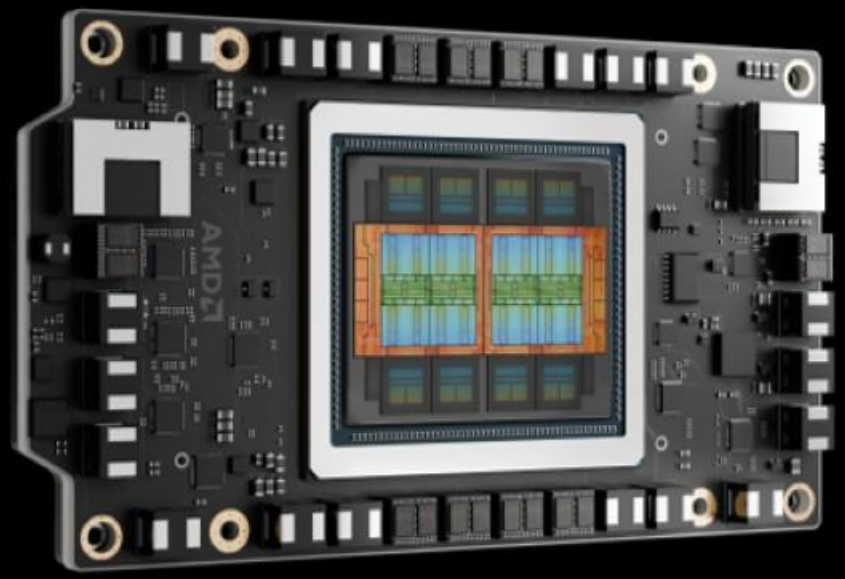
MI350 Series Accelerates Your Gen AI Outcomes

Faster AI Inference & Training	Larger AI Model Support	Rapid AI Infrastructure Deployment
20PF	**288GB**	**UBB8**
FP4 & FP6	HBM3E	Industry Standard GPU Node
4x Gen-on-Gen AI Compute Increase	Supports up to 520B Parameter AI Model	Available in Air Cooled & Direct Liquid Cooled

See endnote: MI350-004

AMD Instinct™

MI350 Series

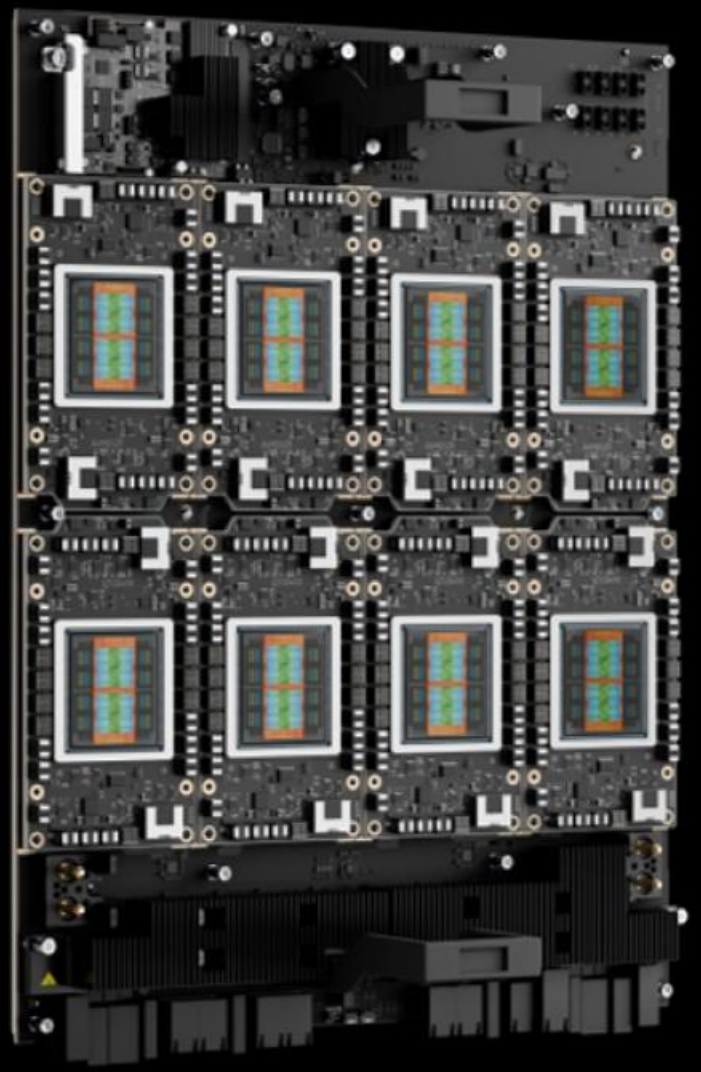


	Instinct™ MI355X
MEMORY	288 GB HBM3E
MEMORY BANDWIDTH	8 TB/s
FP64	79 TF
FP16	5 PF
FP8	10 PF
FP6	20 PF
FP4	20 PF
TBP	1400W

Instinct™ MI350 Series Advantage

	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

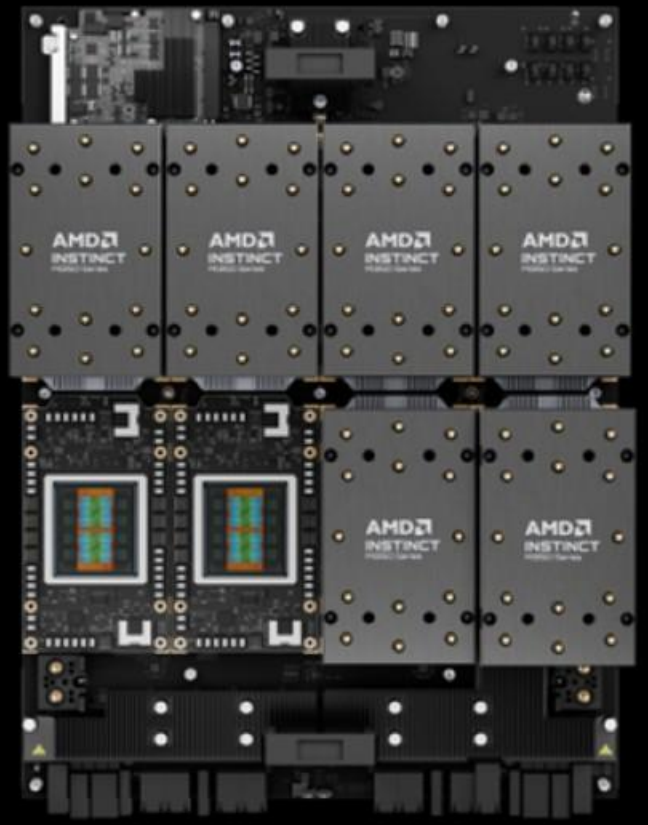
See endnote: MI350-008, 009



AMD Instinct™

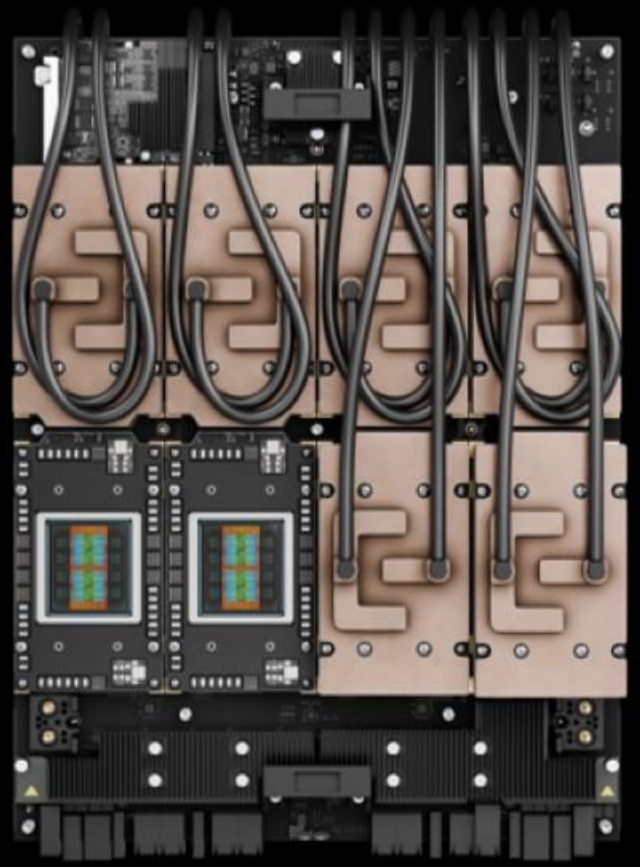
MI350 Series Platform

	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



Air Cooled

AMD Instinct™ **MI350 Series**



Liquid Cooled

AMD Instinct™ **MI350 Series**

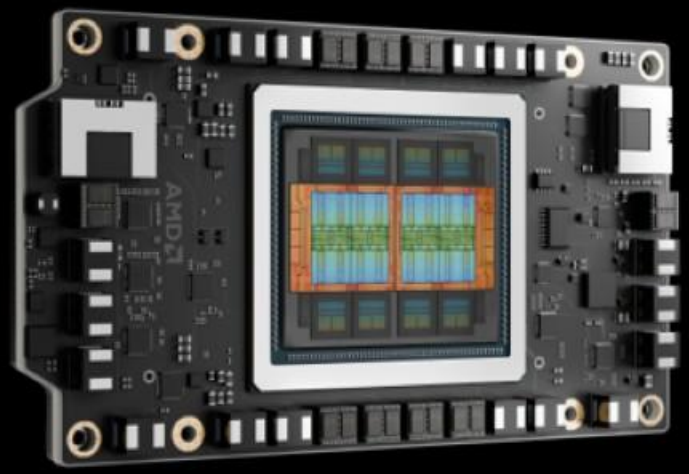
MI355X Delivers the Highest Inference Throughput

For Large Models



Inference performance, throughput

See endnote: MI350-038, 039, 040



Up to 40% More Tokens / $

Using AMD Instinct™ MI355X vs. B200

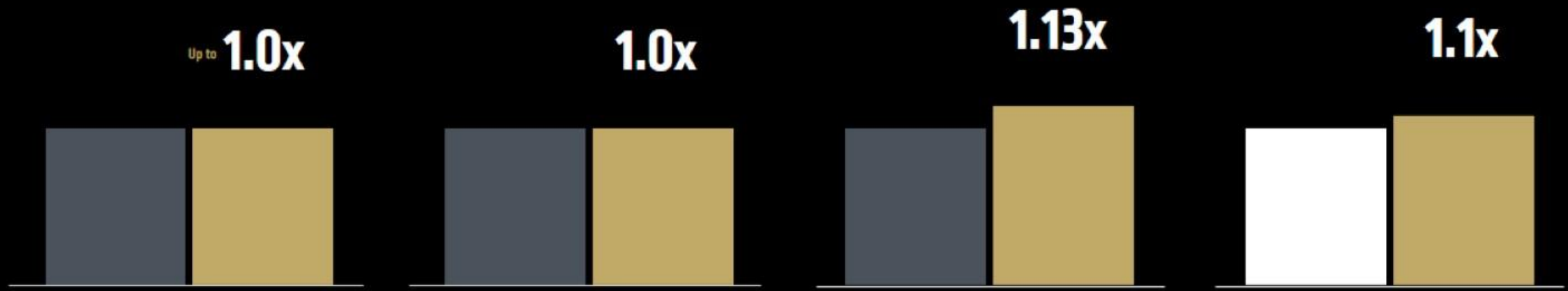
See endnote: MI350-049

In addition to Llama Inference, Meta's AI Recommendation Inference & Training models run on AMD MI300X GPUs

AMD × Meta

World Class Training & Fine-Tuning Performance

Across Models & Data Types



Pre-Training: Throughput • Fine Tuning: Time To Train • Unofficial MLPerf 5.0 MI355X vs. MLPerf 5.0 B200 and GB200

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

AMD Instinct™ MI350 Series

Proven Liquid Cooled Infrastructure

128/96	Up to 36 TB	Up to 2.6 EF	Up to 1.3 EF
GPUs	HBM3E	FP4	FP8

AMD EPYC	"Turin"
AMD INSTINCT	MI350 Series
AMD PENSANDO	Pollara 400

Specs available on 2ou/3ou server



AMD Instinct™ MI350 Series

Proven Air Cooled Infrastructure

64	18 TB	1.3 EF	0.6 EF
GPUs	HBM3E	FP4	FP8

AMD EPYC	"Turin"
AMD INSTINCT	MI350 Series
AMD PENSANDO	Pollara 400

Specs available on 6u server

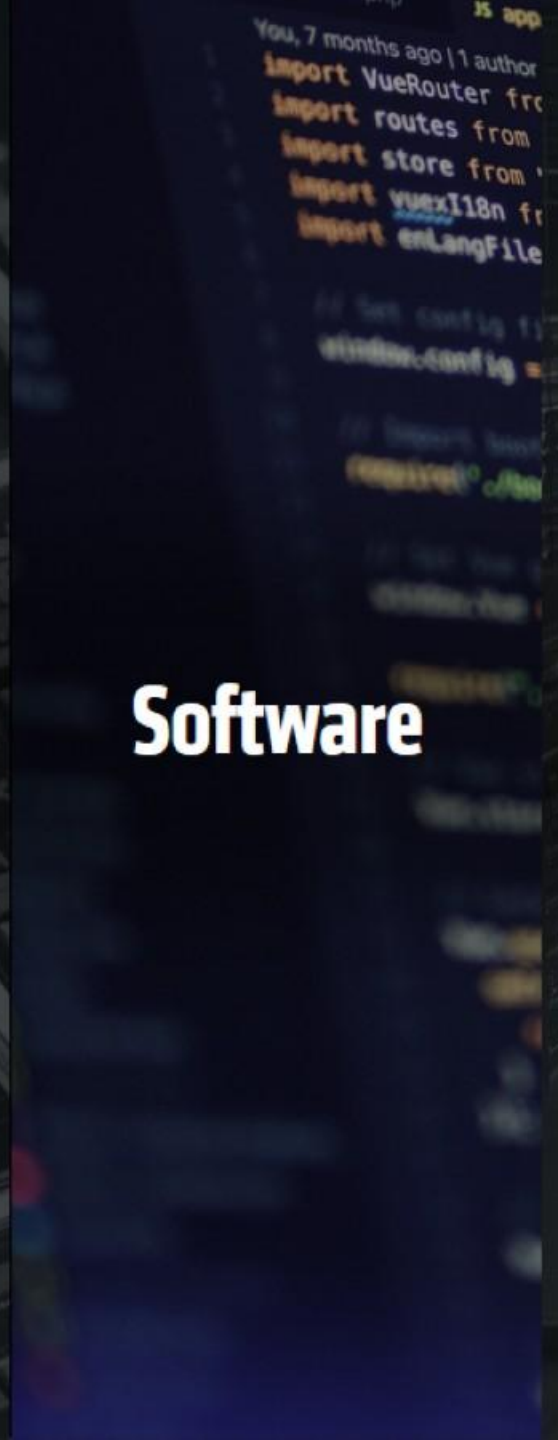


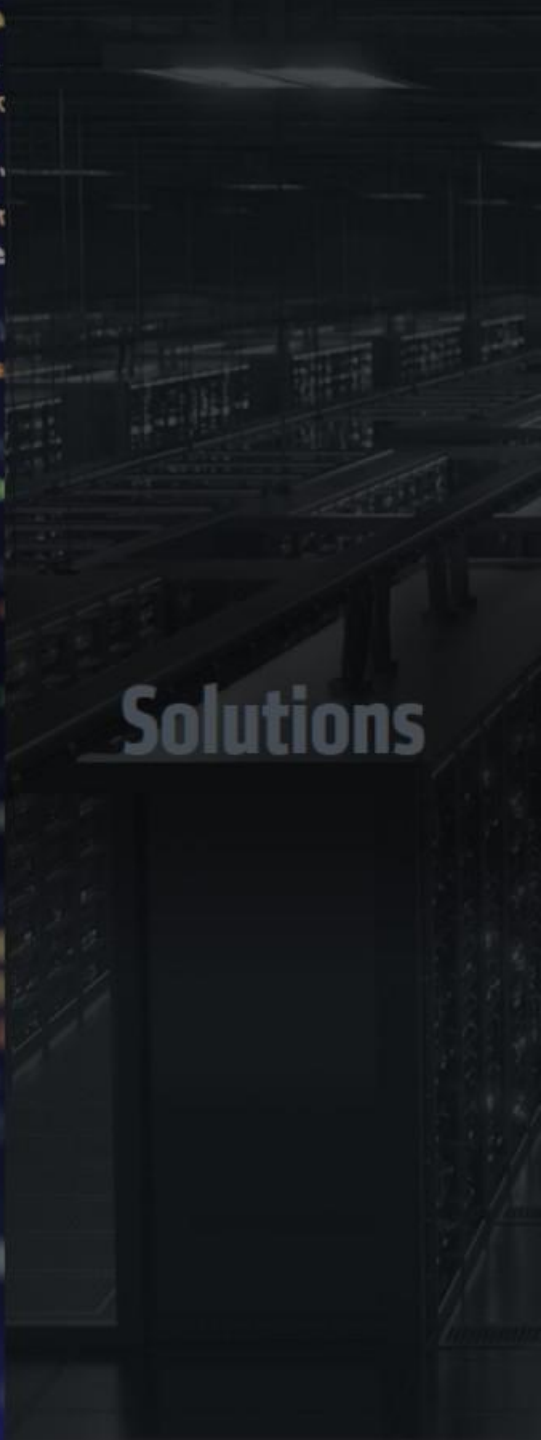
AMD Instinct™ MI350 Series

Generative AI Leadership

AMD AI Solutions







Introducing AMD ROCm™ 7

Accelerating AI Innovation & Developer Productivity

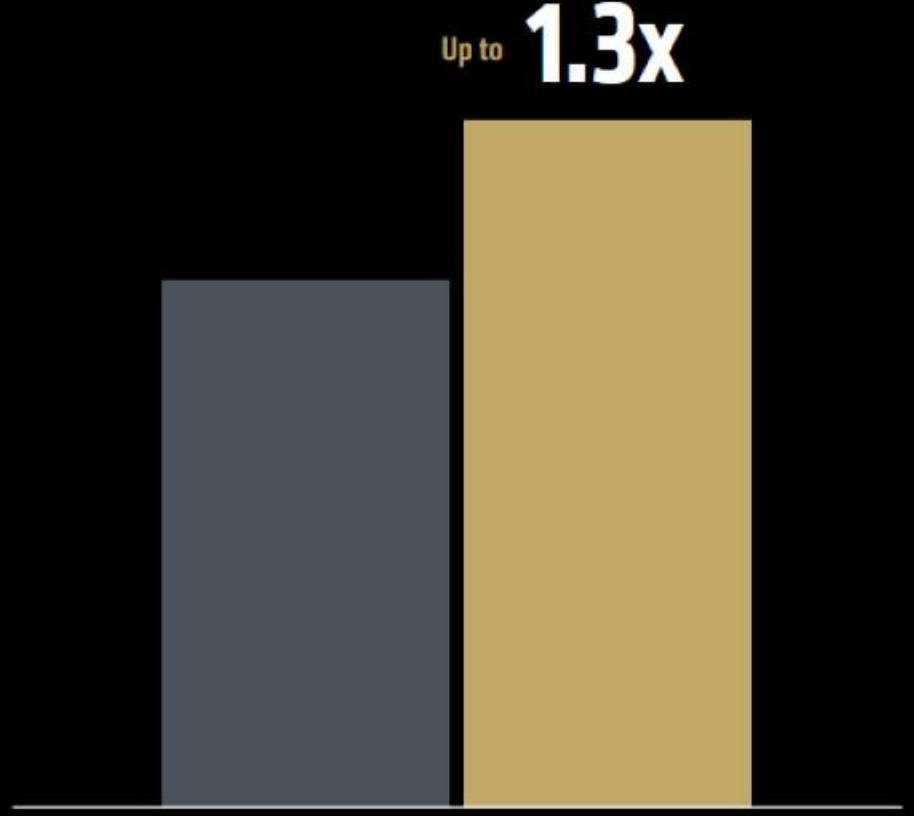
Latest Algorithms & Models	Advanced Features for Scaling AI	MI350 Series Support	Cluster Management	Enterprise Capabilities

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

Open Source: Feature Velocity & Leadership Performance



DeepSeek R1

FP8 Throughput

FP8 Model Support

vLLM ✓ | SGL ✓ | TRT-LLM ✗

Nvidia B200

AMD Instinct MI355X

See endnote: MI350-025

AMD × Microsoft

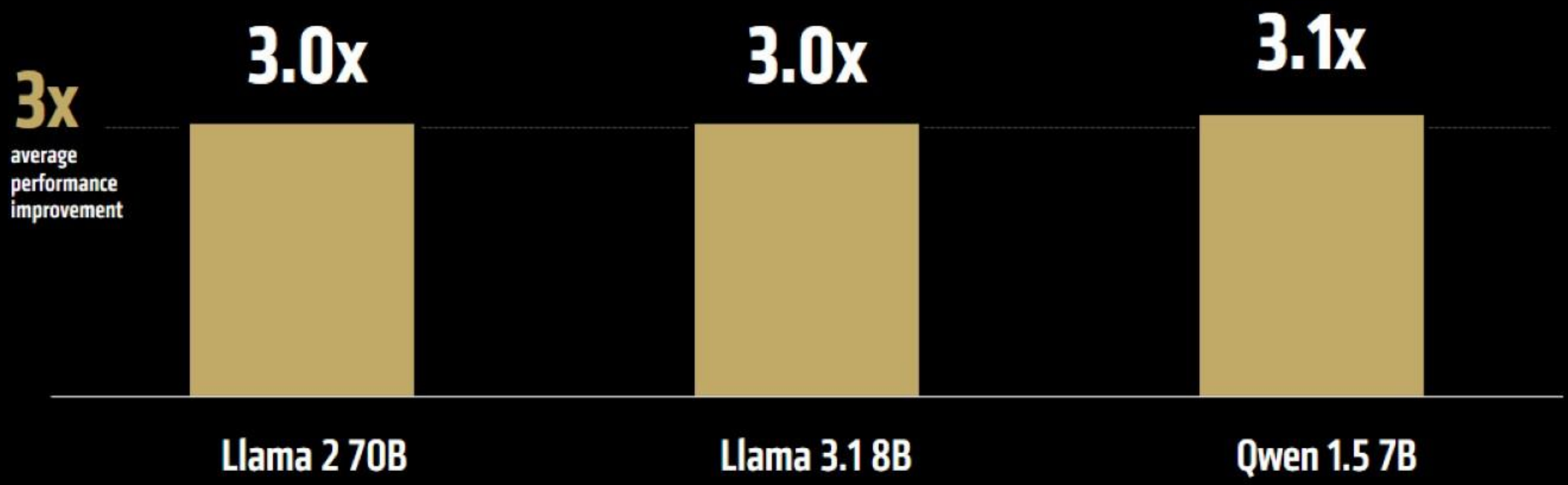
M365 Copilot App	Azure OpenAI Service deepseek	Azure AI Foundry
AI Assistants	Agentic Solutions	Multi-Modal Training

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				

Accelerating Training Performance



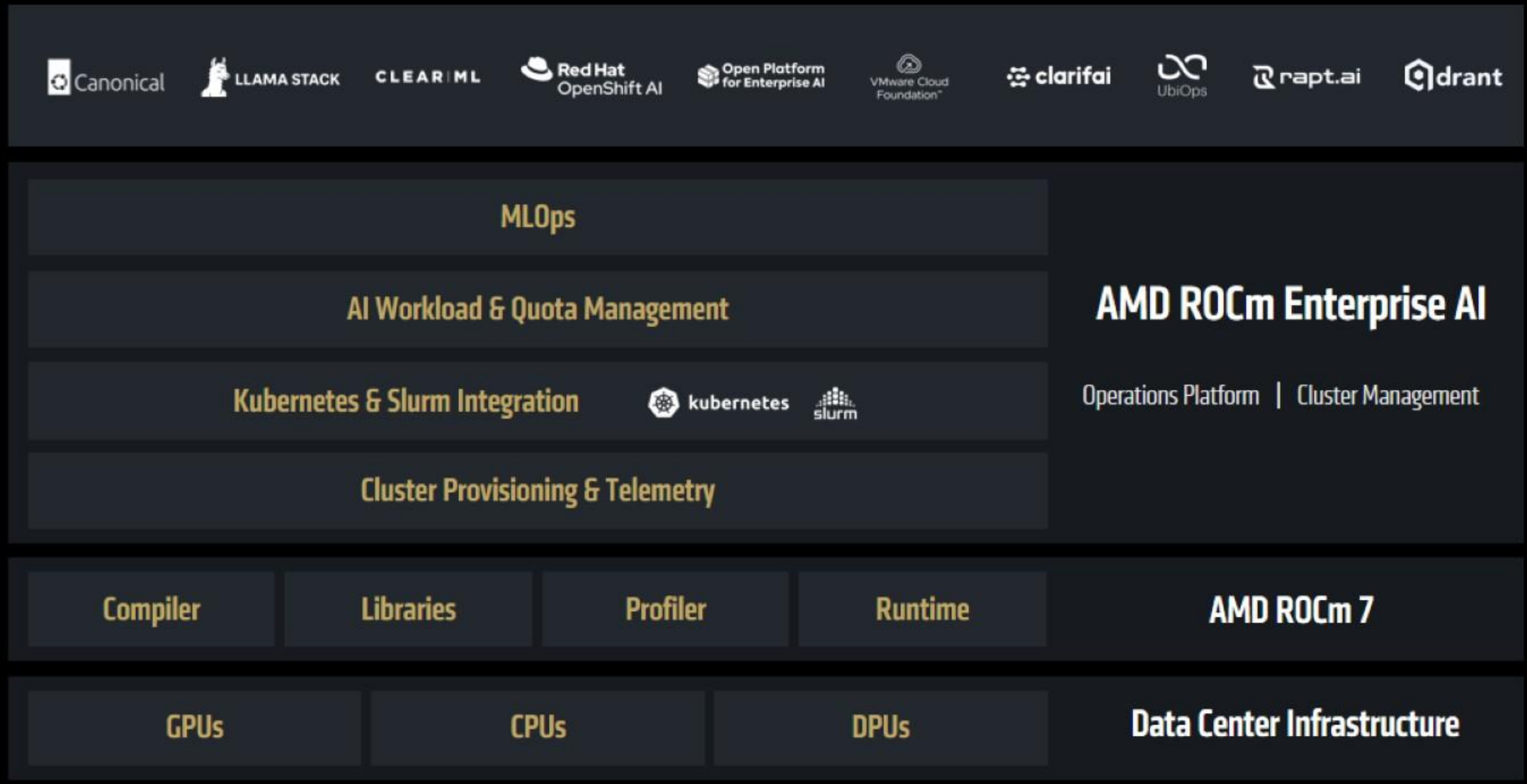
ROCm 7 vs. ROCm 6

See endnote: MI350-081

Distributed Inference at Scale with Open Ecosystem

AI Serving Throughput, Multiplied

Orchestration Framework	vLLM · llm-d · SGL
Key Functions	PD KVcache Transfer | Cross-node Communication | Cross-PD Group Schedule
Key Technologies	Mooncake · GPU Direct Access · DeepEP · Distributed Triton · SHMEM

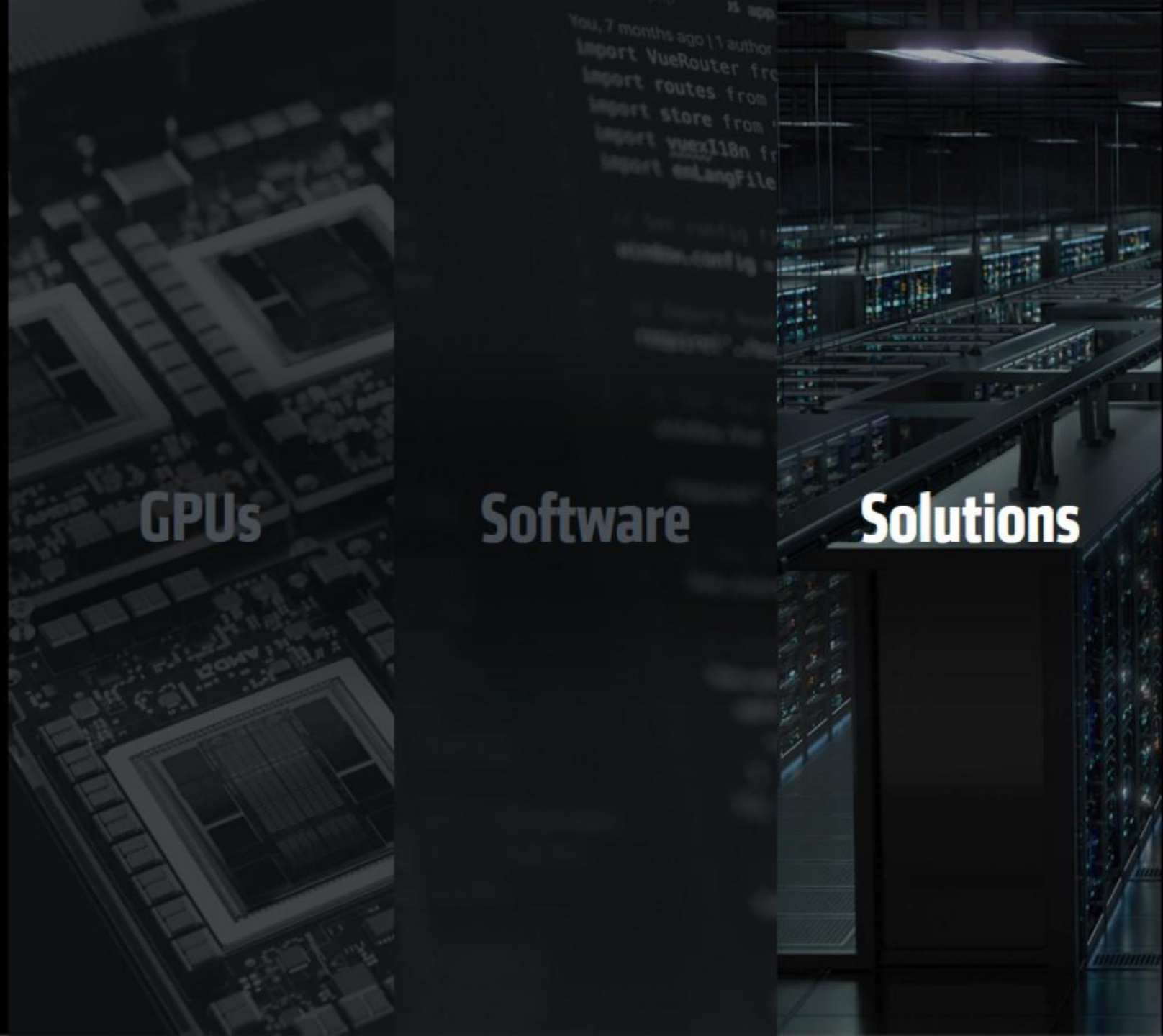

Canonical
LLAMA STACK
CLEAR|ML
Red Hat OpenShift AI
Open Platform for Enterprise AI
VMware Cloud Foundation
clarifai
UbiOps
rapt.ai
drant
MLOps
AI Workload & Quota Management
Kubernetes & Slurm Integration
kubernetes
slurm
Cluster Provisioning & Telemetry
AMD ROCm Enterprise AI
Operations Platform | Cluster Management
Compiler
Libraries
Profiler
Runtime
AMD ROCm 7
GPUs
CPUs
DPUs
Data Center Infrastructure



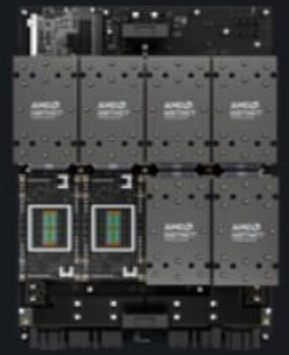
Announcing AMD Developer Cloud & Developer Credits

Available for Developers & Open-Source Contributors

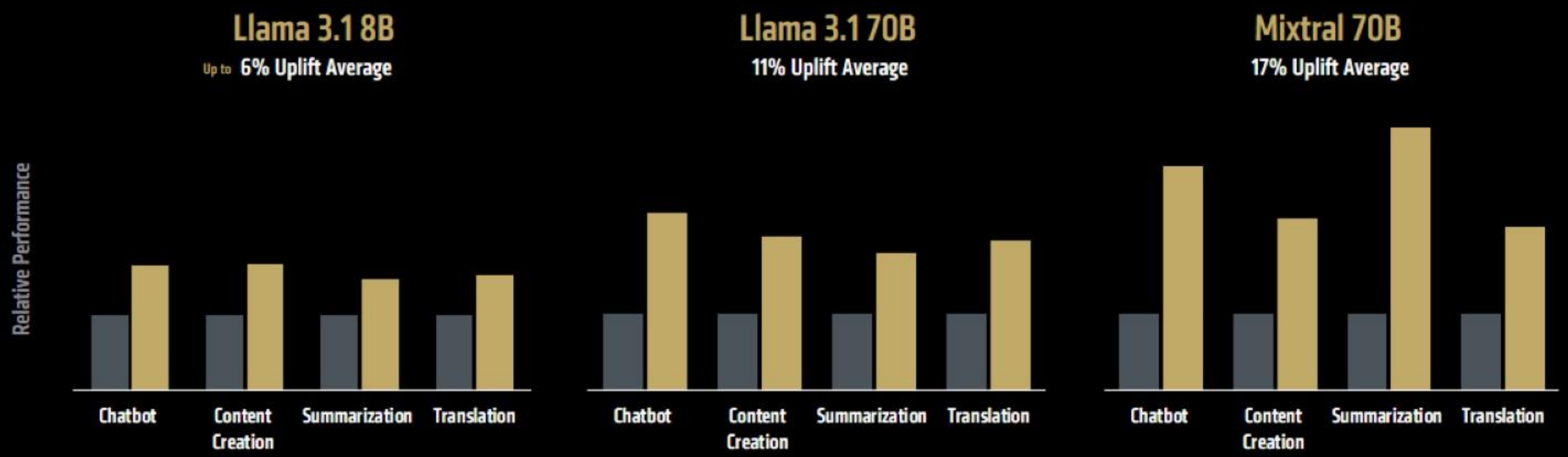
AMD AI Solutions



End-to-End Integrated AI Platform

Front End Network	CPU Node	GPU Node	Scale Up Network	Scale Out Network
			ULTRA ACCELERATOR LINK™	 Ultra Ethernet
Securely Integrate into Enterprise	X86 Applications & AI Execution	AI Model Training & Inference	Connectivity for Training & Distributed Inference	Enable Gigawatt Level Scaling

AMD EPYC™ Driving End-to-End System Performance



AMD Instinct™ MI300X | Intel™ Xeon™ 8592+

AMD Instinct™ MI300X | AMD EPYC™ 9575F

https://www.amd.com/content/dam/amd/en/documents/epyc-business-docs/white-papers/maximize-ai-gpu-efficiency-with-amd-epyc-processors.pdf

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



Any CPU

Any Accelerator

Any Switch

Ultra Ethernet Consortium

Open Architecture for Performant AI Scale-Out

Modern RDMA transport with automatic configuration & tuning

Optimized protocol for high fabric utilization & congestion control

Unparallel scalability to over a million GPUs

World Class Open Platform

Highest memory bandwidth & capacity drives performance

Industry standard x86 for easy integration to enterprise applications

Ready to deploy rack infrastructure

AMD Instinct™ GPU

MI350 SERIES

AMD Server CPU

EPYC™ "TURIN"

AMD Pensando™ NIC

POLLARA 400



96 / 128 GPUs
Up to 1.8x more than GB200 NVL72

36 TB HBM3E
Up to 2.8x more than GB200 NVL72

2.6 EF FP4 **1.3 EF** FP8
Up to 1.6x more than GB200 NVL72

2.6 EF FP6
Up to 3.3x more than GB200 NVL72

x86 CPU
5th Gen EPYC "Turin"

Ultra Ethernet Consortium Scale-Out NIC
AMD Pollara NIC

OCP Compliant
Industry standard design

Ultra Ethernet

OPEN Compute Project

Source: AMD specs based on engineering projections
Nvidia specs based publicly available information

Advancing AI Infrastructure Solutions

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

Previewing Today at **Advancing AI 2025**

AMD “Helios”

Optimized AI Rack Solution

AMD EPYC | AMD INSTINCT | AMD PENSANDO | AMD ROCm

Available in 2026

OPEN Compute Project | ULTRA ACCELERATOR LINK | Ultra Ethernet



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

AMD EPYC™ "Venice"

Highest Performance Server 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

Coming in 2026

See endnote: VEN-003

AMD Instinct™ MI400

Leadership Gen AI Accelerator

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

Coming in 2026

Specs based on engineering projections

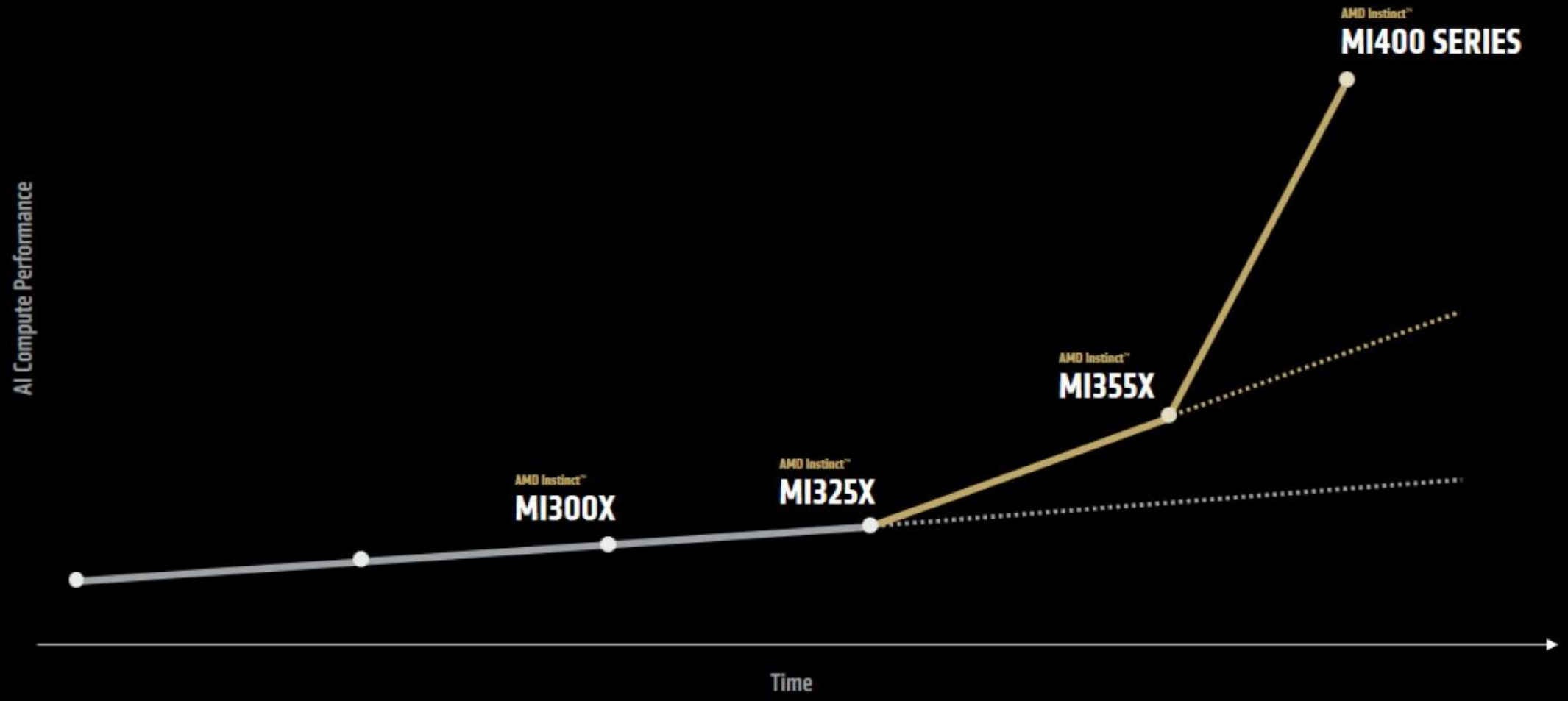
AMD Pensando™ “Vulcano”

Next Gen NIC for AI Clusters

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

Ultra Ethernet Consortium | Coming in 2026

AI Compute Performance



See endnote: MI350-047

Up to 10x More Performance with AMD Instinct™ MI400 Series

Compared to Instinct™ MI355X

See endnote: MI400-001

AMD | Advancing AI Infrastructure on an Annual Cadence

2025

AMD EPYC
"TURIN"

AMD Instinct
MI350 SERIES

AMD Pensando
POLLARA 400



2026

AMD EPYC
"VENICE"

AMD Instinct
MI400 SERIES

AMD Pensando
"VULCANO"



"Helios"

2027

AMD EPYC
"VERANO"

AMD Instinct
MI500 SERIES

AMD Pensando
"VULCANO"

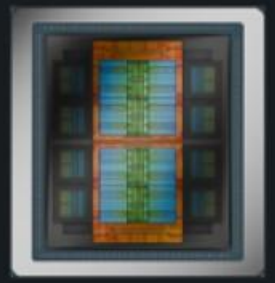


Next Gen AI Rack

Roadmap subject to change

Announced Today

Accelerating Momentum Across Our Broad AI Capabilities

	AMD ROCm	AMD Developer Cloud		
AMD Instinct™ MI350 Series	AMD ROCm 7 Optimized for Reasoning & Agents	Empowering Developers	AMD "Helios" AI Rack	AMD Threadripper™ and Radeon™ AI solutions
Available Now	Available August	Available Now	Coming in 2026	Available Now



AI Innovation is a Global, Collective Effort

Endnotes

SHO-06: Testing as of Dec 2024 using the following benchmark scores compared to Intel Core Ultra 9 288V and Qualcomm Snapdragon X Elite X1E-84-100. Cinebench 2024 nT, 3Dmark Wildlife Extreme, and Blender.. Next gen AI PC defined as a Windows PC with a processor that includes a NPU with at least 40 TOPS. Configuration for AMD Ryzen™ AI Max+ 395 processor: AMD reference board, Radeon™ 8060S graphics, 32GB RAM, 1TB SSD, VBS=ON, Windows 11. Configuration for Qualcomm Snapdragon X Elite X1E-84-100 processor: Samsung Galaxybook, Adreno Graphics, 16GB RAM, Microsoft Windows 11. Configuration for Intel Core Ultra 9 288V: ASUS Zenbook X 14, Intel Arc Graphics, 32GB RAM, 1TBSSD, Microsoft Windows 11 Home. Laptop manufacturers may vary configurations yielding different results.

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.

MI350-004: Based on calculations by AMD Performance Labs in May 2025, to determine the peak theoretical precision performance of eight (8) AMD Instinct™ MI355X and MI350X GPUs (Platform) and eight (8) AMD Instinct MI325X, MI300X, MI250X and MI100 GPUs (Platform) using the FP16, FP8, FP6 and FP4 datatypes with Matrix. Server manufacturers may vary configurations, yielding different results. Results may vary based on use of the latest drivers and optimizations.

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.

MI350-025: Testing by AMD Performance Labs as of May 25, 2025, measuring the inference performance in tokens per second (TPS) of the AMD Instinct MI355X platform with ROCm 7.0 pre-release build 16047, running DeepSeek R1 LLM on SGLang versus NVIDIA Blackwell B200 platform with CUDA version 12.8. Server manufacturers may vary configurations, yielding different results. Performance may vary based on hardware configuration, software version, and the use of the latest drivers and optimizations.

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.

Endnotes

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.

MI350-034: Based on AMD internal testing as of 6/4/2025, using an (8) GPU AMD Instinct™ MI355X Platform for overall GPU-normalized Training Throughput (processed tokens per second) for text generation using the LLaMA3-70B and LLaMA3-8B chat models running Torchtitan or Megatron-LM (FP8 and BF16) as applicable, using a maximum sequence length of 8192 tokens, compared to an (8) GPU AMD Instinct™ MI300X Platform using Megatron-LM (FP8 and BF16). Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI350-035: Based on 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 8 GPU AMD Instinct™ MI300X Platform performance with (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.

MI350-041: Based on AMD internal testing as of 6/5/2025. Using 8 GPU AMD Instinct™ MI355X Platform measuring text generated offline inference throughput for Llama4 Maverick chat model (FP4) compared 8 GPU AMD Instinct™ MI300X Platform performance with (FP8). MI355X ran 8xTP1 (8 copies of model on 1 GPU) compared to MI300X running 2xTP4 (2 copies of model on 4 GPUs). Tests were conducted using a synthetic dataset with different combinations of 128 and 2048 input tokens, and 128 and 2048 output tokens. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

Endnotes

MI350-042: Based on AMD internal testing as of 6/5/2025. Using 8 GPU AMD Instinct™ MI355X Platform measuring text generated offline inference throughput for Llama 3.1-405B chat model (FP4) compared 8 GPU AMD Instinct™ MI300X Platform performance with (FP8). MI355X ran 8xTP1 (8 copies of model on 1 GPU) compared to MI300X running 2xTP4 (2 copies of model on 4 GPUs). Tests were conducted using a synthetic dataset with different combinations of 128 and 2048 input tokens, and 128 and 2048 output tokens. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-043: Based on AMD internal testing as of 6/5/2025. Using 8 GPU AMD Instinct™ MI355X Platform measuring text generated online serving inference throughput for DeepSeek-R1 chat model (FP4) compared 8 GPU AMD Instinct™ MI300X Platform performance with (FP8). Test was performed using input length of 3200 tokens and an output length of 800 tokens with concurrency set to maximize the throughput on each platform, 128 for MI300X and 2048 for MI355X platforms. 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.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

MI350-047: Based on engineering projections by AMD Performance Labs in June 2025, to estimate the peak theoretical precision performance of seventy-two (72) AMD Instinct™ MI400X GPUs (Rack) vs. an 8xGPU AMD Instinct MI355X platform using the FP6 Matrix datatype. Results subject to change when products are released in market.

MI350-048: Based on AMD internal testing as of 6/9/2025. Using 8 GPU AMD Instinct™ MI355X Platform measuring text generated offline inference throughput for Llama 3.3-70B chat model (FP4) compared 8 GPU AMD Instinct™ MI300X Platform performance with (FP8). MI355X ran 8xTP1 (8 copies of model, one per GPU) compared to MI300X running 8xTP1 (8 copies of model, one per GPU). Tests were conducted using a synthetic dataset with different combinations of 128 and 2048 input tokens, and 128 output tokens. 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

MI400-001: Performance projection as of 06/05/2025 using engineering estimates based on the design of a future AMD Instinct MI400 Series GPU compared to the Instinct MI355x, with 2K and 16K prefill with TP8, EP8 and projected inference performance, and using a GenAI training model evaluated with GEMM and Attention algorithms for the Instinct MI400 Series .Results may vary when products are released in market.

Endnotes

MI350-54: Based on calculations by AMD internal testing as of 6/10/2025. 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 H200 Platform performance (FP8). Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-055: Based on engineering projections by AMD Performance Labs in June 2025, to estimate the peak theoretical precision performance of seventy-two (72) AMD Instinct™ MI400X GPUs (Rack) using the FP4 Matrix datatype vs. an 8xGPU AMD Instinct MI00 platform using the FP16 Matrix datatype. Results subject to change when products are released in market.

MI350-056: Based on calculations by AMD Performance Labs in June 2025, to determine the peak theoretical precision performance of 8x GPU AMD Instinct MI355X platform with FP6 Matix datatype vs. an 8x GPU AMD Instinct MI325X/MI300X platforms with FP8 Matrix datatype. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of the latest drivers and optimizations.

MI350-057: *Based on calculations by AMD Performance Labs in June 2025, to determine the peak theoretical precision performance of 8x GPU AMD Instinct MI325X/MI300X platform with FP8 Matix datatype vs. an 8x GPU AMD Instinct MI250X platform with FP16 Matrix datatype. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of the latest drivers and optimizations.

MI350-058: *Based on calculations by AMD Performance Labs in June 2025, to determine the peak theoretical precision performance of 8x GPU AMD Instinct MI325X/MI300X platform with FP8 Matix datatype vs. an 8x GPU AMD Instinct MI250X platform with FP16 Matrix datatype. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of the latest drivers and optimizations.

VEN-003: PCIe Gen comparison based on PCI-SIG published statements, https://pcisig.com/pci-express-6.0-specification. 2P 6th Gen EPYC CPU with 128 lanes of PCIe Gen 6 and 5th Gen EPYC with 128 lanes of PCIe Gen 5 as of 6/3/2025. PCIe is a registered trademark of PCI-SIG Corporation