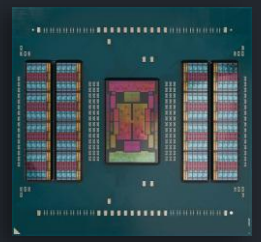
From Datacenter to Client: Deploying LLMs with AMD AI Software, ROCm and NPUs

AMD

together we advance_

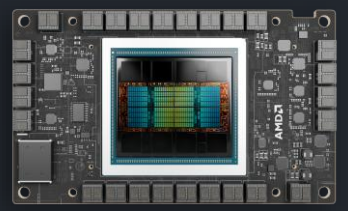
Best End-to-End AI Compute Portfolio in the Industry

AMD EPYC™
Processors



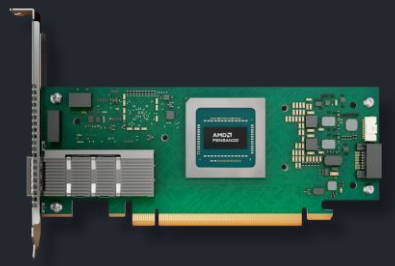
Leading server CPU

AMD Instinct™
Accelerators



World's best GPU accelerator

AMD Pensando™
Networking

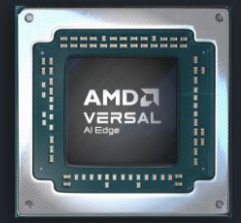


Premier programmable DPUs & AI NICs

AMD Ryzen™ AI
AMD Radeon™ AI
Processors

Most powerful client AI processors

AMD Versal™
Adaptive SOCs



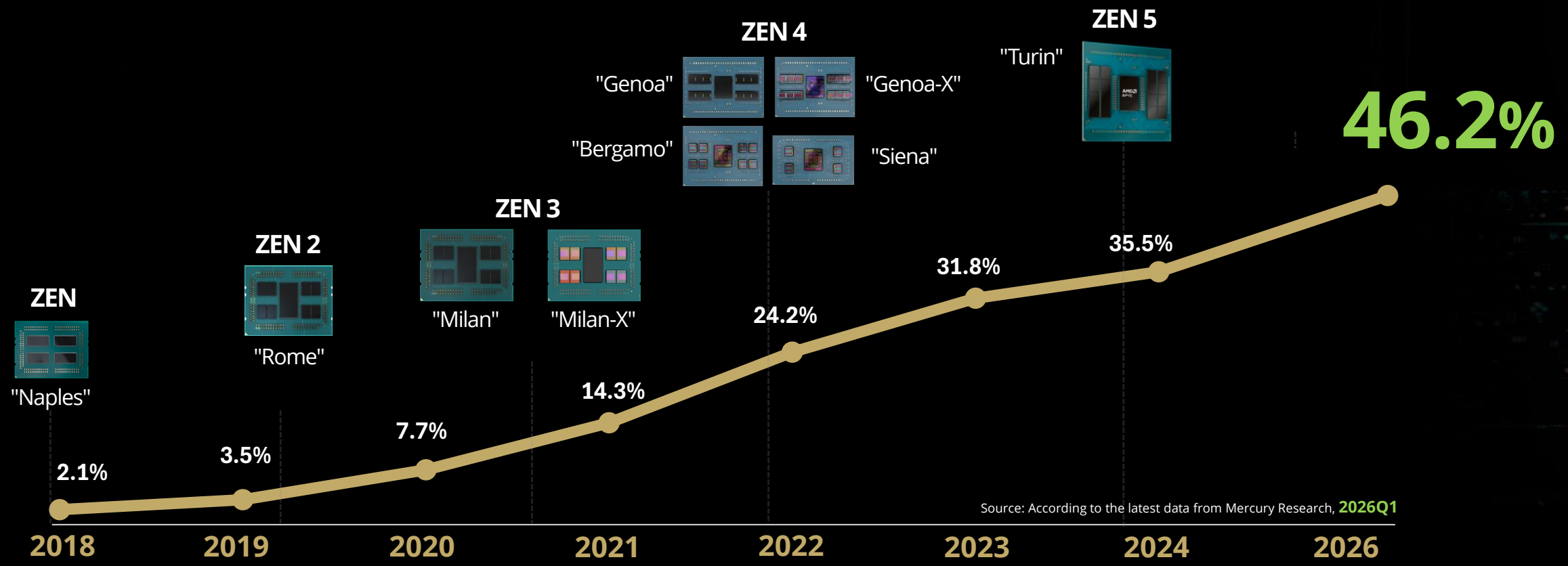
Leadership AI Processing at the edge

See endnote: SHO-000

AMD Zen 5 "Turin" 最新動態 - 解鎖新世代運算效能

Consistent Leadership: Performance and Efficiency

Fueling AMD EPYC™ Growth to Over 40% Share



Source: According to the latest data from Mercury Research, **2026Q1**

11.3x Performance
Across 5 Generations

4.1x Performance / CPU W
Across 5 Generations

AMD
together we advance_

HPE Portfolio

COMPUTE

HPE ProLiant DL325 Gen12
5th Gen AMD EPYC CPUs

HPE ProLiant DL345 Gen12
5th Gen AMD EPYC CPUs

HPE ProLiant DL325 Gen11
4th Gen and 5th Gen AMD EPYC CPUs

HPE ProLiant DL345 Gen11
4th Gen and 5th Gen AMD EPYC CPUs

HPE ProLiant DL365 Gen11
4th Gen and 5th Gen AMD EPYC CPUs

HPE ProLiant DL385 Gen11
4th Gen and 5th Gen AMD EPYC CPUs

HPE ProLiant DL145 Gen11
4th Gen AMD EPYC CPUs

DX385 Gen11
4th Gen AMD EPYC CPUs

DX365 Gen11
4th Gen AMD EPYC CPUs

SUPERCOMPUTING

HPE Cray XD2000
4th Gen AMD EPYC CPUs & MI210

HPE Cray Supercomputing EX4000 and EX2500

EX4252 4th Gen AMD EPYC CPUs

EX255a MI300a

EX235n 3rd Gen AMD EPYC CPUs

EX235a 3rd Gen AMD EPYC CPUs & MI250X

HPE Cray XD665
4th Gen AMD EPYC CPUs

HPE ProLiant Compute XD685
5th Gen AMD EPYC CPUs
8x Instinct MI325X or MI300X

STORAGE

HPE Cray Supercomput g Storage Systems E2000
4th Gen AMD EPYC CPUs

SimpliVity 325 Gen11
4th Gen AMD EPYC CPUs

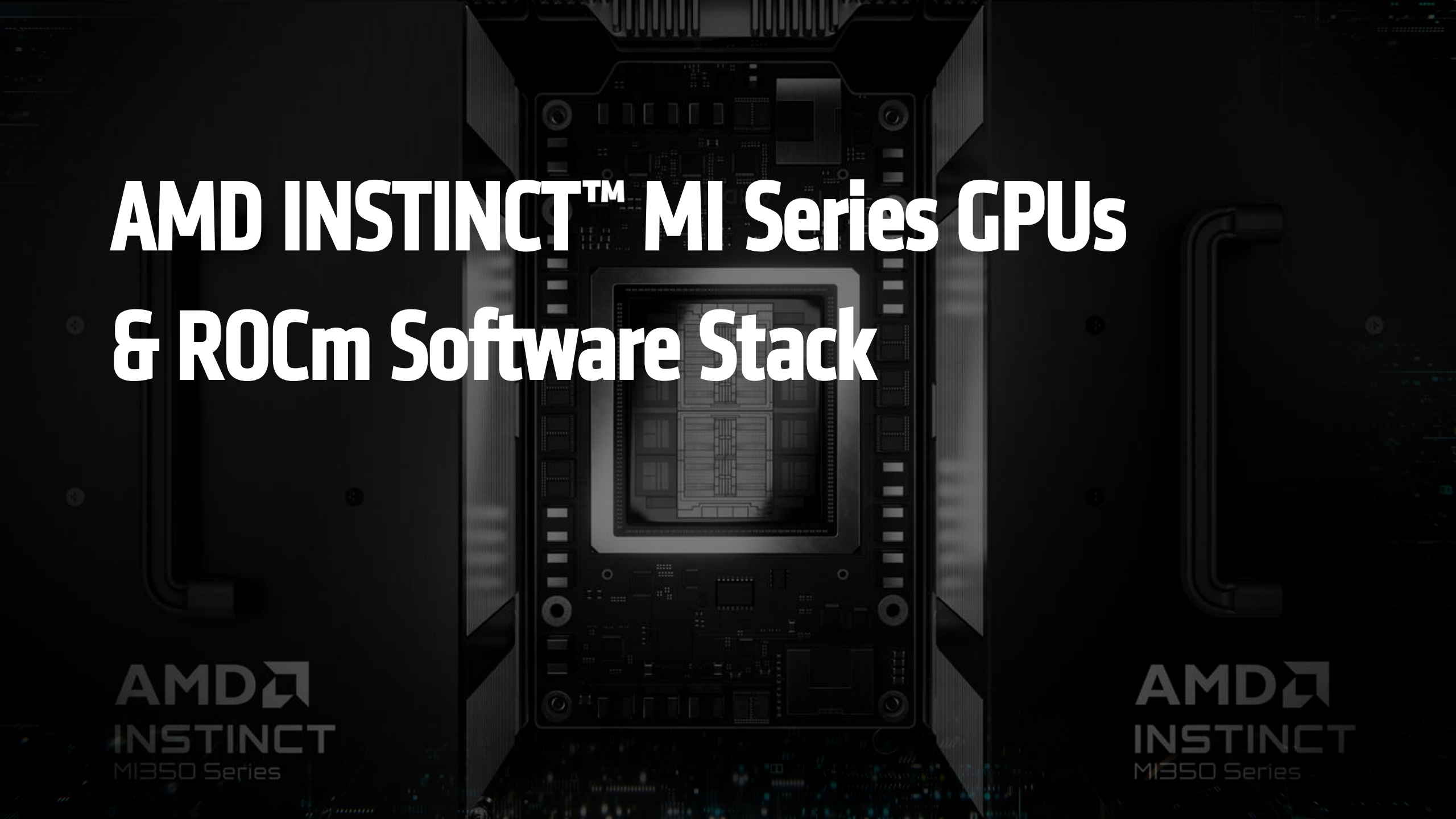
HPE Alletra MP
4th Gen AMD EPYC CPUs

HPE aruba networking

HPE Aruba CX10000
powered by AMD Pensando

AMD
together we advance_

AMD INSTINCT™ MI Series GPUs & ROCm Software Stack



Comprehensive and Open Infrastructure Stack

A decade of innovation, partnerships and collaboration

AMD ROCm

Frameworks
Libraries
Engines
Models

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

Cluster Operations Ecosystem — mlflow

Vertical SDKs - Data & Life Sciences, HPC, Finance

ROCm CORE — AMD ROCm

AMD AI Infrastructure — AMD EPYC | INSTINCT

Core Math & Compute	rocThrust	hipCUB	hipSPARSE	hipSparseLt
	hipFFT	hipBLAS	hipBLASLt	hipRAND
Distributed Compute	RCCL	rocSHMEM		
Media I/O	rocDecode	rocJPEG		
Programming & Runtime	HIP	OpenMP	LLVM	
Profiling & Debugging	rocGDB	rocm-compute	rocm-systems	omnistat
Control & Monitoring	rocm-info	amd-smi		

Pre-optimized, Fully Packaged Dockers

AMD 透過 Container, 預先整合並調校好最佳化的軟體環境, 讓最終用戶能以最少設定成本快速部署各式 LLM 模型於 AMD GPU 上, 大幅簡化開發與上線流程

Inference

- ✅ vLLM Docker
  - User Guide
- ✅ SGLang Docker
  - User guide

Training

- ✅ PyTorch Docker
  - User Guide
- ✅ Megatron-LM Docker
  - User guide
- ✅ JAX MaxText Docker
  - User guide

Additional details available at: https://www.amd.com/en/developer/resources/rocm-hub/dev-ai.html

1. Docker Image

docker pull rocm/vllm:rocm7.0.0_vllm_0.10.2_20251006

- ROCm 7.0.0
- vLLM 0.10.2 (0.11.0rc2.dev160+g790d22168.rocm700)
- PyTorch 2.9.0a0+git1c57644
- hipBLASLt 1.0.0

2. Supported Models

The following models are supported for inference performance benchmarking with vLLM and ROCm. Some instructions, commands, and recommendations in this documentation might vary by model – select one to get started. MXFP4 models are only supported on MI355X and MI350X GPUs.

Model	Meta Llama	DeepSeek	OpenAI GPT OSS
	Mistral AI	Qwen	Microsoft Phi
Variant	Llama 2 70B	Llama 3.1 8B	Llama 3.1 8B FP8
	Llama 3.1 405B	Llama 3.1 405B FP8	Llama 3.1 405B MXFP4
	Llama 3.3 70B	Llama 3.3 70B FP8	Llama 3.3 70B MXFP4
	Llama 4 Scout 17Bx16E	Llama 4 Maverick 17Bx128E	Llama 4 Maverick 17Bx128E FP8

3. Benchmarking Script

Name	Options	Description
`$test_option`	latency	Measure decoding token latency
	throughput	Measure token generation throughput
	all	Measure both throughput and latency
`$num_gpu`	1 or 8	Number of GPUs
`$datatype`	`float16` or `float8`	Data type

技術新知前瞻 - AMD 最新技術論壇與部落格精選

AMD Instinct™ 加速器
擴展對熱門生成式 AI 模型的即時支援

- ✅ 百度文心大模型
- ✅ Qwen 3.6
- ✅ Gemma 4
- ✅ Kimi-K2.5
- ✅ OpenAI
- ✅ MiMo-V2.5-Pro

Additional details available at: AMD Technical Articles & Blogs

OpenAI 最新推出的開放模型 **gpt-oss-120b** 與 **gpt-oss-20b**, 以靈活性與實際應用為核心設計.
我們很高興能在第一時間, 將這些強大能力帶到 AMD 平台上.

- 🔗 在 AMD Instinct™ GPU 上進行推論與微調
- 🔗 在 Radeon™ AI PRO R9700 GPU 上進行推論
- 🔗 在 AMD Ryzen™ AI MAX+ 與 AMD Ryzen™ AI 300 系列處理器上進行推論

How to Run Inference with vLLM using AMD GPUs

AMD has enabled vLLM to support OpenAI gpt-oss-120b and gpt-oss-20b models on AMD GPUs on Day 0, including:

- AMD Instinct MI355X (gfx950),
- AMD Instinct MI300X and MI325X (gfx942)
- AMD Radeon AI PRO R9700 (gfx1201).

Step 1: Select the vLLM Docker Image Based on Your GPU

- Pick the vLLM Docker image according to your choice of GPU (Instinct MI355X, MI325X, MI300X, or Radeon AI PRO R9700).
- Follow these steps to get started.

.....

Step 2: Launch the ROCm vLLM Docker Container

Start a container with the necessary ROCm software, device, network privileges and AMD GPU specific containers:

.....

Step 3: Authenticate with Hugging Face

.....

Step 4: Launch the vLLM Server

.....

AMD Instinct™ MI系列加速器
對熱門生成式 AI 模型的即時支援

GPT-4 GPT-4o Llama 3 Llama 2 Yi

Mistral Mixtral Grok Phi Stable Diffusion

Command R+ DBRX Qwen StarCoder

Zamba Flux Gemma 3 Falcon LLM Kimi-K2

BLOOM MPM4 GPT-NeoX OLMo Aria MPT

Extended support for leading models

1.8M+ models supported out of the box

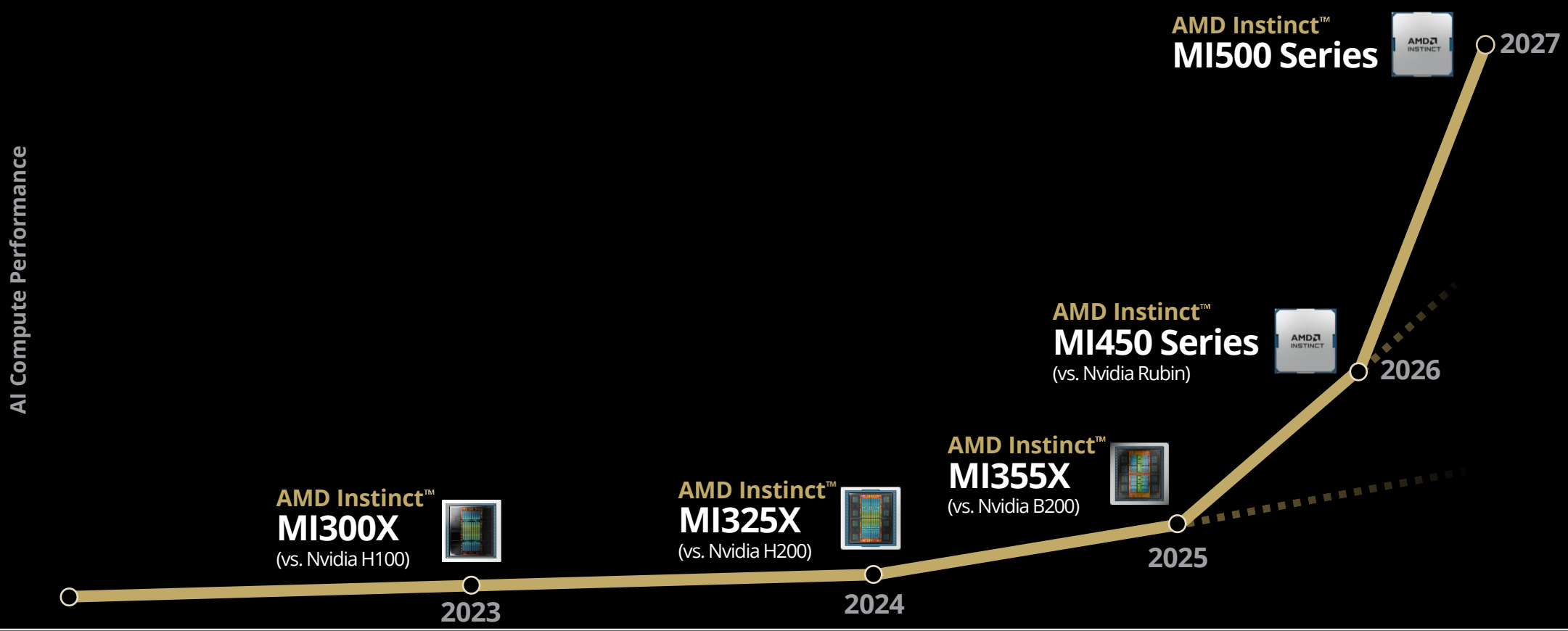
Hugging Face

Day 0 support for AMD GPUs

LLaMA 4 Meta | deepseek | GPT-OSS 120B & 20B

AMD MI Instinct GPUs Overview

Next Big Leap in AI Performance With MI500 Series

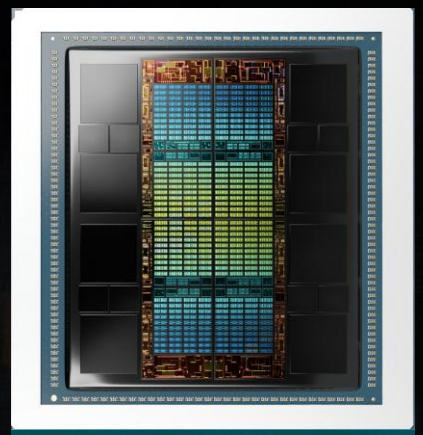


Based on AMD internal projections as of Nov. 2025.

AMD INSTINCT

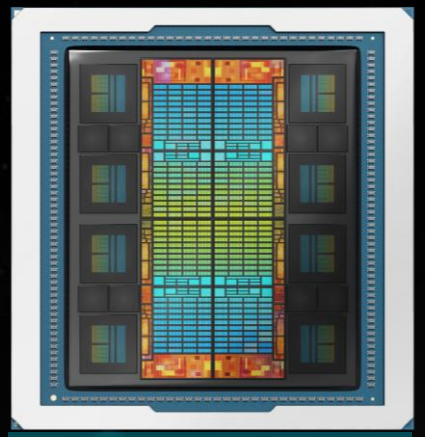
Leadership Roadmap on Annual Cadence

2023
MI300A/X



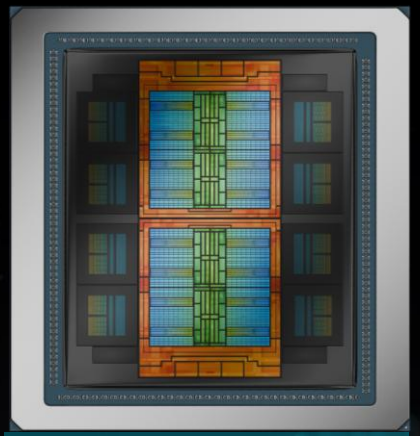
2.4X more HBM
1.3X More FP8
vs. H100

2024
MI325X



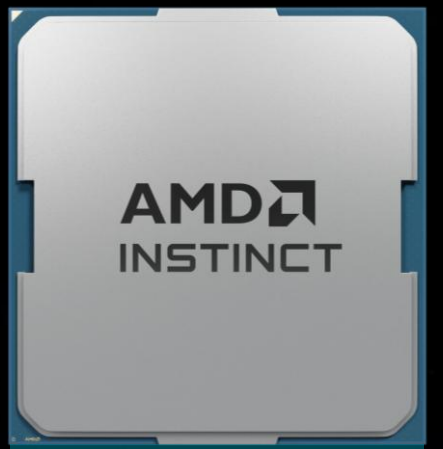
1.8X more HBM
1.3X More FP8
vs. H200

2025
MI350 SERIES



1.6X more HBM
1X FP4 | FP8
vs. B200

2026
MI400 SERIES

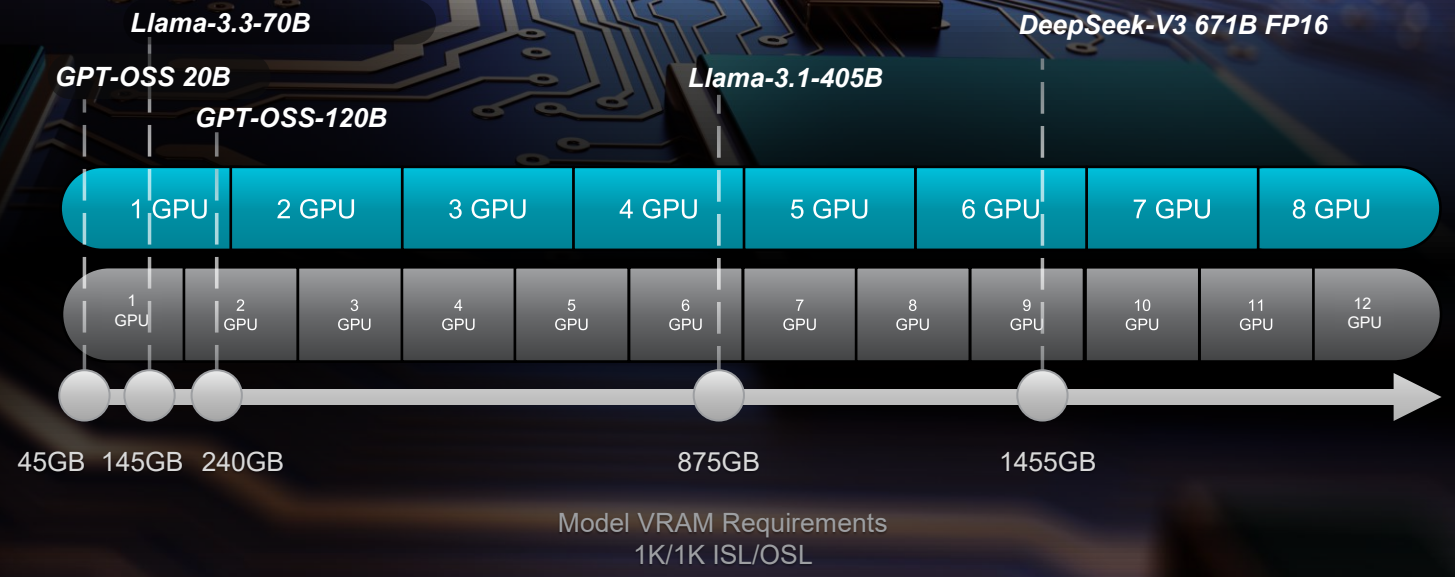


1.5X more HBM
1X FP4 | FP8
vs. Vera Rubin

Roadmap subject to change

AMD Performs:
More Capacity

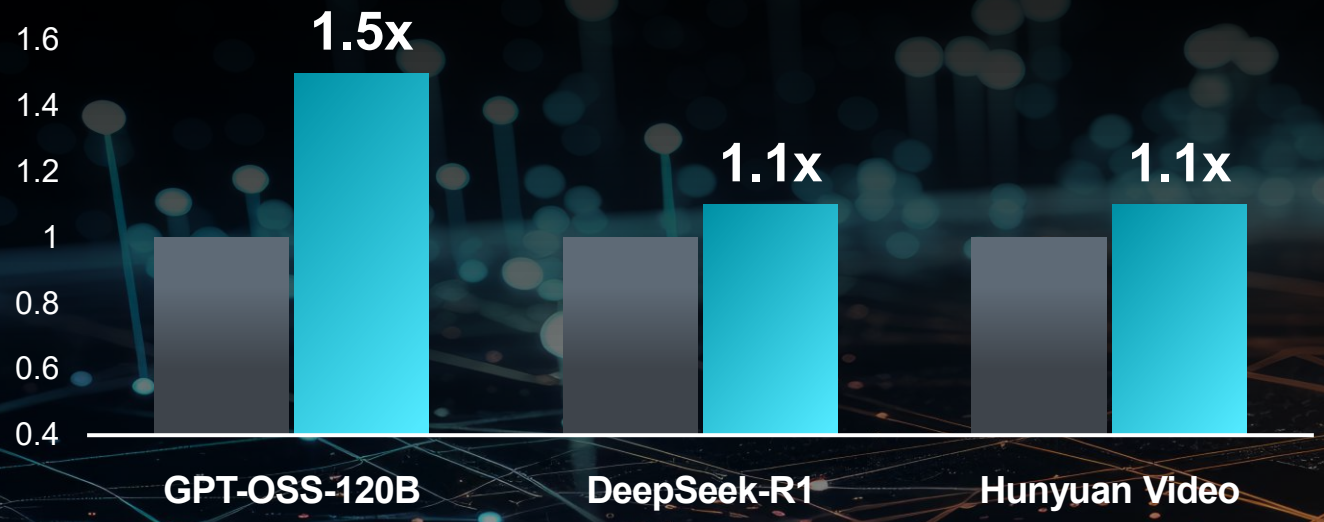
Applications benefit by having more HBM per MI355X GPU



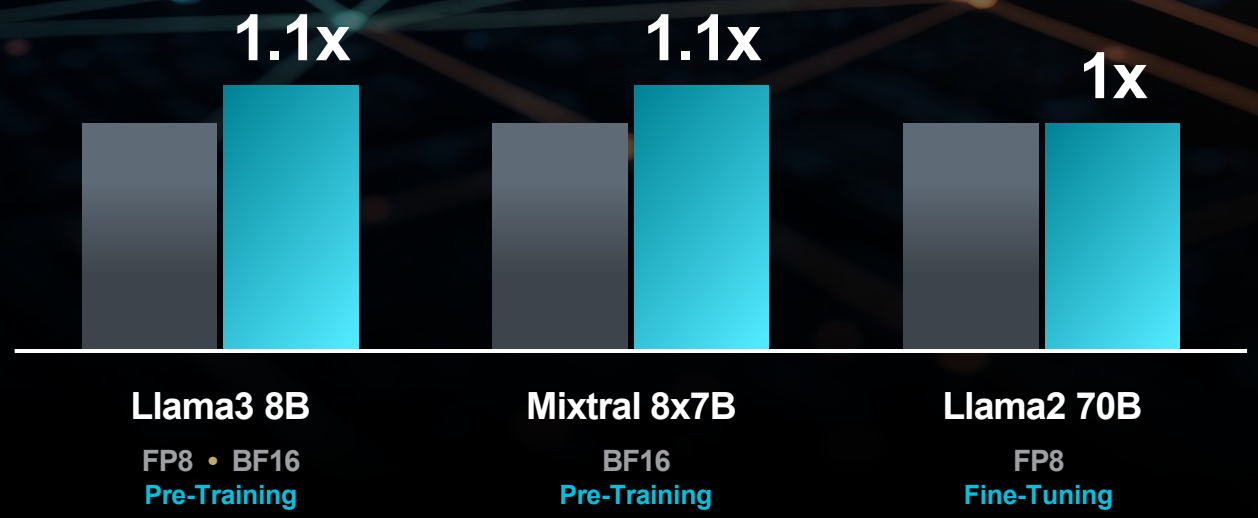
Nvidia	AMD Instinct™
B200	**MI355X**
192GB HBM3e	**288GB HBM3e**

AMD
together we advance_AI

Inference Performance



Training Performance



B200 MI355X

AMD Performs: Competitive Throughput

Production grade performance, without software vendor lock-in

Last Updated: Feb 2026

AMD
together we advance_AI

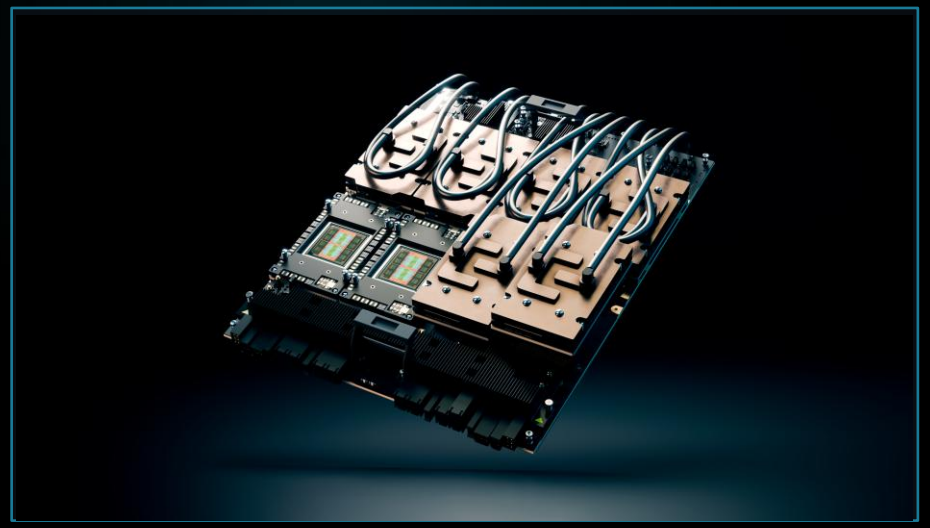
AMD MLPerf Inference 6.0 Partner Ecosystem

9 Partners Tie for the Most AMD Instinct™ GPU Submissions

TIE FOR MOST PARTNER SUBMISSIONS ON AMD INSTINCT HARDWARE

9

CISCO	DELL	GIGA COMPUTING
HPE	MANGOBOOST	MITAC COMPUTING
ORACLE	SUPERMICRO	RED HAT



Partner submissions across AMD Instinct MI300X, MI325X, MI350X and MI355X GPUs show how predictable AMD hardware and ROCm™ software reproduce performance at scale.

4 INSTINCT GPUs ACROSS PARTNER SUBMISSIONS

MI300X | MI325X
MI350X | MI355X

AMD INSTINCT MI355X GPU RESULTS REPRODUCED ACROSS PARTNERS

Average of all AMD Instinct MI355X GPU partner results landed within **4%** of AMD. Some landed within **1%**, even on first-time workloads

*AMD Instinct MI355X GPU reproducibility claims reflect partner comparisons versus AMD Instinct MI355X GPU submissions provided by the user.

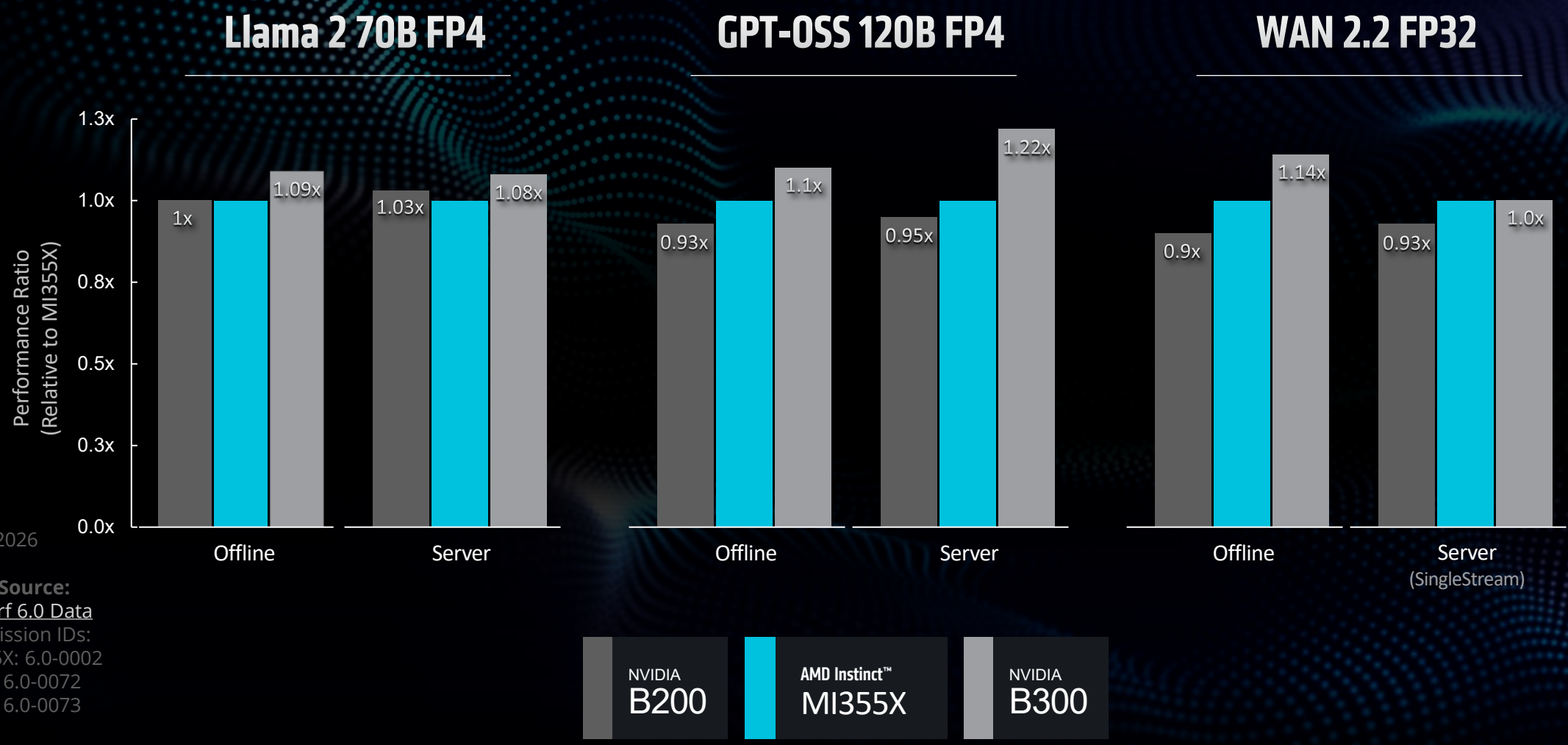
*4% average based on ALL MI355X partner submissions
Dell MI355X submission ID: 6.0-0020
GigaComputing MI355X submission ID: 6.0-0037

HPE MI355X submission ID: 6.0-0049
Cisco MI355X submission ID: 6.0-0086
Supermicro MI355X submission ID: 6.0-0097

NEW
APR '26

MI355X vs B200/B300 Inference Performance

Using Publicly Available Data from MLPerf 6.0 Submissions



Date:
April 2026

Data Source:
MLPerf 6.0 Data
Submission IDs:
MI355X: 6.0-0002
B200: 6.0-0072
B300: 6.0-0073

AMD
together we advance_

AMD MLPerf Inference 6.0 Results

AMD Instinct MI355X Platform

Llama 2-70B | Multi-Node | AMD Closed Submission

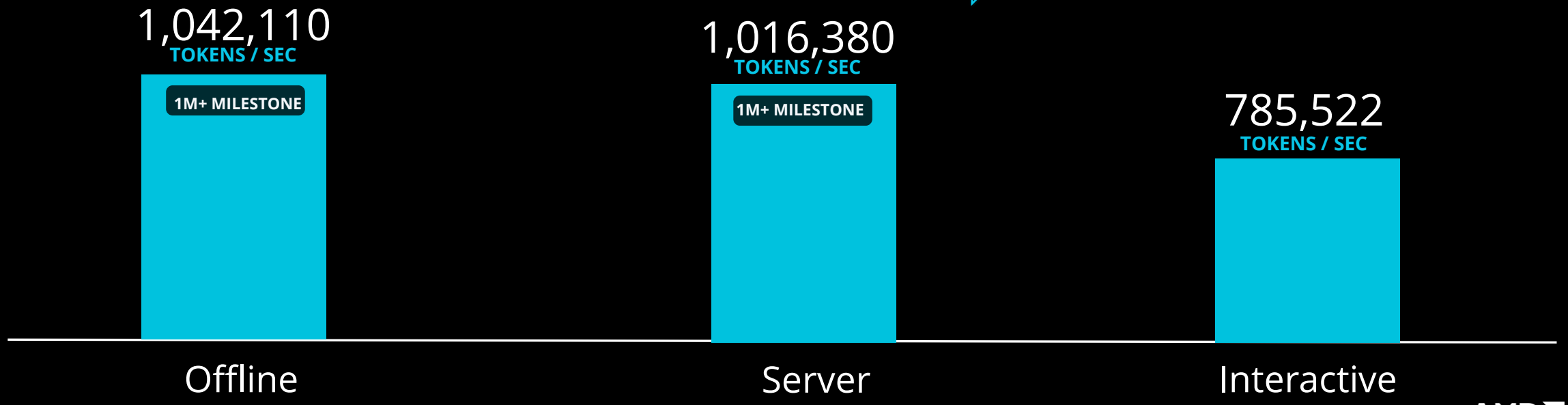
11x SCALE-OUT FROM 8 GPUs TO 87 GPUs

DELIVERING 93% / 93% / 98% OF IDEAL 11x LINEAR SCALING

STARTING POINT
1 NODE
8x MI355X

- OFFLINE SCALE EFFICIENCY 93%
- SERVER SCALE EFFICIENCY 93%
- INTERACTIVE SCALE EFFICIENCY 98%

SCALE OUT
11 NODES
87x MI355X



*AMD Instinct MI355X GPU (11 node) Submission ID: 6.0-0001
*AMD Instinct MI355X GPU (1 node) Submission ID: 6.0-0002

AMD
together we advance_

AMD Instinct™ MI350P PCIe® Card

Deploy and Scale in Existing Data Center Infrastructure

Built for Existing Enterprise Infrastructure

10.5" FHFL Dual-Slot Design
600W Passive Air-cooled
450W Power Cap Configurable
PCIe Gen 5 Host Interface

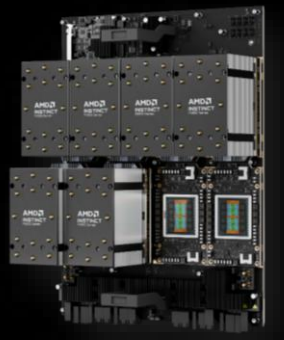
Leadership GPU Architecture

AMD CDNA 4™ Architecture
128 Compute Units
Support for MXFP8, MXFP6, MXFP4
144GB HBM3E Capacity, up to 4.0 TB/s
Dedicated Video and JPEG HW Decode

Open Enterprise Ready Software

Enterprise AI Reference Stack with Inference Microservices
GPU Independence with Cross Platform Interoperability
License-Free, Open Source, Standards based Software
Broad Ecosystem to support Custom Software Stacks

AMD GPU Product Offerings



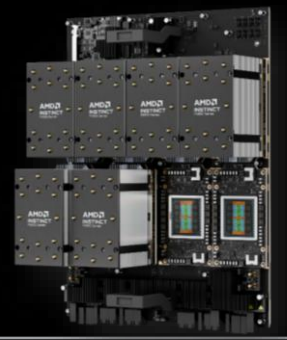
		AMD Radeon™ AI PRO R9700	Instinct™ MI350P	Instinct™ MI355X
HW Specifications	Form Factor	PCIe FHFL 2-slot	PCIe FHFL 2-slot	UBB8
	TBP	300W Air-cooled	600W Air-cooled	1400W Air-cooled
	Compute Units	64	128	256
	Peak Memory	32GB GDDR6 @ 640 GB/s	144GB HBM3e @ 4 TB/s	288GB HBM3e @ 8 TB/s
	GPU Partitions	n/a	Up to 4 @ 36GB each	Up to 8 @ 36GB each
	Multimedia	Video & JPEG Encode and Decode	Video & JPEG Decode	Video & JPEG Decode
	GPU Low-latency High-Speed Interconnect	n/a	n/a	IF @ 537.6 GB/s All-to-All IF @ 76.8 GB/s P2P
AI Performance	Delivered BF16 (TF)	191*	713[4]	1445[1]
	Delivered FP8 (TF)	383*	1529[4]	3000[1]
	Delivered MXFP6[3] (TF)	n/a	1804[4]	3630[2]
	Delivered MXFP4[3] (TF)	n/a	2299[4]	4630[2]

*R9700 Peak Theoretical Flops
[1]Actual measured silicon performance data.
[2]Pre-silicon performance estimates based on emulation. Further update pending final silicon measurements.
[3]AMD Scaled data formats conform to OCP standard MX data types.
[4]Delivered FLOPs performance projections data is preliminary and subject to change.

AMD AI Solution Portfolio - Air Cooled

HPE ProLiant Compute XD685

HPE ProLiant DL345, DL385



	EPYC™ CPU	Radeon AI PRO™	AMD Instinct™ PCIe®	AMD Instinct™ UBB8
Positioning	Mixed Workload General purpose + AI Batch Inference	Mixed Workload Edge AI + Graphics	Low-latency, Memory-intensive Compute	High- throughput AI at Scale for Inference + Training
Use cases	Inference: Classical ML, RecSys, SLM/RAG, NLP	SLM Inference / RAG / MCP / Txt to Image / Video AI	SLM / MLM/ LLM Inference / RAG	Large-scale LLM Inference & Training
Model Size* (Max)	~ 30B per CPU	~ 40-50B per GPU	~ 200-250B per GPU	~up to 500B per GPU
AMD offering	AMD EPYC 5th Gen	R9700S | R9600D	MI350P[1]	MI350X | MI355X
GPUs/node	0	Up to 8GPUs per node	Up to 8 GPUs per node	8 GPU per node
GPU-to-GPU Interconnect	N/A	N/A	N/A	Fully-connected 8-GPU Infinity Fabric
Memory Capacity with Peak Bandwidth	614 GB/s per CPU 6400 MT/s	32G GDDR6 640 GB/s	144GB HBM3e 4 TB/s	2.3TB HBM3e 8TB/s per GPU

*Model size is estimated based on FP4/MXFP4. Estimated by AI

AMD Instinct™ MI400 Series GPUs

Building Strong Customer Momentum

AMD × OpenAI × Meta

Frontier AI Development

6GW

With 1GW Deployment Starting in 2H'2026

AMD × ORACLE

Zetta-Scale Compute

50,000

AMD Instinct™ MI400 Series GPUs Starting in 2H'2026

AMD × Meta

Aligned Open Infrastructure

AMD Helios ORW Rack

Open Rack Scale Infrastructure for Hyperscale

AMD × U.S. DEPARTMENT of ENERGY

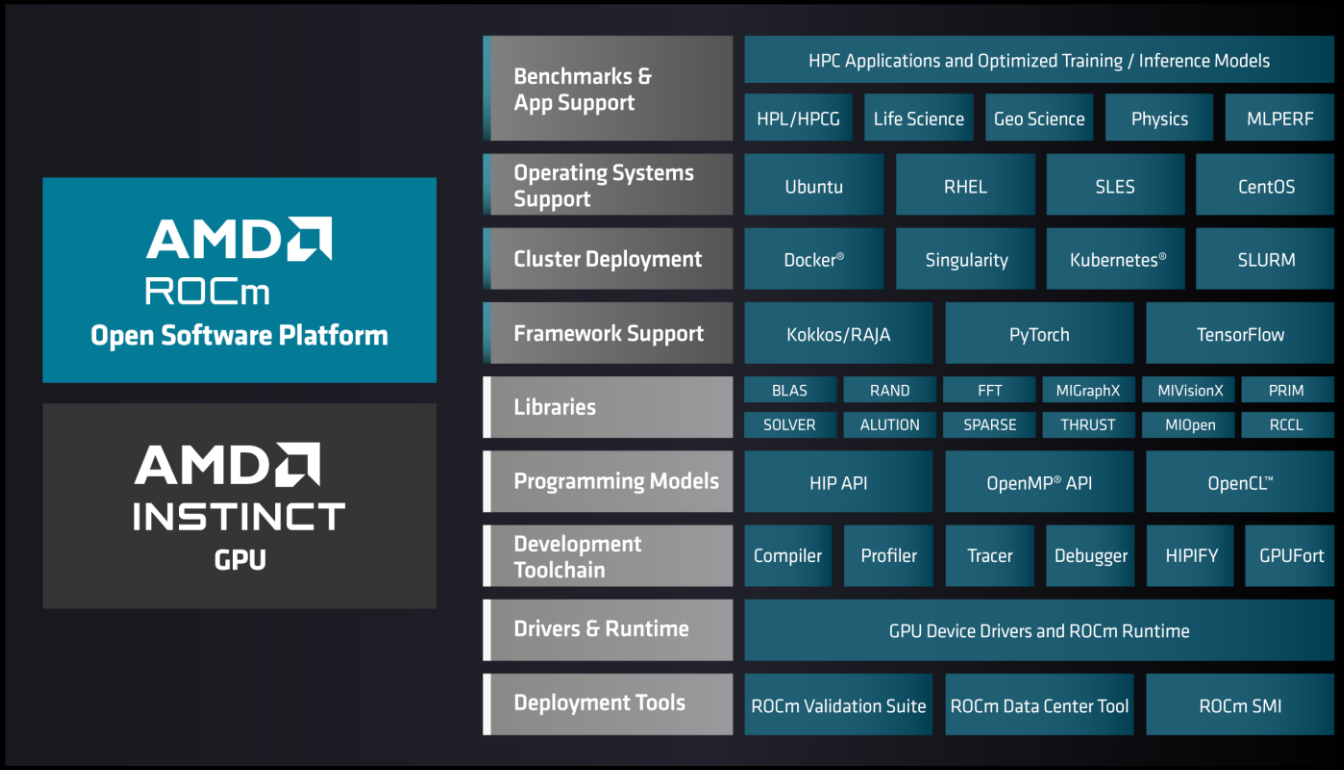
Extending US HPC Leadership

Lux: First US AI Factory
AMD Instinct MI355X Series

Discovery: Flagship AI Supercomputer
AMD Instinct MI430X Series

Using ROCm for HPC

ROCm 已從「GPU runtime / compiler」成長為面向 HPC 的完整軟體堆疊，並有官方 catalog 支援物理研究常見 workload。



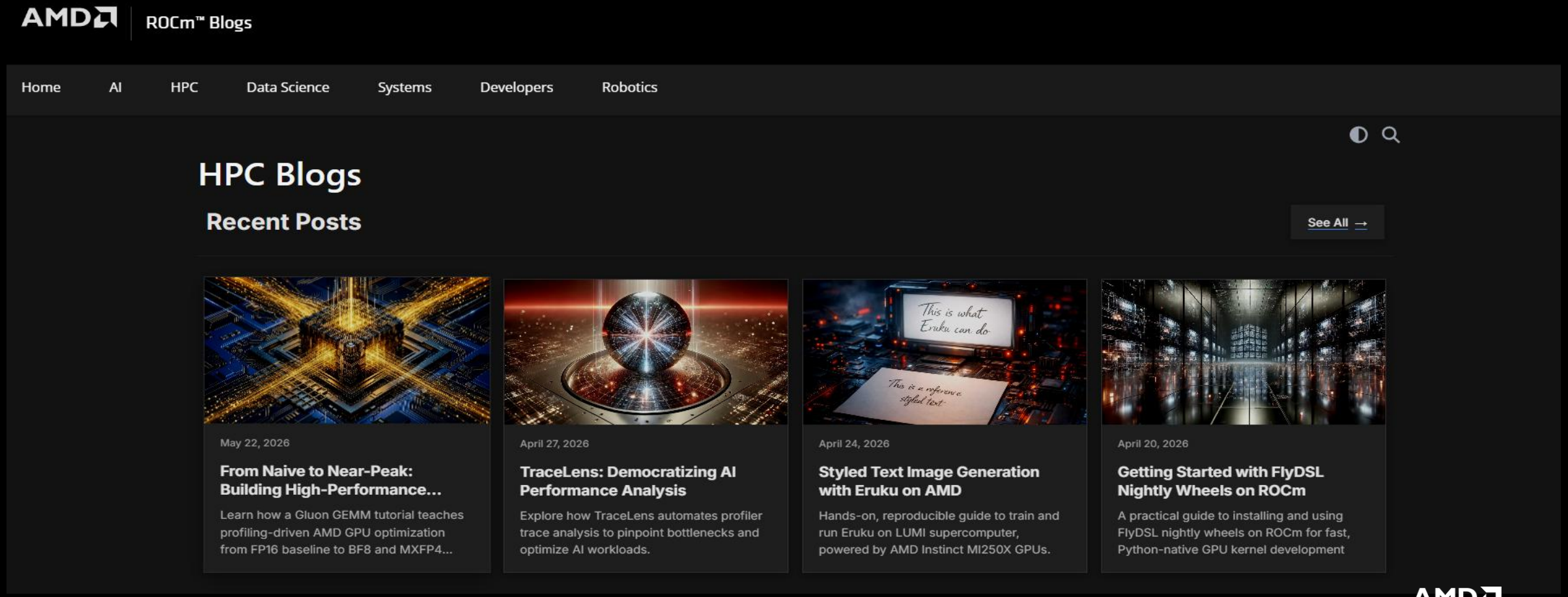
https://rocm.docs.amd.com/en/latest/how-to/rocm-for-hpc/index.html

Application domain	HPC application
Physics	Chroma Grid MILC QUDA PIConGPU
Astrophysics	Cholla
Geophysics	SPECFEM3D Cartesian
Molecular dynamics	Amber GROMACS with HIP — AMD implementation LAMMPS
Computational fluid dynamics	Ansys Fluent NEKO Simcenter Star-CCM+
Quantum Monte Carlo Simulation	QMCPACK
Climate and weather	MPAS
Energy, Oil, and Gas	DevitoPRO
Benchmark	rocHPL rocHPL-MxP HPCG
Tools and libraries	AMD ROCm with OpenMPI container AMD ROCm with MPICH container AMD ROCm with Conda Environment Container Kokkos PyFR RAJA Trilinos VLLM

Using ROCm for HPC

ROCm 近期 blog 已經開始提供物理/HPC 應用的實測案例、建置流程與 tuning 方法.

[GROMACS on AMD Instinct GPUs: A Complete Build Guide]
[Fine-Tuning AI Surrogate Models for Physics Simulations with Walrus on AMD Instinct GPU Accelerators]
[Foundations of Molecular Generation with GP-MoLFormer on AMD Instinct MI300X Accelerators]



https://rocm.docs.amd.com/en/latest/how-to/rocm-for-hpc/index.html

Industry's Highest Performing GPUs for AI and HPC

AMD INSTINCT MI355X GPU VS NVIDIA B200/GB200

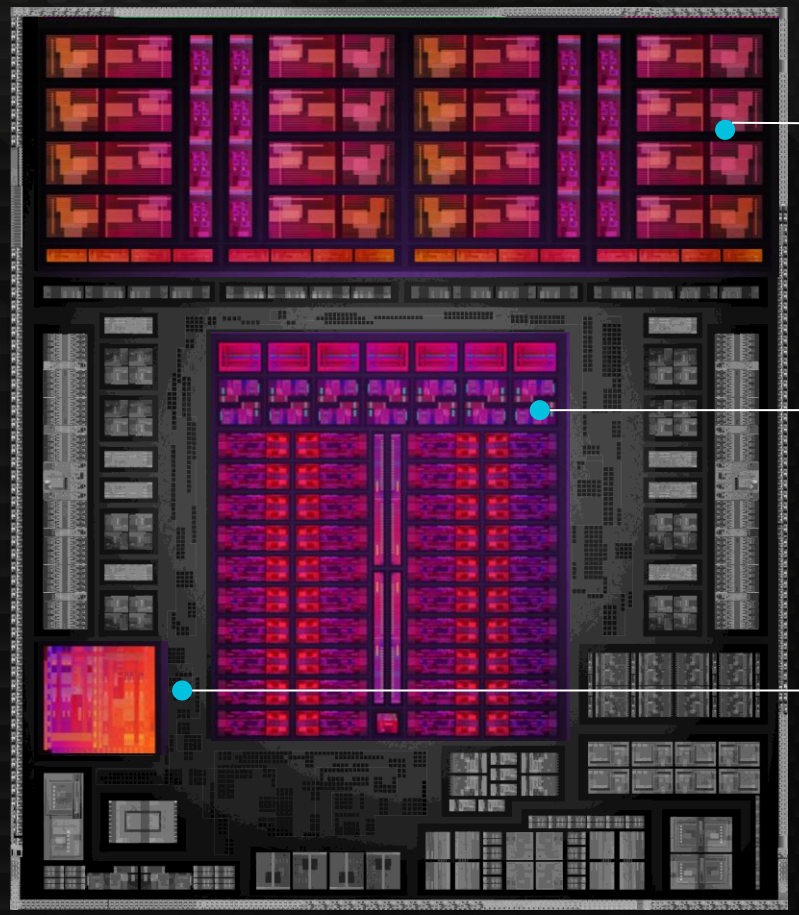
Instinct™ MI355X	vs. GB200	vs. B200
MEMORY CAPACITY	1.6x	1.6x
MEMORY BANDWIDTH	1.0x	1.0x
PEAK FP64	**2.0x**	**2.1x**
PEAK FP16	1.0x	1.1x
PEAK FP8	1.0x	1.1x
PEAK FP6	2.0x	2.2x
PEAK FP4	1.0x	1.1x

See Endnote: MMI350-010A, 18A, 23 and 24

AI PC時代來臨: 如何搶先布局、迎接未來

AMD Ryzen™ AI Max PRO Series Processors

Redefining Performance For Compact Workstations



AMD Zen 5

Up to 16 Core CPU with 32 Threads

For cutting-edge single- and multi-threaded CPU performance

AMD RDNA 3.5

Up to 40 CU GPU with 80 AI Accelerators

For powerful integrated, ISV-certified graphics and GPU compute

AMD XDNA 2

Up to 55 TOPS NPU

Most powerful AI engine for Copilot+ PCs

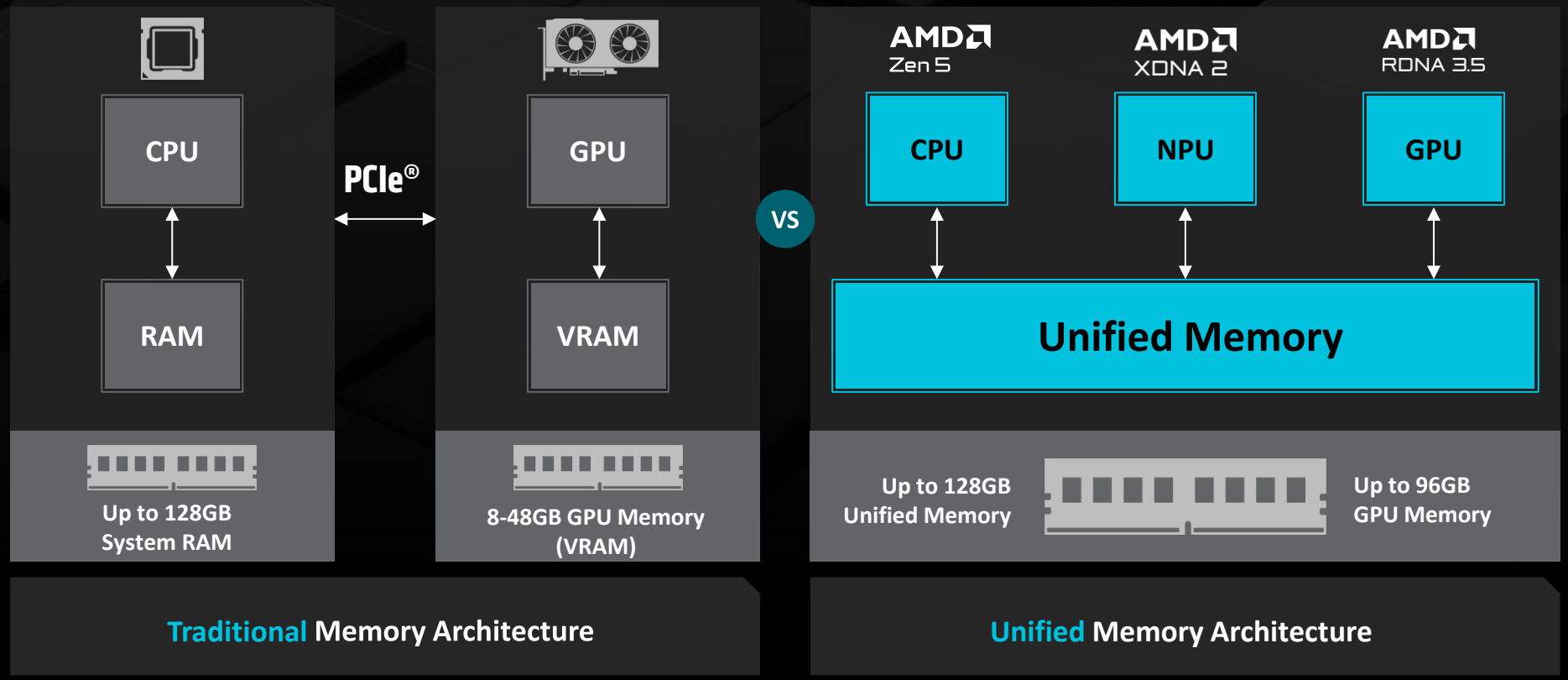
See endnotes GD-243, STX-04a

CU = Compute Units

AMD Ryzen™ AI Max PRO Series Processors

Unified Memory Architecture enables up to 96GB VRAM

- Work with massively large AI models locally
- Run multiple applications simultaneously
- Handle complex 3D data sets interactively



AMD Ryzen™ AI Max PRO Series Processors

Product Stack

		CPU				GPU		
AMD Ryzen™ AI **Max+ PRO**	**395**	16 cores 32 threads	up to **5.1 GHz** max boost	**80 MB** cache	**up to 50+ TOPS NPU*** Copilot+PC	**AMD Radeon™ 8060S Graphics** incl. Radeon PRO driver	**40 CUs** (80 AI accelerators)	Up to **96GB VRAM** With 128GB system RAM
AMD Ryzen™ AI **Max PRO**	**390**	12 cores 24 threads	up to **5.0 GHz** max boost	**76 MB** cache		**AMD Radeon™ 8050S Graphics** incl. Radeon PRO driver	**32 CUs** (64 AI accelerators)	
AMD Ryzen™ AI **Max PRO**	**385**	**8 cores 16 threads**	up to **5.0 GHz** max boost	**40 MB** cache		**AMD Radeon™ 8050S Graphics** incl. Radeon PRO driver	**32 CUs** (64 AI accelerators)	
AMD Ryzen™ AI **Max PRO**	**380**	**6 cores 12 threads**	up to **4.9 GHz** max boost**	**22 MB** cache		**AMD Radeon™ 8040S Graphics** incl. Radeon PRO driver	**16 CUs (32 AI accelerators)**	Up to With 64GB system RAM

* See endnote GD-243 ** See endnote GD-150

AMD Ryzen™ AI Max PRO Series Processors

New Mobile Workstation Designs

HP ZBook Ultra G1a

"The World's Most Powerful 14" Mobile Workstation*"

Available in Q1'2025

Weight: 3.5lbs / 1.587kg
Max height: 18.35mm

AMD RYZEN AI MAX PRO Series

Copilot+PC

* Based on HP's internal analysis of non-gaming 14" mobile workstations with a minimum 3 ISV certs, configurable professional graphics, and a dedicated workstation brand as of September 2024. Most powerful based on highest processor, graphics, memory, storage supported.

AMD × hp
together we advance_

AMD Ryzen™ AI Max PRO Series Processors

New Desktop Workstation Designs

HP Z2 Mini G1a

"Mini Workstation.
Transformative AI Performance."

Available in Q2'2025

Dimensions:
3.4"H x 6.6"W x 7.9"D (8.55cm x 16.8cm x 20cm)
Weight: Starting at 5.07lbs / 2.4kg

Integrated power supply

AMD RYZEN AI MAX PRO Series

Rear view

Copilot+PC

Side view

AMD × hp
together we advance_

AMD Ryzen™ AI Max PRO 400 Series Processors

Product Stack

AMD RYZEN AI Max PRO 400 Series

		CPU			Copilot+PC	GPU		
AMD Ryzen™ AI **Max+ PRO**	**495**	16 cores 32 threads	up to **5.2 GHz** max boost	**80 MB** total cache	up to **55 TOPS** NPU	**AMD Radeon™ 8065S Graphics**	**40 CUs**	Up to **160 GB VRAM** (with 192 GB system memory)
AMD Ryzen™ AI **Max PRO**	**490**	12 cores 24 threads	up to **5.0 GHz** max boost	**76 MB** total cache	up to **50 TOPS** NPU	**AMD Radeon™ 8050S Graphics**	**32 CUs**	
AMD Ryzen™ AI **Max PRO**	**485**	**8 cores 16 threads**	up to **5.0 GHz** max boost	**40 MB** total cache		**AMD Radeon™ 8050S Graphics**	**32 CUs**	

See endnotes GD-150, GD-243

AMD Ryzen™ AI Max PRO 400 Series Processors

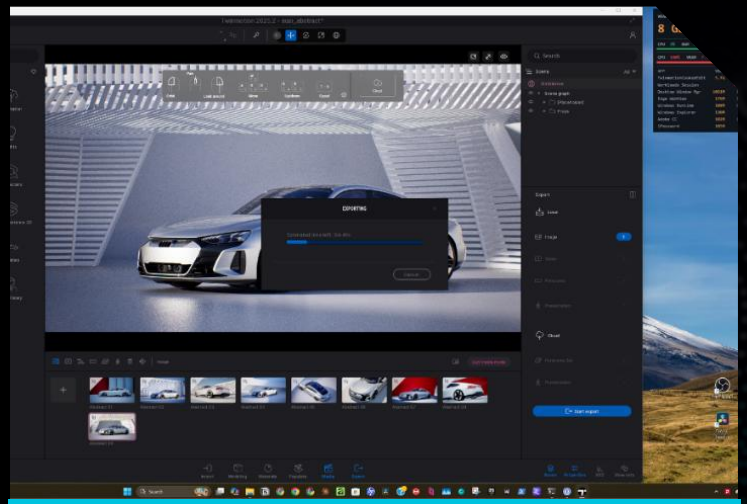
Generational Comparison

	AMD Ryzen™ AI Max PRO Series ("Strix Halo")	AMD Ryzen™ AI Max PRO 400 Series ("Gorgon Halo")
Unified Memory (max)	128GB	192GB
"Zen 5" Cores / Threads	Up to 16C / 32T	Up to 16C / 32T
Max boost frequency*	Up to 5.1GHz	Up to 5.2GHz
Total Cache	Up to 80MB	Up to 80MB
Memory Speed	8000 MT/s	8533 MT/s
Memory Bandwidth	256 GB/s	273 GB/s
Integrated GPU / max VRAM	AMD RDNA™ 3.5 (40 CUs) / 96 GB	AMD RDNA™ 3.5 (40 CUs) / 160 GB
Neural Processing Unit*	AMD XDNA™ 2 / 50 TOPS	AMD XDNA™ 2 / 55 TOPS
AMD Socket / cTDP	Socket FP11 (45-120W)	Socket FP11 (45-120W)

See endnotes GD-150, GD-243

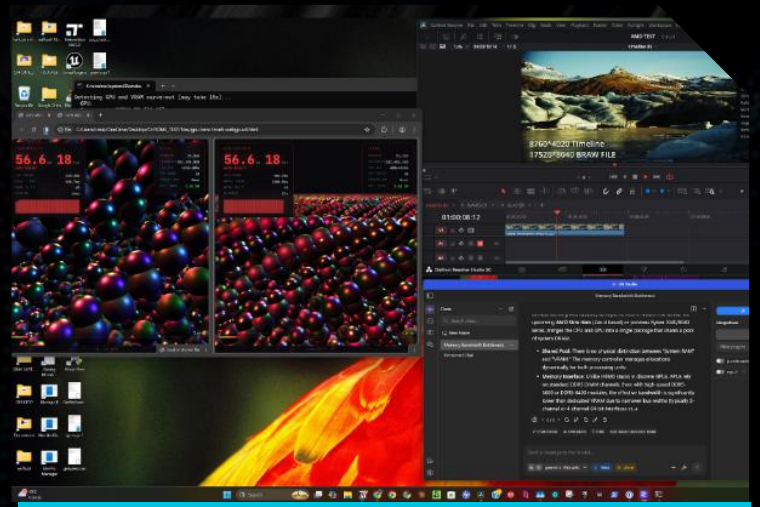
AMD Ryzen™ AI Max PRO 400 Series Processors

The Ultimate AI Processor for Creators & Developers



Agentic Operations

Orchestrate multi-model task execution in Windows® or Linux



Heavy-Scale AI

Execute VRAM-bound, high-parameter models and datasets entirely on-device

Professional Design

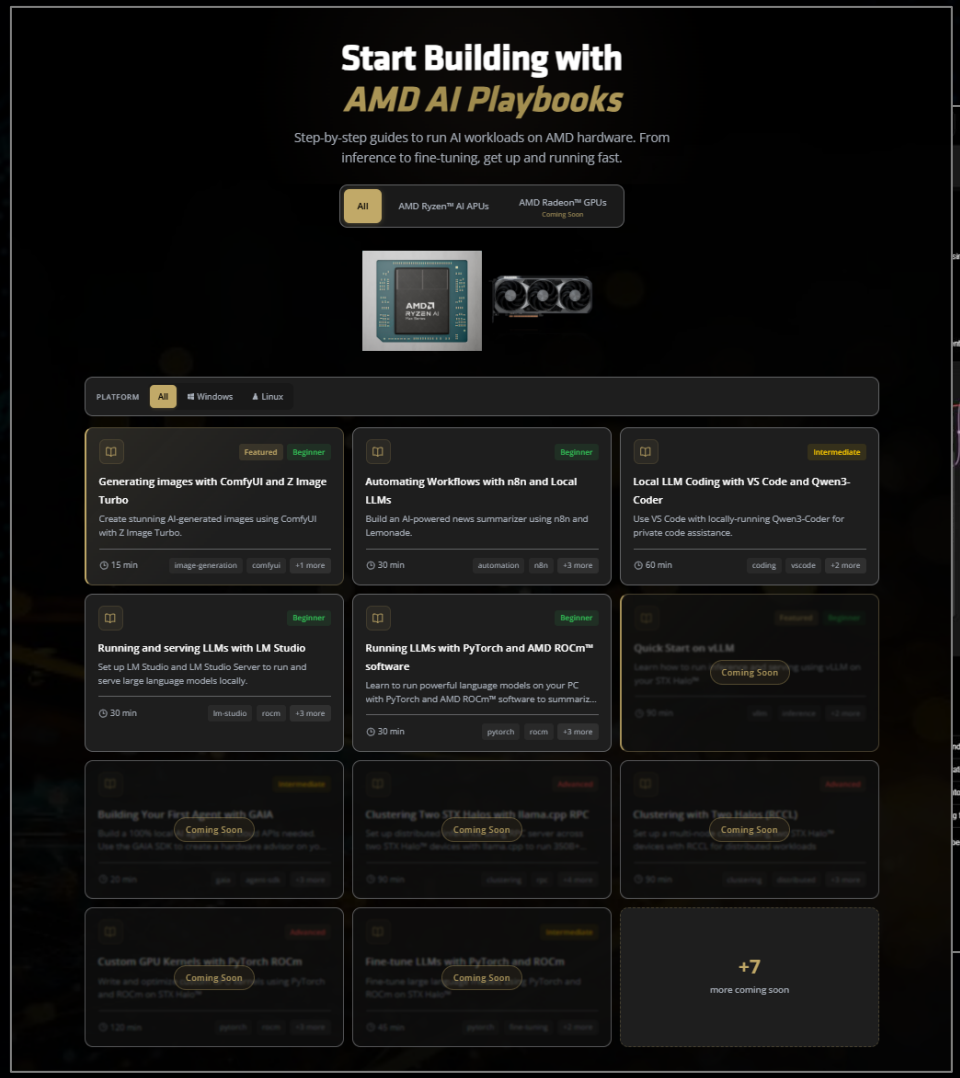
Accelerate advanced 3D design and visual production with local AI

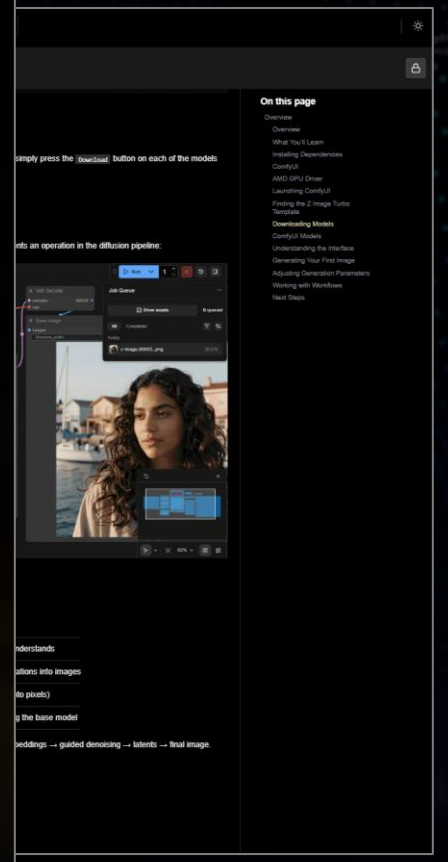
Maximizing compute, graphics, AI and unified memory performance in compact PC designs

AI Playbooks – Getting Started Has Never Been Easier

Playbooks: Guided learning path for developers, including quick and easy setups

- New playbooks released monthly





Windows 11

AMD Software Adrenalin Edition

https://developer.amd.com/playbooks/

AMD together we advance_

Available now in LM Studio

Run MCP-sized context length

MCP tool calls require large context sizes

4096 Context Length

LM Studio Defaults

Not enough for typical MCP use-cases.

Parsing AMD.com's landing page with a Microsoft Playwright navigate_browser tool call returns 9358 tokens as it tokenizes the entire contents of the page and will not fit inside the default context length of 4096.

32,000 Context Length

Flash Attention: ON, KV Cache Q8

Fast and agile: Google Gemma 3n E4b

3 tool calls with a similar token length in-context would require a context size of 28,074.

200,000 Context Length

Flash Attention: ON, KV Cache Q8

Power User: Llama 4 Scout

The 96 GB Variable Graphics Memory on AMD Ryzen™ AI MAX+ 128 GB is sufficient to hold up to 21 such tool calls in-context.

AMD Ryzen™ AI Max+ 395 128GB and AMD Software: Adrenalin Edition™ 25.8.1 WHQL

AMD × LM Studio
together we advance_

Key milestone: Local models are extremely capable

Not every Agent and Workflow needs a frontier model.

Migrate the “grunt work” to local.

Claude Sonnet 4.5

Agentic (Terminal Bench): 50

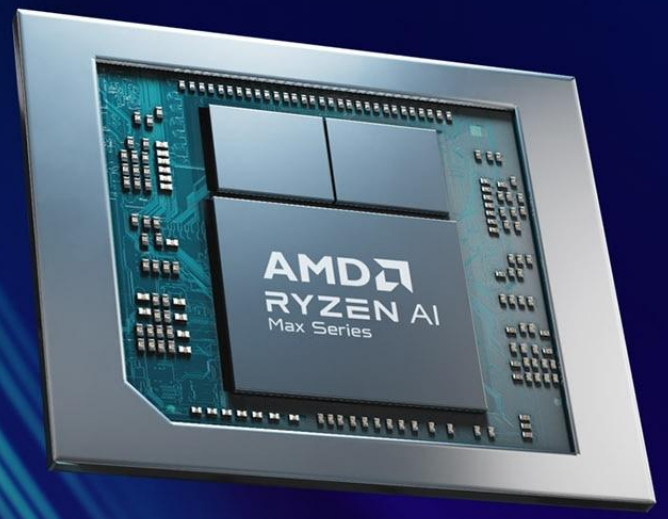
Score as reported by vendor on model card.

Qwen 3.6 35B A3B

Agentic (Terminal Bench): 51.5

Score as reported by vendor on model card.

See endnote: SHO-55



Upto 31 Million Tokens

Max output tokens per month (8 hours/day)

Sustained decode at 128,000 pre-filled context : 36 tokens per second

Upto 385 Million Tokens

Max input tokens per month (8 hours/day)

Sustained prefill throughput at 128,000 tokens: 446 tokens per second

Max Cloud Bill Avoided if Switching

Upto 750 USD per month

Assuming switch from Claude Sonnet 4.5 API at $15/M output and $3/M input tokens

- Scenario: 10:1 input to output ratio. 8 hours per day
- Scenario consumption: 0.573 Million output tokens per day (4.43 hours)
- Scenario consumption: 5.73 Million input tokens per day (3.57 hours)

Power Math

- 150W sustained “nightmare case” draw
- $0.15 USD kWh electricity pricing
- $16.2 USD monthly bill

See endnote: SHO-49


AMD
together we advance_data centers

Disclaimers and Attributions

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’s products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18