AMD INSTINCT™ Accelerator

Compute & Accelerator Forum June 2024

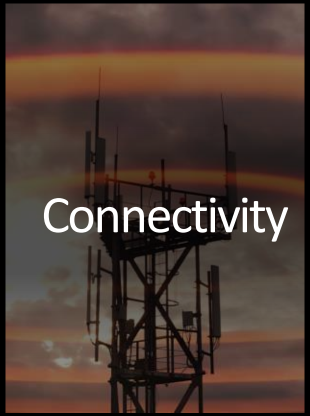
AMD

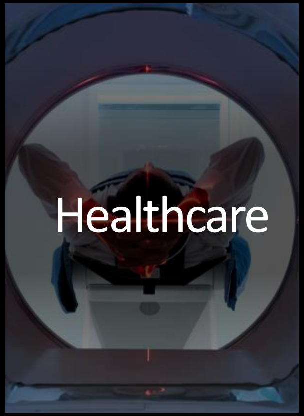
together we advance_

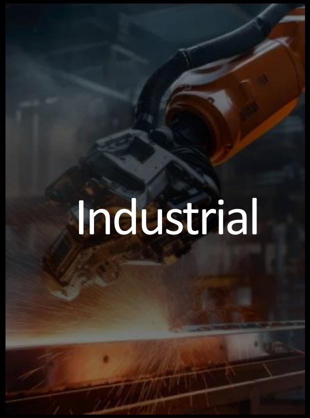
Agenda

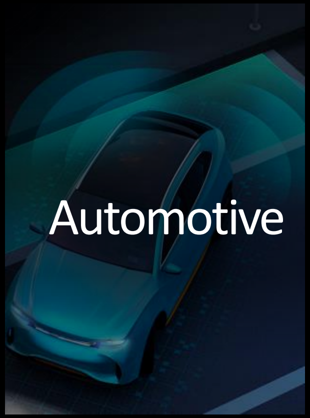
1. AMD AI Vision and Portfolio
2. Instinct DC GPU History and Roadmap
3. Instinct Product Strategy
4. MI300X Product Overview
5. MI300X Performance Proof Points
6. ROCm Software Strategy and Ecosystem
7. Customer proof points

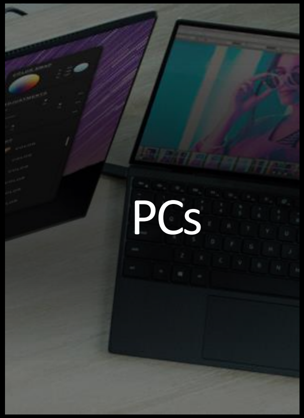
AMD | Powers the daily lives of billions

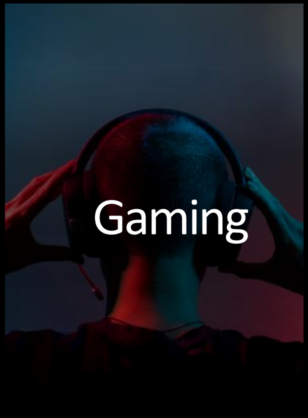


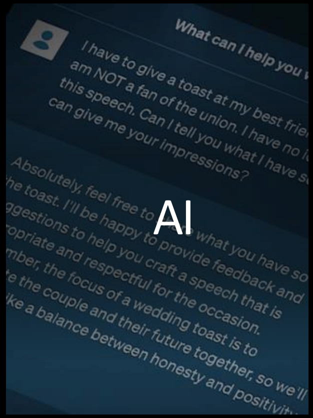












What can I help you

I have to give a toast at my best frie
am NOT a fan of the union. I have no
this speech. Can I tell you what I have
can give me your impressions?

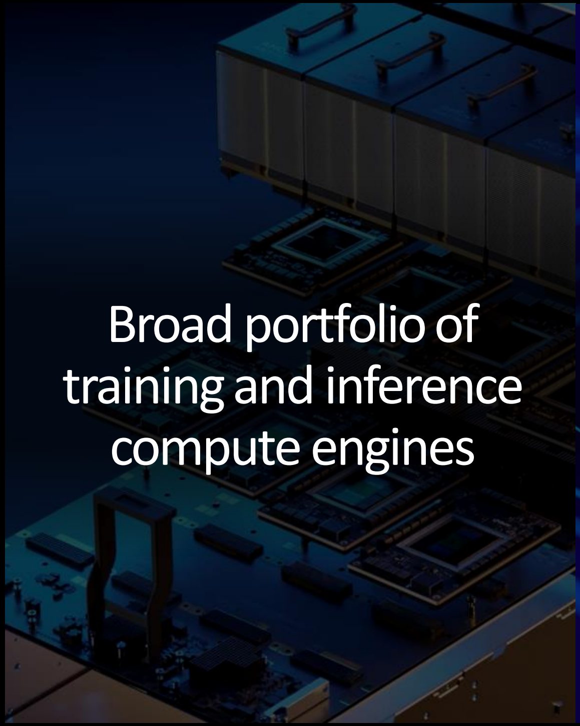
Absolutely, feel free to what you have so
he toast. I'll be happy to provide feedback and
ggestions to help you craft a speech that is
opriate and respectful for the occasion.
mber, the focus of a wedding toast is to
e the couple and their future together, so we'll
ke a balance between honesty and positivit

AMD

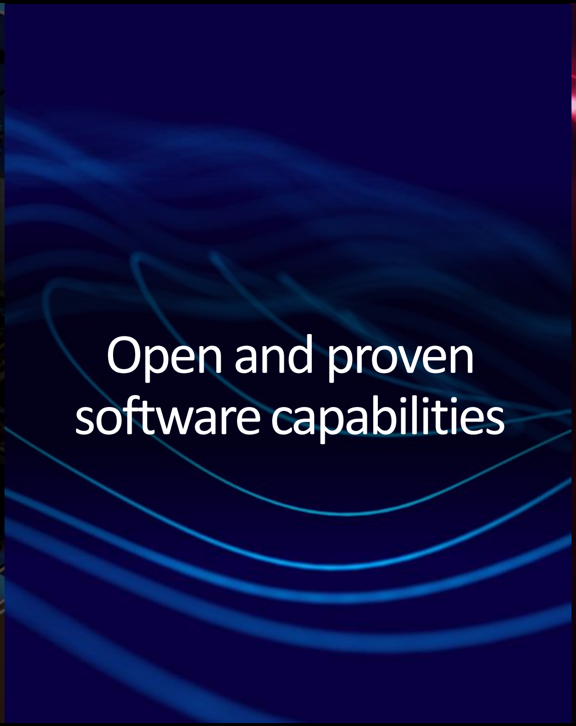
Advancing end-to-end AI infrastructure

Cloud HPC Enterprise Embedded PC

AMD AI Platforms



Broad portfolio of training and inference compute engines

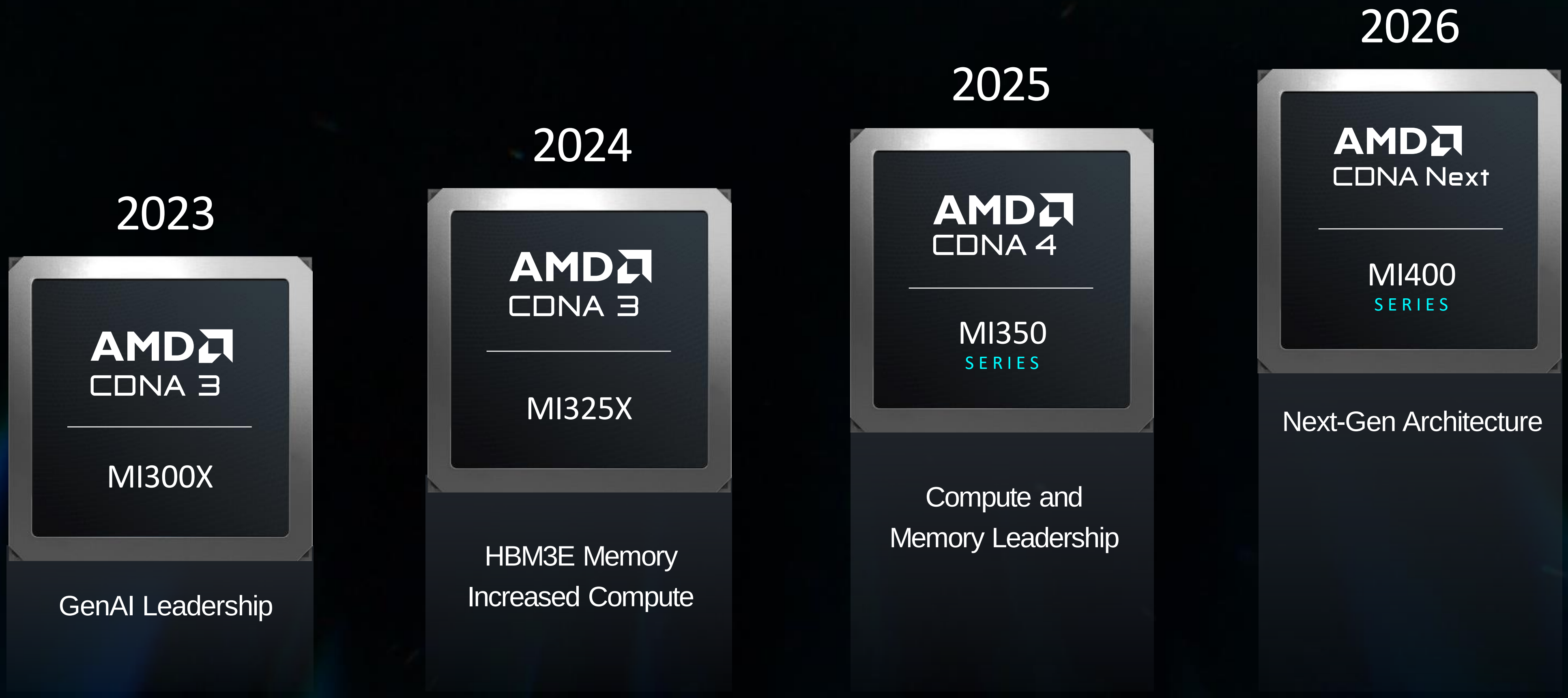


Open and proven software capabilities

AI ecosystem with deep co-innovation

AMD INSTINCT

Leadership roadmap on an annual cadence



2023

AMD CDNA 3

MI300X

GenAI Leadership

2024

AMD CDNA 3

MI325X

HBM3E Memory
Increased Compute

2025

AMD CDNA 4

MI350 SERIES

Compute and
Memory Leadership

2026

AMD CDNA Next

MI400 SERIES

Next-Gen Architecture

All roadmaps are subject to change.

Only AMD powers the full range of data center workloads

Traditional IT workloads

Max performance and efficiency

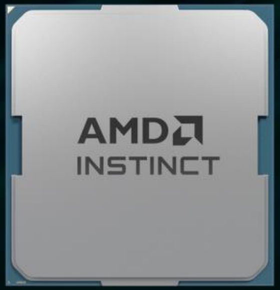
General Purpose

CPU

Mix of AI and traditional workloads

Smaller model sizes

AI Inference



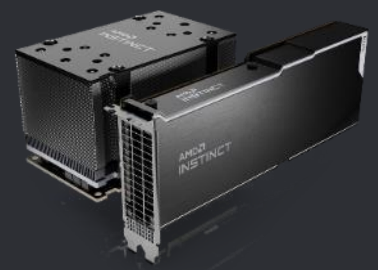
GPU

AI at scale and dedicated infrastructure

Larger model sizes

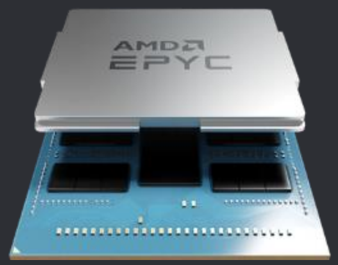
AI Training

AMD Product Portfolio
from cloud to client



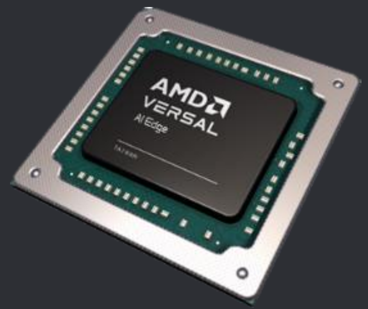
AMD Instinct GPU Accelerators

Data center HPC and AI solutions



4th Gen AMD EPYC

Industry-leading x86 CPU server solution

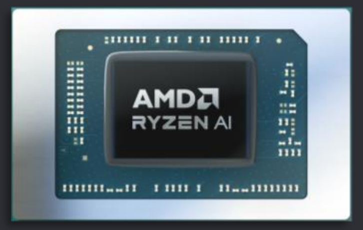


Embedded Versal and Alveo

AI + sensor fusion for embedded, FPGA

Radeon GPU

GPU for AI in gaming and AI developers

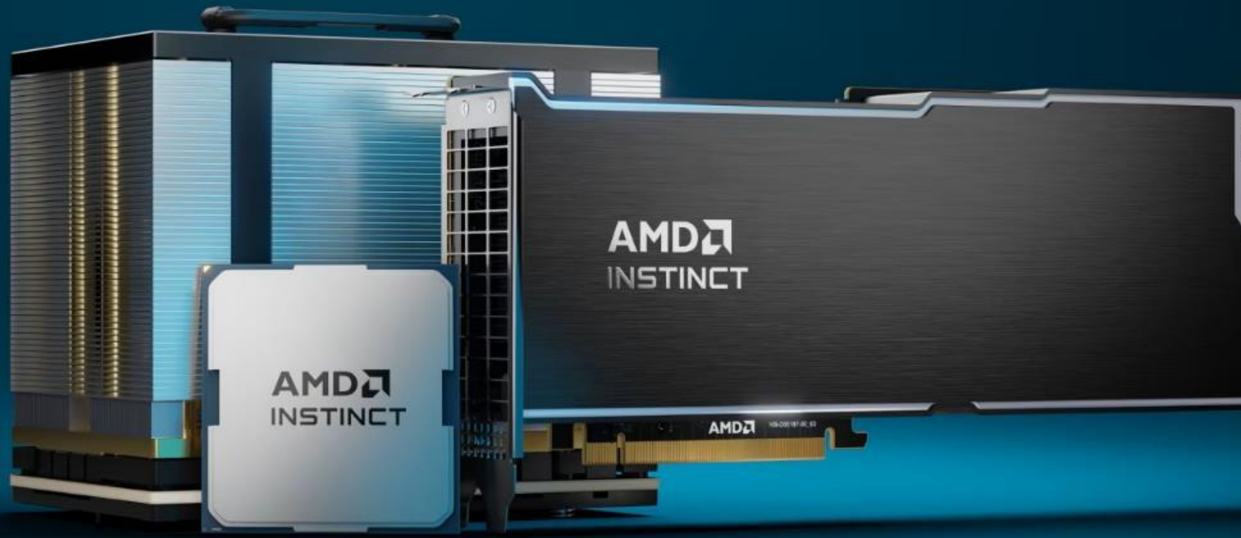


Ryzen Mobile Processors

x86 with integrated GPU and Ryzen AI accelerator

AMD Instinct™ Accelerators

Data center GPU for the most demanding AI and HPC workloads



AMD Instinct Strategic Pillars

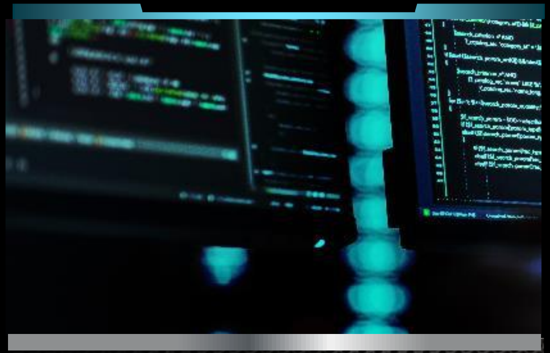
Enabling customer success

Ease of Migration

Drop-in compatible with existing infrastructure for hardware and software

Performance Leadership

Leading performance without compromise



Commitment to Openness

Investment and participation in open standards across the entire ecosystem

Customer Focused

Roadmap and support structure geared towards customer success

AMD
together we advance_

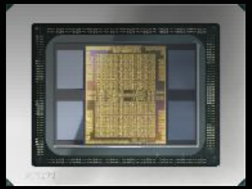
The AMD Instinct™ Accelerator Journey

Multiple generations of architecture focused advancing HPC & AI compute

MI100
AMD CDNA™

ECOSYSTEM GROWTH

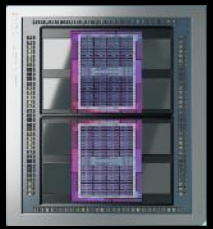
First purpose-built GPU architecture to accelerate FP64 and FP32 HPC workloads



MI200
AMD CDNA™ 2

DRIVING HPC AND AI TO A NEW FRONTIER

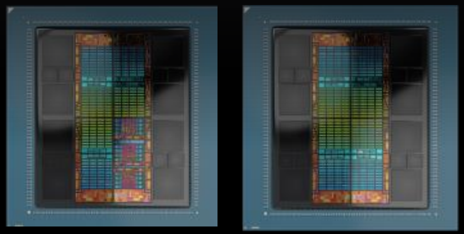
Denser compute architecture with leading memory capacity/bandwidth



MI300
AMD CDNA™ 3

DATA CENTER APU & DISCRETE GPU

Focused improvements on Unified memory, AI data format performance and in-node networking

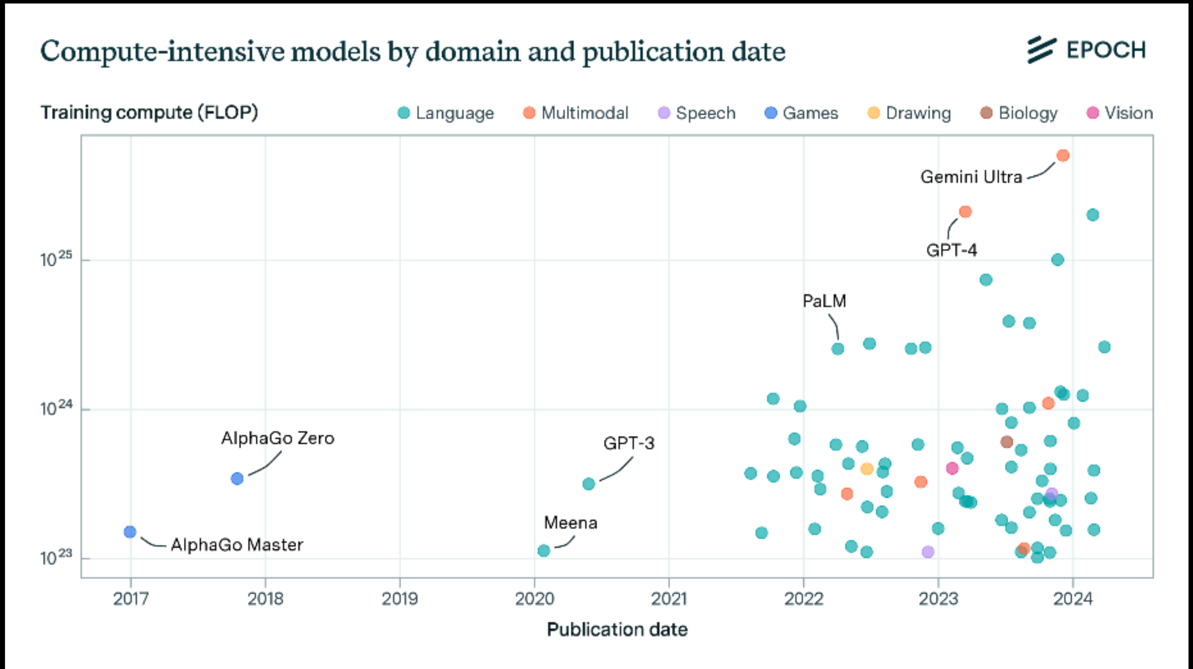


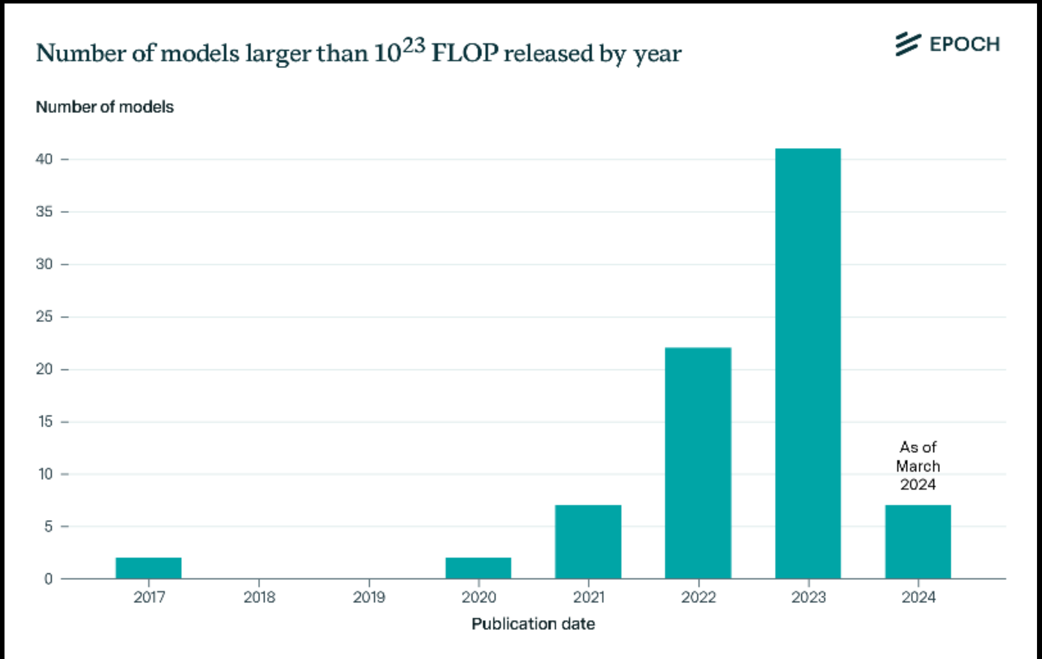
2020 → 2023

Model Evolution Accelerating Rapidly

AI performance needs driving GPU demand &cluster growth

- The pace of compute intensive model releases is accelerating with frontier models advancing rapidly
- Majority of the compute intensive models are LLM but newer multi modal and other domain models are emerging
- In 2020, only two models were trained with more than 10^{23} FLOP. This increased exponentially over the subsequent three years, and over 40 models trained at this scale were released in 2023



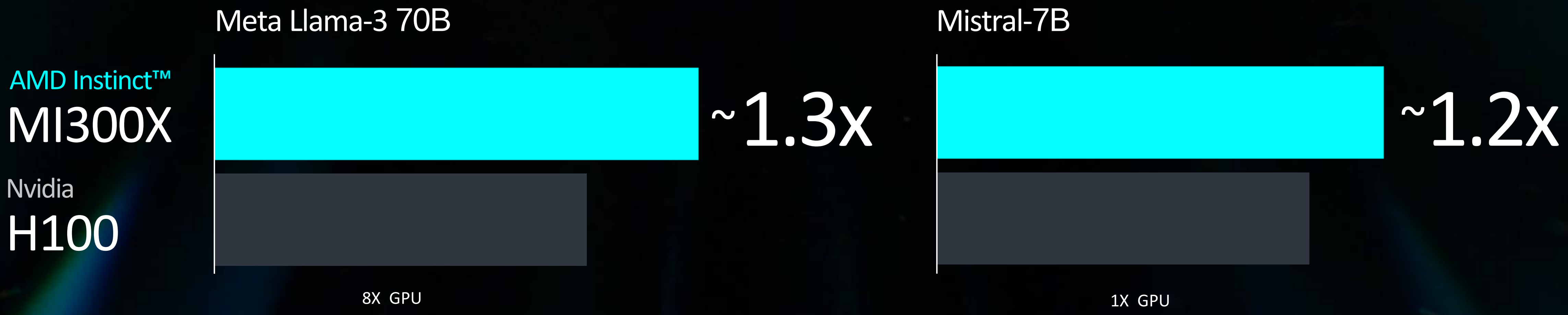


Source: Robi Rahman, David Owen and Josh You (2024), "Tracking Compute-Intensive AI Models". Published online at epochai.org. Retrieved from: 'https://epochai.org/blog/tracking-compute-intensive-ai-models' [online resource]

AMD
ROCm

ROCm™ 6 Software

Leadership performance for generative AI



Token Generation Throughput

MI300-53, MI300-54

AMD Instinct™ MI300X GPU vs. Competition

		MI300X (Up to)	H100 SXM	AMD Instinct™ Advantage (Up to)
Hardware Specifications	TBP	750W	**700W**	**-**
	Memory Capacity	192 GB HBM3	80GB HBM3	**2.4x**
	Memory Bandwidth (Peak Theoretical)	5.3 TB/s	3.3TB/s	**1.6x**
HPC Performance (Peak Theoretical)	FP64 Matrix | Vector (TFLOPS)	163.4 | 81.7	66.9 | 33.5	**2.4x | 2.4x**
	FP32 Matrix | Vector (TFLOPS)*	163.4 | 163.4	N/A | 66.9	**N/A | 2.4x**
AI Performance (Peak Theoretical)	TF32 (TFLOPS)	653.7	494.7	**1.3x**
	FP16 (TFLOPS)	1307.4	989.4	**1.3x**
	BFLOAT16 (TFLOPS)	1307.4	989.4	**1.3x**
	FP8 (TFLOPS)	2614.9	1978.9	**1.3x**
	INT8 (TFLOPS)	2614.9	1978.9	**1.3x**

See endnotes: MI300-05A, MI300-17, MI300-18

- Nvidia H100 GPUs don't support FP32 Tensor.
- Nvidia H100 source: https://resources.nvidia.com/en-us-tensor-core/

AMD Instinct™ Platform

8x MI300X in a ready to deploy OCP form factor

8x
MI300X

21 PF
BF16 | FP16

1.5 TB/s
HBM3

896 GB/s
Infinity Fabric™ Bandwidth

Industry-Standard
OCP Design

AMD
together we advance_

AMD Instinct™ MI300X Platform

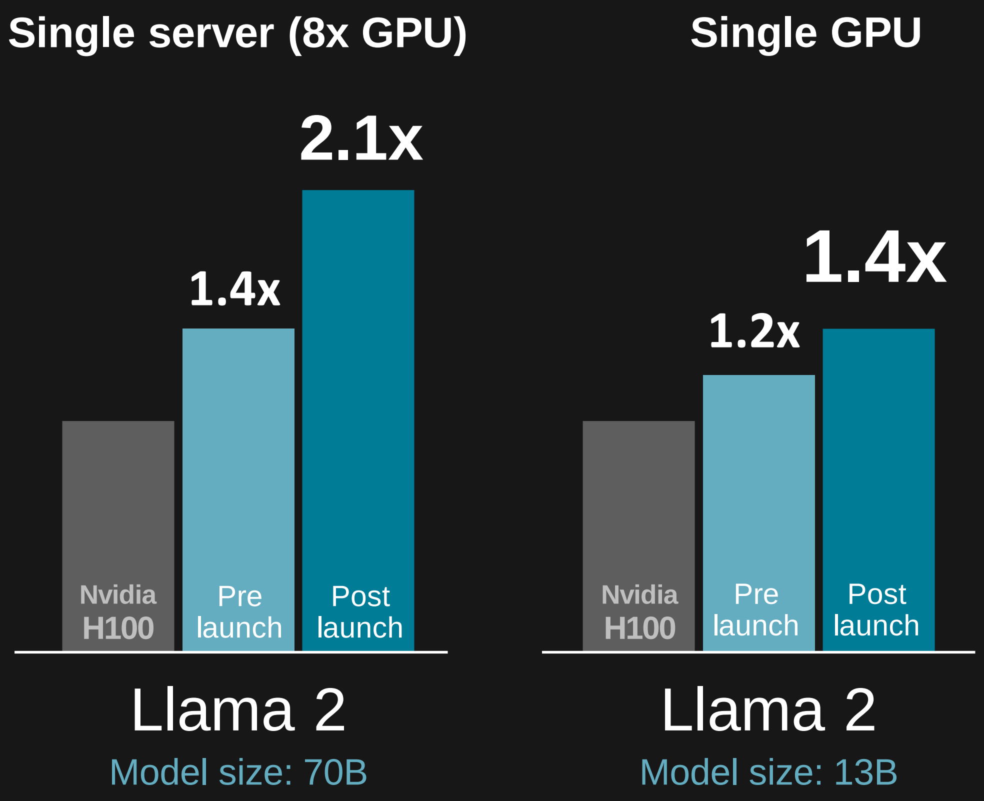
Infrastructure performance

AMD Instinct™ MI300X Platform	Nvidia H100 HGX	AMD Instinct™ MI300X Advantage
1.5 TB HBM3 memory	**640** GB HBM3 memory	**2.4X** More memory
~10.4 PF FP16 / BF16 FLOPS	**7.9** PF FP16 / BF16 FLOPS	**~1.3X** More Compute
~896 GB/s Aggregate bi-directional bandwidth	**900** GB/s Aggregate bi-directional bandwidth	Comparable
448 GB/s Single node ring bandwidth	**450** GB/s Single node ring bandwidth	Comparable
Up to **400** GbE NIC / GPU	Up to **400** GbE NIC / GPU	Equivalent
PCIe® Gen 5 128 GB/s	**PCIe® Gen 5** 128 GB/s	Equivalent

See endnotes:MI300-25
Nvidia H100 source: https://resources.nvidia.com/en-us-tensor-core/

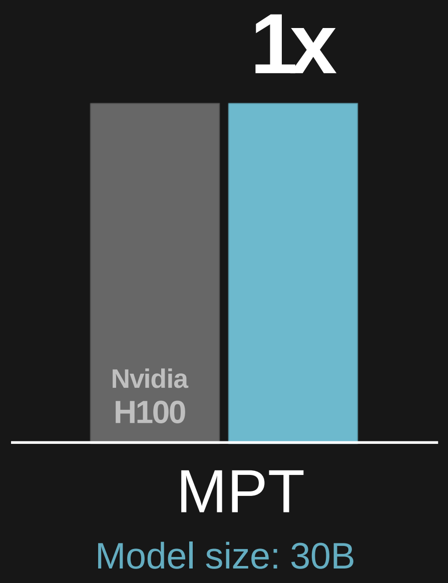
MI300X AI Performance Leadership

MI300X Inference



Latency improvement (ms)

MI300X Training



Throughput

Results may vary, see endnotes:MI300-40

AMD
together we advance_

AMD Instinct™ Platform: Performance Advantage

1 Nvidia H100 HGX

640 GB HBM3 | 26.4 TB/s

1 AMD Instinct™ MI300X Platform

1.5 TB HBM3 | 42.4 TB/s

Training & Inference		Training	Inference
1x	Performance per system	1x MPT-30B	2.1x Llama 70B
1x	Models per system	2x ~30B	2x ~70B
1x	Max LLM model size per system	2x ~70B vs.~30B	2x ~680B vs 290B

Results may vary. See endnotes:MI300-34, MI300-40, MI300-39, MI300-42

Delivering Exceptional Value to AI leaders

Microsoft Azure

MI300X enables to serve larger AI models with fewer GPUs

"With MI300X's larger memory capacity and bandwidth, we can serve larger models with fewer GPUs. We have already got GPT-4 up and running on MI300X"

Satya Nadella
CEO, Microsoft
November 2023

Meta

Ecosystem growth over the years has made ROCm a highly competitive software platform

"We have had a great experience with ROCm and the performance it has been able to deliver with MI300X. The optimizations and the ecosystem growth over the years have made ROCm a highly competitive software platform. We see great performance numbers which we believe will benefit the industry"

Ajit Mathews
Sr. Director, Meta
December 2023

databricks

ROCm runs out of the box from day one

ROCm runs out of the box from day one. It was was very easy to run and include ROCm in our stack . Many of the generative AI tools today are open source like PyTorch, Triton, Huggingface and these tools can run today on AMD ROCm software stack and this makes ROCm another key component of the open source ecosystem

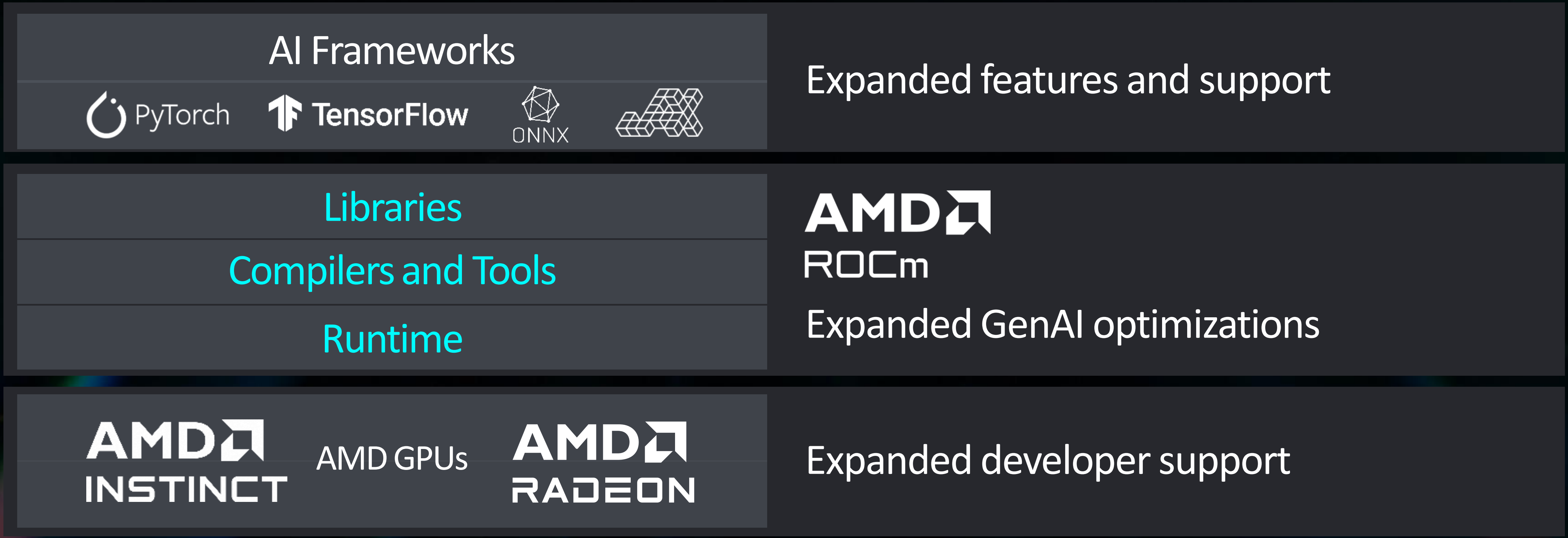
Ion Stoica
Co-Founder and Executive Chairman, Databricks
December 2023

AMD ROCm

Open | Proven | Ready

AMD ROCm

Open software ecosystem



AI Frameworks

PyTorch TensorFlow ONNX

Expanded features and support

Libraries

Compilers and Tools

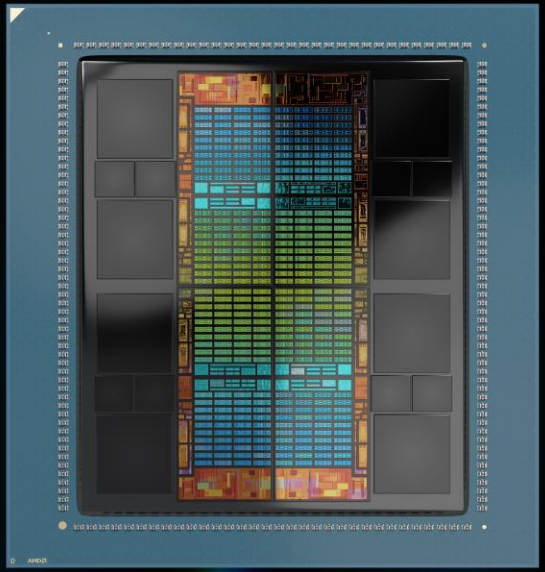
Runtime

AMD ROCm

Expanded GenAI optimizations

AMD INSTINCT AMD GPUs AMD RADEON

Expanded developer support



AMD Instinct™ MI300X Accelerator

Leadership performance
Out-of-box support on popular GenAI models

GPT-4 Llama 2 Llama 3 Hugging Face StarCoder Yi Phi

Falcon LLM BLOOM Qwen MPT MISTRAL AI_ Stable Diffusion

GPT-NeoX DBRX AlphaFold 2 OPT Databricks' Dolly 2.0 Gemma VICUNA

AMD
ROCm

Committed to Open-Source Innovation

Hugging Face

700,000+ models run

out-of-box on
AMD ROCm™ platform

OpenAI Triton

Fully upstreamed AMD ROCm™ platform support

Used for key LLM kernel generation

PyTorch

Fully upstreamed AMD ROCm™ platform support

Continuous Integration

JAX

vLLM

Tensor Flow

MLIR | IREE

ONXX Runtime

OpenXLA

Use or mention of third-party marks, logos, products, services, or solutions herein is for informational purposes only and no endorsement by AMD is intended or implied. GD-83

Frameworks Support Status

Key frameworks fully upstreamed and optimized for AMD Instinct™ Accelerators

PyTorch

- Full Feature Support on Day 0 since Pytorch 2.0

TensorFlow

- Upstream Tensorflow Version Optimized For AMD Instinct (2.13, 2.14)

JAX

- Upstreamed JAX version optimized for AMD Instinct
- JAX supported w/ OpenXLA & Triton backends

OpenAI Triton

- AMD is the “top” 3rd party hardware contributor to OpenAI Triton
- Upstreamed support for AMD Instinct
- FP8 datatype supported on MI300X
- Available Now: Docker pull rocm/oai-triton

OpenXLA

- OpenXLA project “founding member”
- AMD support functional (and upstreamed)
  - Focused on maintaining current AMD support for Tensorflow while code bases are being refactored
- Available Now: https://github.Com/openxla/xla

Transitioning Workloads to Instinct & ROCm

Low friction softward porting for existing Nvidia users to AMD

DROP IN OUT-OF-THE-BOX SUPPORT

For Existing Code

PyTorch TensorFlow

ONNX JAX

OpenXLA OpenAI Triton

PORT & OPTIMIZE

For Custom Kernels

Leverage AMD HIPIFY tool for large custom kernels or code re-write if smaller number of lines of code (typical)

EQUIVALENT LIBRARIES

For New Code Dev

ROCm Libraries Developed to Mirror CUDA-based libraries

- rocBLAS, rocSparse, rocFFT, RCCL, MIOpen…

PURPOSELY DESIGNED TO LEVERAGE EXISTING CUSTOMER CODE WITH MINIMAL CHANGES

- Vast majority of AI end users engaged by AMD are programing at the framework level and their code functions out of the box with no edits
- Performance optimizations for common models and customer driven asks underway to ensure out of the box performance
- Foundational model builders with custom CUDA kernels have the option to use AMD HIPIFY to convert CUDA code, but often find it to be a low lift to re-write that small portion of code for AMD GPUs

Training: Case Studies

Published AMD Instinct™ training runs

OAK RIDGE National Laboratory

- 1T GPT model
- 3072 MI250s
- 87% strong scaling eff

kt cloud

- 221B T5 based model
- 1200 MI250s
- Pre-tests outperformed A100

UNIVERSITY OF TURKU

- 13B Finnish model
- 768 MI250s
- Utilized Megatron DeepSpeed

AI2

- Olmo 7B (65B in progress)
- 1024 MI250s
- Utilized PyTorch FSDP

SILOGEN

- Poro 34B model
- 512 MI250s

Microsoft

- 6.7B RetNet
- 512 MI250s
- Reported “decent throughput”

MosaicML

- MPT-1B,3B,7B
- 32 MI250
- Proved interoperability between AMD and NV GPUs

LAMINI

- Fine-tuning of open-source models
- Utilizes MI210
- Able to host 200B model in single server

Announced last week

Ultra Accelerator Link

Partner group of innovators for scale up AI infrastructure

AMD BROADCOM cisco Google Hewlett Packard Enterprise intel Meta Microsoft

High Performance | Open | Scalable

Ultra Ethernet is the answer for scale out AI infrastructure

Ultra Ethernet Consortium

AMD ARISTA BROADCOM CISCO EVIDEN

Hewlett Packard Enterprise intel Meta Microsoft ORACLE

AMD

Advancing the AI Data Center

AMD EPYC™ CPUs + AMD Instinct™ GPUs + UALink and Ultra Ethernet Networking

AMD