Cautionary Statement

This presentation contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) such as the data center accelerator total addressable market ; the features, functionality, performance, availability, timing and expected benefits of AMD products and technology including the AMD Instinct™ MI300X and MI 300A accelerators, AMD ROCm™ 6 open software stack, and the AMD Ryzen™ 8040 Series processors; expected number of GPUs required to train models; the expectation that El Capitan will be the world's first two-exaflop supercomputer; and the AMD Ryzen AI roadmap, which are made pursuant to the Safe Harbor provisions of the Private Securities Litigation Reform Act of 1995. Forward-looking statements are commonly identified by words such as "would," "may," "expects," "believes," "plans," "intends," "projects" and other terms with similar meaning. Investors are cautioned that the forward-looking statements in this presentation are based on current beliefs, assumptions and expectations, speak only as of the date of this presentation and involve risks and uncertainties that could cause actual results to differ materially from current expectations. Such statements are subject to certain known and unknown risks and uncertainties, many of which are difficult to predict and generally beyond AMD's control, that could cause actual results and other future events to differ materially from those expressed in, or implied or projected by, the forward-looking information and statements. Material factors that could cause actual results to differ materially from current expectations include, without limitation, the following: Intel Corporation's dominance of the microprocessor market and its aggressive business practices; economic uncertainty; cyclical nature of the semiconductor industry; market conditions of the industries in which AMD products are sold; loss of a significant customer; impact of the COVID-19 pandemic on AMD's business, financial condition and results of operations; competitive markets in which AMD's products are sold; quarterly and seasonal sales patterns; AMD's ability to adequately protect its technology or other intellectual property; unfavorable currency exchange rate fluctuations; ability of third party manufacturers to manufacture AMD's products on a timely basis in sufficient quantities and using competitive technologies; availability of essential equipment, materials, substrates or manufacturing processes; ability to achieve expected manufacturing yields for AMD's products; AMD's ability to introduce products on a timely basis with expected features and performance levels; AMD's ability to generate revenue from its semi-custom SoC products; potential security vulnerabilities; potential security incidents including IT outages, data loss, data breaches and cyber-attacks; potential difficulties in operating AMD's newly upgraded enterprise resource planning system; uncertainties involving the ordering and shipment of AMD's products; AMD's reliance on third-party intellectual property to design and introduce new products in a timely manner; AMD's reliance on third-party companies for design, manufacture and supply of motherboards, software, memory and other computer platform components; AMD's reliance on Microsoft and other software vendors' support to design and develop software to run on AMD's products; AMD's reliance on third-party distributors and add-in-board partners; impact of modification or interruption of AMD's internal business processes and information systems; compatibility of AMD's products with some or all industry-standard software and hardware; costs related to defective products; efficiency of AMD's supply chain; AMD's ability to rely on third party supply-chain logistics functions; AMD's ability to effectively control sales of its products on the gray market; impact of government actions and regulations such as export regulations, tariffs and trade protection measures; AMD's ability to realize its deferred tax assets; potential tax liabilities; current and future claims and litigation; impact of environmental laws, conflict minerals-related provisions and other laws or regulations; impact of acquisitions, joint ventures and/or investments on AMD's business and AMD's ability to integrate acquired businesses; impact of any impairment of the combined company's assets; restrictions imposed by agreements governing AMD's notes, the guarantees of Xilinx's notes and the revolving credit facility; AMD's indebtedness; AMD's ability to generate sufficient cash to meet its working capital requirements or generate sufficient revenue and operating cash flow to make all of its planned R&D or strategic investments; political, legal and economic risks and natural disasters; future impairments of technology license purchases; AMD's ability to attract and retain qualified personnel; and AMD's stock price volatility. Investors are urged to review in detail the risks and uncertainties in AMD's Securities and Exchange Commission filings, including but not limited to AMD's most recent reports on Forms 10-K and 10-Q.

Advancing AI

Dr. Lisa Su
Chair and CEO, AMD


AI
Most transformational technology in 50 years

Healthcare

Climate Research

AI Assistants

Robotics

Security

Content Creation

$30B

2023

~50% CAGR

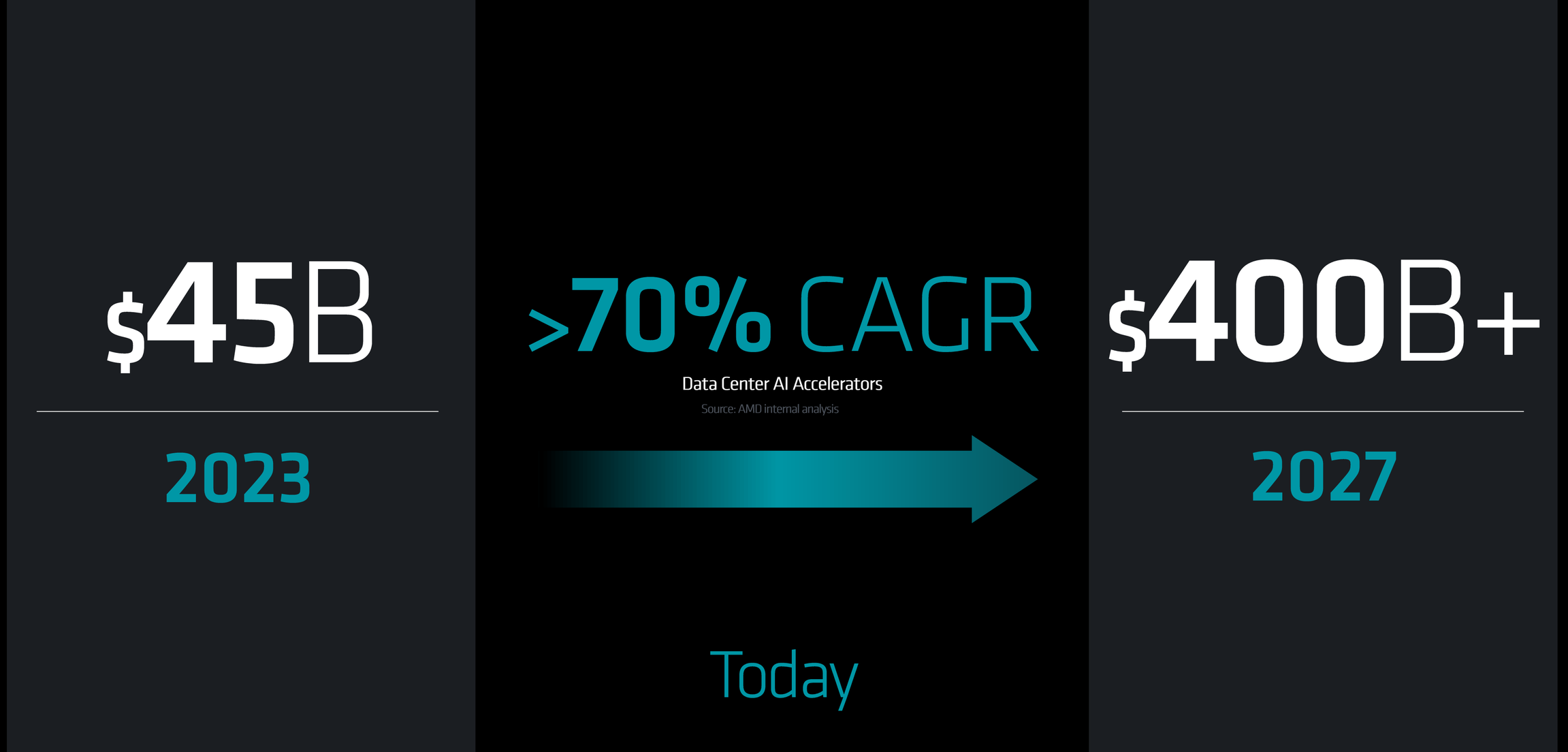
Data Center AI Accelerators

Source: AMD internal analysis

A year ago

$150B+

2027


$45B
2023
>70% CAGR
Data Center AI Accelerators
Source: AMD internal analysis
Today
$400B+
2027


AMD
Advancing end-to-end AI infrastructure
Cloud | HPC | Enterprise | Embedded | PC


Broad portfolio of training and inference compute engines


Open and proven software capabilities

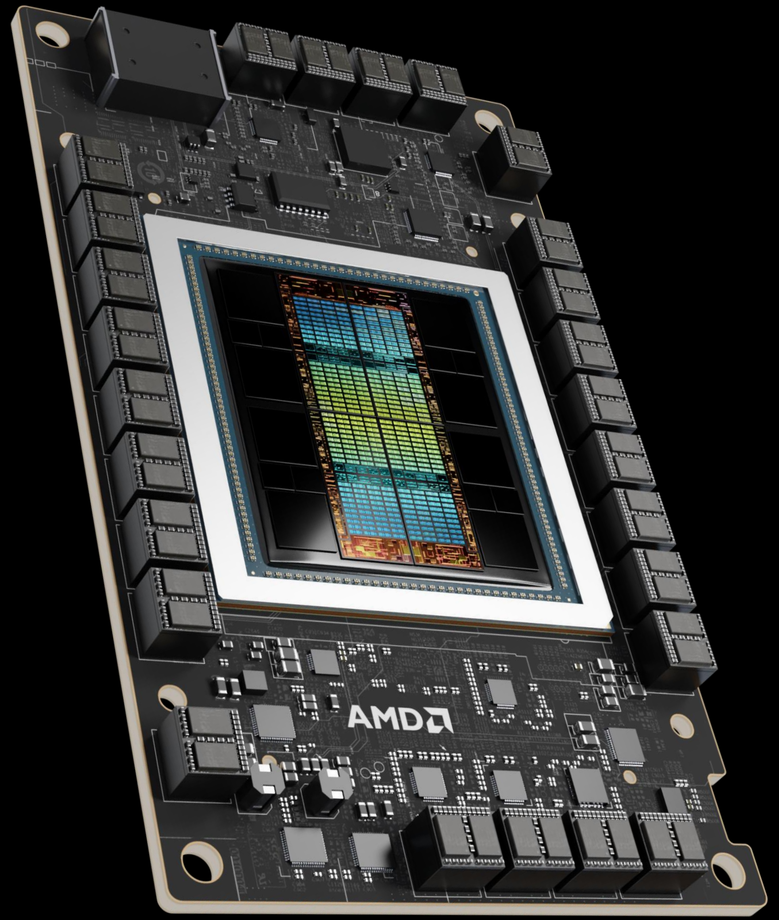

AI ecosystem with deep co-innovation

Launching Today

AI infrastructure solutions for

Cloud | Enterprise | HPC | PCs

Generative AI is the most
demanding data center workload



Launching today

AMD Instinct™ MI300X Accelerator

Advancing Generative AI

AMD
CDNA 3

3.4x FP16 | **BF16**

6.8x INT8

Up to **1.5x** more memory

1.6x memory bandwidth

Support for **TF32, FP8, Sparsity**

256 MB AMD Infinity Cache™ Technology

AMD Instinct MI300X vs AMD Instinct MI250X. See endnotes:MI300-11, MI300-13. Actual results may vary.

AMD Instinct™

MI300X Accelerator

4X IOD

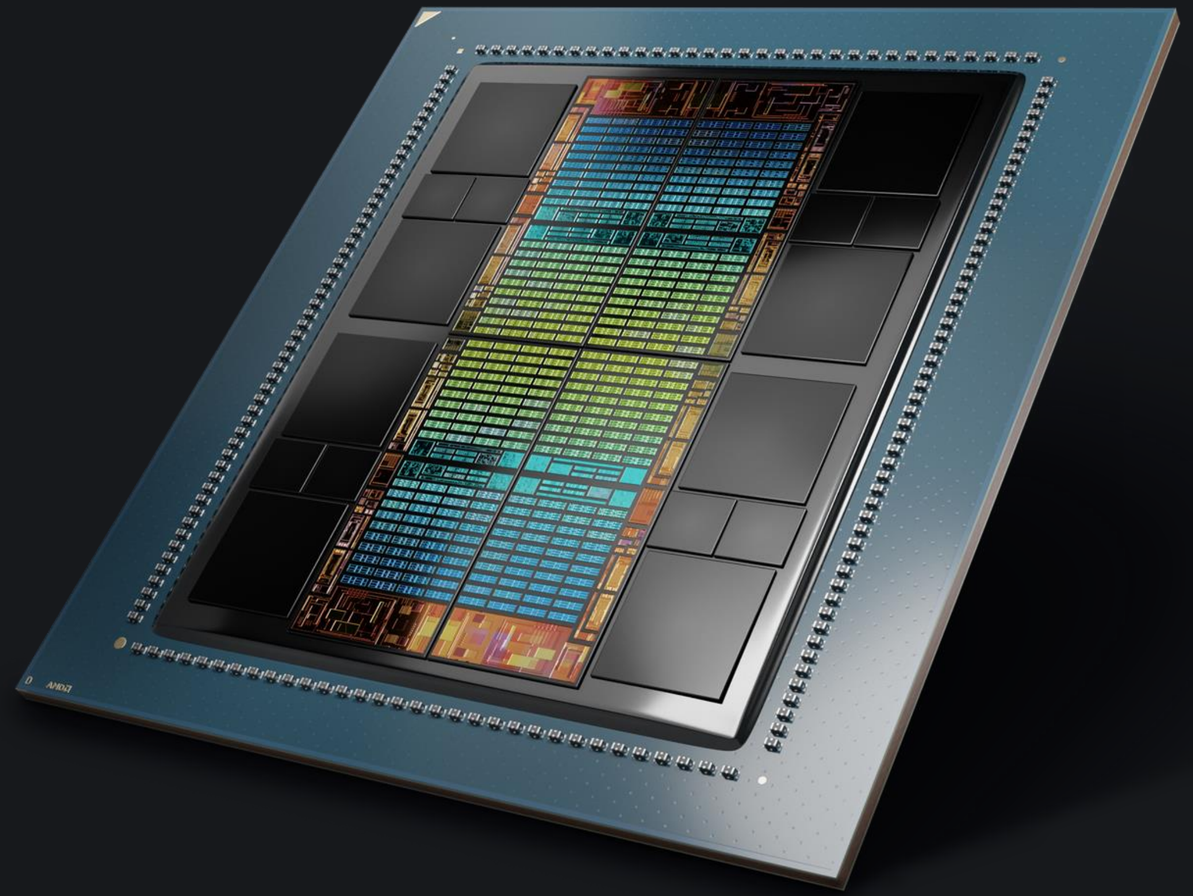
256 MB AMD INFINITY CACHE™
TECHNOLOGY NEXT GEN I/O

8X XCD

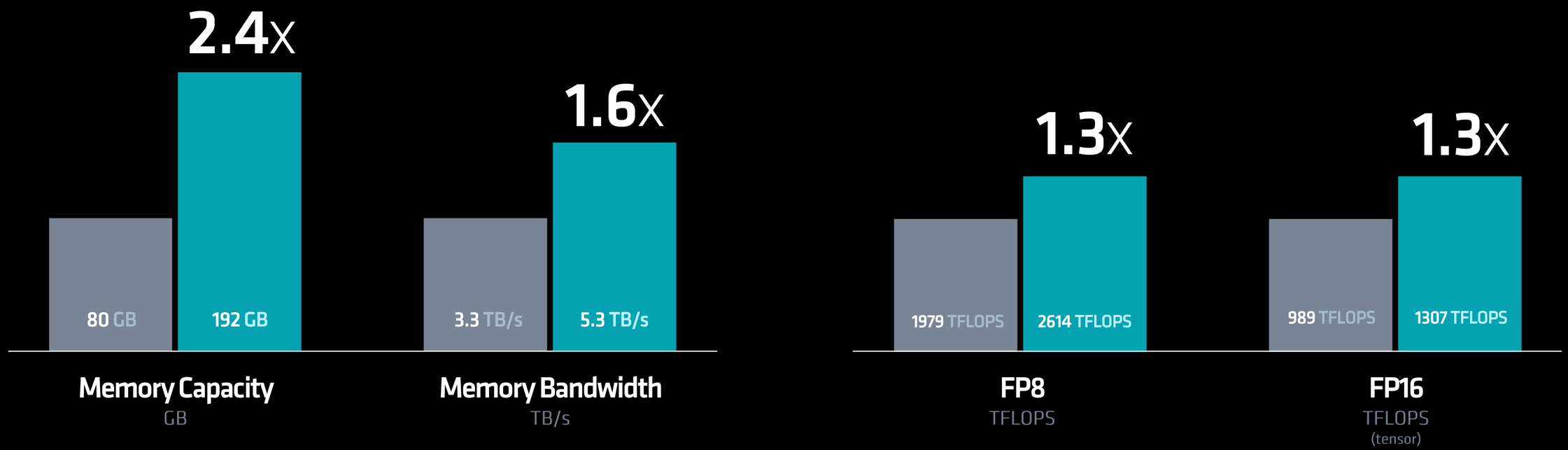
304 AMD CDNA™ 3
COMPUTE UNITS

8X HBM3

192 GB ~5.3 TB/S
PEAK BANDWIDTH



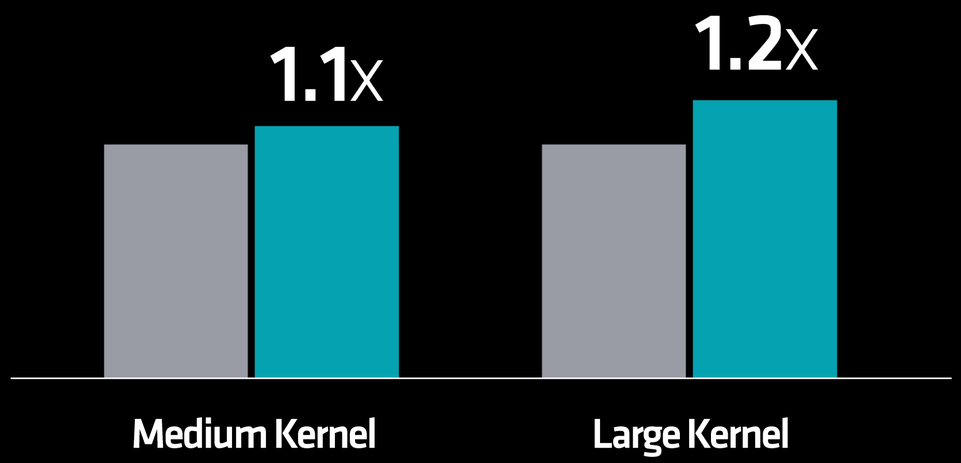
Generative AI Leadership



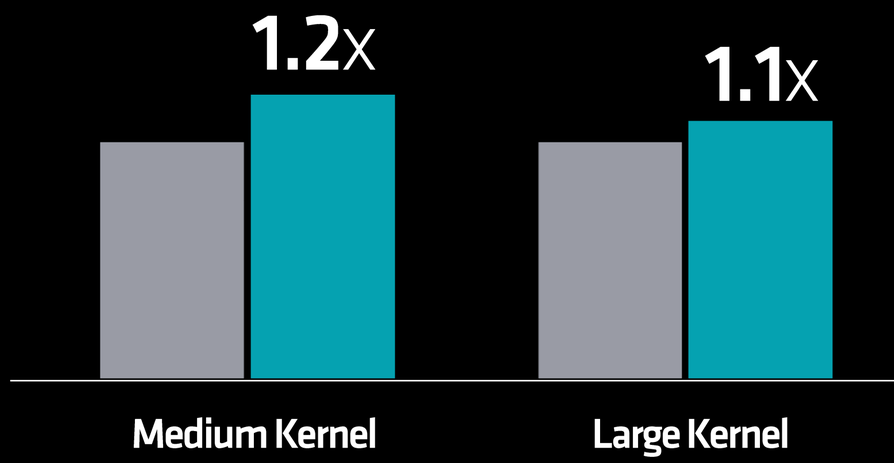
Theoretical peak See endnotes:MI300-05A, MI300-18

Key AI Kernel Performance Leadership

Common LLM kernels (TFlops)



FlashAttention-2



Llama 2-70B

Nvidia
H100

AMD Instinct™
MI300X

Results may vary. See endnotes:MI300-35, MI300-37

World-class Training Performance

Single server (8x GPU)

MPT

Model size: **30B**

Model Fine Tuning

Throughput

Tokens per second

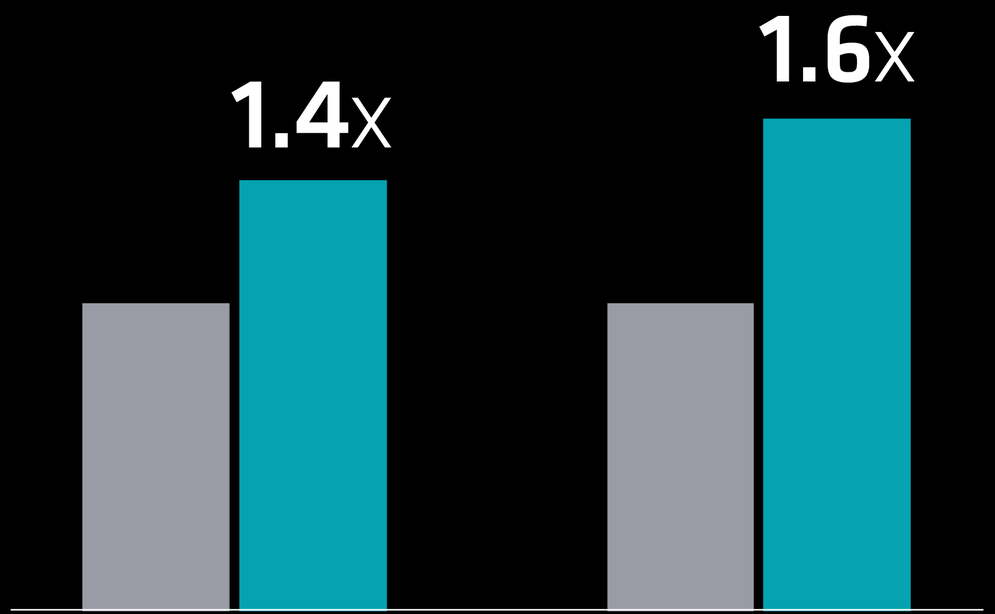
Results may vary, see endnotes:MI300-40

Inference Performance Leadership

Single server (8x GPU)

Nvidia
H100
HGX

AMD Instinct™
MI300X
Platform



Figures are median. Results may vary. See endnotes: MI300-34, MI300-38

AMD Instinct™ Platform

Industry leading generative AI platform

8x AMD Instinct MI300X Accelerator	**Leadership memory capacity**	**4th Gen AMD Infinity Fabric™ Technology**	**Industry-standard design**

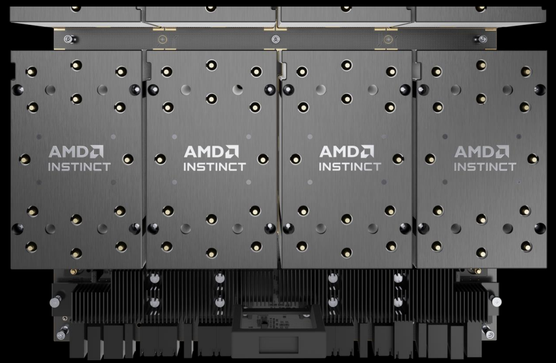
See endnotes:MI300-25

Nvidia H100 HGX	AMD Instinct™ MI300X Platform	AMD Instinct™ MI300X Platform Advantage
640 GB HBM3 memory	**1.5** TB HBM3 memory	**2.4x** more memory
7.9 PF FP16 / BF16 FLOPS	**10.4** PF FP16 / BF16 FLOPS	**1.3x** more compute
900 GB/s Aggregate bi-directional bandwidth	**896** GB/s Aggregate bi-directional bandwidth	Comparable
450 GB/s Single node ring bandwidth	**448** 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/

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	1.6X Bloom 176B
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

AMD Instinct™ MI300X

Leadership generative AI accelerator

Delivering on our software vision

Momentum in the AMD Instinct™ software ecosystem

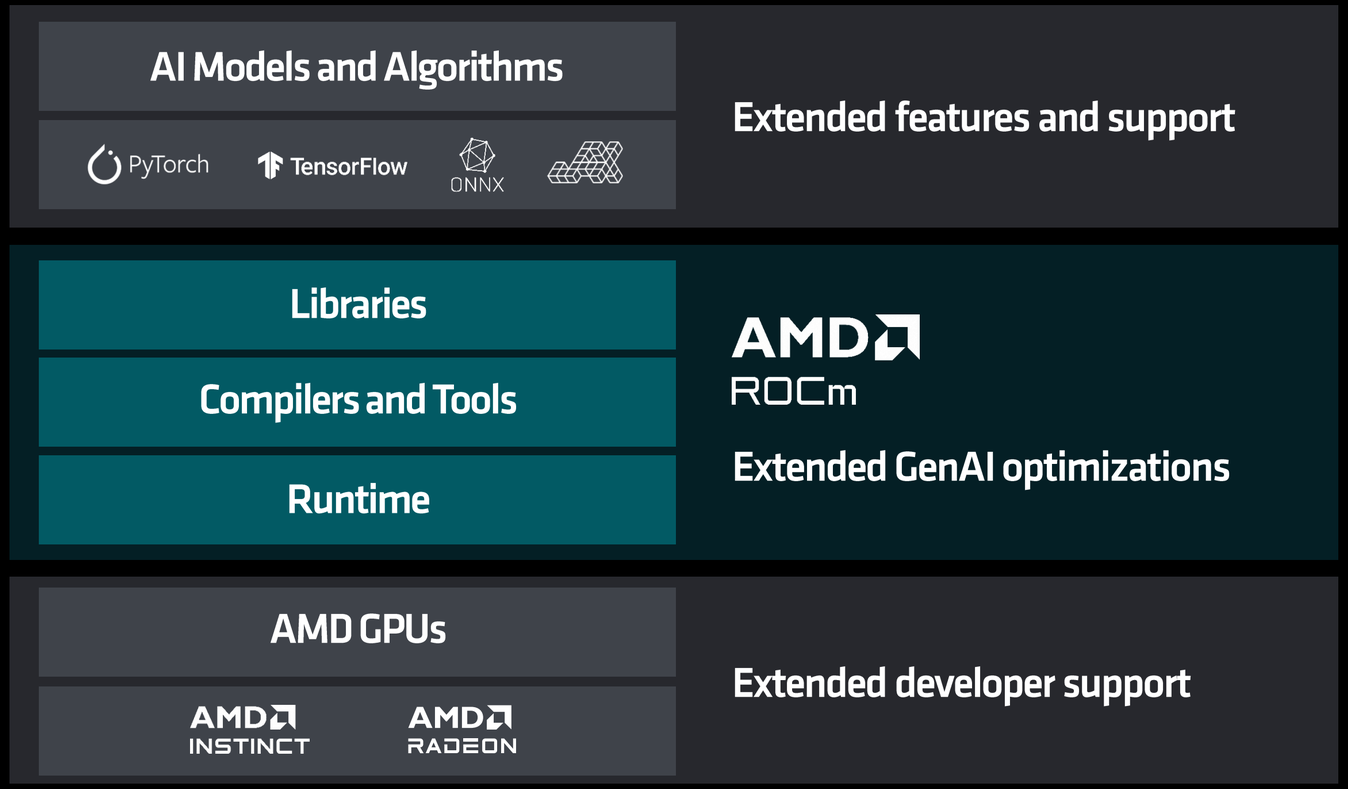
Victor Peng
President, AMD

Cloud to end point AI solutions

ROCm	ZenDNN	Vitis AI
Data center GPU	Data center CPU	Client, Embedded edge/End points

Any model | **Generative AI end to end** | **Broad AI ecosystem**

ROCm™ and software momentum



Introducing ROCm™ 6

Delivering new capabilities for generative AI

Advanced LLM optimizations

Dynamic FP16, BF16, FP8 utilization | Advanced attention algorithms and kernels | Collective communications | HIPGraph

Performant AI libraries

Compute and bandwidth operators | Structured sparsity | Quantization libraries

Expanded ecosystem support

Frameworks | Models | ML pipelines

LLM performance optimizations

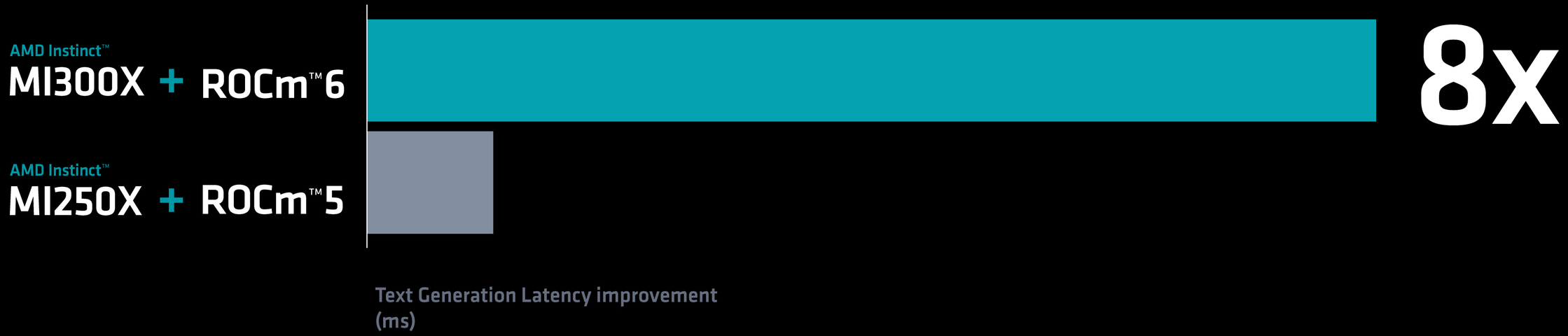
AMD ROCm™ feature improvements

2.6X	1.4X	1.3X
vLLM Optimized Inference Libraries	**HIP Graph** Optimized Runtime	**Flash Attention** Optimized Kernels

See endnotes: MI300-44 Results may vary.
Geomean performance uplifts over baseline

AMD Instinct™ GPU + ROCm™ Platform Gen on Gen Performance

Llama 2 70B inference



Results may vary. See endnotes: MI300-33

Llama 2

Model size: **13B**

Inference Performance Leadership

Single GPU

Chat Latency improvement
(ms)

Results may vary. See endnotes: MI300-36

Expanding software investment and the ecosystem

Strengthening software capabilities

Enhancing developer experience

nod	Open source AI compiler
Mipsology	High efficiency inference

Strong developer ecosystem momentum

Hugging Face

62,000+ models running nightly
Fully integrated optimum library

PyTorch

From ‘port-to’ to ‘develop-on’
with latest platforms


Tensor Flow


Dynamo Inductor


JAX


OpenAI Triton


ONNX Runtime


OpenXLA


DeepSpeed


MLIR | IREE

Increasing open-source contributions
and expanding footprint

"OpenAI is working with AMD in support of an open ecosystem. We plan to support AMD's GPUs including MI300 in the standard Triton distribution starting with the upcoming 3.0 release."

Philippe Tillet, OpenAI

AI Innovators

databricks

Ion Stoica

Co-Founder and Executive Chairman

MosaicML's Generative AI platform for enterprise AI development

essential AI

Ashish Vaswani

Co-Founder and CEO

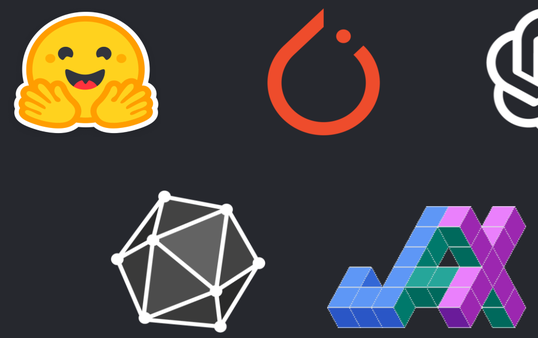
Essential's mission is to accelerate human-machine collaboration for enterprises

LAMINI

Sharon Zhou

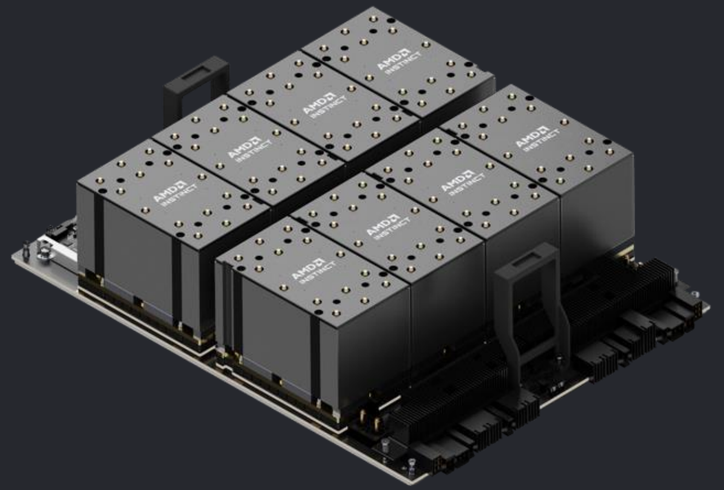
Founder and CEO

Enterprise LLM platform for building customized private models



Expanding ecosystem

Hugging Face, PyTorch, Jax, OAI Triton, ONNX and more



MI300X + ROCm™ 6

Delivering leadership generative AI performance

Advancing AI

Customer Momentum

Dr. Lisa Su
Chair and CEO, AMD

Advancing AI

OEM Partner Innovation

Launching Today

AI infrastructure solutions for

Cloud | **Enterprise** | HPC | PCs

AMD Instinct™ MI300X Accelerator

OEM and solution partners

Broadening infrastructure choice

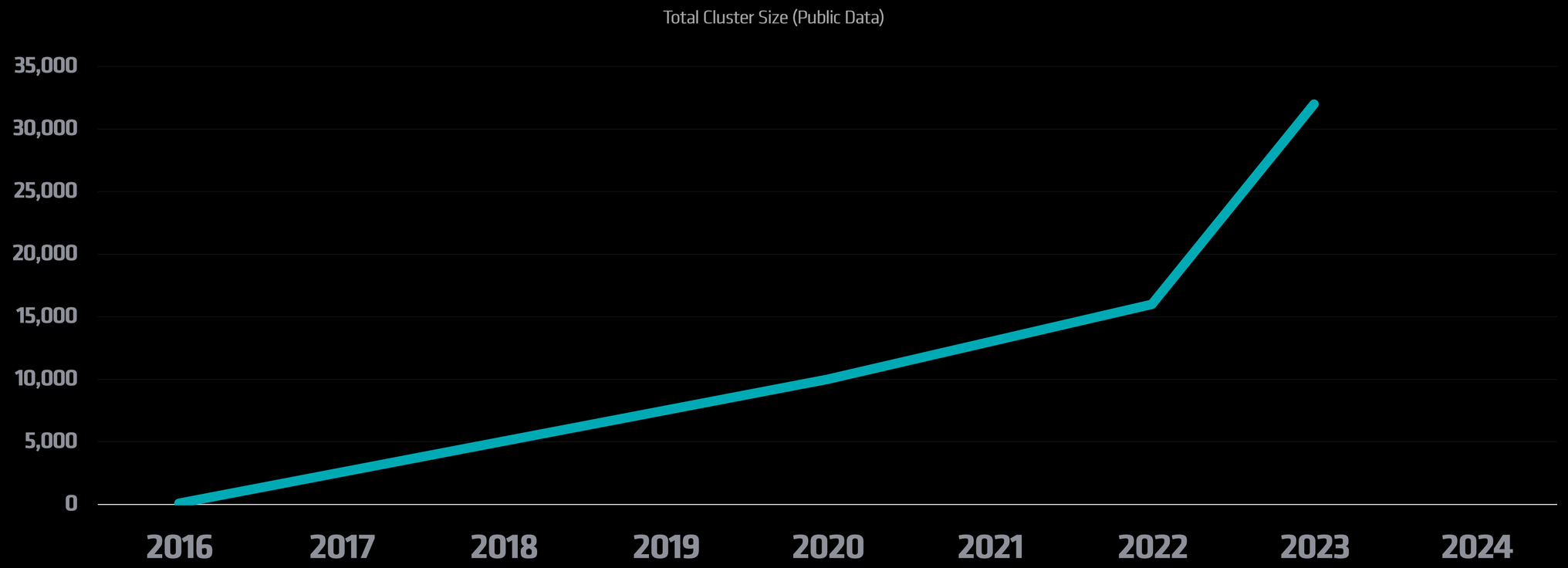
Greater AMD Instinct™ access

Open networking for AI infrastructure

Forrest Norrod
EVP and GM
Data Center Solutions Business Group

AI performance needs driving cluster growth

Number of GPUs required to train models



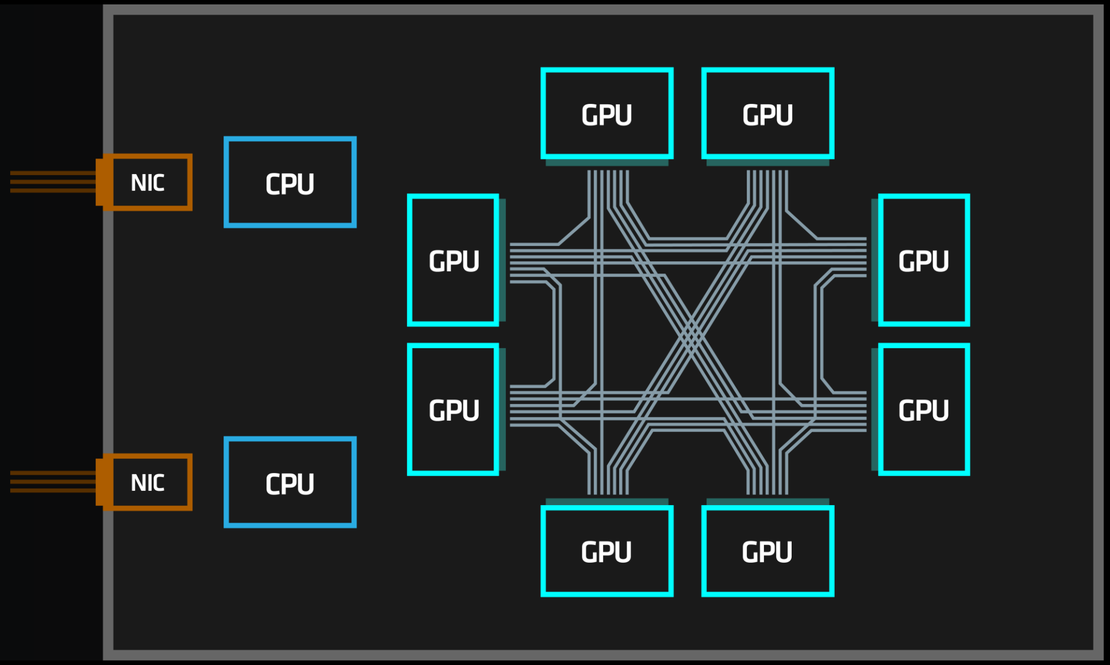
Data from published sources

Scaling within the server

Tightly interconnected GPUs in each server

Front-end network

Data center connection



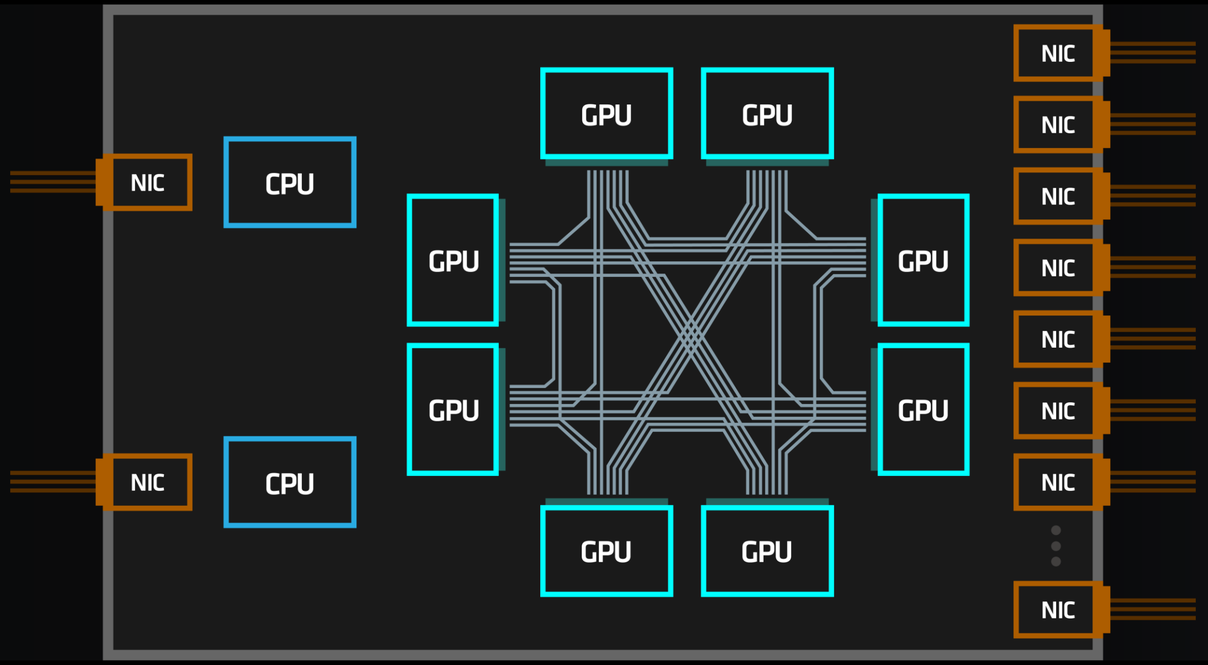
extends access to the AMD Infinity Fabric™ ecosystem to innovators and strategic partners

Scaling beyond the server

Extending the GPU workspace

Front-end network

Data center connection

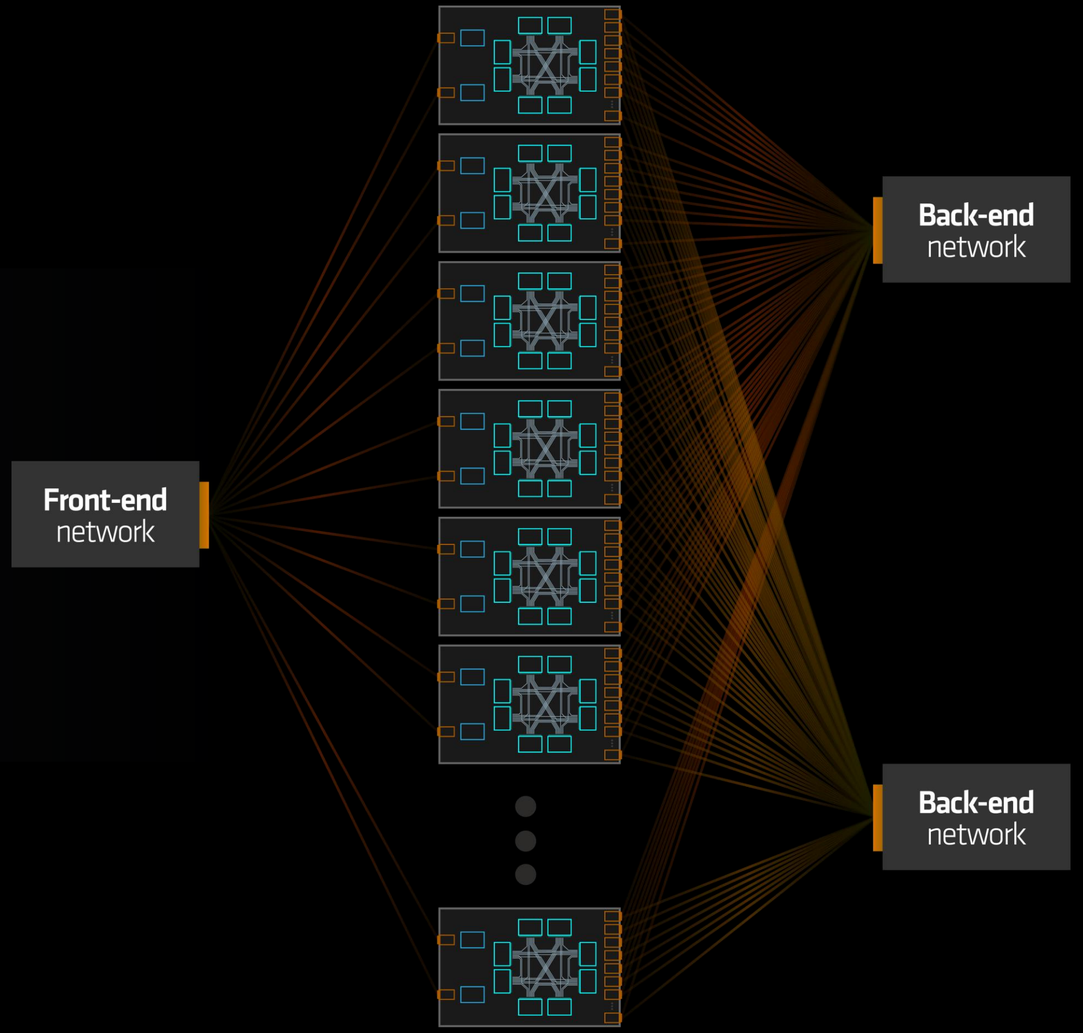


Back-end network

Cluster scale-out

The network is critical

High performance, scalable, open



Ethernet is the answer

High performance	Scalable	Open

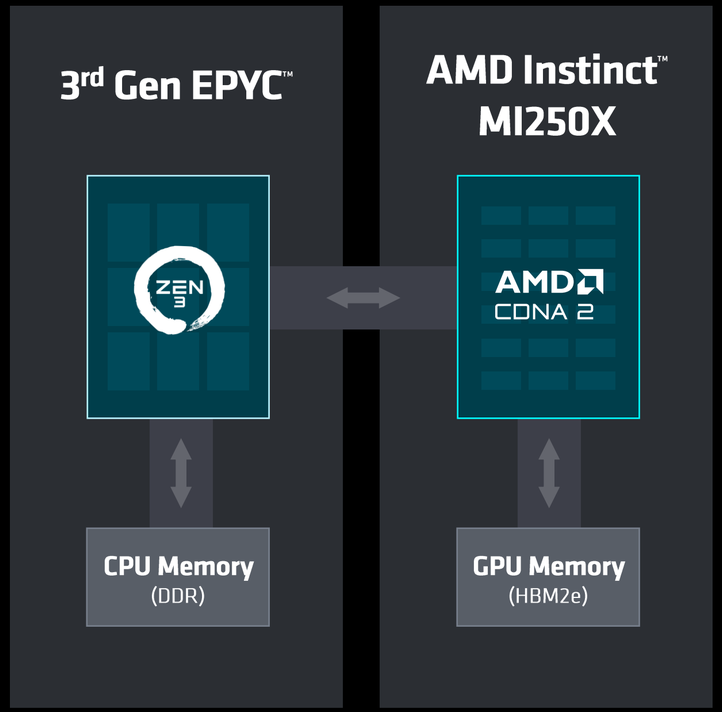
AI networking leaders

Launching Today

Open AI solutions for

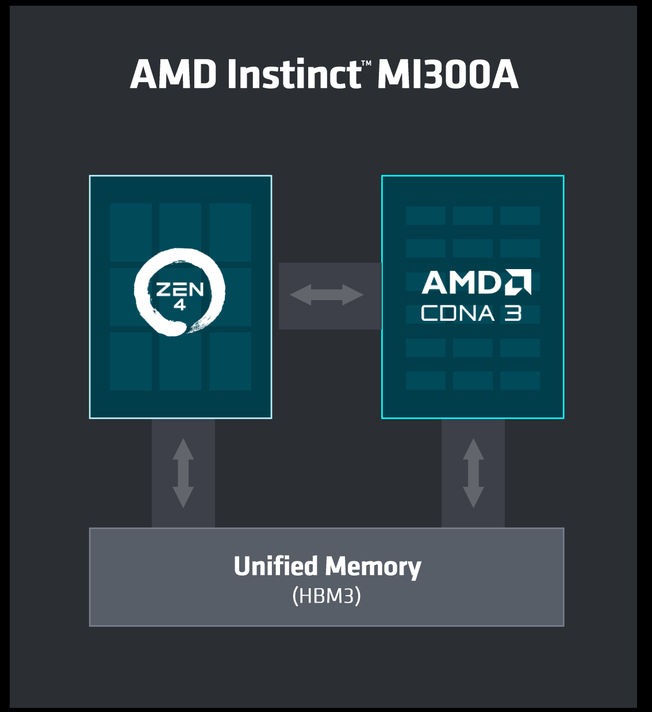
Cloud | Enterprise | **HPC** | PCs

3rd Generation
Infinity Architecture

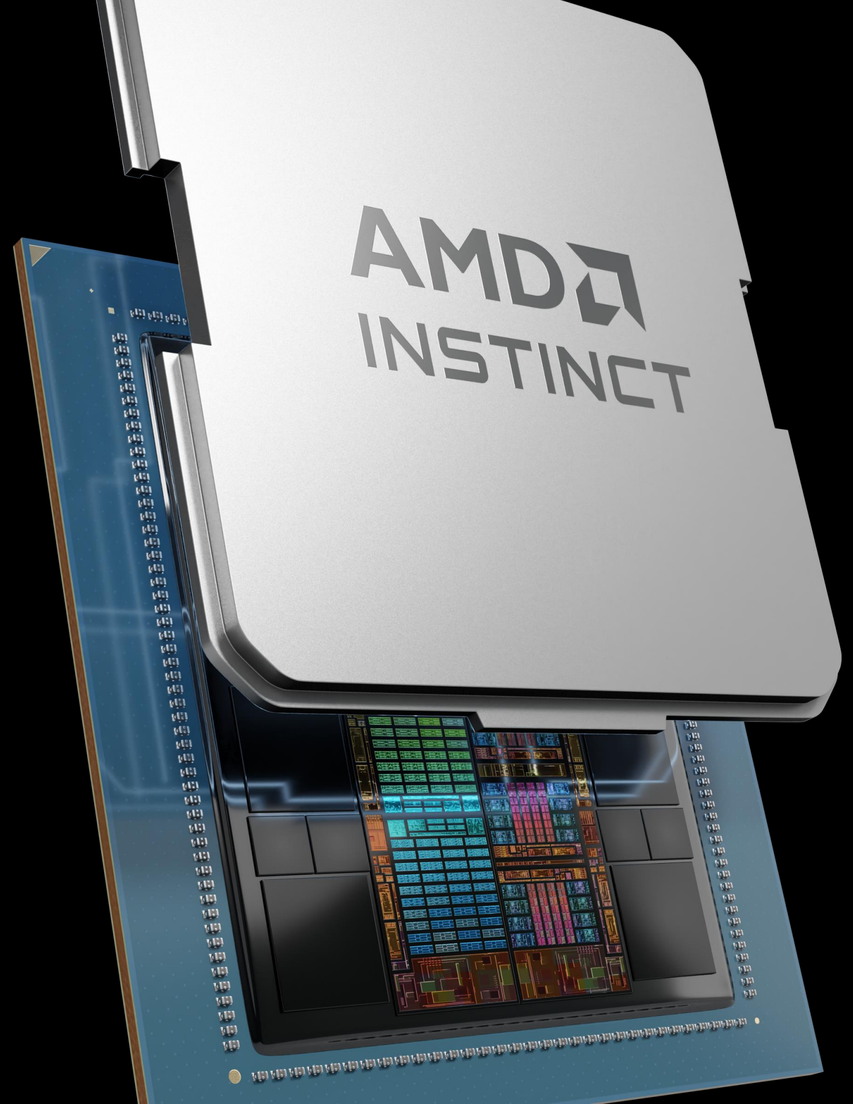


2021

4th Generation
Infinity Architecture



2023



In Volume Production

AMD Instinct™ MI300A Accelerator

World's first data center APU for HPC and AI

AMD Instinct™

MI300A Accelerator

4X IOD

256 MB AMD INFINITY CACHE™
TECHNOLOGY NEXT GEN I/O

6X XCD

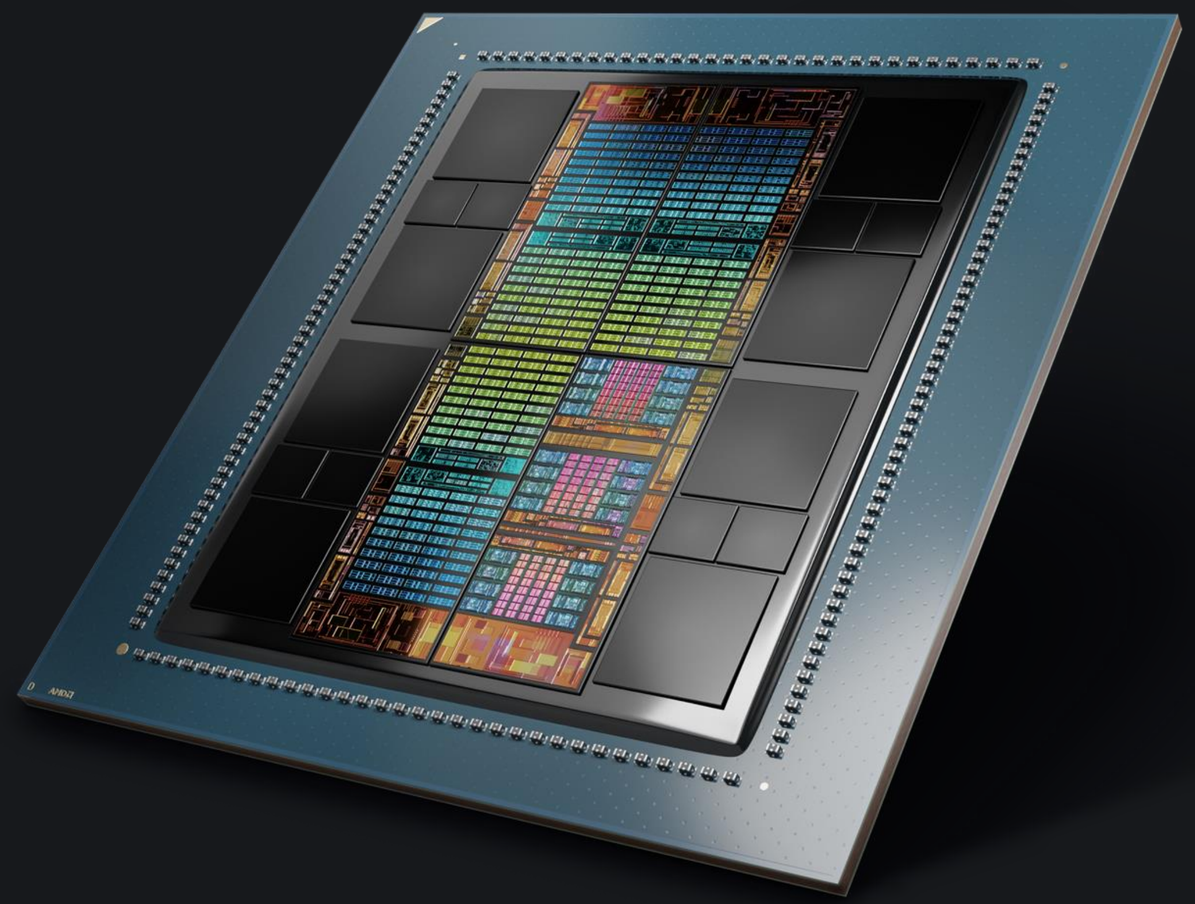
228 AMD CDNA™ 3 COMPUTE UNITS

3X CCD

24 “ZEN 4” X86 CORES

8X HBM3

128 GB ~5.3 TB/S PEAK BANDWIDTH



APU advantage

Unlocking new performance capabilities

Unified memory	AMD Infinity Cache™ Technology	Dynamic power sharing	Streamlined programming

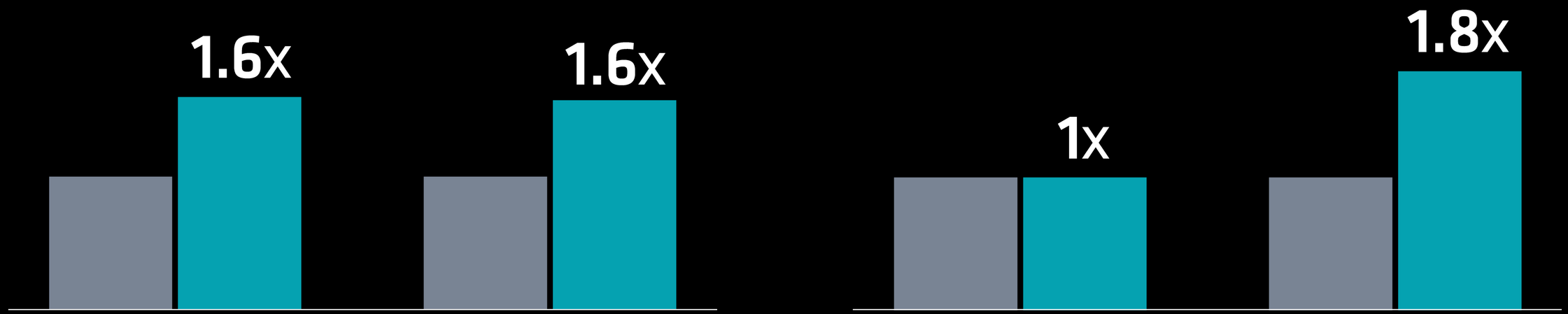
AMD Instinct™ MI300A

World's first data center APU for HPC and AI

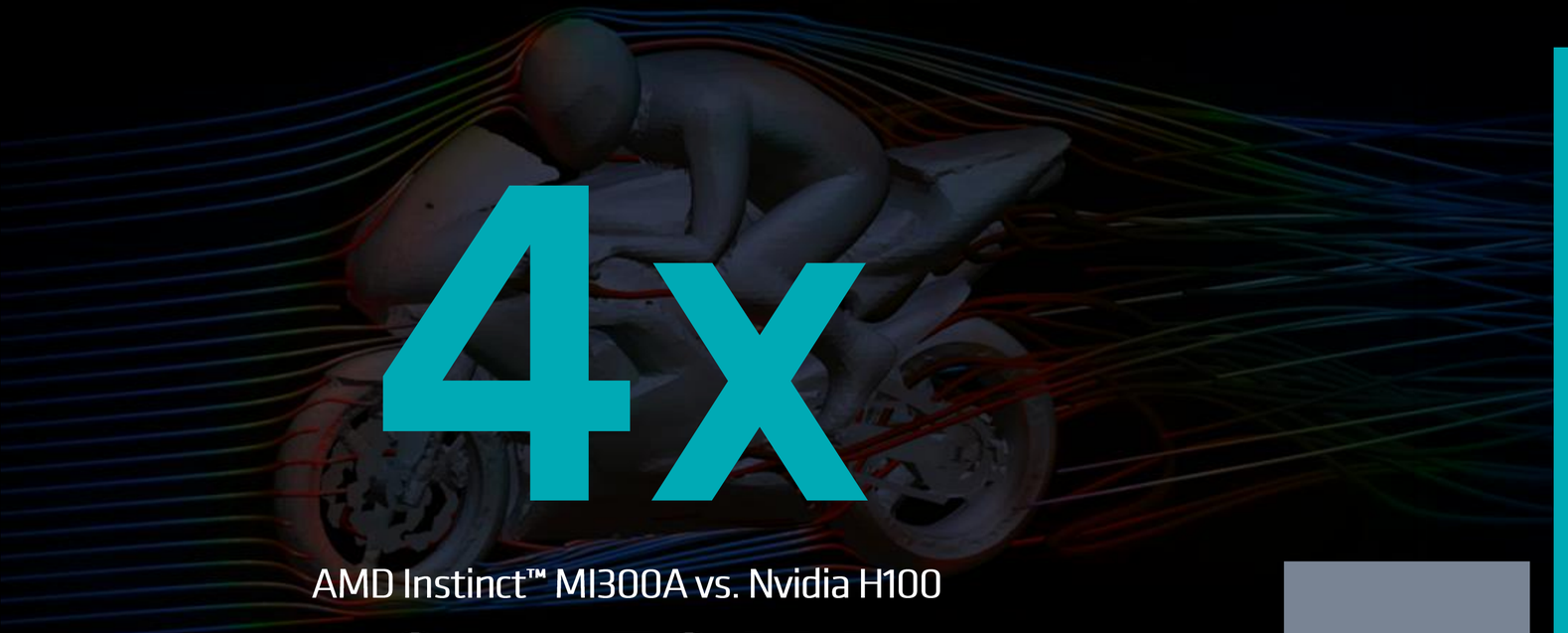
61 TF	122 TF	128 GB	~5.3 TB/s	146B
FP64 (Peak)	FP32 (Peak)	HBM3	Memory Bandwidth (Peak)	Transistors

See endnotes: MI300-10, MI300-14

Leadership Performance



Theoretical peak. See endnotes: MI300-12, MI300-20. *NVIDIA H100 GPUs don't support FP32 Tensor cores.

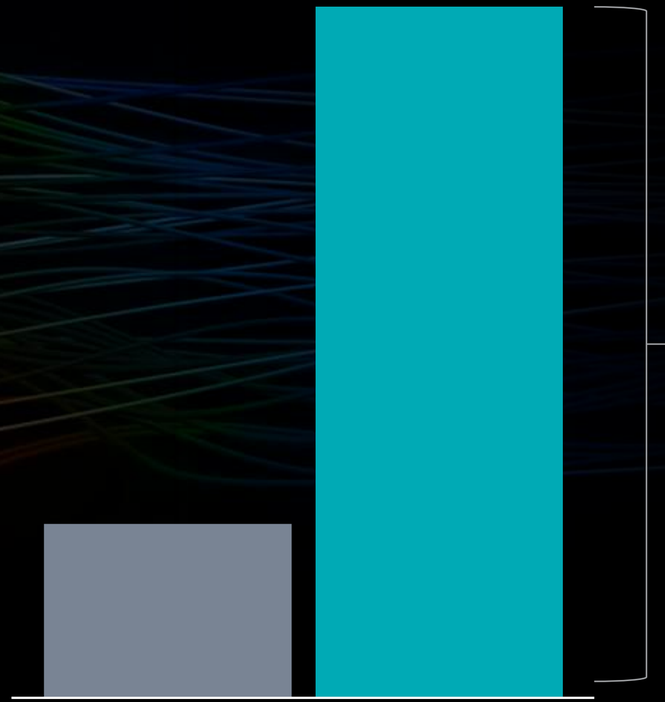


4X

AMD Instinct™ MI300A vs. Nvidia H100

OpenFOAM

HPC Motorbike



unified memory
+
memory capacity & bandwidth
+
GPU performance

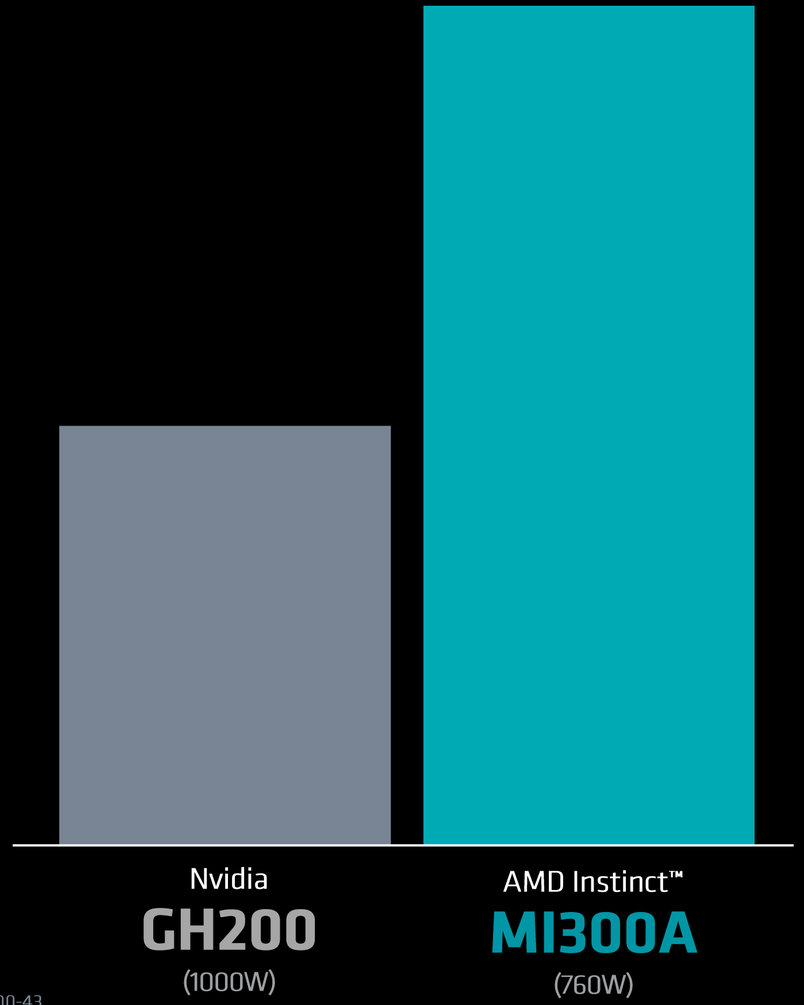
Illustrative use only,

Results may vary. See endnotes: MI300-32

2x

AMD Instinct™ MI300A vs. GH200

Peak HPC Performance per watt



Results may vary. See endnotes: MI300-43

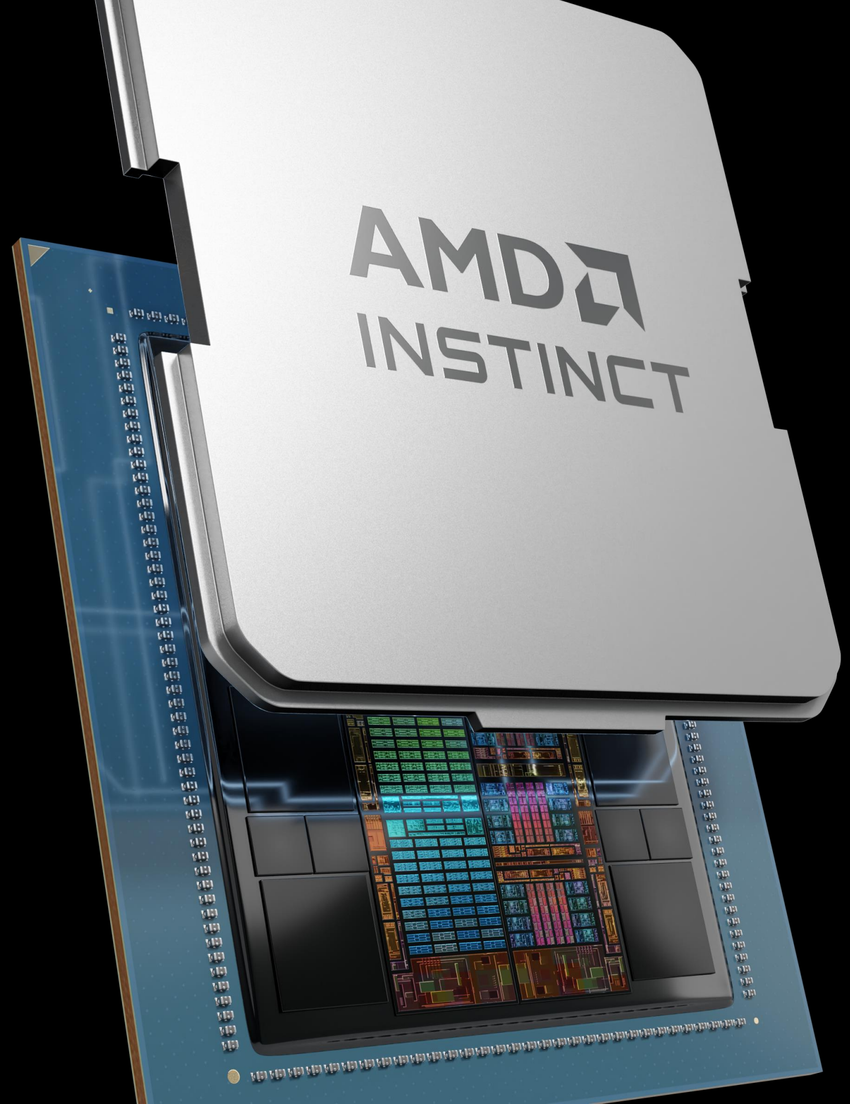
Advancing research

Accelerating the convergence of AI and HPC

MI300A enables deep learning with CosmoFlow to learn the universe at scale

El Capitan

Expected to be the World's first two-exaflop supercomputer



AMD Instinct™ MI300A

A new level of high-performance leadership

AMD Instinct™ MI300A APU

OEM and Solution Partners

Hewlett Packard Enterprise

EVIDEN an atos business

GIGABYTE™

SUPERMICRO

Advancing AI PCs

Dr. Lisa Su
Chair and CEO, AMD

Launching Today

AI infrastructure solutions for

Cloud | Enterprise | HPC | PCs

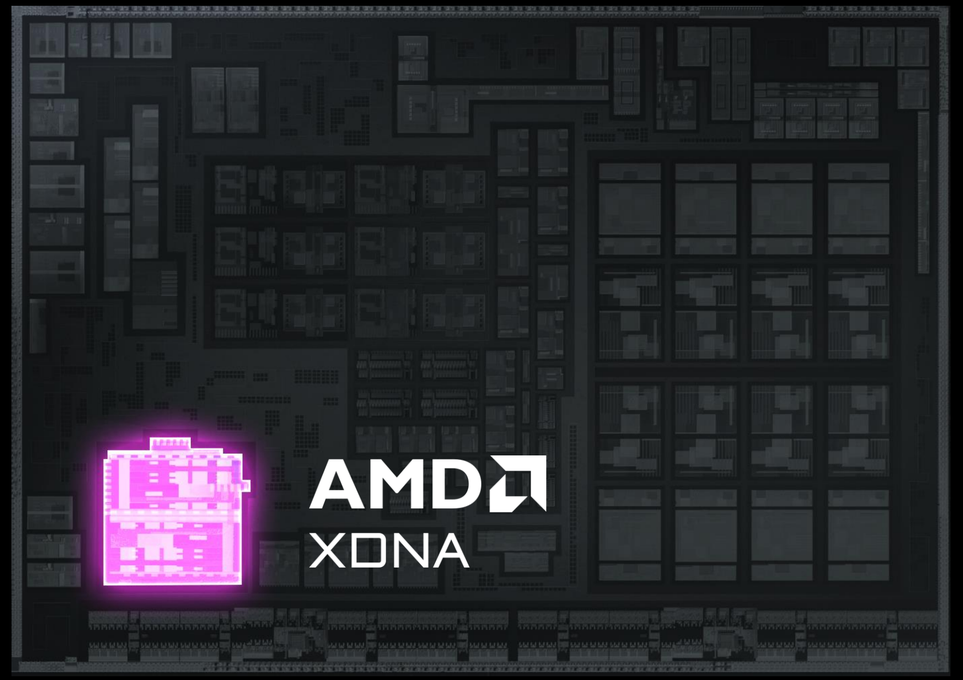
AMD RYZEN AI

Dedicated NPUs reimagine the PC to enable a truly intelligent and personal experience

AMD
RYZEN AI

Leads the AI PC era

World's first x86 processor with integrated NPU



AMD Ryzen™ Mobile 7040 Series Processors

See endnotes: PHX-3

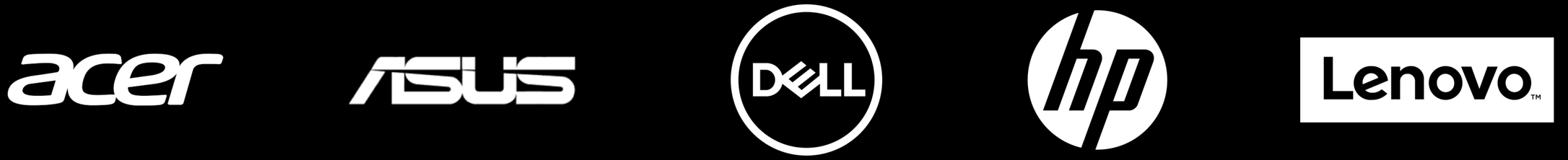
AMD XDNA

Adaptive AI architecture

Integrated NPU with AMD Ryzen™ 7040 Series Processors

Scalable NPU architecture	Integrated on-die for lower latency	High performance and energy efficient	Up to 4 concurrent dedicated AI streams

Millions of Ryzen™ AI PCs shipped in 2023



AMD Ryzen™ AI PC
Microsoft and ISV solutions ecosystem

100+ AI-powered experiences

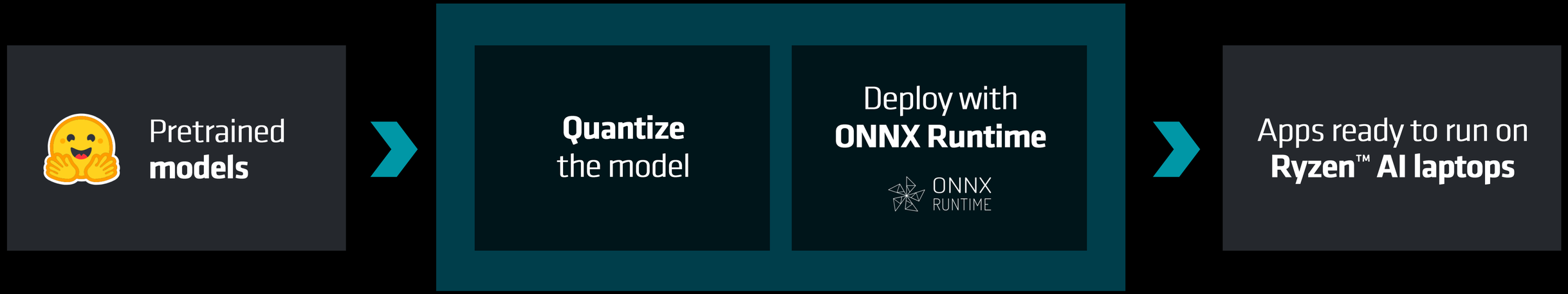
Adobe, Black Magic, Topaz Labs and more for AMD Ryzen™ AI compatible PCs

First to reach market scale with NPU-powered Windows 11 Studio Effects

AMD Ryzen™ AI software accelerates generative AI adoption on AI PCs

AMD Ryzen™ AI 1.0 Software

Quick and easy to deploy. Get started in minutes.



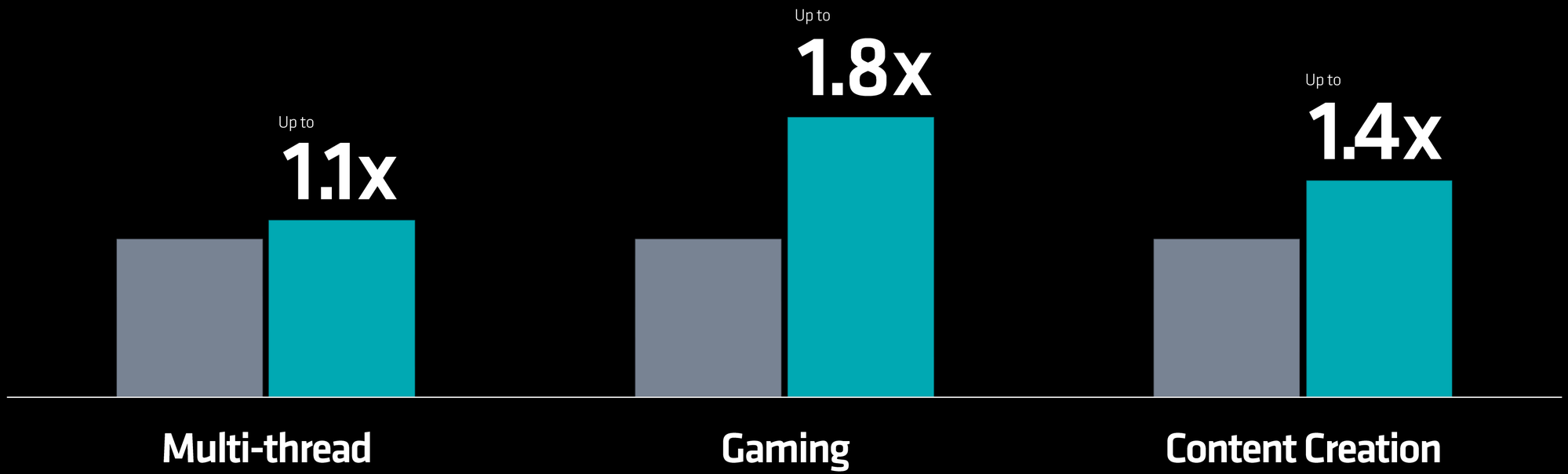
Now Available

“Hawk Point” AI PC processors now shipping

AMD Ryzen™ 8040 Series Processors

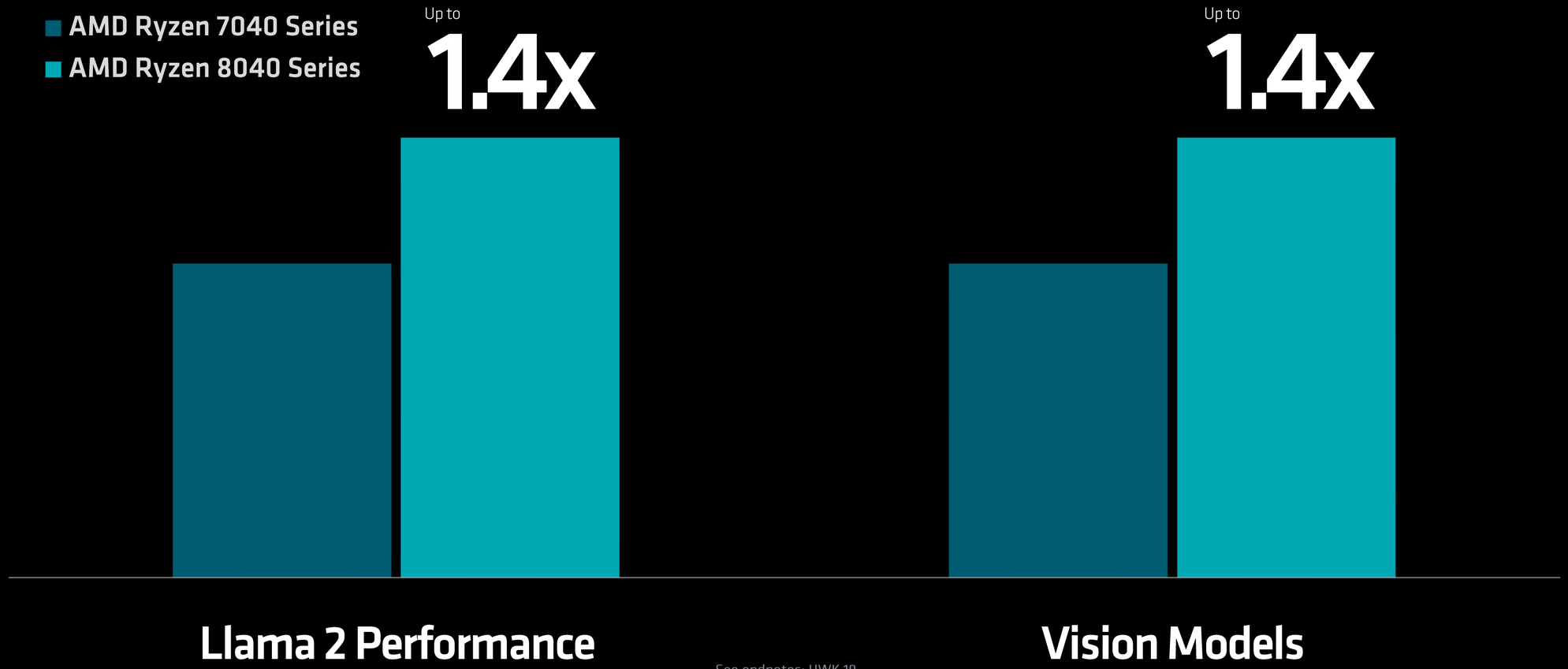
ZEN 4	AMD RDNA 3	AMD XDNA	Up to 39 TOPS
8 Core | 16 Threads	Graphics	16 TOPS	Total processor performance

AMD Ryzen™ 8040 Series processors performance leadership



See endnotes: HWK-01, HWK-02, and HWK-03

Ryzen™ 8040 Series NPU Performance Uplift

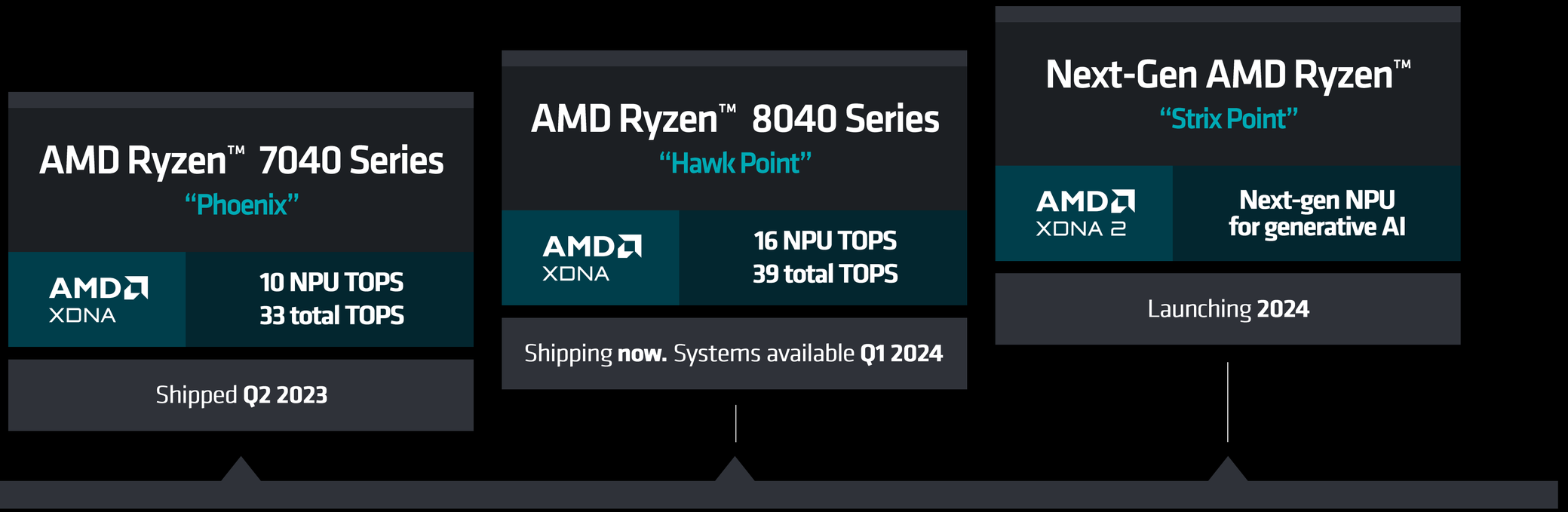


See endnotes: HWK-19
Vision models are based on ResNet-50, Yolov3, ESRGAN

Next-gen AMD Ryzen™ AI processors

Ready for generative AI in the PC

AMD Ryzen™ AI Roadmap



Roadmap Subject to change.

AMD XDNA 2

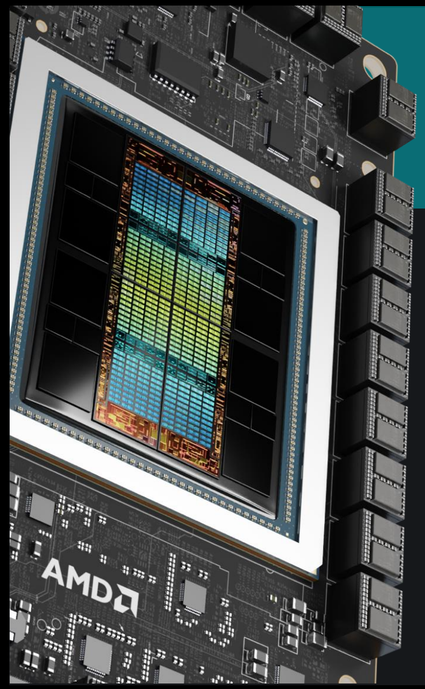
Designed for more than 3x generative AI NPU performance

(as compared to the previous generation)

Coming with “Strix Point” AMD Ryzen™ AI processors in 2024

See endnotes: STX-1

Advancing end-to-end AI infrastructure



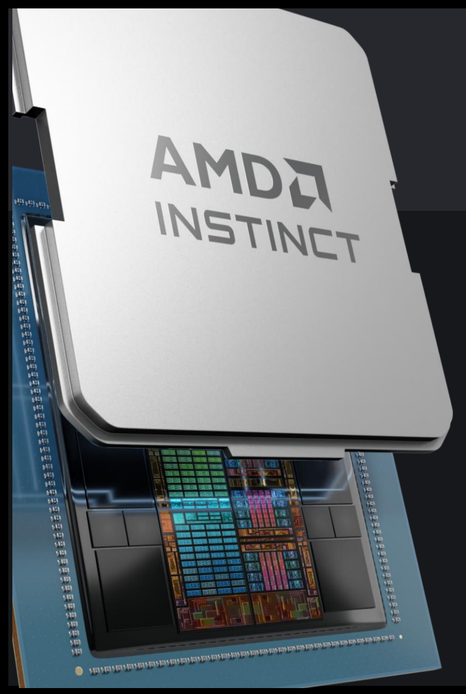
AMD Instinct™

MI300X

Shipping Today

Broad AI Solutions Ecosystem

Microsoft · Meta · ORACLE · DELL Technologies · Hewlett Packard Enterprise · Lenovo · SUPERMICRO · ASUS · GIGABYTE · Ingrasys · Inventec · QCT · Wistron · wiwynn · Aligned · ARKON ENERGY · Cirrascale · Crusoe · DENVR dataworks · TENSORWAVE · CISCO · BROADCOM · ARISTA



AMD Instinct™

MI300A

In Volume Production

Hewlett Packard Enterprise · EVIDEN an atos business · GIGABYTE · SUPERMICRO

Powering LLNL

El Capitan

AMD ROCm

ROCm™ 6

Available December

AI Software Ecosystem and Innovators

databricks · DeepLearning.AI · essential AI · IBM · LAMINI · ONNX · PyTorch · TensorFlow

AMD Ryzen™ 8040 Series

Shipping Now

ASUS · Windows 11 · hp · acer · Lenovo · DELL


AMD

DISCLAIMER

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

-

ENDNOTES

- MI300-05A: Calculations conducted by AMD Performance Labs as of November 17, 2023, for the AMD Instinct™ MI300X OAM accelerator 750W (192 GB HBM3) designed with AMD CDNA™ 3 5nm FinFet process technology resulted in 192 GB HBM3 memory capacity and 5.325 TFLOPS peak theoretical memory bandwidth performance. MI300X memory bus interface is 8,192 and memory data rate is 5.2 Gbps for total peak memory bandwidth of 5.325 TB/s (8,192 bits memory bus interface * 5.2 Gbps memory data rate/8). The highest published results on the NVidia Hopper H100 (80GB) SXM5 GPU accelerator resulted in 80GB HBM3 memory capacity and 3.35 TB/s GPU memory bandwidth performance. https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet
- MI300-10: Measurements conducted by AMD Performance Labs as of November 11th, 2023 on the AMD Instinct™ MI300A (760W) APU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 122.6 TFLOPS peak theoretical double precision (FP64 Matrix), 61.3 TFLOPS peak theoretical double precision (FP64), 122.6 TFLOPS peak theoretical single precision matrix (FP32 Matrix), 122.6 TFLOPS peak theoretical single precision (FP32), 490.3 TFLOPS peak theoretical TensorFloat-32 (TF32), 980.6 TFLOPS peak theoretical half precision (FP16), 980.6 TFLOPS peak theoretical Bfloat16 format precision (BF16), 1961.2 TFLOPS peak theoretical 8-bit precision (FP8), 1961.2 TOPs INT8 floating-point performance. The results calculated for the AMD Instinct™ MI250X (560W) 128GB HBM2e OAM accelerator designed with AMD CDNA™ 2 5nm FinFET process technology at 1,700 MHz peak boost engine clock resulted in 95.7 TFLOPS peak theoretical double precision (FP64 Matrix), 47.9 TFLOPS peak theoretical double precision (FP64), 95.7 TFLOPS peak theoretical single precision matrix (FP32 Matrix), 47.9 TFLOPS peak theoretical single precision (FP32), TF32* (N/A), 383.0 TFLOPS peak theoretical half precision (FP16), 383.0 TFLOPS peak theoretical Bfloat16 format precision (BF16), FP8* (N/A), 383.0 TOPs INT8 floating-point performance.Server manufacturers may vary configuration offerings yielding different results. * MI200 Series GPUs don’t support TF32, FP8 or sparsity
- MI300-11: Measurements conducted by AMD Performance Labs as of November 11th, 2023 on the AMD Instinct™ MI300X (750W) GPU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 163.4 TFLOPS peak theoretical double precision (FP64 Matrix), 81.7 TFLOPS peak theoretical double precision (FP64), 163.4 TFLOPS peak theoretical single precision matrix (FP32 Matrix), 163.4 TFLOPS peak theoretical single precision (FP32), 653.7 TFLOPS peak theoretical TensorFloat-32 (TF32), 1307.4 TFLOPS peak theoretical half precision (FP16), 1307.4 TFLOPS peak theoretical Bfloat16 format precision (BF16), 2614.9 TFLOPS peak theoretical 8-bit precision (FP8), 2614.9 TOPs INT8 floating-point performance. The results calculated for the AMD Instinct™ MI250X (560W) 128GB HBM2e OAM accelerator designed with AMD CDNA™ 2 5nm FinFET process technology at 1,700 MHz peak boost engine clock resulted in 95.7 TFLOPS peak theoretical double precision (FP64 Matrix), 47.9 TFLOPS peak theoretical double precision (FP64), 95.7 TFLOPS peak theoretical single precision matrix (FP32 Matrix), 47.9 TFLOPS peak theoretical single precision (FP32), TF32 (N/A), 383.0 TFLOPS peak theoretical half precision (FP16), 383.0 TFLOPS peak theoretical Bfloat16 format precision (BF16), FP8 (N/A), 383.0 TOPs INT8 floating-point performance.AMD TFLOPS and TOPS calculations conducted with the following equation for AMD Instinct MI300X and MI250X GPUs: FLOPS calculations are performed by taking the engine clock from the highest DPM state and multiplying it by xx CUs per GPU. Then, multiplying that number by xx stream processors, which exist in each CU. Then, for MI300X that number is multiplied by 256 FLOPS per clock/CU for FP64 Matrix to determine TFLOPS, 128 FLOPS per clock/CU for FP64 to determine TFLOPS, 256 FLOPS per clock/CU for FP32 Matrix to determine TFLOPS, 256 FLOPS per clock/CU for FP32 to determine TFLOPS, 1024 FLOPS per clock/CU for TF32 to determine TFLOPS, 2048 FLOPS per clock/CU for FP16 to determine TFLOPS, 2048 FLOPS per clock/CU for BF16 to determine TFLOPS, 4096 FLOPS per clock/CU for FP8 to determine TFLOPS, 4046 FLOPS per clock/CU for INT8 to determine TOPS. Divide results by 100,000 to get TFLOPS.Then, for MI250X that number is multiplied by 256 FLOPS per clock/CU for FP64 Matrix to determine TFLOPS, 128 FLOPS per clock/CU for FP64 to determine TFLOPS, 256 FLOPS per clock/CU for FP32 Matrix to determine TFLOPS, 128 FLOPS per clock/CU for FP32 to determine TFLOPS, 1024 FLOPS per clock/CU for FP16 to determine TFLOPS, 1024 FLOPS per clock/CU for BF16 to determine TFLOPS. 1024 FLOPS per clock/CU for INT8 to determine TOPS. Divide results by 100,000 to get TFLOPS (TF32, FP8 Not Available). Divide results by 100,000 to get TFLOPS.Calculations:FP64 Matrix: 163.4 / 95.7 = 1.71x (71% faster)FP64: 81.7 / 47.9 = 1.71x (71% faster)FP32 Matrix: 163.4 / 95.7 = 1.71x (71% faster)FP32: 163.4 / 47.9 = 3.41x (241% faster)FP16: 1307.4 / 383.0 = 3.41x (241% faster)BF16: 1307.4 / 383.0 = 3.41x (241% faster)FP8: 2614.9 (FP8) / 383.0 (FP16) = 6.83x (583% faster)*INT8: 2614.9 / 383.0 = 6.83x (583% faster) * MI200 Series GPUs don’t support TF32, FP8 or sparsity
- MI300-12: Calculations conducted by AMD Performance Labs as of November 7, 2023, for the AMD Instinct™ MI300A APU accelerator 760W (128 GB HBM3) designed with AMD CDNA™ 3 5nm FinFet process technology resulted in 128 GB HBM3 memory capacity and 5.325 TFLOPS peak theoretical memory bandwidth performance. MI300A memory bus interface is 8,192 (1024 bits x 8 die) and memory data rate is 5.2 Gbps for total peak memory bandwidth of 5.325 TB/s (8,192 bits memory bus interface * 5.2 Gbps memory data rate/8). The highest published results on the NVidia Hopper H100 (80GB) SXM GPU accelerator resulted in 80GB HBM3 memory capacity and 3.35 TB/s GPU memory bandwidth performance. https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet Memory Capacity:MI300A APU: 128GB HBM3 / H100 SXM5: 80GB HBM3 = 1.6X (60% more)Memory Bandwidth: MI300A OAM: 5.325 TB/s / H100 SXM5: 3.352 TB/s = ~1.589X (up to 59% more)

ENDNOTES

- MI300-13: Calculations conducted by AMD Performance Labs as of November 7, 2023, for the AMD Instinct™ MI300X OAM accelerator 750W (192 GB HBM3) designed with AMD CDNA™ 3 5nm FinFet process technology resulted in 192 GB HBM3 memory capacity and 5.325 TFLOPS peak theoretical memory bandwidth performance. MI300X memory bus interface is 8,192 (1024 bits x 8 die) and memory data rate is 5.2 Gbps for total peak memory bandwidth of 5.325 TB/s (8,192 bits memory bus interface * 5.2 Gbps memory data rate/8). The AMD Instinct™ MI250 (500W) / MI250X (560W) OAM accelerators (128 GB HBM2e) designed with AMD CDNA™ 2 6nm FinFet process technology resulted in 128 GB HBM3 memory capacity and 3.277 TFLOPS peak theoretical memory bandwidth performance. MI250/MI250X memory bus interface is 8,192 (4,096 bits times 2 die) and memory data rate is 3.20 Gbps for total memory bandwidth of 3.277 TB/s ((3.20 Gbps*(4,096 bits*2))/8). Memory Capacity:MI300X OAM: 192GB HBM3 / MI250/MI250X OAMs: 128GB HBM2e = 1.5X (50% more)Memory Bandwidth:MI300X OAM: 5.325 TB/s / MI250/MI250X OAMs : 3.2 TB/s = ~1.66X (up to 66% more)
- MI300-14: Calculations conducted by AMD Performance Labs as of November 7, 2023, for the AMD Instinct™ MI300A APU accelerator 760W (128 GB HBM3) designed with AMD CDNA™ 3 5nm FinFet process technology resulted in 128 GB HBM3 memory capacity and 5.325 TFLOPS peak theoretical memory bandwidth performance. MI300A memory bus interface is 8,192 (1024 bits x 8 die) and memory data rate is 5.2 Gbps for total peak memory bandwidth of 5.325 TB/s (8,192 bits memory bus interface * 5.2 Gbps memory data rate/8). The AMD Instinct™ MI250 (500W) / MI250X (560W) OAM accelerators (128 GB HBM2e) designed with AMD CDNA™ 2 6nm FinFet process technology resulted in 128 GB HBM3 memory capacity and 3.2 TFLOPS peak theoretical memory bandwidth performance. MI250/MI250X memory bus interface is 8,192 (4,096 bits times 2 die) and memory data rate is 3.20 Gbps for total memory bandwidth of 3.277 TB/s ((3.20 Gbps*(4,096 bits*2))/8). Server manufacturers may vary configuration offerings yielding different results.
- MI300-16: Measurements conducted by AMD Performance Labs as of November 11th, 2023 on the AMD Instinct™ MI300X (750W) GPU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 653.7 TFLOPS peak theoretical TensorFloat-32 (TF32), 1307.4 TFLOPS peak theoretical half precision (FP16), 1307.4 TFLOPS peak theoretical Bfloat16 format precision (BF16), 2614.9 TFLOPS peak theoretical 8-bit precision (FP8), 2614.9 TOPs INT8 floating-point performance.The MI300X is expected to be able to take advantage of fine-grained structure sparsity providing an estimated 2x improvement in math efficiency resulting 1,307.4 TFLOPS peak theoretical TensorFloat-32 (TF32), 2,614.9 TFLOPS peak theoretical half precision (FP16), 2,614.9 TFLOPS peak theoretical Bfloat16 format precision (BF16), 5,229.8 TFLOPS peak theoretical 8-bit precision (FP8), 5,229.8 TOPs INT8 floating-point performance with sparsity. The results calculated for the AMD Instinct™ MI250X (560W) 128GB HBM2e OAM accelerator designed with AMD CDNA™ 2 5nm FinFET process technology at 1,700 MHz peak boost engine clock resulted in TF32* (N/A), 383.0 TFLOPS peak theoretical half precision (FP16), 383.0 TFLOPS peak theoretical Bfloat16 format precision (BF16), FP8* (N/A), 383.0 TOPs INT8 floating-point performance. *AMD Instinct MI200 Series GPUs don’t support TF32, FP8 or sparsity.
- MI300-18: Measurements conducted by AMD Performance Labs as of November 11th, 2023 on the AMD Instinct™ MI300X (750W) GPU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 163.4 TFLOPs peak theoretical double precision Matrix (FP64 Matrix), 81.7 TFLOPs peak theoretical double precision (FP64), 163.4 TFLOPs peak theoretical single precision Matrix (FP32 Matrix), 163.4 TFLOPs peak theoretical single precision (FP32), 653.7 TFLOPS peak theoretical TensorFloat-32 (TF32), 1307.4 TFLOPS peak theoretical half precision (FP16), 1307.4 TFLOPS peak theoretical Bfloat16 format precision (BF16), 2614.9 TFLOPS peak theoretical 8-bit precision (FP8), 2614.9 TOPs INT8 floating-point performance.Published results on Nvidia H100 SXM (80GB) GPU resulted in 66.9 TFLOPs peak theoretical double precision tensor (FP64 Tensor), 33.5 TFLOPs peak theoretical double precision (FP64), 66.9 TFLOPs peak theoretical single precision (FP32), 494.7 TFLOPs peak TensorFloat-32 (TF32)*, 989.4 TFLOPS peak theoretical half precision tensor (FP16 Tensor), 133.8 TFLOPS peak theoretical half precision (FP16), 989.4 TFLOPS peak theoretical Bfloat16 tensor format precision (BF16 Tensor), 133.8 TFLOPS peak theoretical Bfloat16 format precision (BF16), 1,978.9 TFLOPS peak theoretical 8-bit precision (FP8), 1,978.9 TOPs peak theoretical INT8 floating-point performance.Nvidia H100 source: https://resources.nvidia.com/en-us-tensor-core/ * Nvidia H100 GPUs don’t support FP32 Tensor.

ENDNOTES

- MI300-20: Measurements conducted by AMD Performance Labs as of November 11th, 2023 on the AMD Instinct™ MI300A (760W) GPU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 122.6 TFLOPs peak theoretical double precision Matrix (FP64 Matrix), 61.3 TFLOPs peak theoretical double precision (FP64), 122.6 TFLOPs peak theoretical single precision Matrix (FP32 Matrix), 122.6 TFLOPs peak theoretical single precision (FP32), 490.29 TFLOPS peak theoretical TensorFloat-32 (TF32), 980.58 TFLOPS peak theoretical half precision (FP16), 980.58 TFLOPS peak theoretical Bfloat16 format precision (BF16), 1,961.16 TFLOPS peak theoretical 8-bit precision (FP8), 1,961.16 TOPs INT8 floating-point performance.Published results on Nvidia H100 SXM (80GB) 700W GPU resulted in 66.9 TFLOPs peak theoretical double precision tensor (FP64 Tensor), 33.5 TFLOPs peak theoretical double precision (FP64), 66.9 TFLOPs peak theoretical single precision (FP32), 494.7 TFLOPs peak TensorFloat-32 (TF32), 989.4 TFLOPS peak theoretical half precision tensor (FP16 Tensor), 133.8 TFLOPS peak theoretical half precision (FP16), 989.4 TFLOPS peak theoretical Bfloat16 tensor format precision (BF16 Tensor), 133.8 TFLOPS peak theoretical Bfloat16 format precision (BF16), 1,978.9 TFLOPS peak theoretical 8-bit precision (FP8), 1,978.9 TOPs peak theoretical INT8 floating-point performance.Nvidia H100 source: https://resources.nvidia.com/en-us-tensor-core/ AMD TFLOPS and TOPS calculations conducted with the following equation for AMD Instinct MI300A APUs: FLOPS calculations are performed by taking the engine clock from the highest DPM state and multiplying it by xx CUs per APU. Then, multiplying that number by xx stream processors, which exist in each CU. Then, for MI300A that number is multiplied by 1024 FLOPS per clock/CU for TF32 to determine TFLOPS, 2048 FLOPS per clock/CU for FP16 to determine TFLOPS, 2048 FLOPS per clock/CU for BF16 to determine TFLOPS, 4096 FLOPS per clock/CU for FP8 to determine TFLOPS, 4046 FLOPS per clock/CU for INT8 to determine TOPS. Divide results by 100,000 to get TFLOPS or TOPs.Calculations:FP64 Matrix | Tensor: MI300A 122.57 / H100 66.9 = 1.832x (83% faster) the floating-point performanceFP64: MI300A 61.3 / H100 33.5 = 1.830x (83% faster) the floating-point performanceFP32 Matrix: MI300A 122.57 / H100 (FP32) 66.9 = 1.832x (83% faster) the floating-point performance*FP32: MI300A 122.57 / H100 66.9 = 1.832x (83% faster) the floating-point performanceTF32: MI300A 490.29 / H100 494.7 = 0.991x (0.009% slower) the floating-point performanceFP16 (Tensor): MI300A 980.58 / H100 (FP64 Tensor) 989.4 = 0.991x (0.009% slower) the floating-point performanceFP16: MI300A 980.58 / H100 133.8 = 7.329x (633% faster) the floating-point performanceBF16 (Tensor): MI300A 980.58 / H100 (Tensor) 989.4 = 0.991x (0.009% slower) the floating-point performance BF16: MI300A 980.58 / H100 133.8 = 7.329x (633% faster) the floating-point performance FP8: MI300A 1,961.16 / H100 1,978.9 = 0.991x (0.009% slower) the floating-point performance INT8: MI300A 1,961.16 / H100 1,978.9 = 0.991x (0.009% slower) the floating-point performance* Nvidia H100 GPUs don’t support FP32 Tensor.
- MI300-25: Measurements conducted by AMD Performance Labs as of November 18th, 2023 on the AMD Instinct™ MI300X (192 GB HBM3) 750W GPU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 1307.4 TFLOPS peak theoretical half precision (FP16), 1307.4 TFLOPS peak theoretical Bfloat16 format precision (BF16).The MI300X is expected to be able to take advantage of fine-grained structure sparsity providing an estimated 2x improvement in math efficiency resulting 2,614.9 TFLOPS peak theoretical half precision (FP16), 2,614.9 TFLOPS peak theoretical Bfloat16 format precision (BF16 floating-point performance with sparsity.Published results on Nvidia H100 SXM (80GB HBM3) 700W GPU resulted in 989.4 TFLOPS peak theoretical half precision (FP16 Tensor), 989.4 TFLOPS peak theoretical Bfloat16 format precision (BF16 Tensor), 1,978.9 TFLOPS peak theoretical half precision (FP16 Tensor) with sparsity, 1,978.9 TFLOPS peak theoretical Bfloat16 format precision (BF16 Tensor) with sparsity floating-point performance. Nvidia H100 source: https://resources.nvidia.com/en-us-tensor-core/ AMD Instinct™ MI300X AMD CDNA 3 technology-based accelerators include up to eight AMD Infinity Fabric links providing up to 1,024 GB/s peak aggregate theoretical GPU peer-to-peer (P2P) transport rate bandwidth performance per GPU OAM module.
- MI300-29: GROMACS STMV comparison based on AMD internal testing as of 11/18/2023 and published Nvidia data.Configurations: AMD Instinct™ MI300A bring-up platform with 1x AMD Instinct MI300A (128GB, 550W) APU, Pre-release build of ROCm® 6, Ubuntu® 22.04.2. GROMACS version 2022.0.Vs.Nvidia public claims https://developer.nvidia.com/hpc-application-performance, as of 11/17/2023. GROMACS version 2023.2. Only 1 GPU on each system was used in this test.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

ENDNOTES

- MI300-32: OpenFOAM® v2206 HPC Motorbike comparison based on AMD internal testing as of 11/15/2023.Configurations: AMD Instinct™ MI300A bring-up platform with 4x AMD Instinct MI300A (128GB, 550W) APU, pre-release build of ROCm™ 6.0, Ubuntu 22.04.2Vs.Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, CUDA 12.0, Ubuntu 22.04.3.Only 1 GPU on each system was used in this test.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.
- MI300-33: Text generated with Llama2-70b chat using input sequence length of 4096 and 32 output token comparison using custom docker container for each system based on AMD internal testing as of 11/17/2023.Configurations: 2P Intel Xeon Platinum CPU server using 4x AMD Instinct™ MI300X (192GB, 750W) GPUs, ROCm® 6.0 pre-release, PyTorch 2.2.0, vLLM for ROCm, Ubuntu® 22.04.2.Vs.2P AMD EPYC 7763 CPU server using 4x AMD Instinct™ MI250 (128 GB HBM2e, 560W) GPUs, ROCm® 5.4.3, PyTorch 2.0.0., HuggingFace Transformers 4.35.0, Ubuntu 22.04.6.4 GPUs on each system was used in this test.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.
- MI300-34: Token generation throughput using DeepSpeed Inference with the Bloom-176b model with an input sequence length of 1948 tokens, and output sequence length of 100 tokens, and a batch size tuned to yield the highest throughput on each system comparison based on AMD internal testing using custom docker container for each system as of 11/17/2023.Configurations: 2P Intel Xeon Platinum 8480C CPU powered server with 8x AMD Instinct™ MI300X 192GB 750W GPUs, pre-release build of ROCm™ 6.0, Ubuntu 22.04.2.Vs.An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 80GB 700W GPUs, CUDA 12.0, Ubuntu 22.04.3.8 GPUs on each system were used in this test. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.
- MI300-35: Flash Attention v2 forward kernel for inference, head_dim=128 and causal=false, comparison based on AMD internal testing as of 11/29/2023.Configurations: AMD MI300X bring-up platform with 1x AMD Ryzen™ 9 7950X, 1x AMD Instinct™ MI300X (192GB, 750W) GPU, Ubuntu® 22.04.3, Pre-release build of ROCm™ 6.0, Flash attention v2 forward kernel using an internal containerVsAn Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, Ubuntu® 22.04.3, CUDA® 12.2.2Flash attention v2 forward kernel using nvcr.io/nvidia/pytorch:23.10-py3 container.Only 1 GPU on each system was used in this test.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.
- MI300-36: Overall latency for text generation using the Llama2-13b chat model with vLLM comparison based on AMD internal testing as of 11/29/2023. Tests were performed using an input sequence length of 2048 input tokens and 128 output tokens.Configurations: AMD MI300X bring-up platform with 1x AMD Ryzen 9 7950X, 1x AMD Instinct™ MI300X (192GB, 750W) GPU, ROCm® 6.0 pre-release, Ubuntu® 22.04.2, AMD port of vLLM for ROCm.Vs.An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, CUDA 12.1, Ubuntu 22.04.3, vLLM v.0.2.2 (most recent).Only 1 GPU on each system was used in this test.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.
- MI300-37: Llama2-70b inference comparison, with Key GEMM kernels used, based on AMD internal testing as of 11/17/2023.Configurations: AMD MI300X bring-up platform with 1x AMD Ryzen 9 7950X, 1x AMD Instinct™ MI300X (192GB, 750W) GPU, ROCm® 6.0 pre-release, Ubuntu® 22.04.2.Vs.An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, CUDA 12.2.2, Ubuntu 22.04.3Only 1 GPU on each system was used in this test.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.
- MI300-38: Overall latency for text generation using the Llama2-70b chat model with vLLM comparison using custom docker container for each system based on AMD internal testing as of 11/23/2023. Sequence length of 2048 input tokens and 128 output tokens.Configurations: 2P Intel Xeon Platinum 8480C CPU server with 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, ROCm® 6.0 pre-release, PyTorch 2.2.0, vLLM for ROCm, Ubuntu® 22.04.2.Vs.An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, CUDA 12.1., PyTorch 2.1.0., vLLM v.02.2.2 (most recent), Ubuntu 22.04.3.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.
- MI300-39: Number of simultaneous text generating copies of the Llama2-70b chat model, using vLLM, comparison using custom docker container for each system based in AMD internal testing as of 11/26/2023.Configurations: 2P Intel Xeon Platinum 8480C CPU server with 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, ROCm® 6.0 pre-release, PyTorch 2.2.0, vLLM for ROCm, Ubuntu 22.04.2.Vs.An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, CUDA 12.1., PyTorch 2.1.0. vLLM v.02.2.2 (most recent), Ubuntu 22.04.3.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

ENDNOTES

- MI300-40: Testing completed 11/28/2023 by AMD performance lab using MosaicML vllm-foundry to fine tune the MPT-30b model for 2 epochs using the MosaicML instruct-v3 dataset and a max sequence length of 8192 tokens using custom docker container for each system .Configurations: 2P Intel Xeon Platinum 8480C CPU server with 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, ROCm® 6.0 pre-release, PyTorch 2.0.1, MosaicML llm-foundry pre-release, Ubuntu 22.04.2.Vs.An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, CUDA 11.8, PyTorch 2.0.1., MosaicML llm-foundry, Ubuntu 22.04.3.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.
- M300-42: Measurements by internal AMD Performance Labs as of December 1, 2023 on current specifications and/or internal engineering calculations. Inference and training Large Language Model (LLM) run comparisons with FP16 precision to determine the largest Large Language model size that is expected to run on the 8x AMD Instinct™ MI300X (192GB) accelerator platform and on the Nvidia 8x H100 (80GB) GPUs DGX platform. Calculated estimates based on GPU-only memory size versus memory required by the model at defined parameters plus 10% overhead. Calculations rely on published and sometimes preliminary model memory sizes. Multiple LLMs and parameter sizes were analyzed. Max size determined by memory capacity of 8x platform. Configurations: 8x AMD Instinct™ MI300X (192GB HBM3, OAM Module) 750W accelerator at 2,100 MHz peak boost engine clock designed with 3rd Gen AMD CDNA™ 3 5nm FinFET process technology. Vs.8x Nvidia HGX H100 (80GB HBM3, SXM5) platform Nvidia memory specification at https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet. Results for Inferencing:Largest parameter size for 8X H100: MI300X GPUs H100 GPUs Gopher Deepmind (290B) 4 Calculated 8 Calculated Largest parameter size for 8x MI300X: MI300X GPUs H100 GPUsPaLM-1 (680B) 8 Calculated 19 Calculated Results for Training: Largest parameter size for 8X H100: MI300X GPUs H100 GPUs Mosiac MPT-30B parameter 4 Calculated 8 CalculatedLargest parameter size for 8x MI300X: MI300X GPUs H100 GPUsMosiac MPT-70B parameter 7 Calculated 16 CalculatedAssumptions:FP16 Datatype Batchsize 1Memory needs for model = 2Bytes per Parameter Memory size needs for activations and others = +10% Actual maximum LLM parameter size that can run on each platform may vary upon performance testing with physical servers.
- MI300-43: Measurements conducted by AMD Performance Labs as of December 4th, 2023 on the AMD Instinct™ MI300A (760W) APU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in:• 122.6 TFLOPs peak theoretical double precision Matrix (FP64 Matrix), • 61.3 TFLOPs peak theoretical double precision (FP64), • 122.6 TFLOPs peak theoretical single precision Matrix (FP32 Matrix), • 122.6 TFLOPs peak theoretical single precision (FP32), floating-point performance. Published results on Nvidia GH200 1000W GPU: • 67 TFLOPs peak theoretical double precision tensor (FP64 Tensor), • 34 TFLOPs peak theoretical double precision (FP64), • N/A FP 32 Tensor - Nvidia GH200 GPUs don't support FP32 Tensor. Regular FP32 number used as proxy. • 67 TFLOPs peak theoretical single precision (FP32), floating-point performance. Nvidia GH200 source: https://resources.nvidia.com/en-us-grace-cpu/grace-hopper-superchip GH200 TFLOPs per Watt Calculations (peak wattage of 1000W used):• FP64 Matrix: 67 TFLOPs / 1000W = 0.067 TFLOPs per Watt• FP64: 34 TFLOPs / 1000W = 0.034 TFLOPs per Watt• FP32: 67 TFLOPs / 1000W = 0.067 TFLOPs per Watt* Nvidia GH200 GPUs don't support FP32 Tensor. Actual performance and performance per watt may vary on production systems.
- MI300-44: Llama2-70b model vLLM, hip Graph and Flash Attention LLM Performance Optimization comparison using custom docker containers across sequence lengths from 512 to 7168 based on AMD internal testing as of 11/22/2023.Testing done by comparing baseline (LLama2-70b model vLLM, hip Graph and Flash Attention LLM) performance optimizations off. This performance was measured against the performance with each optimization turned on to determine the performance impact of the optimization. Configurations: 2P Intel Xeon Platinum 8480C CPU server using 4x AMD Instinct™ MI300X (192GB, 750W) GPUs, ROCm® 6.0 pre-release, PyTorch 2.2.0, Ubuntu® 22.04.2.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

ENDNOTES

PHX-3: As of May 2023, AMD has the first available dedicated AI engine on an x86 Windows processor, where 'dedicated AI engine' is defined as an AI engine that has no function other than to process AI inference models and is part of the x86 processor die. For detailed information, please check:https://www.amd.com/en/products/ryzen-ai.

HWK-01: Testing as of Oct 2023 by AMD Performance Labs on the following game titles at 1080p low settings, VBS enabled: Borderlands 3, F1 2022, Far Cry 6, Grand Theft Auto 5, Hitman 3, League of Legends, Shadow of the Tomb Raider, Tiny Tinas Wonderland, WoTenCore. Configuration for AMD reference system: Ryzen 9 8945HS, integrated Radeon 780M graphics, 16GB RAM, Samsung 980 Pro 1TB NVMe, Windows 11 Pro. Configuration for Intel system: Core i9-13900H, integrated Iris Xe Graphics, 16GB RAM, 1TB SSD, Windows 11 Pro. Both with VBS enabled. PC manufacturers may vary configurations yielding different results. Results may vary.

HWK-02: Testing as of Oct 2023 by AMD Performance Labs using the following benchmarks: Blender, POVRay, Handbrake, LAME, Puget Davinci Resolve, Puget Adobe Photoshop, PCMark 10. Configuration for AMD reference system: Ryzen 9 8945HS, integrated Radeon 780M graphics, 16GB RAM, Samsung 980 Pro 1TB NVMe, Windows 11 Pro. Configuration for Intel system: Core i9-13900H, integrated Iris Xe Graphics, 16GB RAM, 1TB SSD, Windows 11 Pro. Both with VBS enabled. PCMark is a registered trademark of Futuremark Corporation. PC manufacturers may vary configurations yielding different results. Results may vary.

HWK-03: Testing as of Oct 2023 by AMD Performance Labs using the following benchmarks: Cinebench R23 and Geekbench 6. Configuration for AMD reference system: Ryzen 9 8945HS, integrated Radeon 780M graphics, 16GB RAM, Samsung 980 Pro 1TB NVMe, Windows 11 Pro. Configuration for Intel system: Core i9-13900H, integrated Iris Xe Graphics, 16GB RAM, 1TB SSD, Windows 11 Pro. Both with VBS enabled. PC manufacturers may vary configurations yielding different results. Results may vary..

HWK-19: Based on testing by AMD as of 12/2023, measuring AI LLM performance on Llama2-7B model, Pytorch in Eager Mode, ONNX Runtime and ONNX perf tool, quantized to INT8 with Vitis™ AI ONNX quantizer. AI Vision Model performance measured based on Resnet50,Yolov3, ESRGAN models. ONNX Runtime ONNX perf tool, quantized to INT8 using Vitis™ AI ONNX quantizer. System configurations: AMD Ryzen 9 7940HS with Radeon 780M Graphics on an AMD reference platform with 16GB DDR5, SSD, Windows 11 Pro vs. a similarly configured AMD Ryzen 7 8840HS processor with Radeon 780M graphics. System manufacturers may vary configuration, yielding different results. Results may vary.

STX-01: An AMD Ryzen “Strix point” processor is projected to offer 3x faster NPU performance for AI workloads when compared to an AMD Ryzen 7040 series processor. Performance projection by AMD engineering staff. Engineering projections are not a guarantee of final performance. Specific projections are based on reference design platforms and are subject to change when final products are released in market.