Journée AuDACES 2025

Apostolos Tsetoglou
Field Applications Engineer
apostolos.tsetoglou@amd.com

Philippe Grégoire
BDE Public Sector
philippe.gregoire@amd.com

AMD
together we advance_

55 years

Founded May 1, 1969
Headquartered in Santa Clara, CA

28,000+ employees

Accelerating next-generation computing

$25.8B annual revenue in 2024

Over 25% reinvested towards research and development

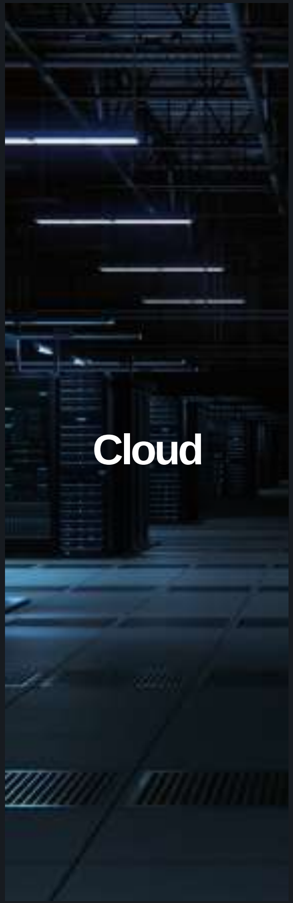
3x market cap growth in 5 years

Top 100 most valuable companies in the world

100+ locations

Around the world

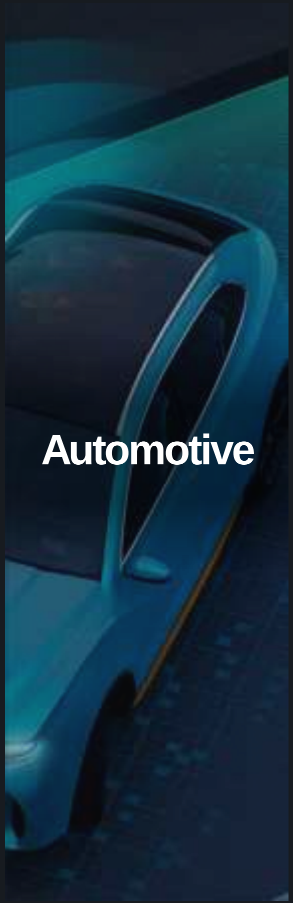
AMD powers the daily lives of billions



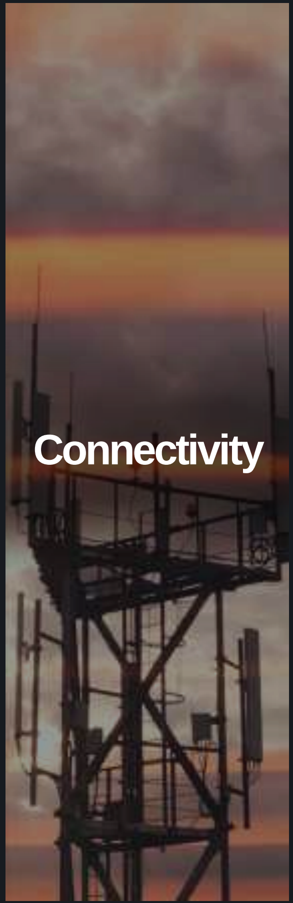
Cloud

Healthcare

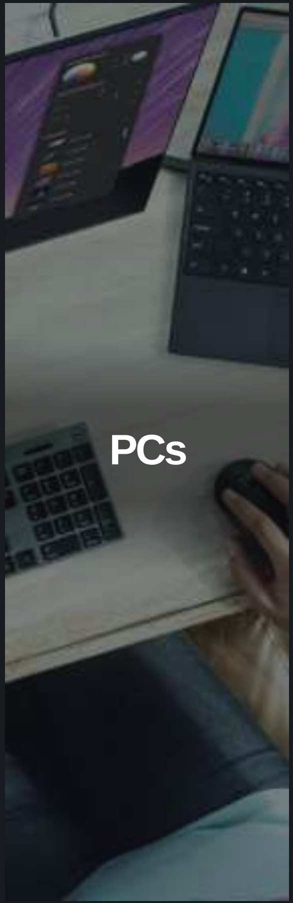
Industrial



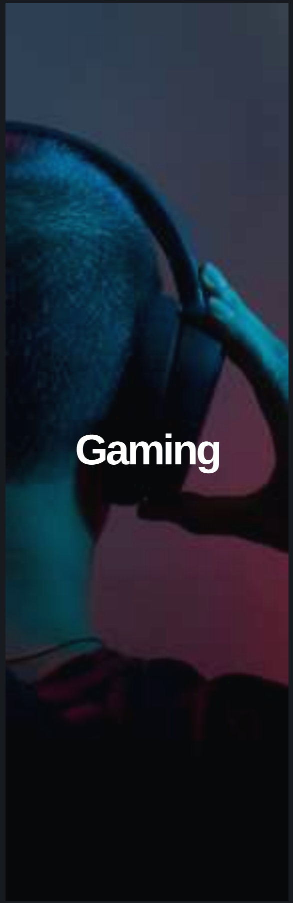
Automotive



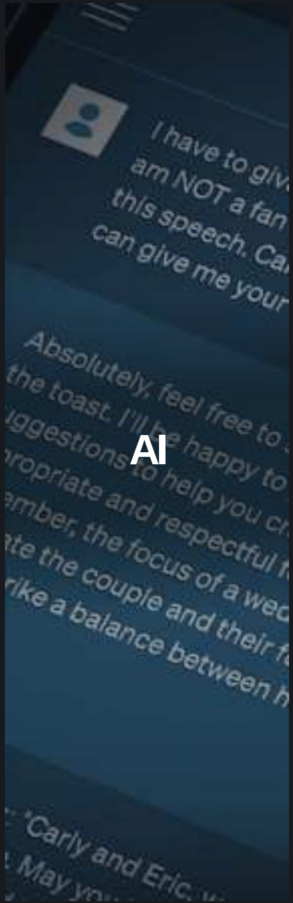
Connectivity



PCs



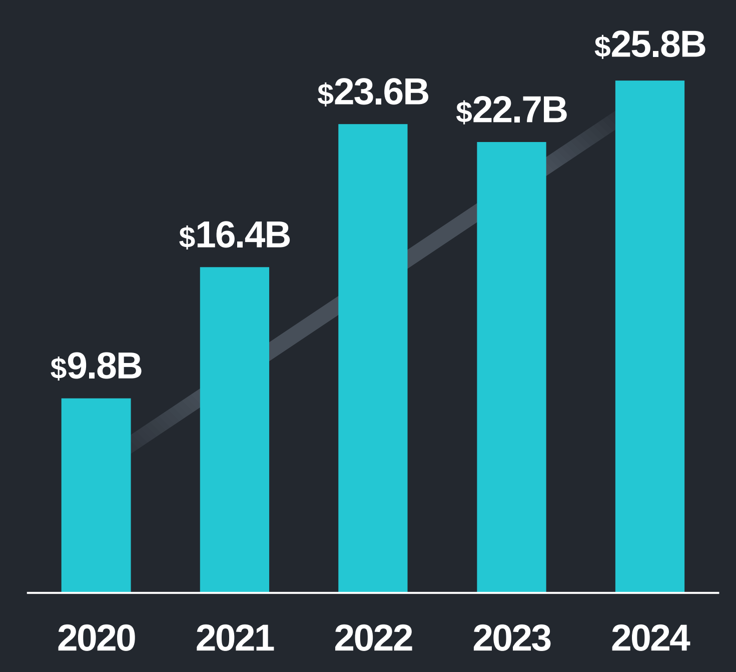
Gaming



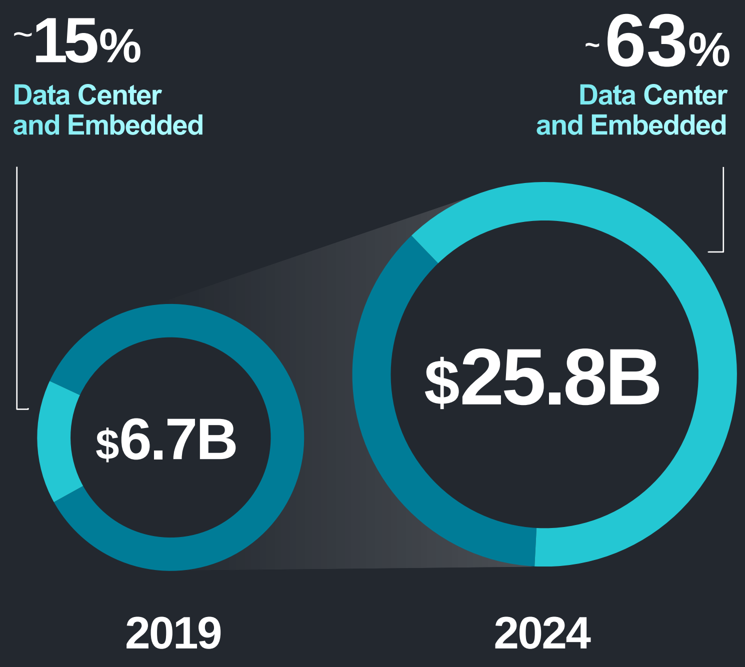
AI

Driving Growth from a Strong Financial Foundation

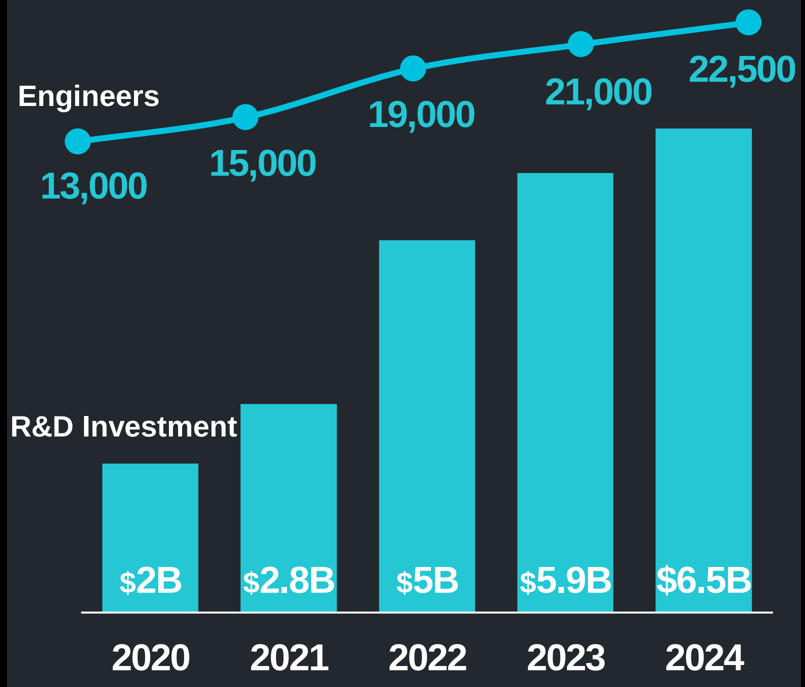
Accelerating Revenue Growth



Transforming Revenue Mix



Investing for Leadership



AMD
together we advance_

AMD

Advancing the AI Data Center

CPUs

AMD EPYC™

+

GPUs

AMD Instinct™

+

AMD PENSANDO

Ultra Ethernet

UA Link | Ultra Accelerator Link

Networking

DPUs, UALink + Ultra Ethernet

+

AMD ROCm

Software Solutions

Open Software Stack

+

zt Systems

Cluster Level Systems Design

Announced acquisition expected to close 1H 2025

AMD EPYC™ Processors

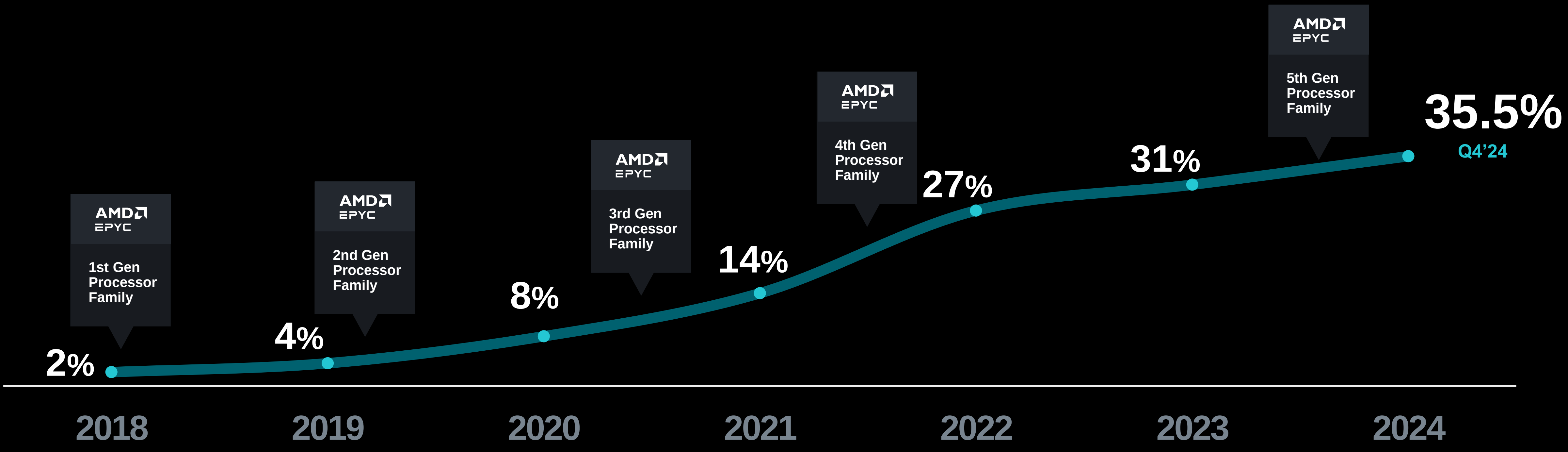
AMD EPYC™ trusted to power over one-third the world's servers

Alibaba Cloud, aws, Google Cloud, IBM Cloud, Meta, Microsoft Azure, ORACLE, Tencent Cloud

Largest and most discerning hyperscale data center customers

ASUS, Hewlett Packard Enterprise, DELL, Lenovo, cisco, SUPERMICRO, Inventec, GIGABYTE, msi, TYAN

Broad range of platforms from all major OEMs and support from 150+ leading ODMs



Source: Mercury Research Sell-in Revenue Shipment Estimates

#1 CPU for hyperscalers

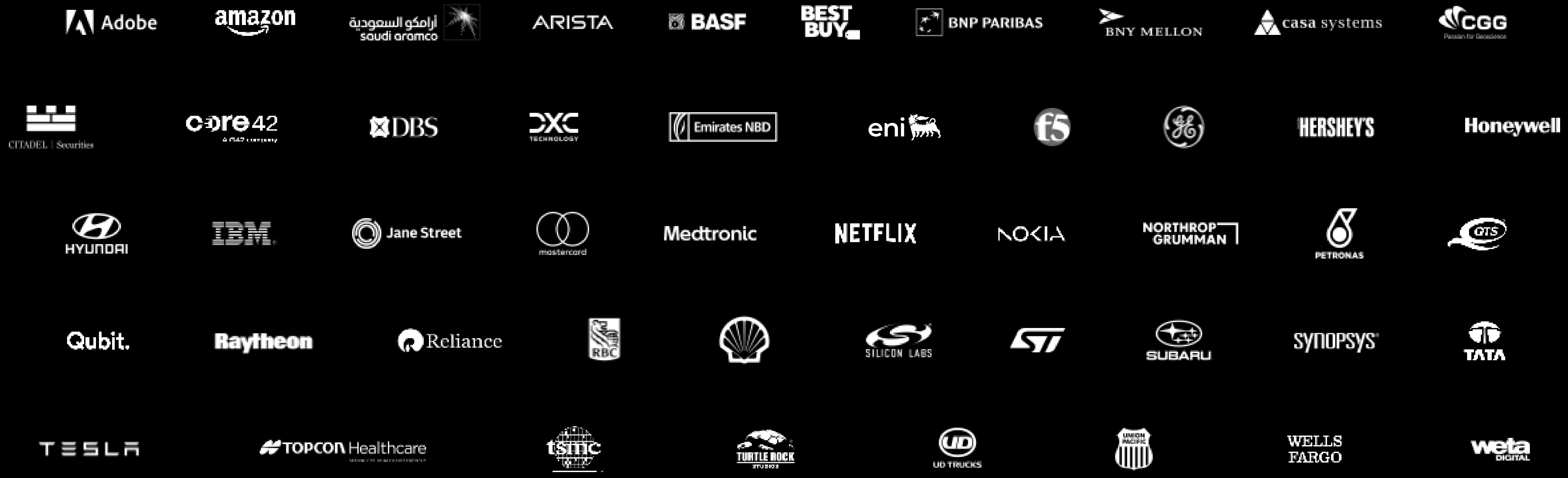
aws · Alibaba Cloud · Microsoft Azure · Google Cloud · IBM Cloud · ORACLE · Meta · Tencent

Hyperscale leaders power internal workloads with AMD, serving billions worldwide

NETFLIX · Office 365 · ORACLE EXADATA · salesforce · SAP · Uber · zoom

Use of third-party marks/logos/products is for informational purposes only and no endorsement of or by AMD is intended or implied GD-83

Trusted by industry leaders on-prem



Use of third-party marks/logos/products is for informational purposes only and no endorsement of or by AMD is intended or implied GD-83

AMD EPYC™ Processors

SUSTAINED LEADERSHIP | ECOSYSTEM COLLABORATION | INDUSTRY STANDARDS

Business Value

Accelerate Productivity

Realize energy-efficiency

Outstanding IT ROI

Data Security

Proven Solutions

Large-scale Enterprise deployments

World's #1 Supercomputer*

Leading Cloud Providers

Seamless Innovation

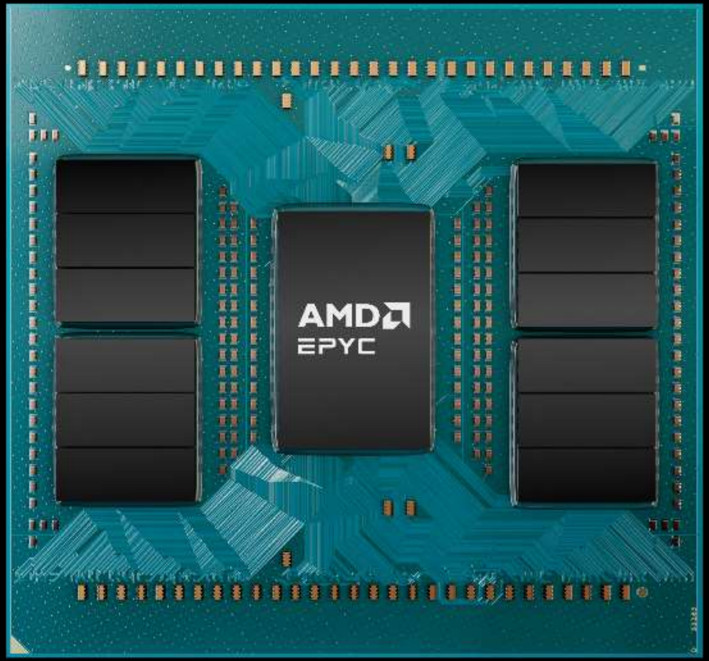
Leadership Technology and Data Center Portfolio

AMD EPYC AMD INSTINCT AMD ALVEO AMD VERSAL AMD PENSANDO

*TOP500 list as of the last publication date of Jun 2024. https://www.top500.org/lists/top500/2024/06/

5th Gen AMD EPYC™ Processors

Formerly codenamed "Turin"



World's best CPU for cloud, enterprise & AI

ZEN 5

TSMC 3/4nm

Up to **192 cores**
Up to **384 threads**

Up to **5GHz**

AVX512
full 512b data path

17%
Enterprise IPC Uplift

37%
HPC/AI IPC Uplift

SP5 Platform
Compatible with "Genoa"

See endnote: 9xx5-001, 048

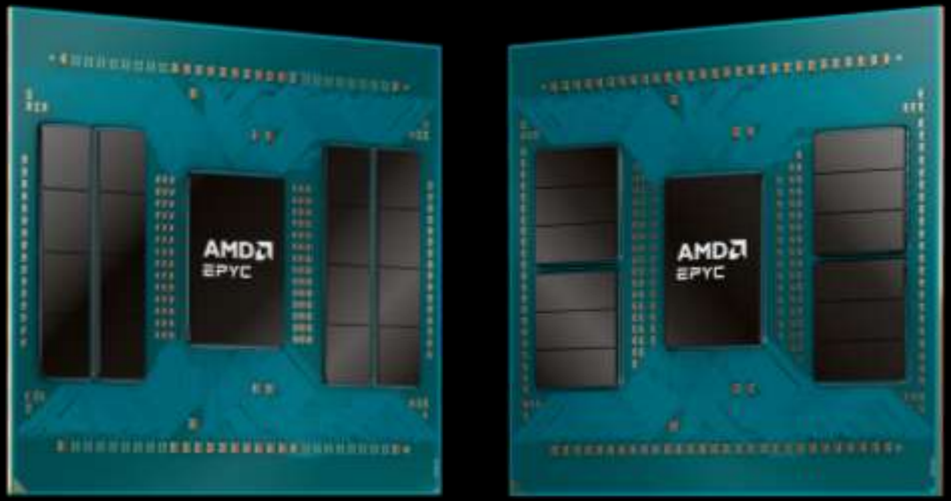
5th Gen AMD EPYC 9005 Series Processors

Continuing to Deliver Technology Leadership

Scale-Up

Up to

16 “Zen 5” CCDs
128 Cores / 256 Threads



Scale-Out

Up to

12 “Zen 5c” CCDs
192 Cores / 384 Threads

Consistent features,
ISA, & IPC uplift

SP5 Socket
“Genoa” Compatible

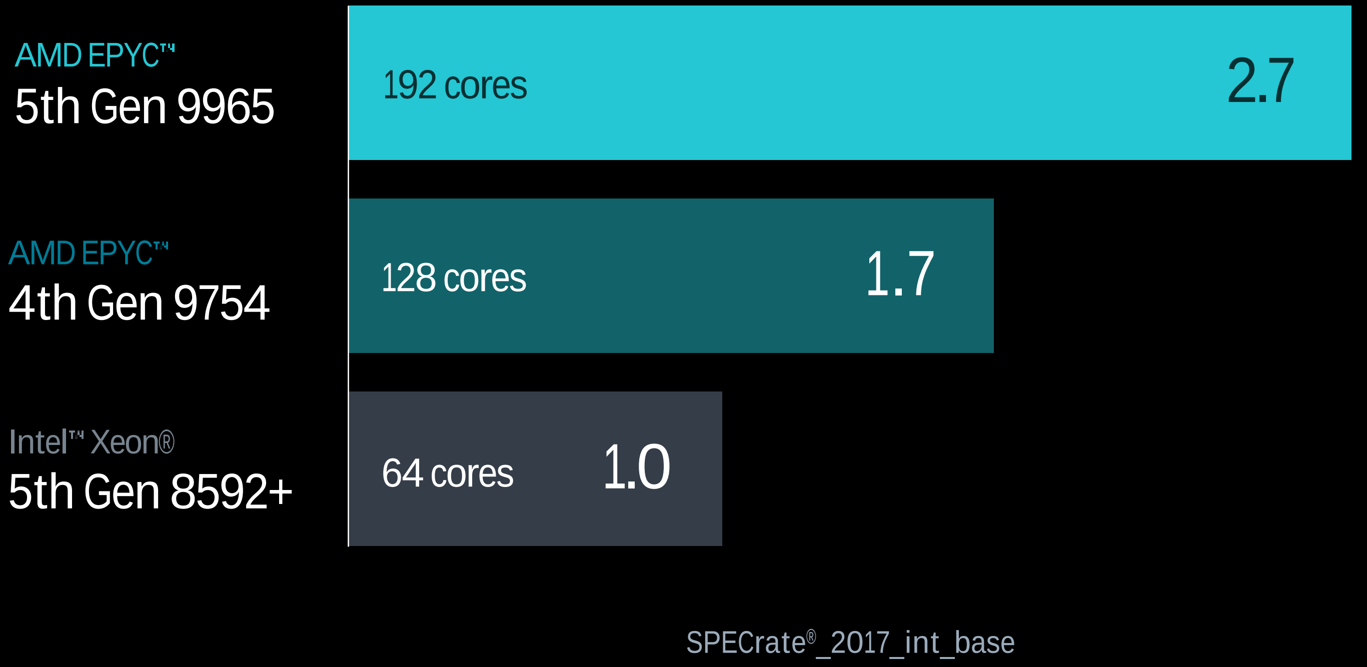
8 to 192 Cores
155W to 500W

Up to
12Ch DDR5-6400
128 PCIe 5.0/CXL 2.0

Confidential Compute
with Trusted I/O

As of 10/1/2024. See endnote 9xx5-048.

Industry's Highest Performing Server CPU



2.7x

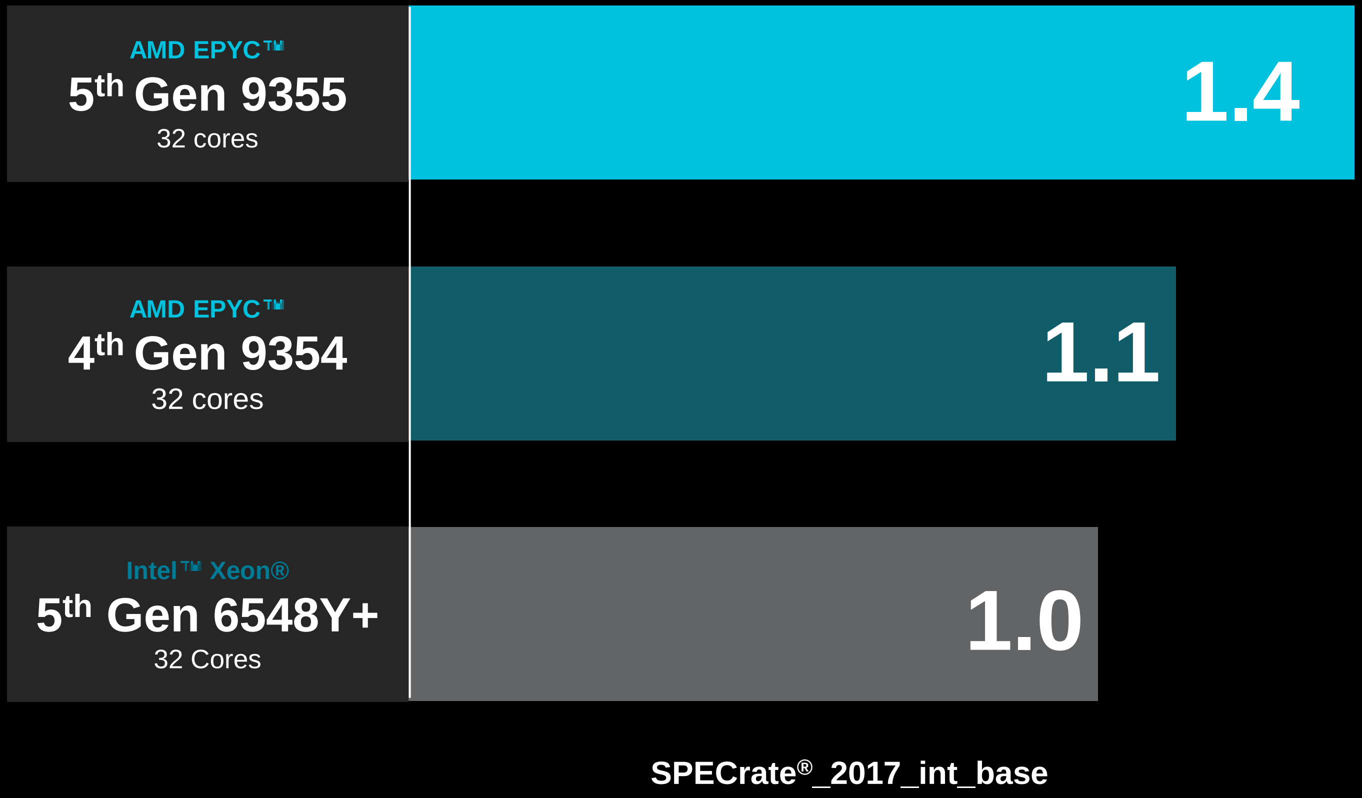
vs. top-of-stack "Emerald Rapids"

As of 10/01/2024. See endnote 9xx5-002c.

Per Core Performance Leadership

1.4X

the throughput per core

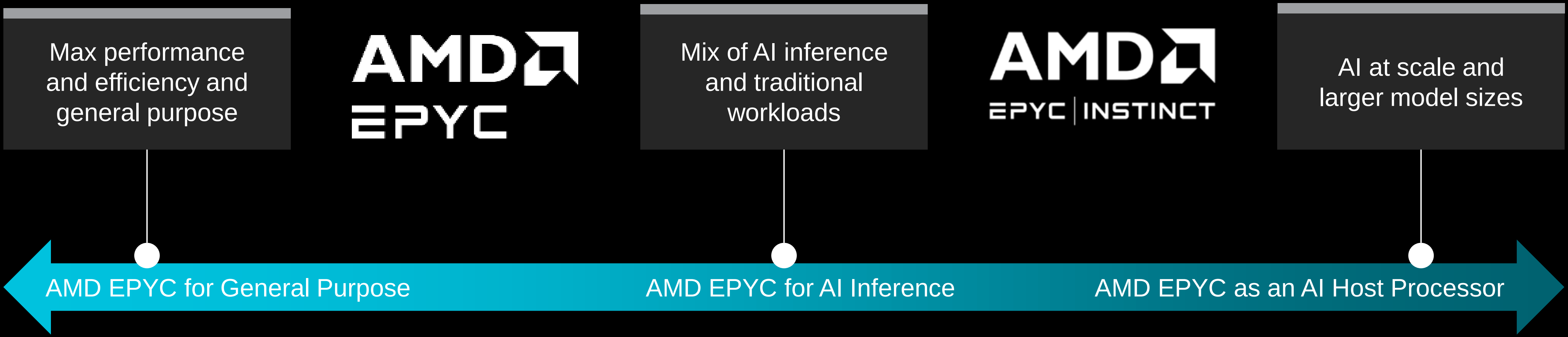


See endnotes 9xx5-003B

AMD EPYC™ for AI

AMD EPYC™ CPUs Enable Customer AI Initiatives

Spanning traditional compute, mixed AI & AI at Scale with optimized CPU + GPU solutions



Max performance and efficiency and general purpose

Mix of AI inference and traditional workloads

AI at scale and larger model sizes

AMD EPYC for General Purpose

AMD EPYC for AI Inference

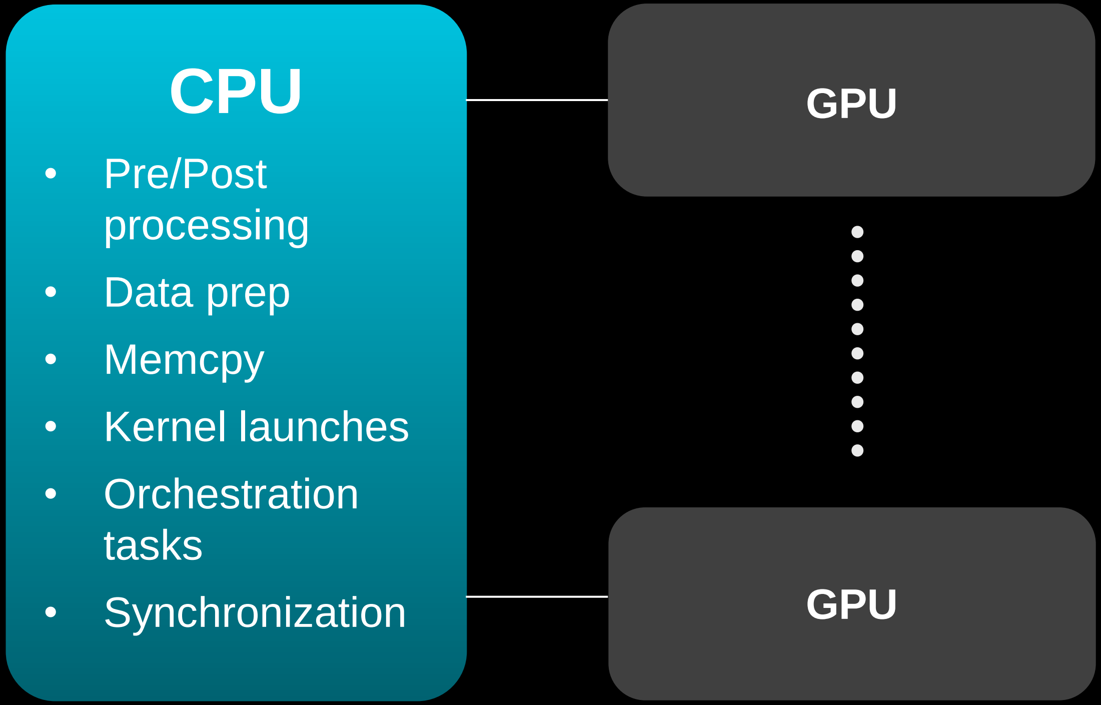
AMD EPYC as an AI Host Processor

5th Gen AMD EPYC™ 9575F – 5.0 GHz High Frequency 64 Core SKU

Designed for GPU Accelerated AI Inference & Training

28%

Faster Processing for GPU orchestration tasks



CPU

- Pre/Post processing
- Data prep
- Memcpy
- Kernel launches
- Orchestration tasks
- Synchronization

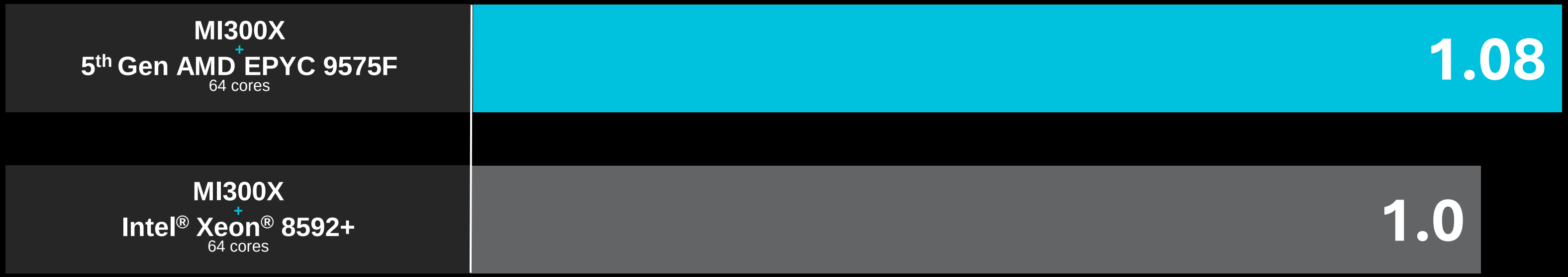
AMD EPYC™ 9575F, 5.0 GHz max frequency vs Intel® Xeon® 8592+, 3.9 GHz max frequency
See endnotes GD-150

5th Gen AMD EPYC™ 9575F

Enabling Maximum GPU System Performance as an AI Host Processor

+8%

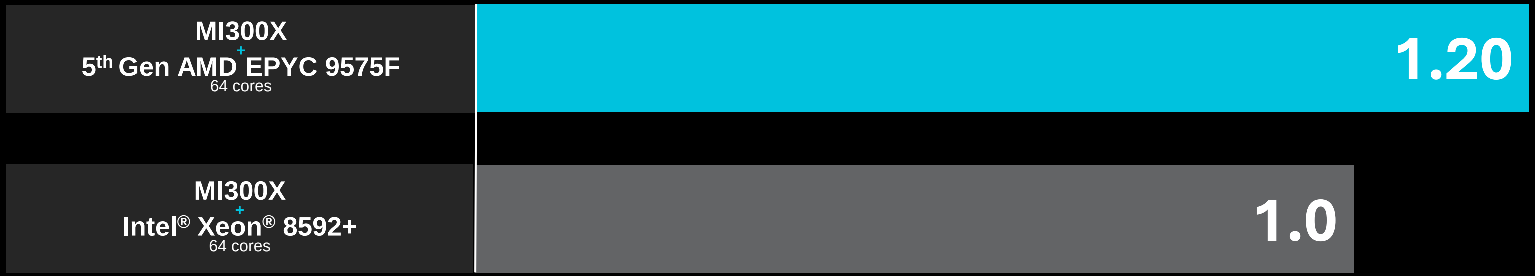
GPU System Performance Inference



Llama3.1-70B Inference Benchmark (8xGPU)

+20%

GPU System Performance Training



Stable Diffusion XL v2 Training Benchmark (8xGPU)

700K More Tokens/Second From 1K Node AI Cluster for Inference

See endnotes 9xx5-056A, 059, 087

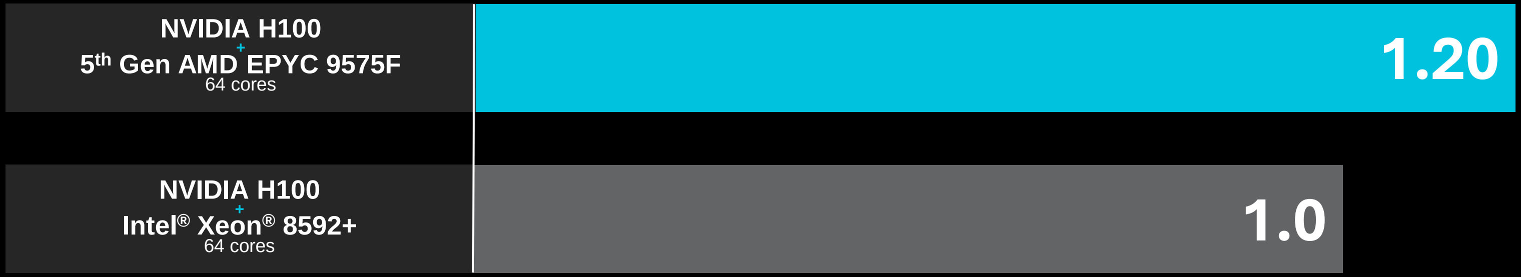
AMD
together we advance_

5th Gen AMD EPYC™ 9575F

Enabling Maximum GPU System Performance as an AI Host Processor

+20%

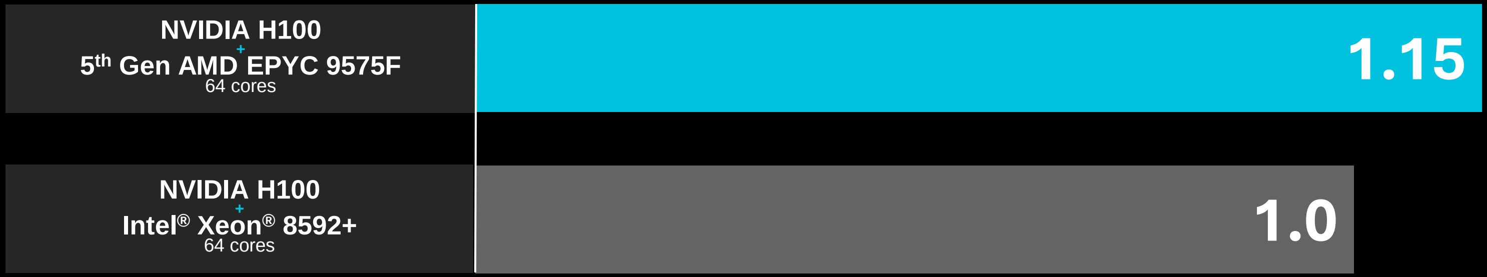
GPU System Performance Inference



Llama3.1-70B Inference Benchmark (8xGPU)

+15%

GPU System Performance Training



Llama3.1-8B Training Benchmark (8xGPU)

Up to 20% more requests and 15% better time to train with AMD EPYC™ 9575F

See endnotes 9xx5-014, 015

AMD INSTINCT™ ACCELERATORS

AMD DRIVING GPU LEADERSHIP

Supercomputer

Frontier (US)
LUMI (Finland/EU)
Adastra (France)
Setonix (Australia)

Data Center

AMD Instinct™ MI250
AMD Instinct™ MI210
Radeon™ PRO V620
AWS, Microsoft Azure

PC

Radeon™ RX 7000 Series
Radeon™ W7000 Series
Ryzen™ 7000 Series

Console

PlayStation 5
Xbox Series X | S
Steam Deck

Embedded

Magic Leap
Tesla

Mobile

Samsung Exynos

What about my already trained models?

Is switching to AMD expensive?

Is MI300X better than H100?

Can I expect lower TCO?

Do they support Tensor Core?

Can I trust AMD software?

What's AMD value prop?

Compatibility?

Compatibility!

Transitioning AI Workloads to AMD GPUs

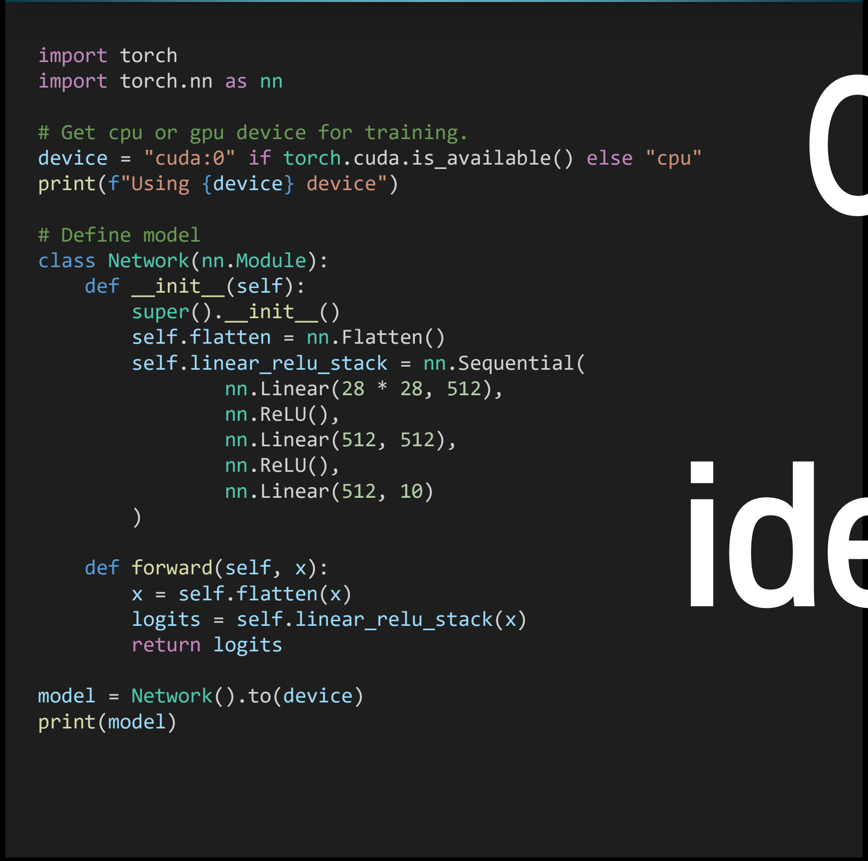
	NVIDIA		AMD ROCm™
ML Frameworks Python	PyTorch ONNX TensorFlow	Drop-in (Out-of-the-box) Support	PyTorch ONNX TensorFlow
ML Kernel Development C++, Triton IR	TRITON KERNELS → CUDA Triton Backend CUDA KERNELS → NVCC	Drop-in Port / Optimize	TRITON KERNELS → ROCM Triton Backend CUDA KERNELS → *HIP* → HIPCC
ML Libraries C++	cuBLAS, cuSparse, cuFFT, NCCL, cuDNN…	MIRROR Equivalent Libraries	rocBLAS, rocSparse, rocFFT, RCCL, MIOpen…

ROCm™ Software: Can You Spot a Difference?

NVIDIA CUDA

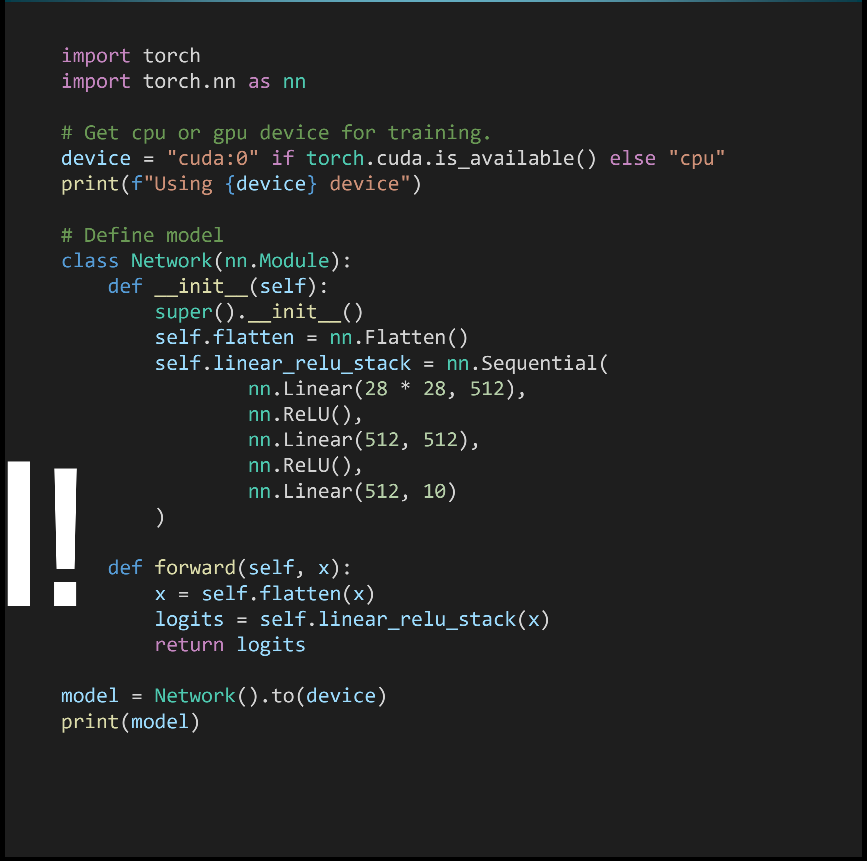
→

AMD ROCm™ Software



```
import torch
import torch.nn as nn

# Get cpu or gpu device for training.
device = "cuda:0" if torch.cuda.is_available() else "cpu"
print(f"Using {device} device")

# Define model
class Network(nn.Module):
    def __init__(self):
        super().__init__()
        self.flatten = nn.Flatten()
        self.linear_relu_stack = nn.Sequential(
                nn.Linear(28 * 28, 512),
                nn.ReLU(),
                nn.Linear(512, 512),
                nn.ReLU(),
                nn.Linear(512, 10)
        )

    def forward(self, x):
        x = self.flatten(x)
        logits = self.linear_relu_stack(x)
        return logits

model = Network().to(device)
print(model)
```

Codes are identical!



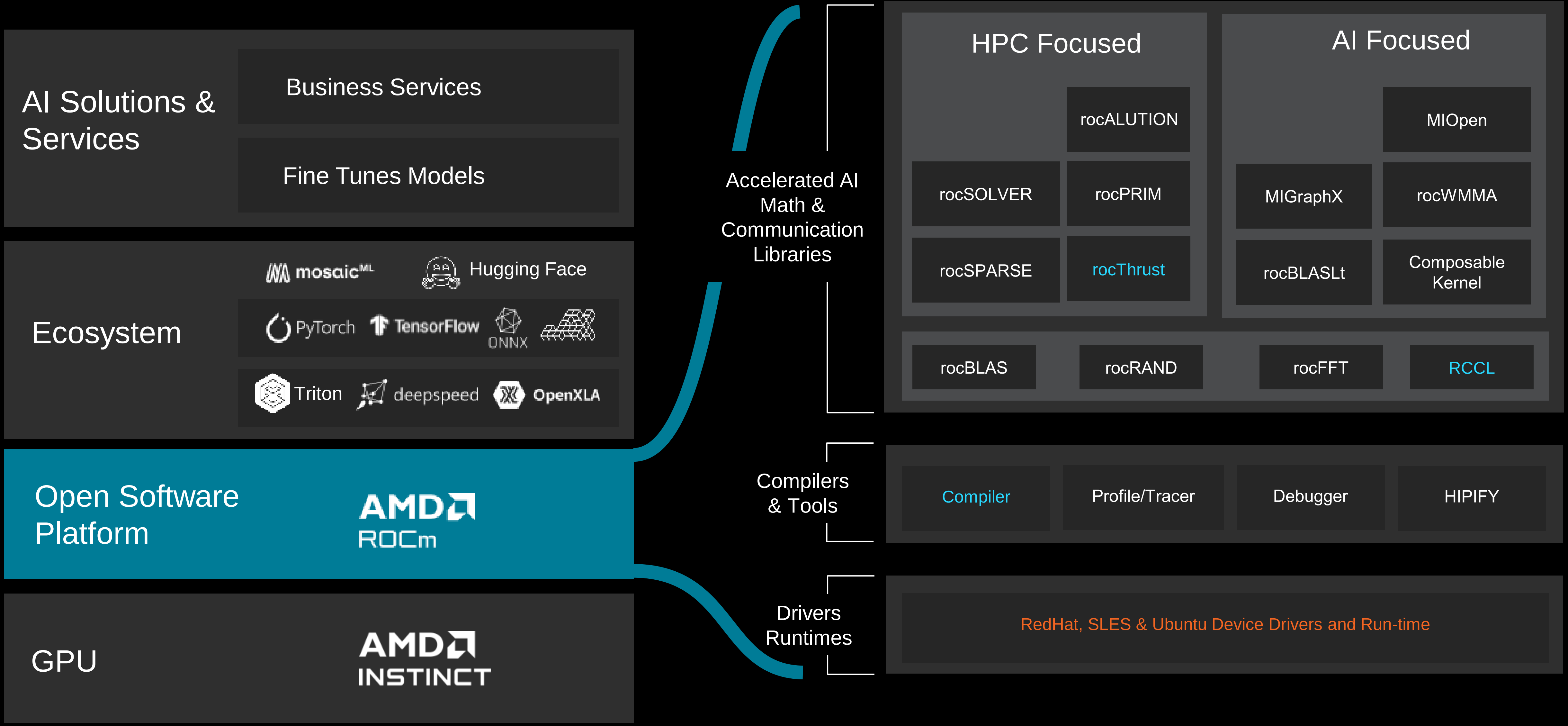
```
import torch
import torch.nn as nn

# Get cpu or gpu device for training.
device = "cuda:0" if torch.cuda.is_available() else "cpu"
print(f"Using {device} device")

# Define model
class Network(nn.Module):
    def __init__(self):
        super().__init__()
        self.flatten = nn.Flatten()
        self.linear_relu_stack = nn.Sequential(
                nn.Linear(28 * 28, 512),
                nn.ReLU(),
                nn.Linear(512, 512),
                nn.ReLU(),
                nn.Linear(512, 10)
        )

    def forward(self, x):
        x = self.flatten(x)
        logits = self.linear_relu_stack(x)
        return logits

model = Network().to(device)
print(model)
```

MIT/BSD License

Apache License

GPL License

AMD ROCm™ Software Stack



AI Solutions & Services

- Business Services
- Fine Tunes Models

Ecosystem

- mosaicML
- Hugging Face
- PyTorch
- TensorFlow
- ONNX
- Triton
- deepspeed
- OpenXLA

Open Software Platform

AMD ROCm

GPU

AMD INSTINCT

Accelerated AI Math & Communication Libraries

HPC Focused

- rocALUTION
- rocSOLVER
- rocPRIM
- rocSPARSE
- rocThrust

AI Focused

- MIOpen
- MIGraphX
- rocWMMA
- rocBLASLt
- Composable Kernel

- rocBLAS
- rocRAND
- rocFFT
- RCCL

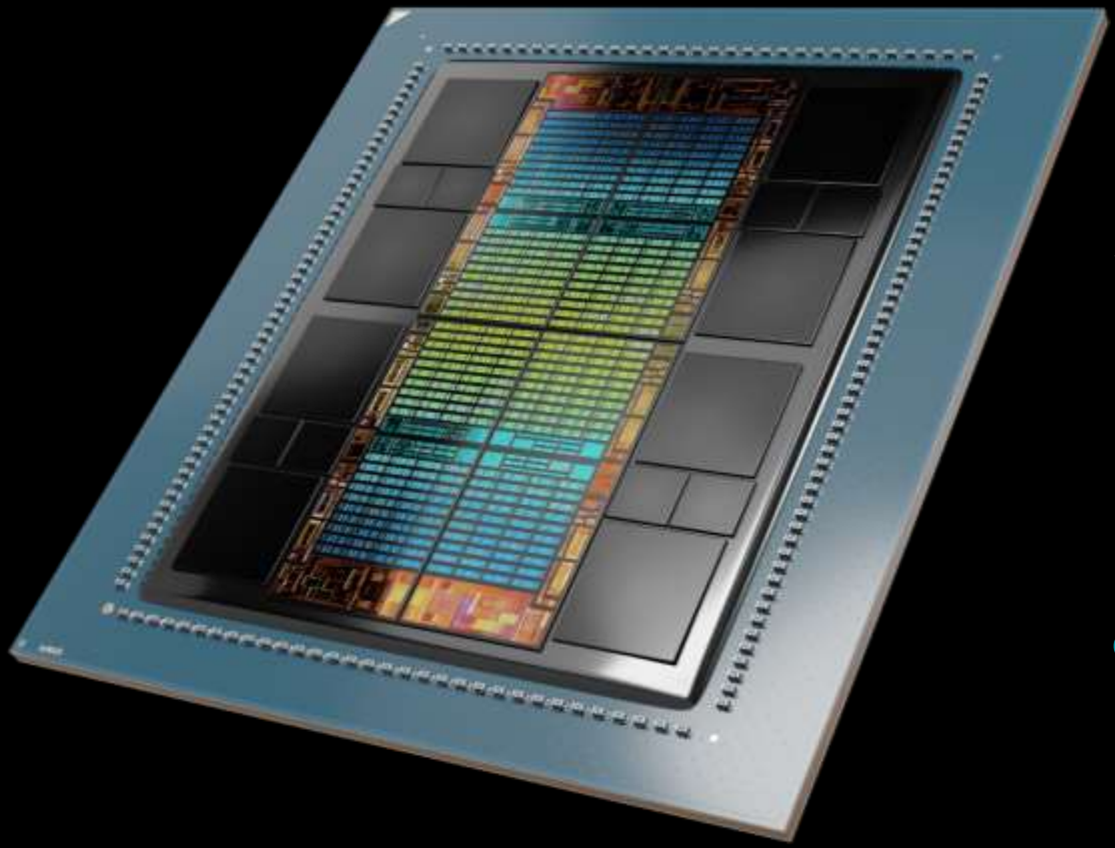
Compilers & Tools

- Compiler
- Profile/Tracer
- Debugger
- HIPIFY

Drivers Runtimes

RedHat, SLES & Ubuntu Device Drivers and Run-time

Use of third party marks/logos/products is for informational purposes only and no endorsement of or by AMD is intended or implied GD-83



AMD Instinct™ MI300X

Leadership generative AI accelerator

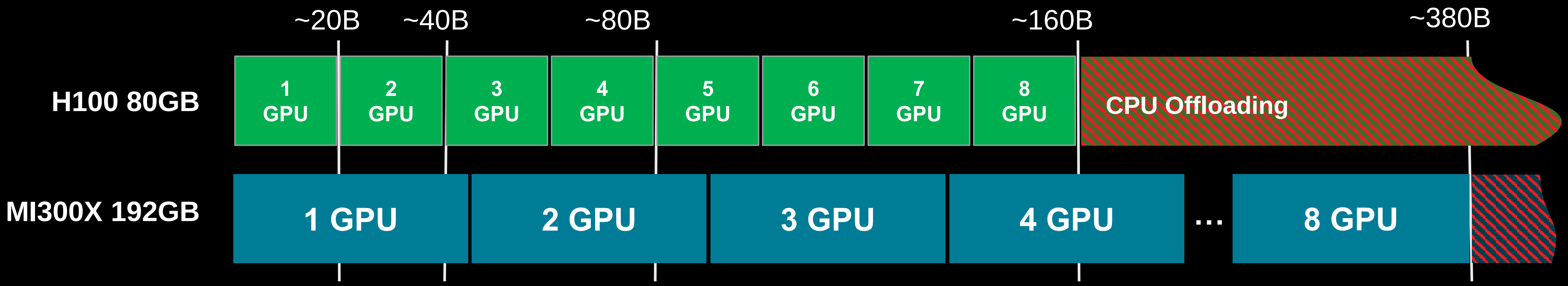
AMD CDNA 3

192 GB
HBM3

~5.3 TB/s
Memory Bandwidth
(Peak Theoretical)

Up to 896 GB/s
AMD Infinity Fabric™ Bandwidth

More Memory Enables Larger Model Size in Single Node



- MI300X provides large TCO benefit by containing the same parameters in half the number of GPUs needed of H100
- MI300X supports up to 2.4x larger model sizes within a node to prevent CPU offloading

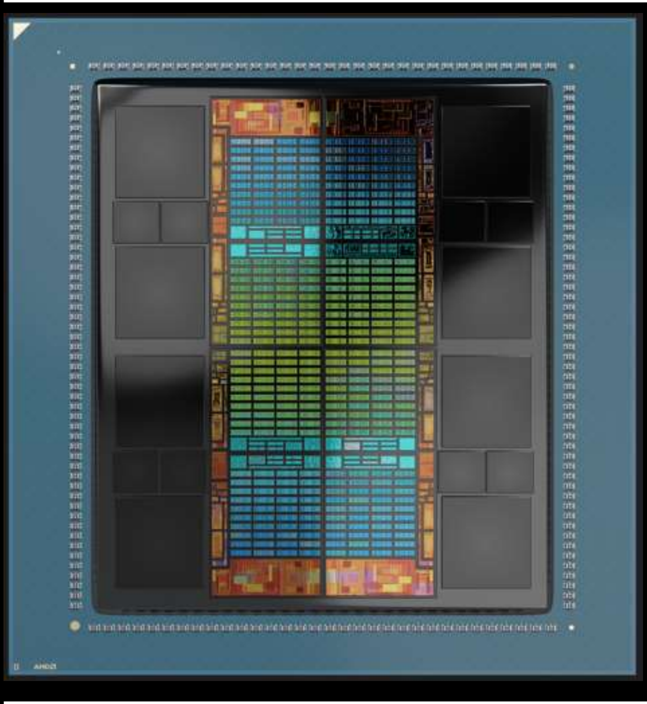
Need for two DGX H100 to match one XE9680 MI300X

Parameter limits based on FP32 w/o Quantization

AMD
together we advance_

AMD Instinct™ MI300X GPU Partitioning

Enabled multiple workloads for optimal GPU utilization

GPU partition options	1	2	4	8
	GPU Instance 192GB	GPU Instance 96GB	GPU Instance 48GB	GPU Instance 24GB
				GPU Instance 24GB
			GPU Instance 48GB	GPU Instance 24GB
				GPU Instance 24GB
		GPU Instance 96GB	GPU Instance 48GB	GPU Instance 24GB
				GPU Instance 24GB
			GPU Instance 48GB	GPU Instance 24GB
				GPU Instance 24GB

Partitioning mode selected applies to all MI300X GPUs on UBB8 Node

No AMD License Fee

Launching Today

AMD Instinct™ MI325X GPU

Extending generative AI leadership

256GB HBM3E	6TB/s	1.3 PF	2.6 PF	AMD CDNA 3
1.8x memory	1.3x bandwidth	1.3x FP16	1.3x FP8	

Compared to Nvidia H200. See endnotes MI325-001a, MI325-002.

Previewing today

AMD Instinct™ MI350 Series

Continued Gen AI Leadership

3nm
Process Node

Up to 288GB
HBM3E

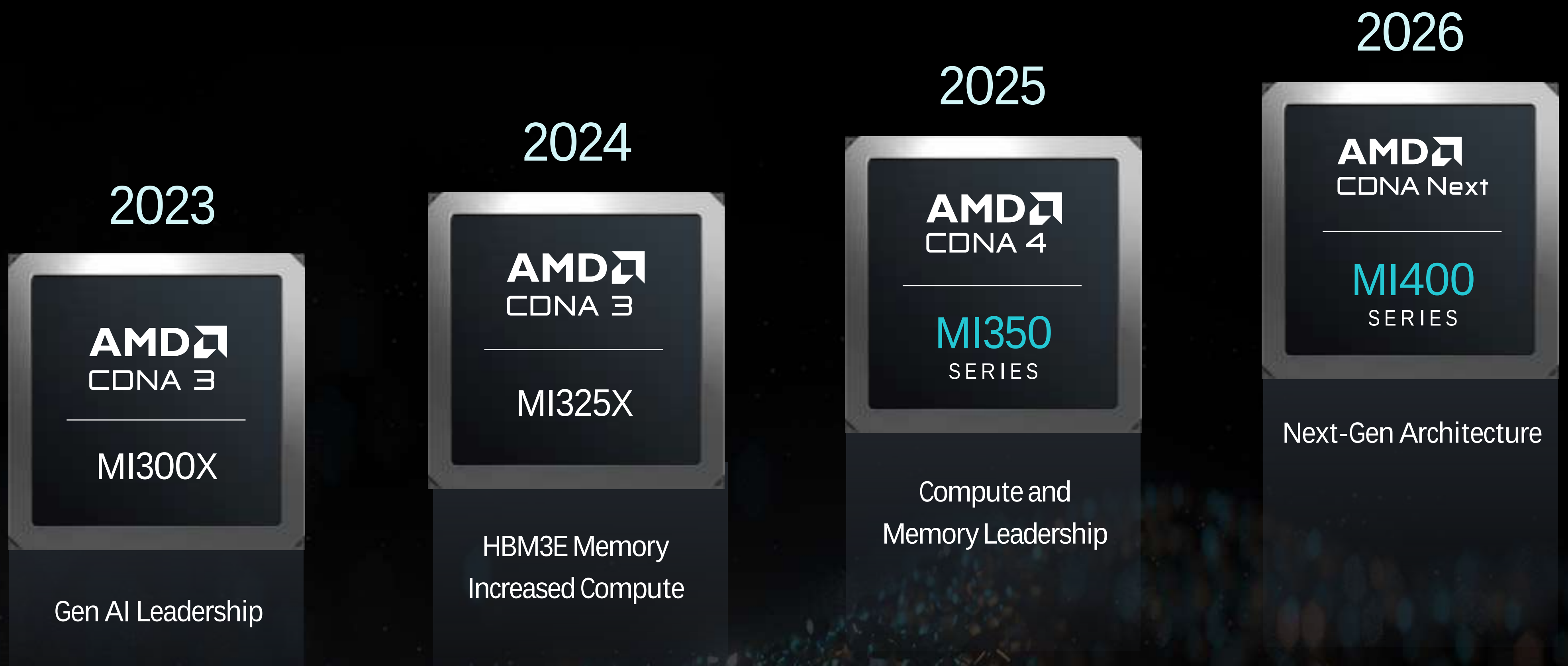
FP4 / FP6
Datatype Support

AMD
CDNA 4

Planned availability 2H 2025

AMD INSTINCT

Leadership roadmap commitment continues



2023

AMD CDNA 3

MI300X

Gen AI 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

Roadmap subject to change

AMD