Cautionary Statement

This presentation contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) such as the features, functionality, performance, availability, timing and expected benefits of AMD products and product roadmaps, the evolving AI landscape, AMD’s ability to advance AI, and the growing AMD EPYC™ market share, 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. 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.

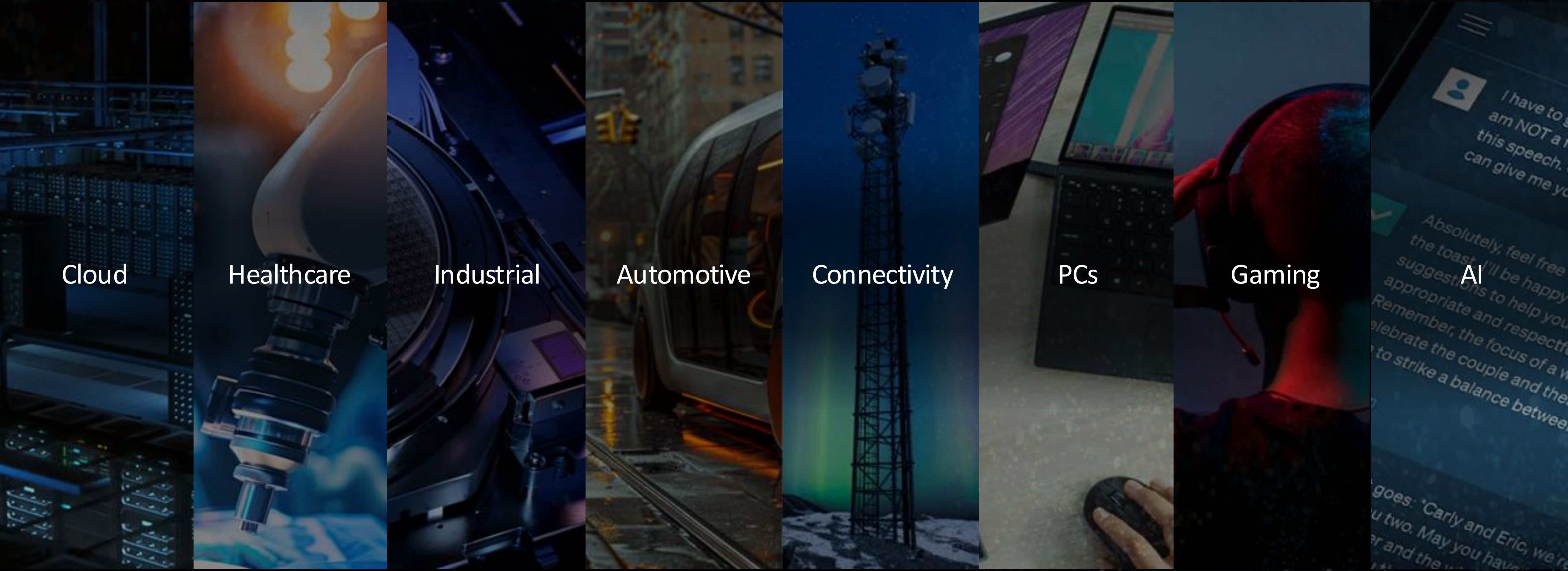
AMD does not assume, and hereby disclaims, any obligation to update forward-looking statements made in this presentation, except as may be required by law.

ADVANCING AI 2024

AMD

together we advance_AI

Computing at the heart of modern life



Cloud

Healthcare

Industrial

Automotive

Connectivity

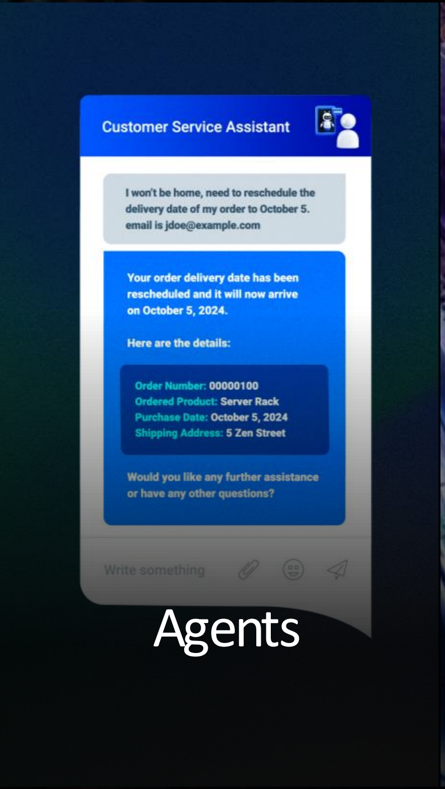
PCs

Gaming

AI

AI

Most transformational technology in 50 years



Customer Service Assistant

I won't be home, need to reschedule the delivery date of my order to October 5. email is jdoe@example.com

Your order delivery date has been rescheduled and it will now arrive on October 5, 2024.

Here are the details:

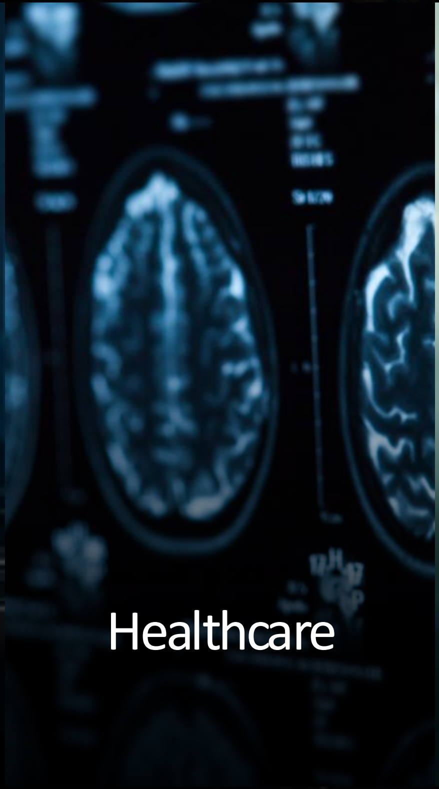
Order Number: 00000100
Ordered Product: Server Rack
Purchase Date: October 5, 2024
Shipping Address: 5 Zen Street

Would you like any further assistance or have any other questions?

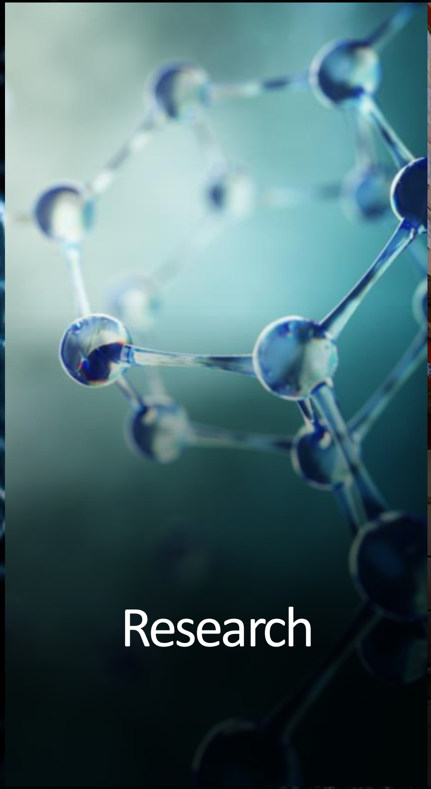
Agents

Smarter Cities

Robotics



Healthcare



Research

Supply Chain

AMD AI Platforms

Unmatched portfolio of training and inference compute engines

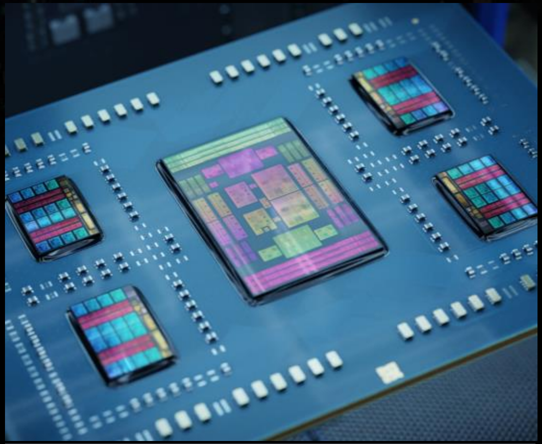
Open software solutions

AI ecosystem with deep co-innovation

Cluster level systems design

Evolving AI Landscape

Innovation moving from silicon to nodes to racks to clusters



Silicon

Server

Rack

Data Center

AMD | Advancing the AI Data Center

+

CPUs
AMD EPYC™

+

AMD PENSANDO

Ultra Ethernet Consortium

UA Link | Ultra Accelerator Link

Networking
DPUs, UALink + Ultra Ethernet

+

AMD ROCm

Software Solutions
Open Software Stack

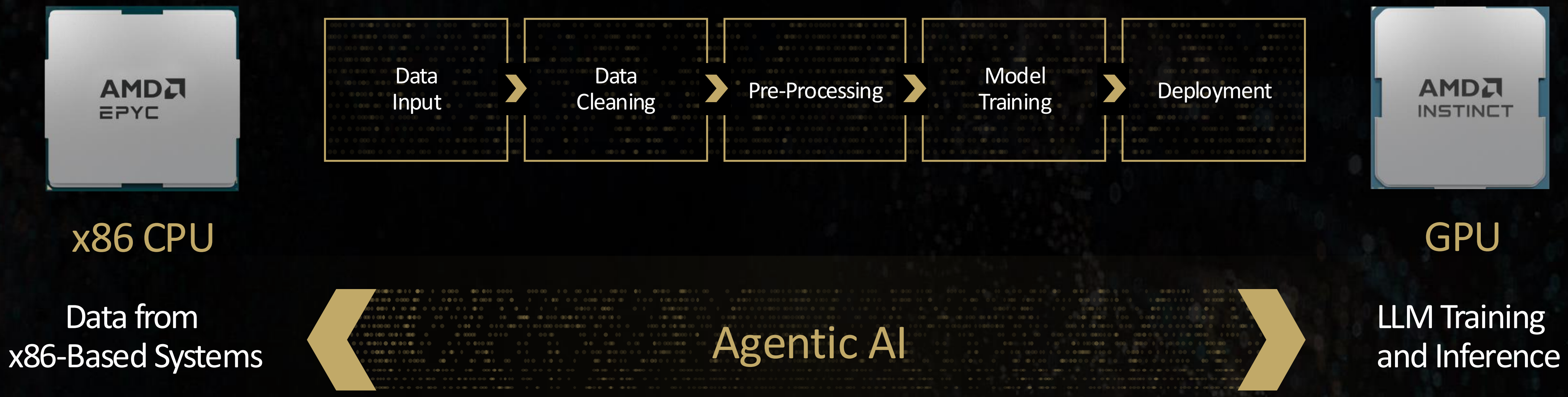
+

zt Systems

Cluster Level Systems Design

Announced acquisition expected to close 1H 2025

Leadership Engines for Enterprise AI Workloads



Data Input

Data Cleaning

Pre-Processing

Model Training

Deployment

x86 CPU

Data from
x86-Based Systems

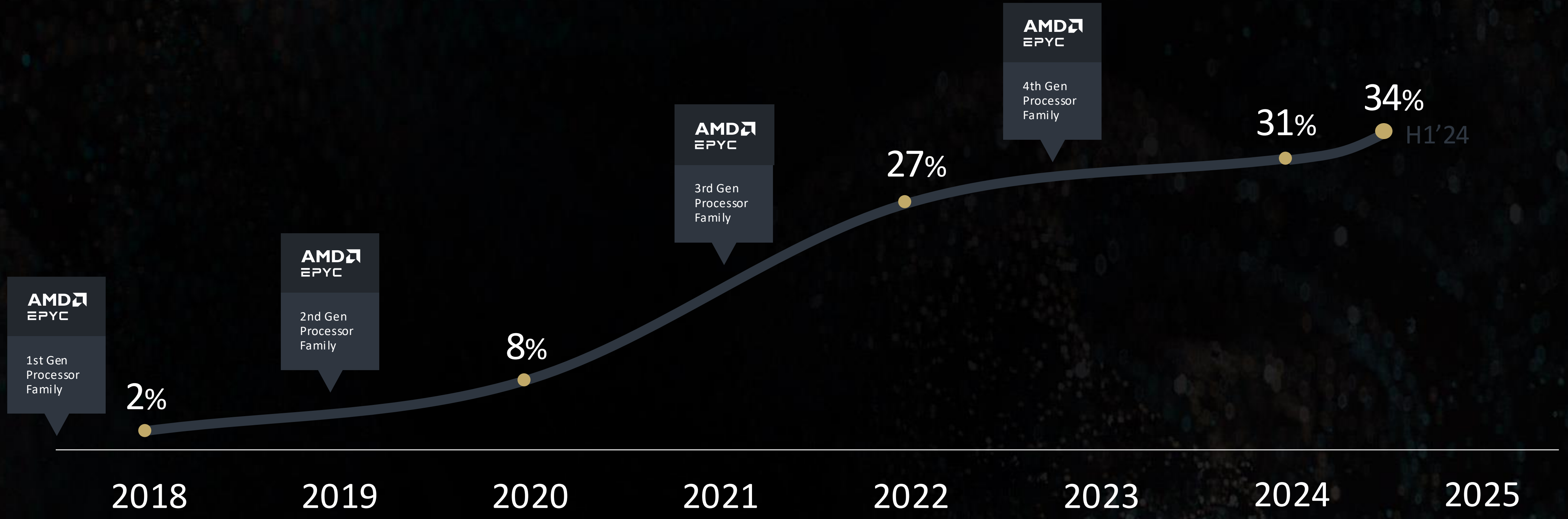
Agentic AI

GPU

LLM Training
and Inference

From analytics to generative AI to agentic AI

AMD EPYC™ record market share…and growing



>350 OEM Platforms | >950 Cloud Instances

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

Trusted by industry leaders on-prem

Adobe
amazon
أرامكو السعودية saudi aramco
ARISTA
BASF
BEST BUY
BNP PARIBAS
BNY MELLON
casa systems
CGG Passion for Geoscience

CITADEL | Securities
core42 A G42 company
DBS
DXC TECHNOLOGY
Emirates NBD
eni
f5
GE
HERSHEY'S
Honeywell

HYUNDAI
IBM
Jane Street
mastercard
Medtronic
NETFLIX
NOKIA
NORTHROP GRUMMAN
PETRONAS
QTS

Qubit.
Raytheon
Reliance
RBC
SILICON LABS
ST
SUBARU
SYNOPSYS
TATA

TESLA
TOPCON Healthcare SEEING EYE HEALTH DIFFERENTLY
tsmc
TURTLE ROCK STUDIOS
UD TRUCKS
UNION PACIFIC
WELLS FARGO
weta DIGITAL

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™ MI300 Series

Powering the most popular Gen AI platforms

Microsoft OpenAI Meta

cohere stability.ai essential AI LAMINI Reka LUMA AI Lepton AI Fireworks AI

databricks 310.AI scale MOREH World Labs anyscale clarifai

UbiOps FlexAI OPEN INNOVATION ZYPHRA Rhymes rapt.ai CLEAR ML NEURAL MAGIC

Right and Left Screen

AMD INSTINCT

Solutions from leading OEMs and cloud

Microsoft ORACLE DELL Technologies Hewlett Packard Enterprise Lenovo SUPERMICRO

EVIDEN an atos business CISCO ASUS GIGABYTE TYAN PENGUIN COMPUTING AIVRES ASRock Rack

Aligned Elio MiTAC QCT NSCALE VULTR Crusoe

Cirrascale CLOUD SERVICES TENSORWAVE COMPAL Inventec ingrasys wiwynn wistron

Today at Advancing AI 2024

AMD end-to-end AI infrastructure leadership

Data Center CPUs

5th Generation
AMD EPYC™ “Turin”

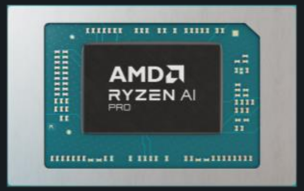
Data Center GPUs

AMD Instinct™
MI325X and MI350 Series

Networking

AMD Pensando™
Pollara 400/Salina 400

AI PCs



3rd Generation
AMD Ryzen™ AI PRO

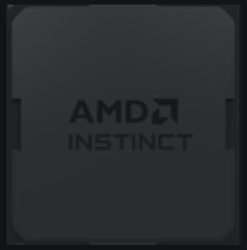
Today at Advancing AI 2024

AMD end-to-end AI infrastructure leadership

Data Center CPUs

5th Generation
AMD EPYC™ "Turin"

Data Center GPUs

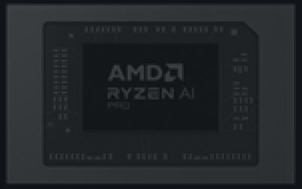


AMD Instinct™
MI325X and MI350 Series

Networking

AMD Pensando™
Pollara 400/Salina 400

AI PCs



3rd Generation
AMD Ryzen™ AI PRO

Launching Today: “Turin”

5th Gen AMD EPYC™

World’s best CPU for Cloud, Enterprise and AI

3nm
4nm

150 billion
transistors

Up to 192 cores
384 threads

17% IPC uplift
Full AVX512

Up to 5 GHz

*~17% Across 36 cloud and enterprise workloads
As of 10/1/2024. See endnotes 9xx5-001, EPYC-029C

"Turin" Continues AMD EPYC™ Leadership

Consistent x86 ISA
Consistent IPC

SP5 Socket
"Genoa" Compatible

8 to 192 cores
125W to 500W

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

Confidential Compute
with Trusted I/O

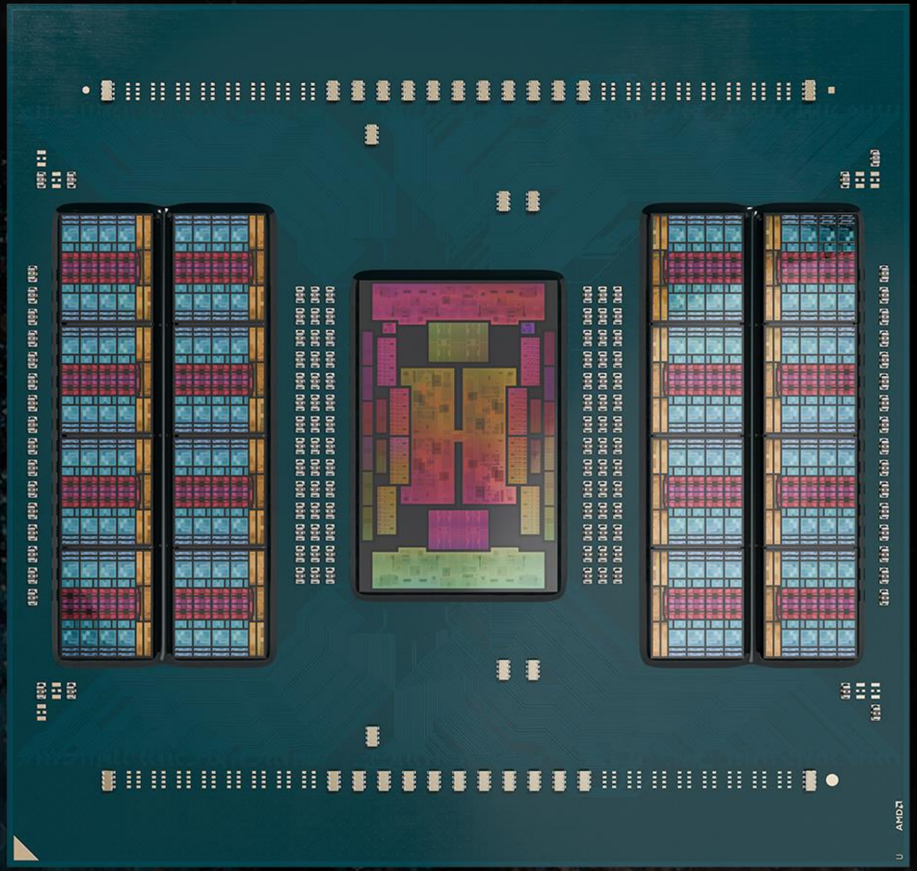
As of 10/1/2024. See endnotes 9xx5-048, 9xx5-083: 5th Gen EPYC processors support DDR5-6400 MT/s for targeted customers and configurations. 5th Gen production SKUs support up to DDR5-6000 MT/s to enable a broad set of DIMMs across all OEM platforms and maintain SP5 platform compatibility

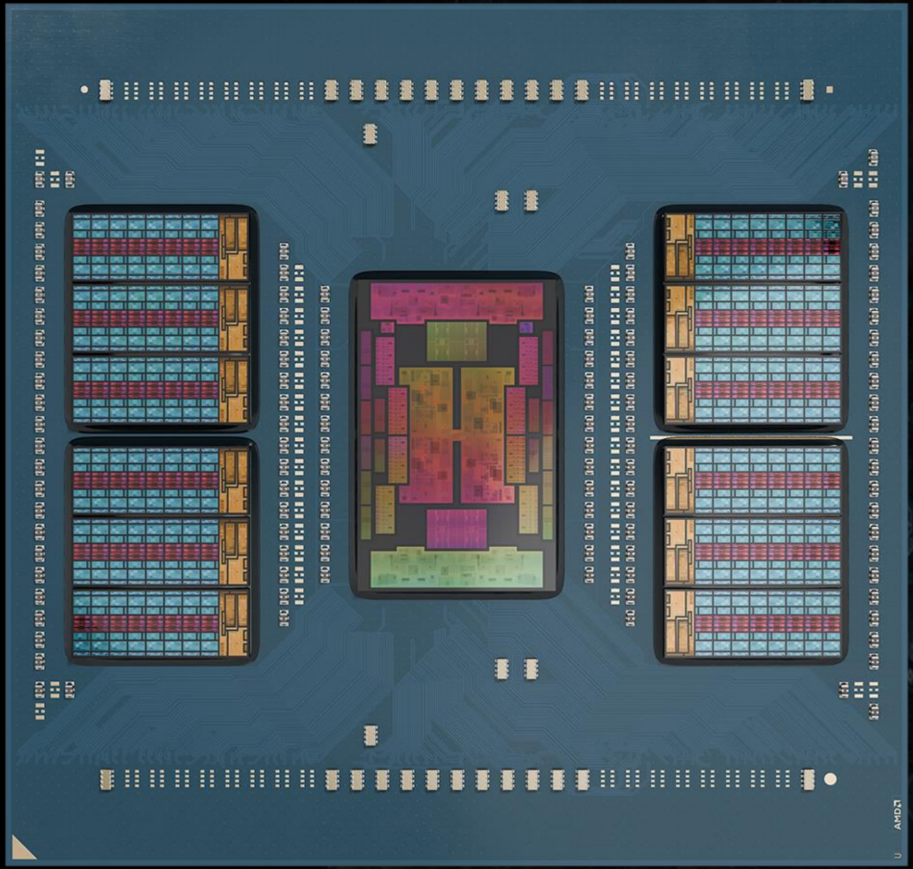
Scale-Up

Up to

16 “Zen 5” CCDs

128 Cores · 256 Threads





Scale-Out

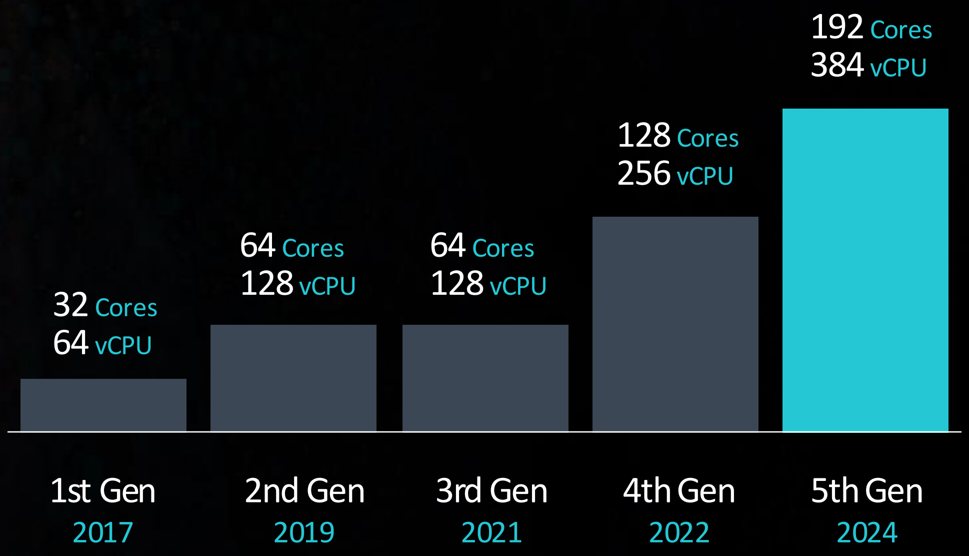
Up to

12 “Zen 5c” CCDs

192 Cores · 384 Threads

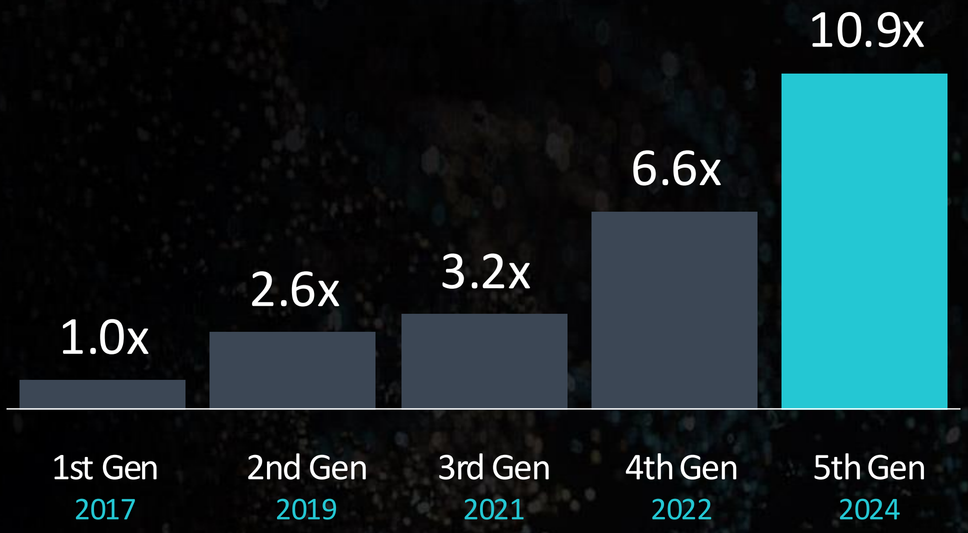
Fifth Generation of CPU Leadership Starts Today

6x Core Count
Across 5 Generations



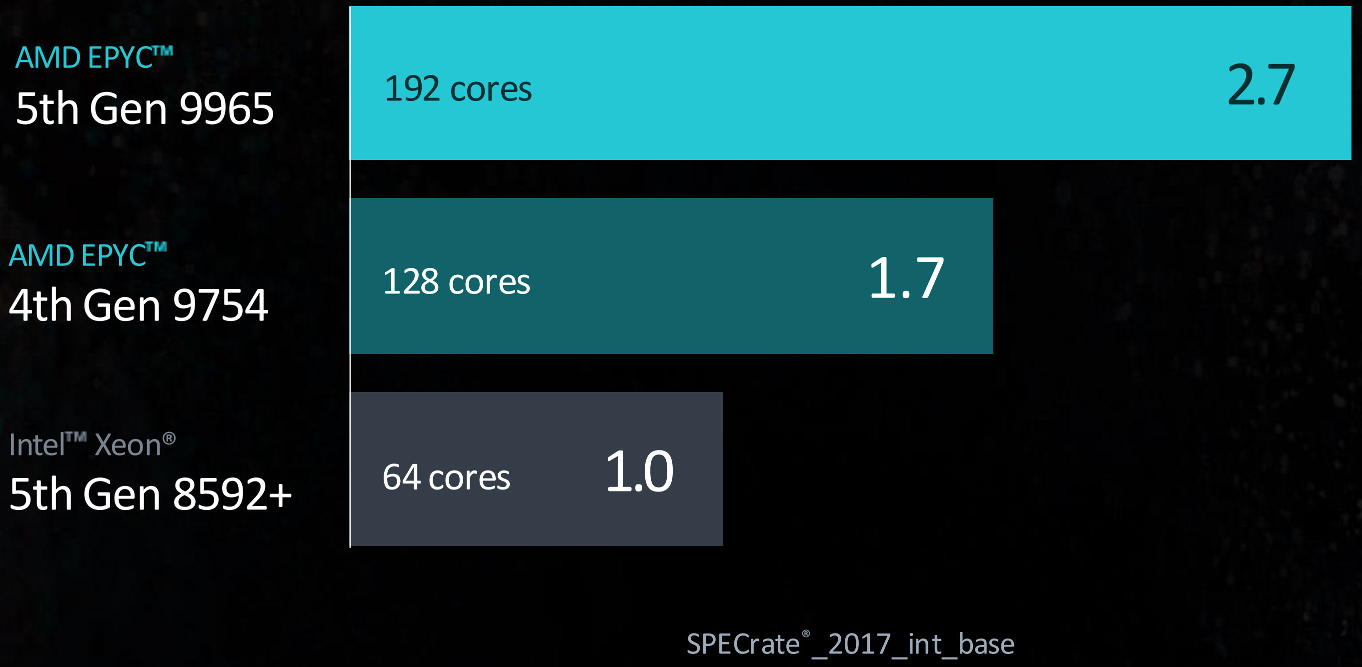
10.9x Performance
Across 5 Generations

Relative SPECrate2017_int_base best score by end of launch year



As of 10/1/2024. See endnote 9xx5-069a

Industry's Highest Performing Server CPU

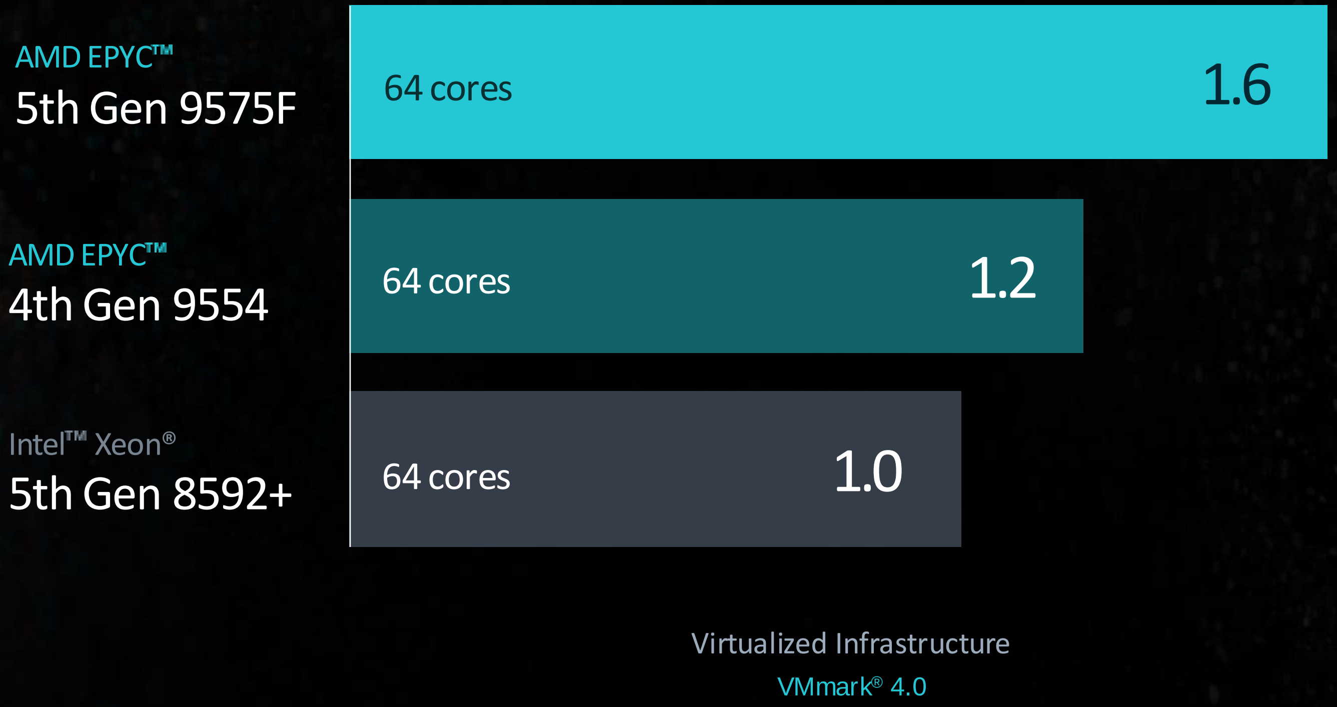


2.7x

vs. top-of-stack "Emerald Rapids"

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

60% More Performance at the Same Licensing Cost

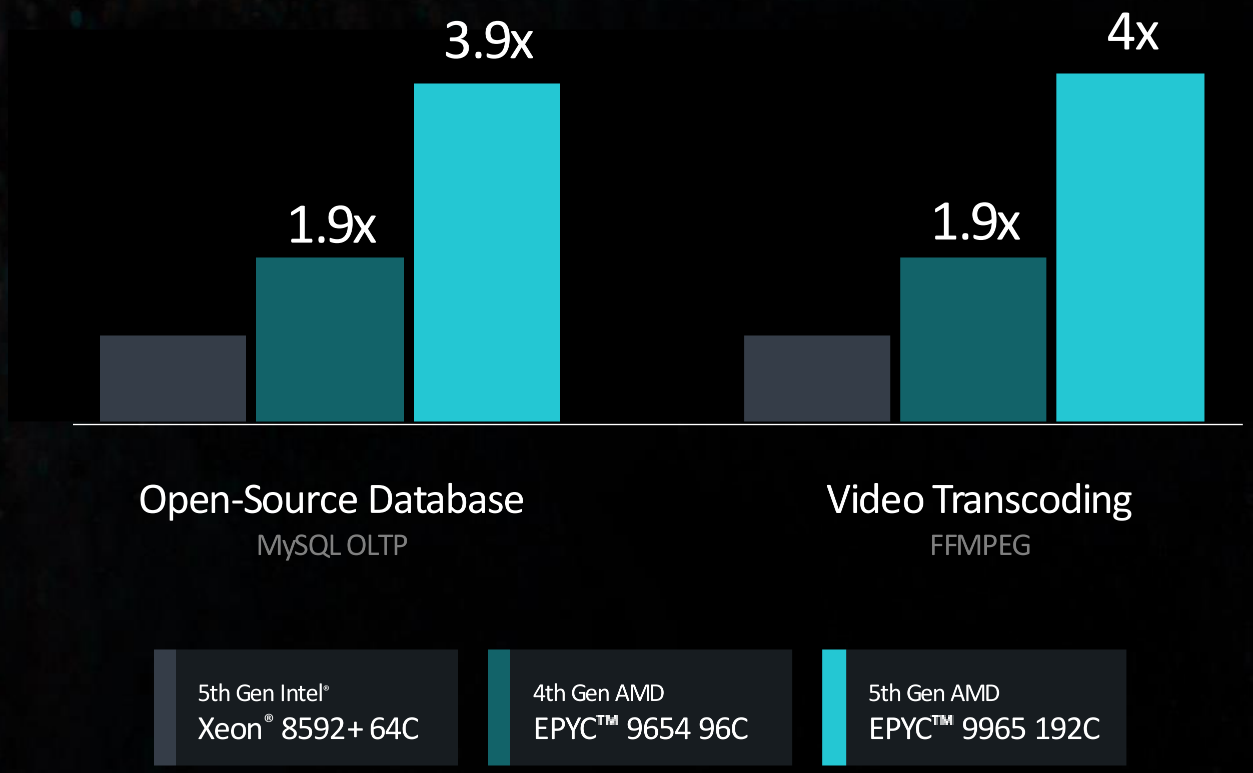


up to 1.6x

Performance per core in virtualized infrastructure

As of 10/4/2024. See endnote 9xx5-071.

Grow Your Database and Media Processing Capabilities

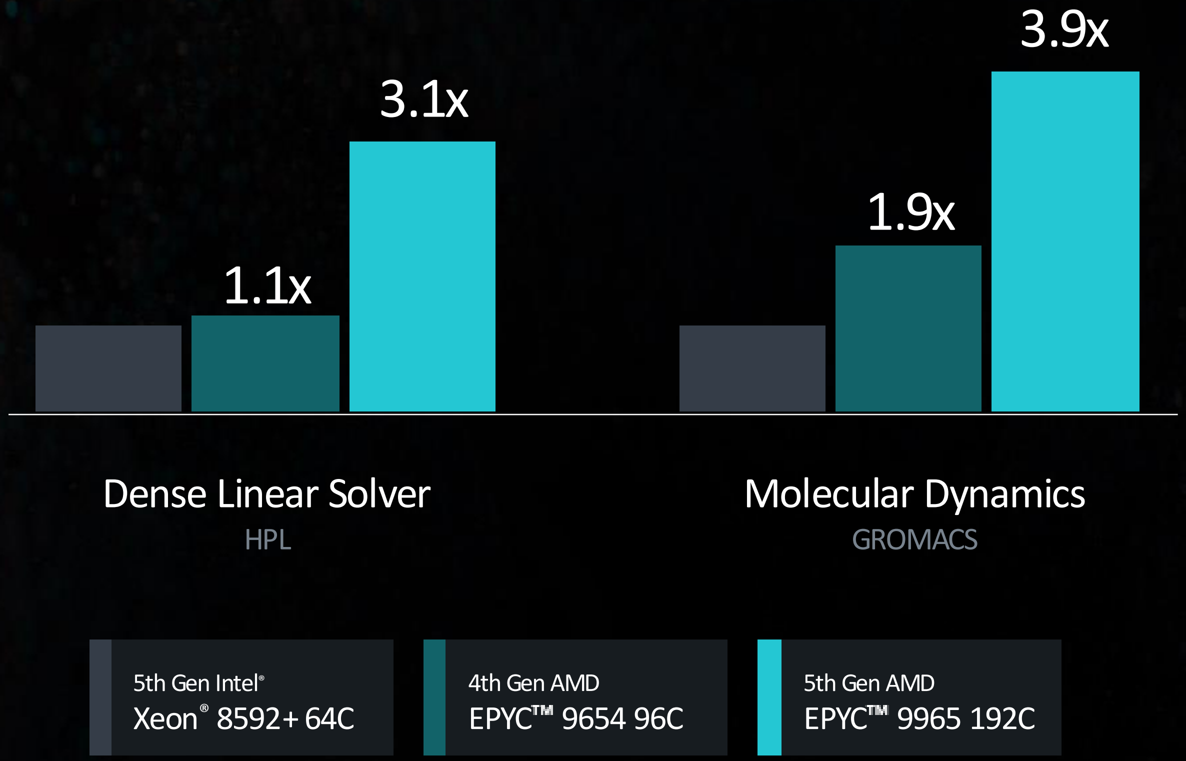


up to **4x**

Throughput performance

As of 10/1/2024. See endnote 9xx5-005A and 9xx5-006.

Fastest CPU for the Most Challenging HPC Problems

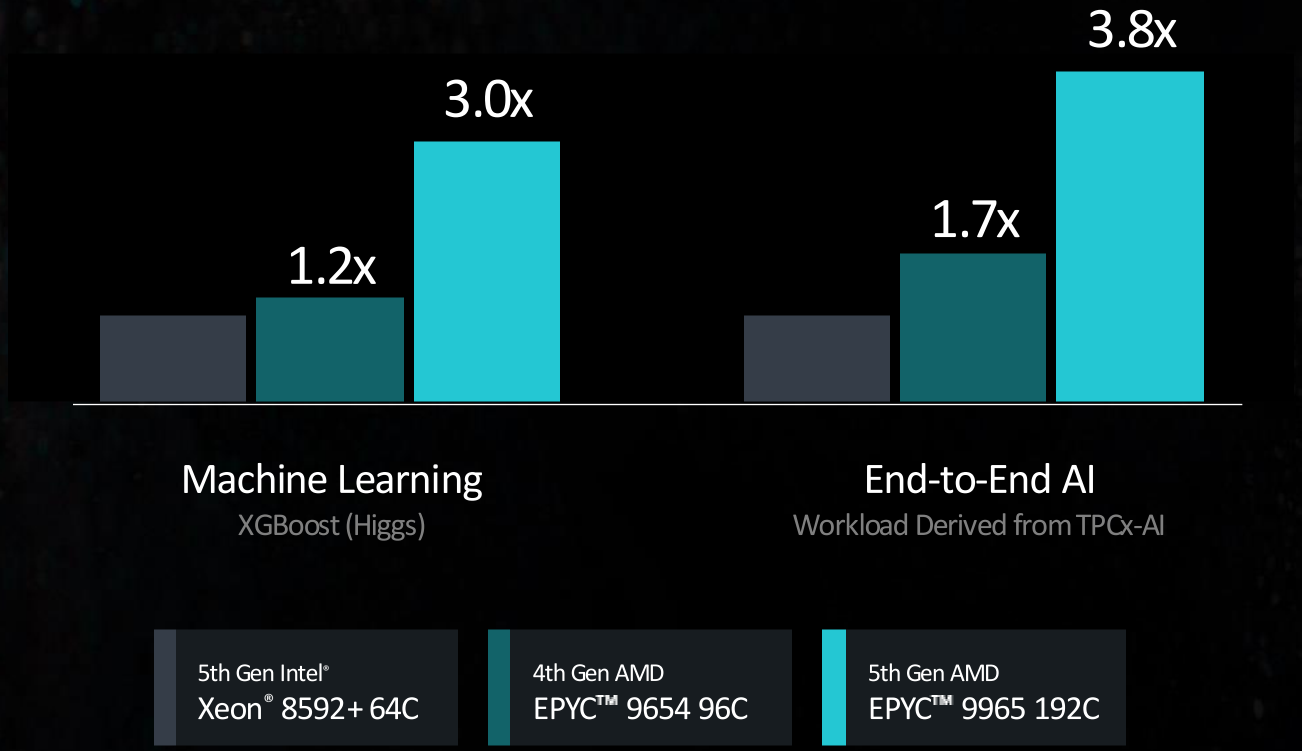


up to 3.9x

Improved time to insight

As of 9/30/2024. See endnote 9xx5-038, 9xx5-039 and 9xx5-022.

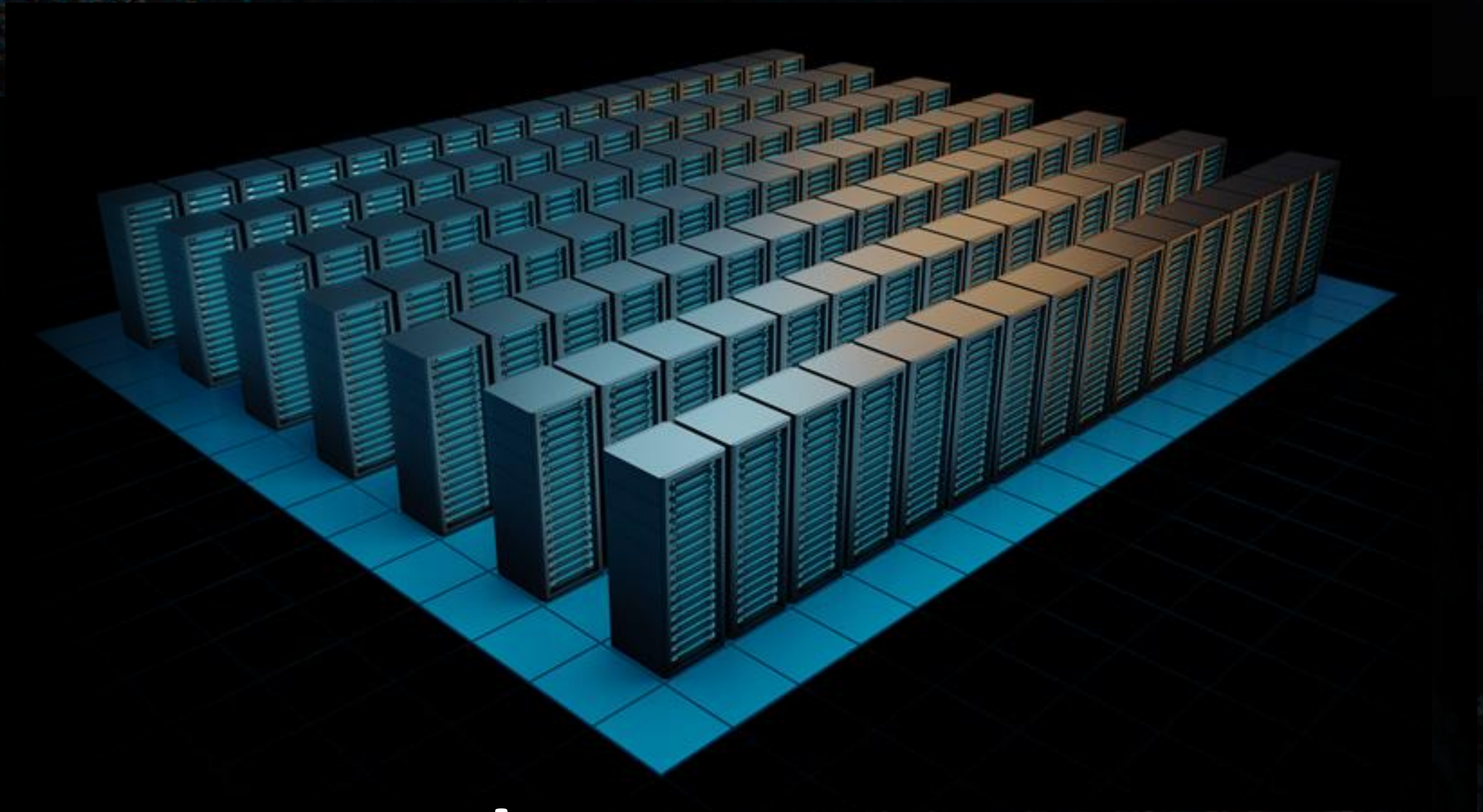
End-to-End AI and Inference Performance



up to **3.8x**

AI performance on CPU

As of 9/30/2024. See endnote 9xx5-040a and 9xx5-012

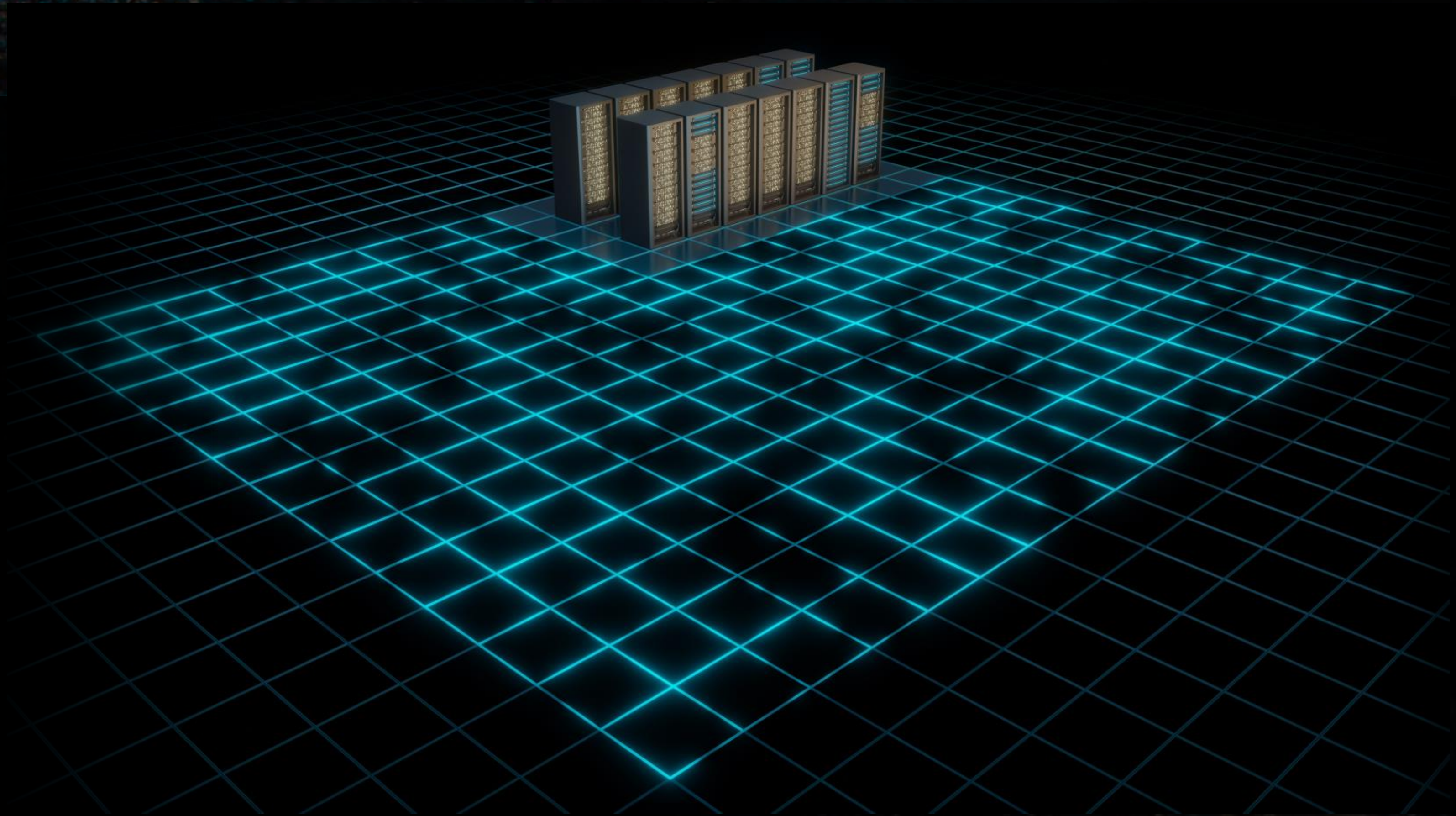


1,000 legacy servers

2P Intel® Xeon® Platinum 8280 servers

To deliver 391,000 unit of integer performance

Servers required to achieve a total of 391,000 SPEC®rate_2017_ int_base performance score. AMD EPYC 9965 SPEC®rate_2017_int_base score is estimated



131 modern servers

2P AMD EPYC™ 9965

To deliver 391,000 unit of integer performance

7:1 consolidation

Use the savings, space and power to grow your business

~87% fewer servers | ~67% lower TCO | ~68% less power

2P EPYC™ 9965 vs 2P Intel Xeon® 8280 to deliver 391,000 unit of integer performance

As of 10/4/2024. See endnote 9xx5TCO-002a.

Today at Advancing AI 2024

AMD end-to-end AI infrastructure leadership

Data Center CPUs

5th Generation
AMD EPYC™ "Turin"

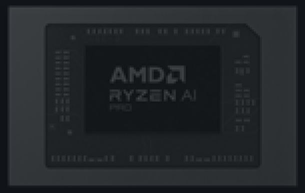
Data Center GPUs

AMD Instinct™
MI325X and MI350 Series

Networking

AMD Pensando™
Pollara 400/Salina 400

AI PCs



3rd Generation
AMD Ryzen™ AI PRO

$45B

2023

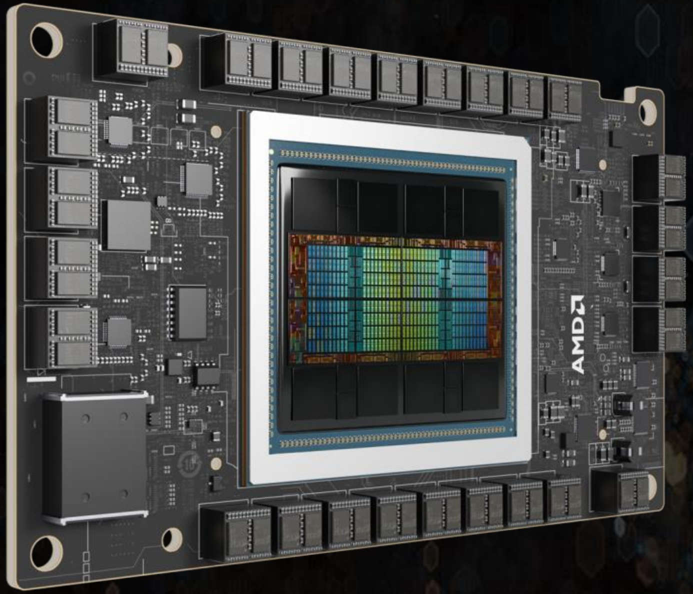
>60% CAGR

Data Center AI Accelerators

Source: AMD

$500B

2028



AMD Instinct™ MI300 GPU

Fastest ramping product in AMD history

Advancing ROCm™ performance and ecosystem

~2x

Improvement in inference and training performance

~3x*

ROCm now supports 1M+ models out of box*

PyTorch ONNX

TensorFlow

Deepening partnership with AI ecosystem

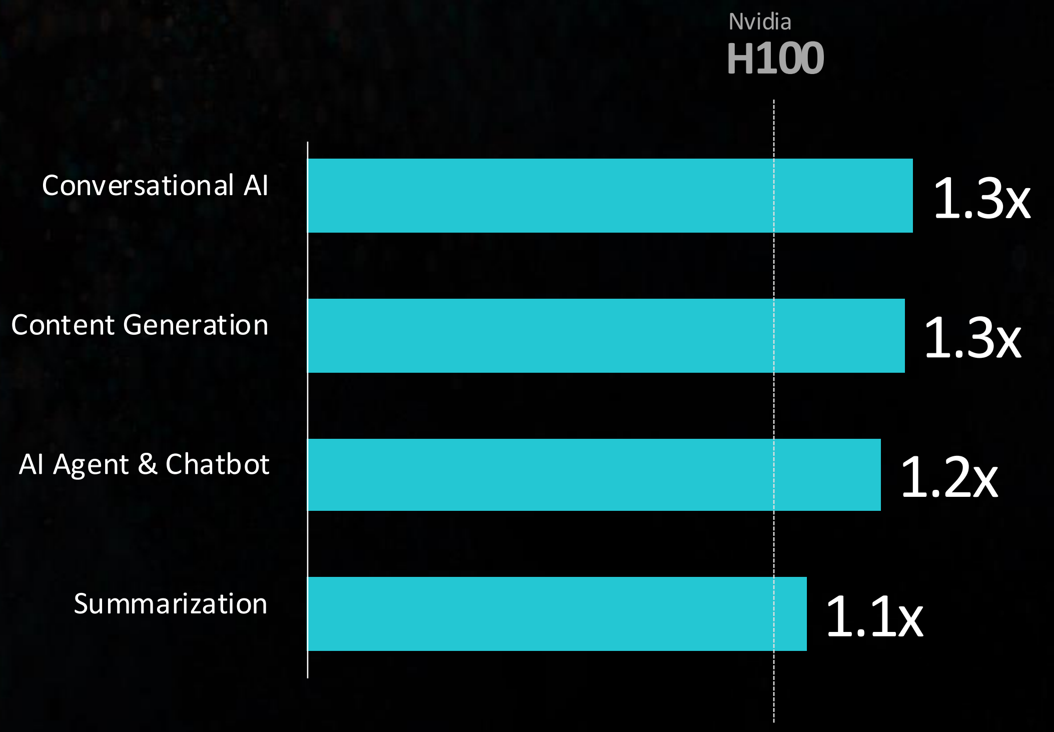
AMD

SILO AI

Growing AI expertise and customer implementations

See endnote MI300-62. ROCm 6.2 vs. ROCm 6.0. HuggingFace Models today vs. Dec 2023

Demonstrated inference leadership at key customers



Llama 3.1 • 405B Latency Improvement

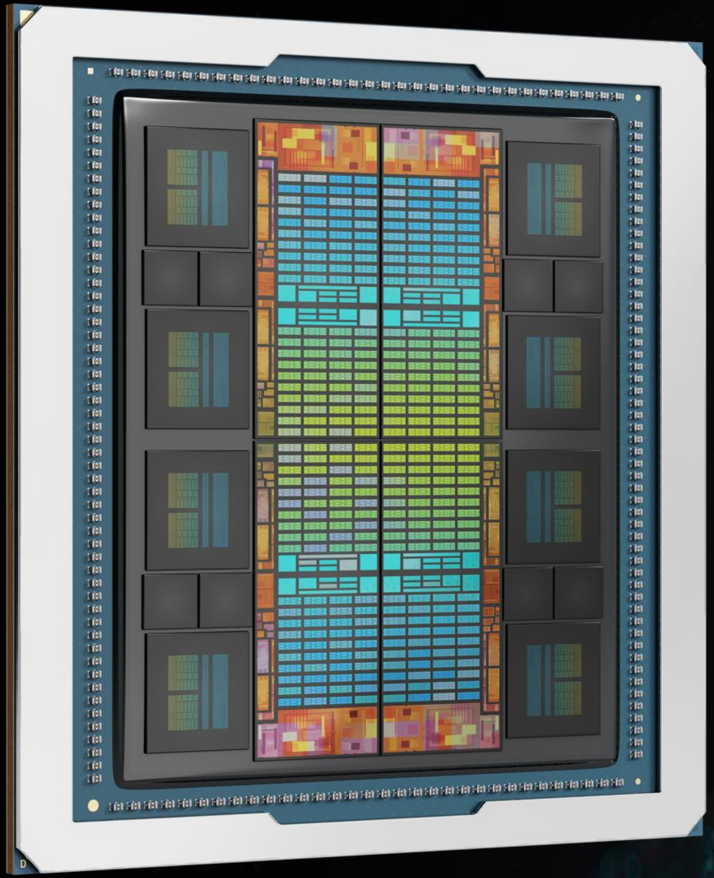
up to **1.3x**

Higher performance across key workloads

Llama 3 · Llama 2 · Mistral · Mixtral

Qwen · CommandR · Stable Diffusion

See endnotes MI300-053, MI300-054, MI300-064.



Launching Today

AMD Instinct™ MI325X GPU

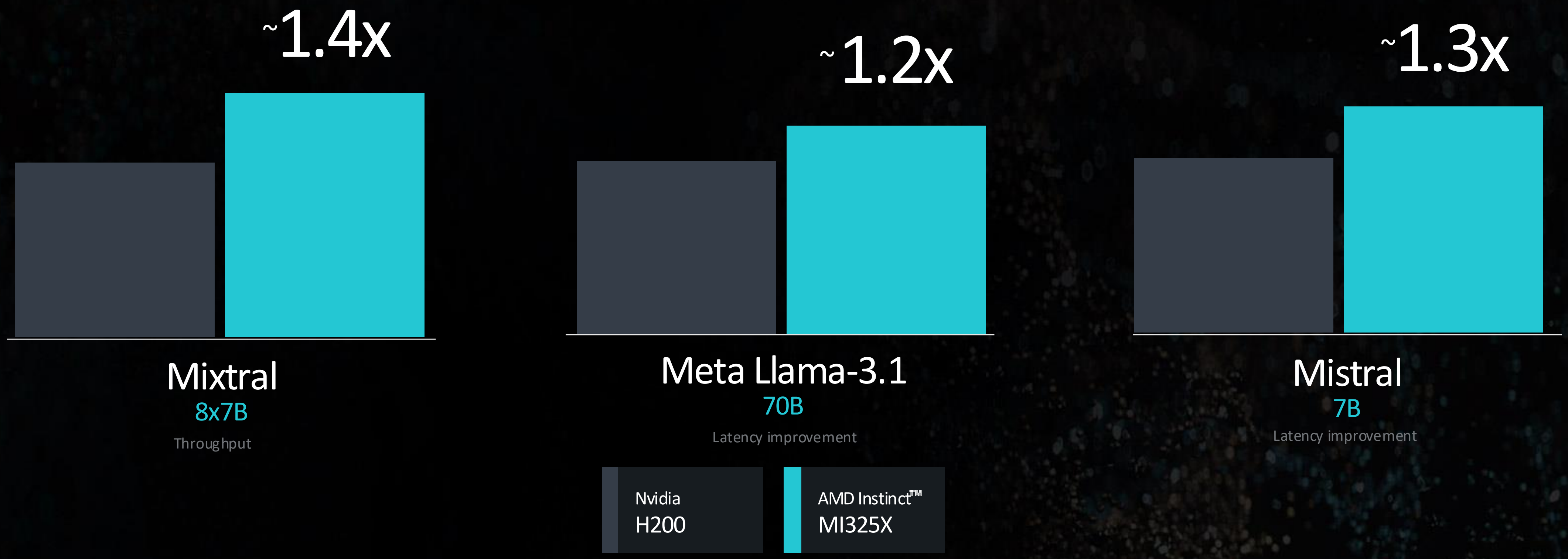
Extending generative AI leadership

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	

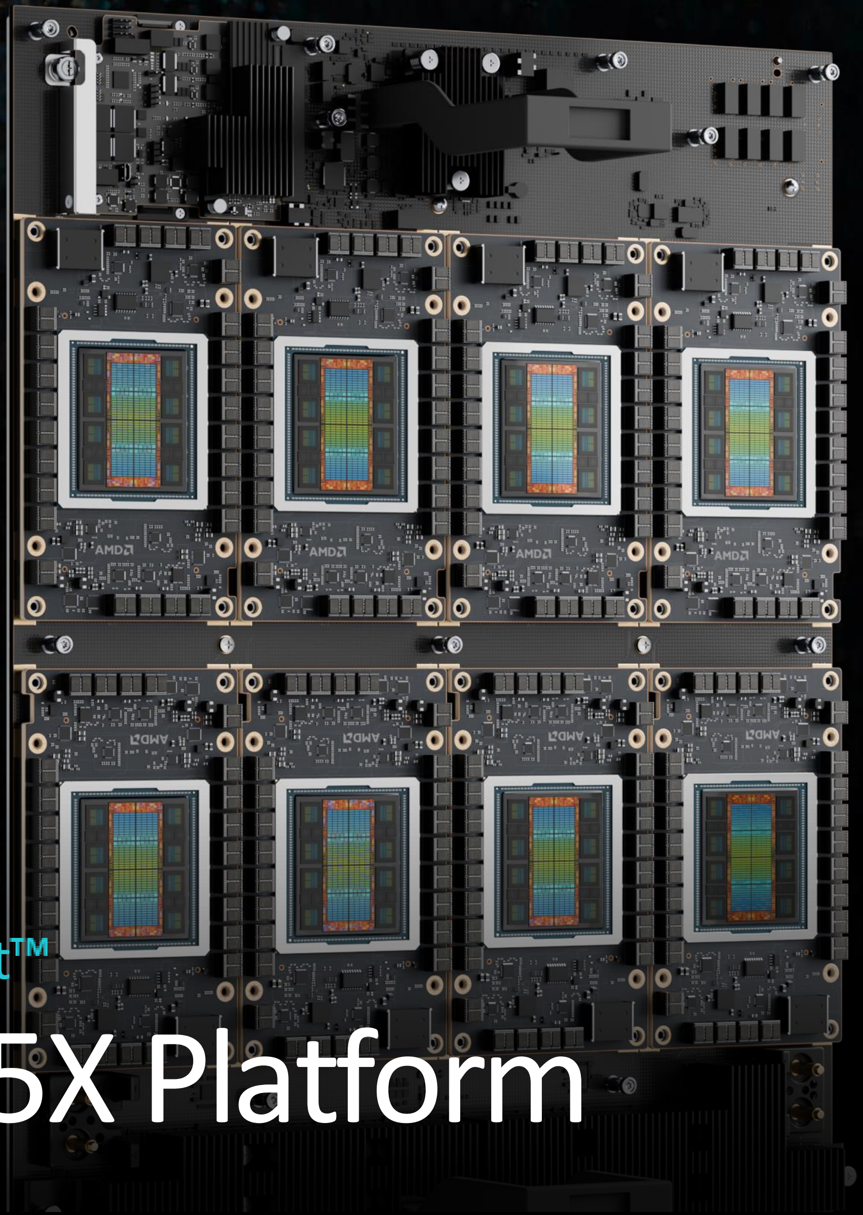
Compared to Nvidia H200. See endnotes MI325-001a, MI325-002.

AMD Instinct™ MI325X GPU

Leadership inference performance



See endnotes:MI325-004, MI325-005, MI325-006



AMD Instinct™

MI325X Platform

2 TB | HBM3E

1.8x memory vs. H200 HGX

48 TB/s | Memory Bandwidth

1.3x memory bandwidth vs. H200 HGX

10.4 PF | FP16

1.3x compute flops vs. H200 HGX

20.8 PF | FP8

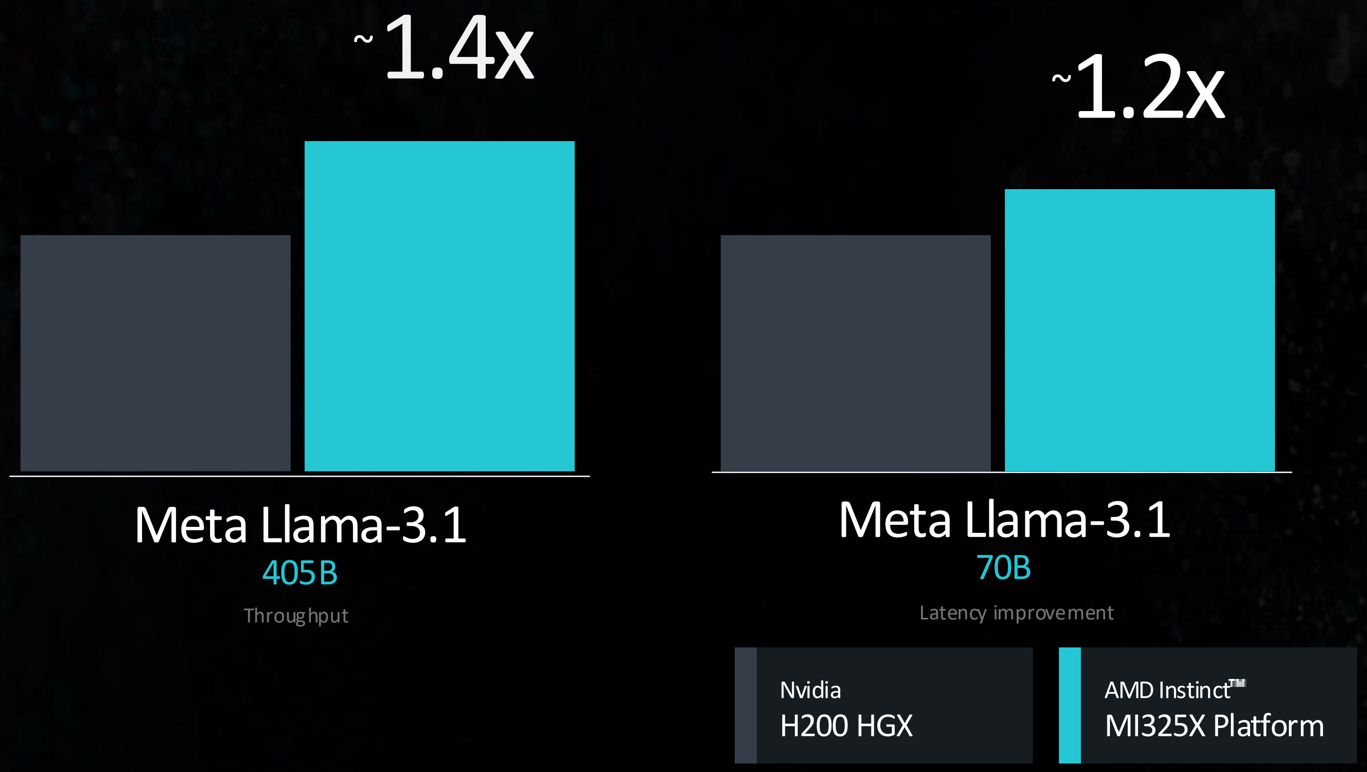
1.3x compute flops vs. H200 HGX

See endnotes MI325-001A, MI325-002

*Dense flops

AMD Instinct™ MI325X Platform

Leadership Inference performance using 8x MI325X



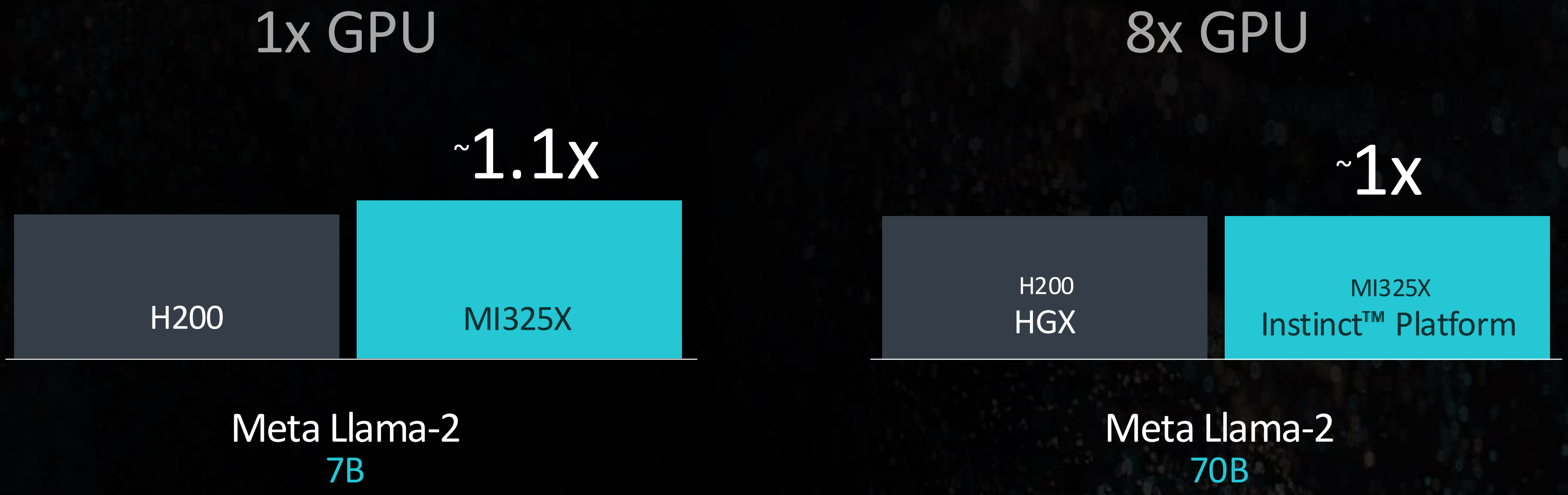
up to **1.4x**

Inference performance

See endnotes MI325-015, MI325-014

World-Class Training Performance

Single GPU and 8 GPU Training



Nvidia H200 HGX

AMD Instinct™ MI325X Platform

See endnotes MI325-013, MI325-012.

AMD Instinct™ MI325X GPU

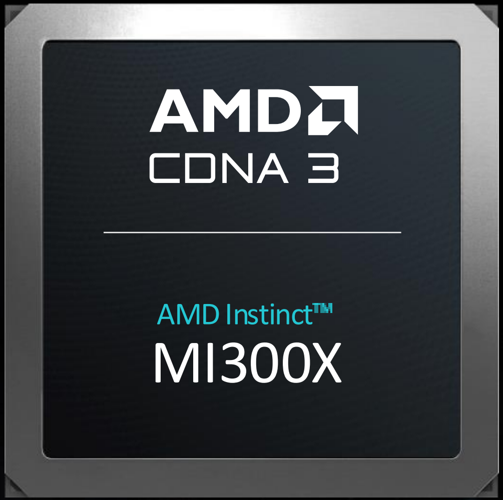
Production starting in Q4 2024

DELL Technologies · EVIDEN an atos business · GIGABYTE™ · Hewlett Packard Enterprise · Lenovo · SUPERMICRO

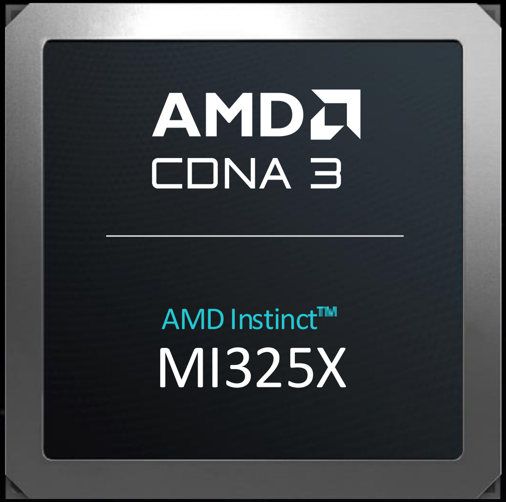
AIVRES · ASRock Rack · ASUS · Celestica · ingrasys · Inventec · MiTAC · QCT · wistron · wiwynn

Available from leading system and infrastructure solution partners starting Q1 2025

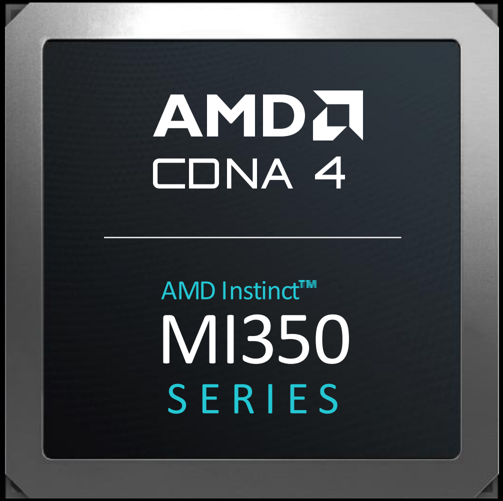
AMD Instinct™ Annual Roadmap Cadence



2023



2024



2025

Roadmap subject to change

Previewing today

AMD Instinct™ MI350 Series

Continued Gen AI Leadership

3nm
Process Node

Up to 288GB
HBM3E

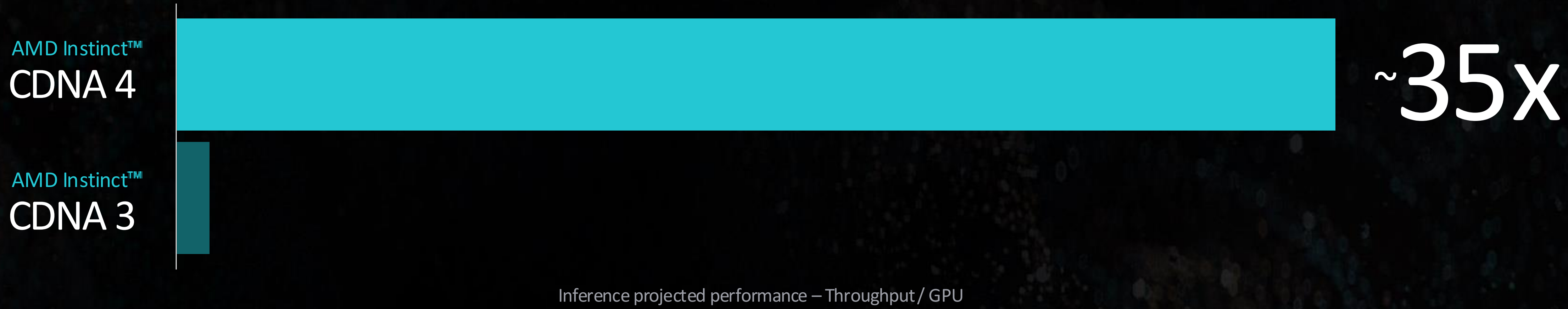
FP4 / FP6
Datatype Support

AMD CDNA 4

Planned availability 2H 2025

AMD CDNA 4

Generational performance leap



~7x
AI Compute

~1.5x
Memory Capacity / Bandwidth

FP4 and FP6
New Datatypes

Improved
Network Efficiency

7x AI compute compares MI300x FP16 to MI355x FP4. See endnotes MI355-004, MI355-002

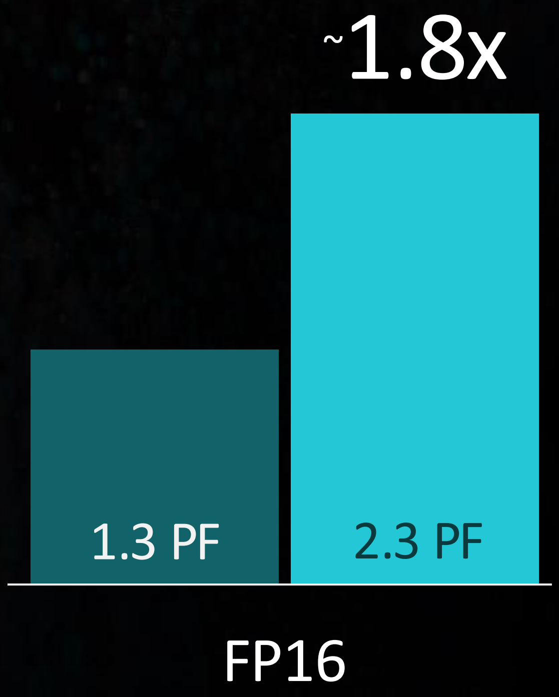
AMD
CDNA 4

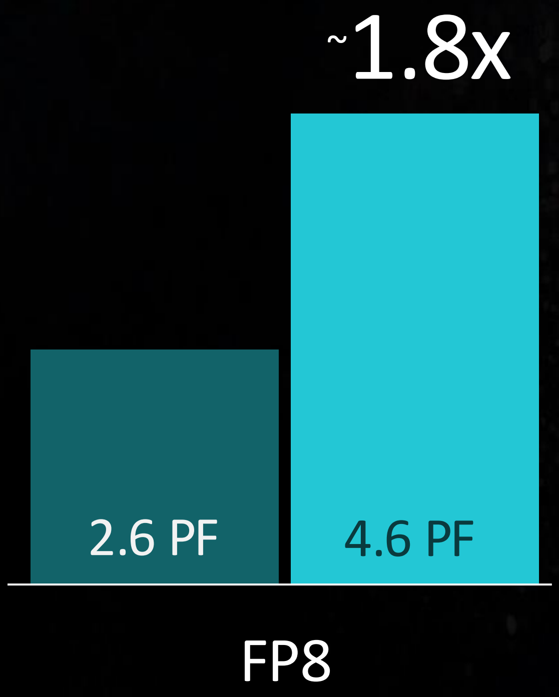
New Datatypes	~3.5x AI Flops vs. FP8
AI Compute	~1.8x FP16 / FP8
HBM3E Memory	~1.5x memory capacity memory bandwidth

AMD Instinct MI355X vs AMD Instinct MI300X. See endnotes: MI355-005, MI355-002. Actual results may vary.

AMD Instinct™ MI355X Accelerator

Leadership performance for Gen AI





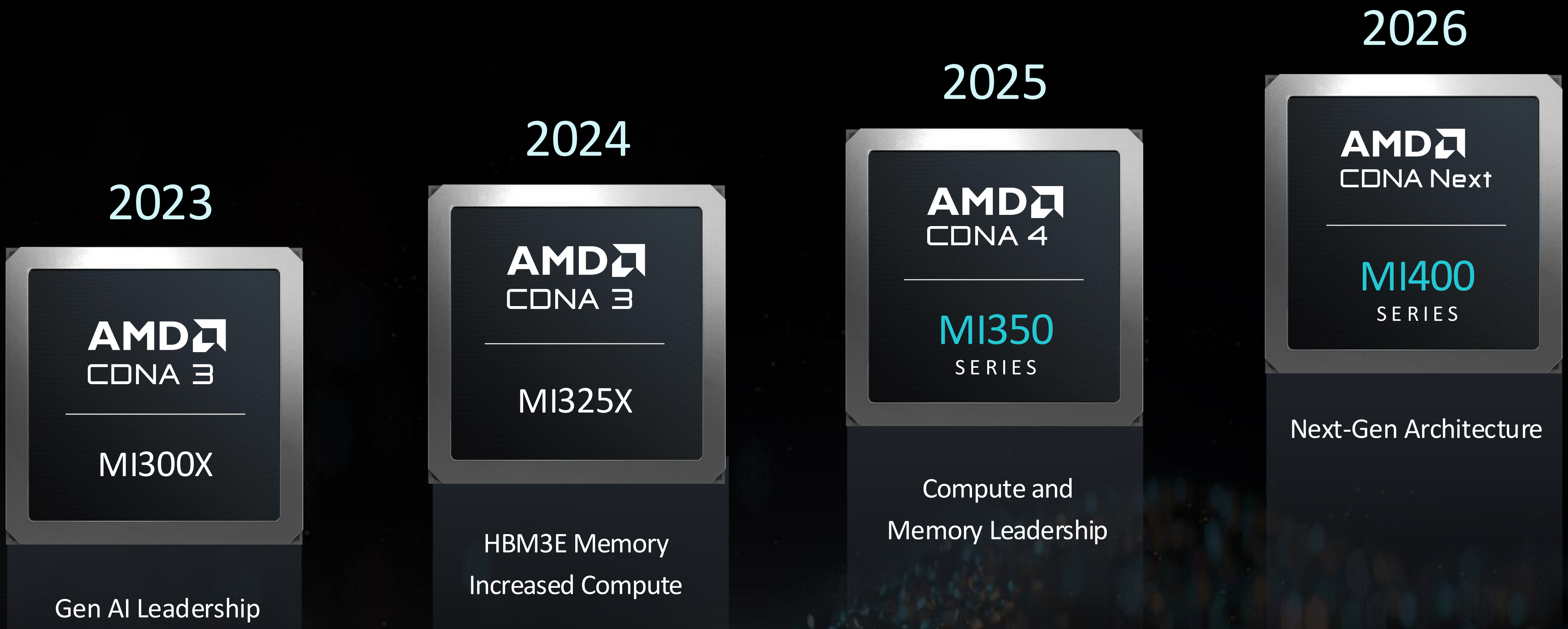
9.2 PF

Introducing FP6 and FP4

See endnotes:MI355-005

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

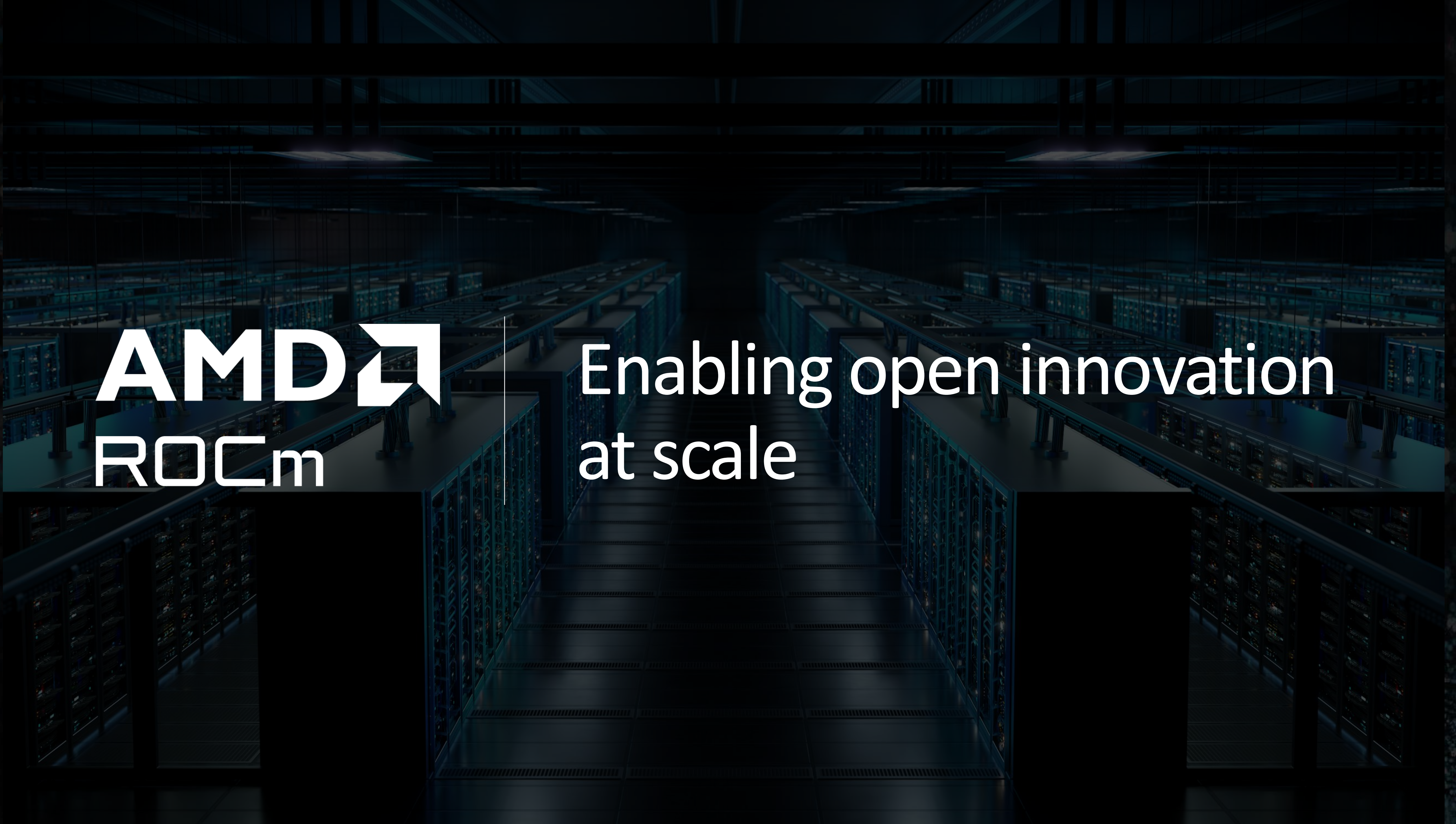
2026

AMD CDNA Next

MI400 SERIES

Next-Gen Architecture

Roadmap subject to change

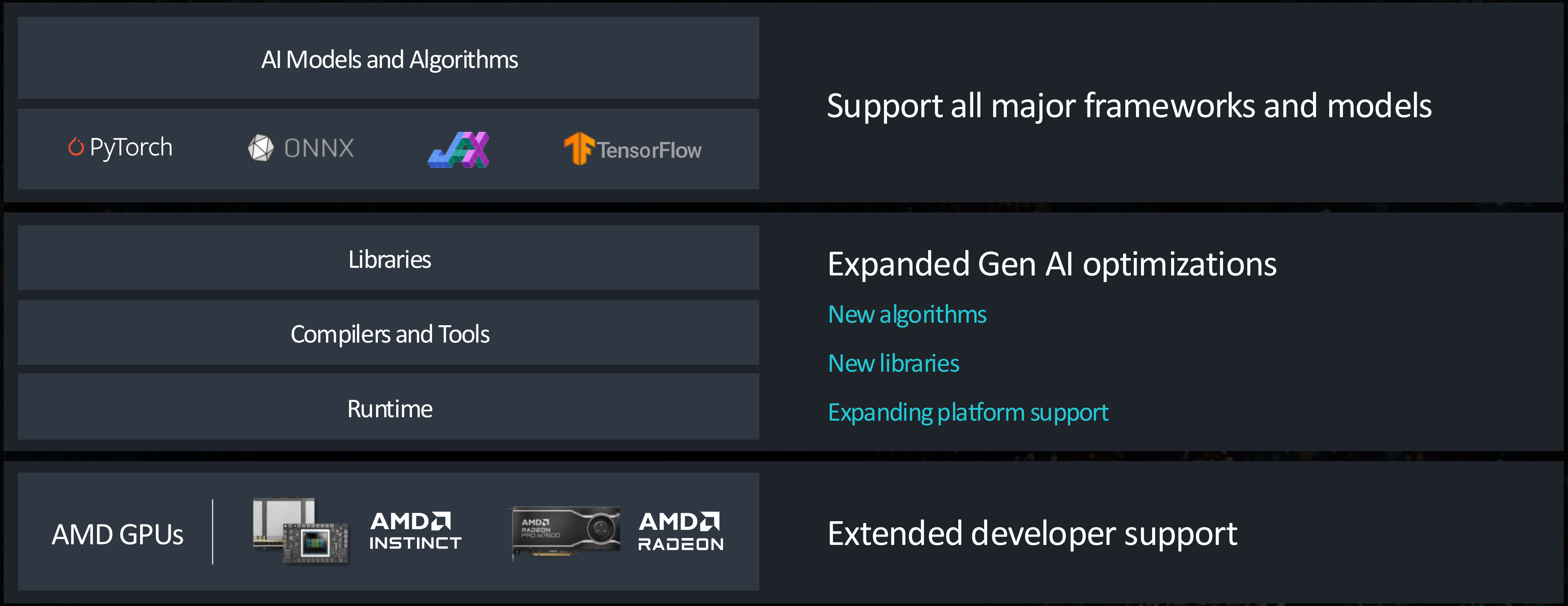


AMD ROCm

Enabling open innovation at scale

AMD ROCm

Open, modular software stack



AI Models and Algorithms

PyTorch ONNX JAX TensorFlow

Support all major frameworks and models

Libraries

Compilers and Tools

Runtime

Expanded Gen AI optimizations

New algorithms

New libraries

Expanding platform support

AMD GPUs | AMD INSTINCT AMD RADEON

Extended developer support

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

AMD ROCm

Deep collaboration with developer community

PyTorch

Day 0 support for latest features

Triton

Vendor agnostic compiler support

Hugging Face

Nightly CI/CD ensuring all models work out-of-box

- vLLM
- SGLang
- JAX
- Tensor Flow
- ONNX Runtime
- OpenXLA
- DeepSpeed
- MLIR | IREE

Increasing open-source contributions and expanding footprint

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

AMD Instinct™ MI300X Accelerator

Performant out-of-box support on popular generative AI models

1M+

models supported out of the box

Hugging Face

Extended support for leading models

Meta Llama 3.1

Meta Llama 3.2

Stable Diffusion 3

Day 0 support for AMD GPUs

Llama 3 405B latency improvement

MI300X vs. H100

Leadership performance on popular models

AMD Instinct™ MI300X Accelerator

Expanding out-of-box support on popular generative AI models

GPT-4
GPT-4o
Llama 2
Llama 3
MISTRAL AI_
Mixtral
Grok 1
Stable Diffusion

DBRX
Qwen
Zamba
Yi
StarCoder
Phi
Flux
Hugging Face

CommandR
Gemma
BLOOM
MPM4
GPT-NeoX
OLMo
Falcon LLM
Aria
MPT

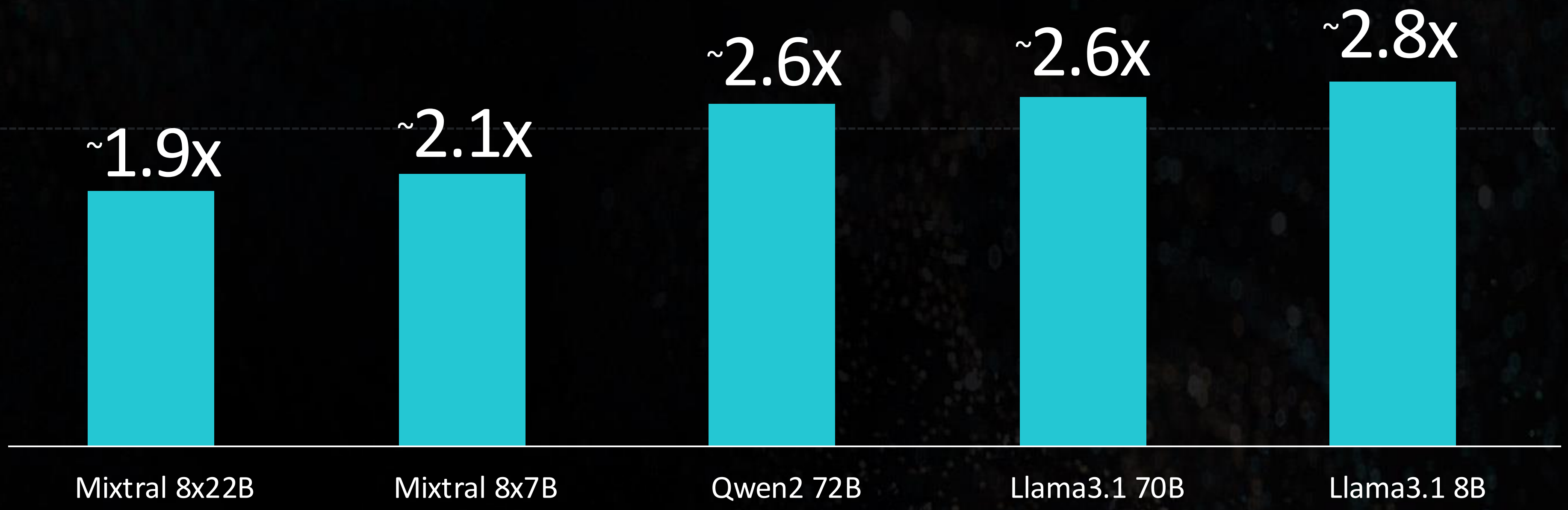
Use or mention of third-party marks, logos, products, services, or solutions herein is for informational purposes only and no endorsement by AMD is intended or implied. GD-83

Generational inference improvement

ROCm™ 6.2 vs. ROCm 6.0

~2.4x

average performance improvement



Runtime Optimization | Kernel Fusion | Collective Communication | Subgraph

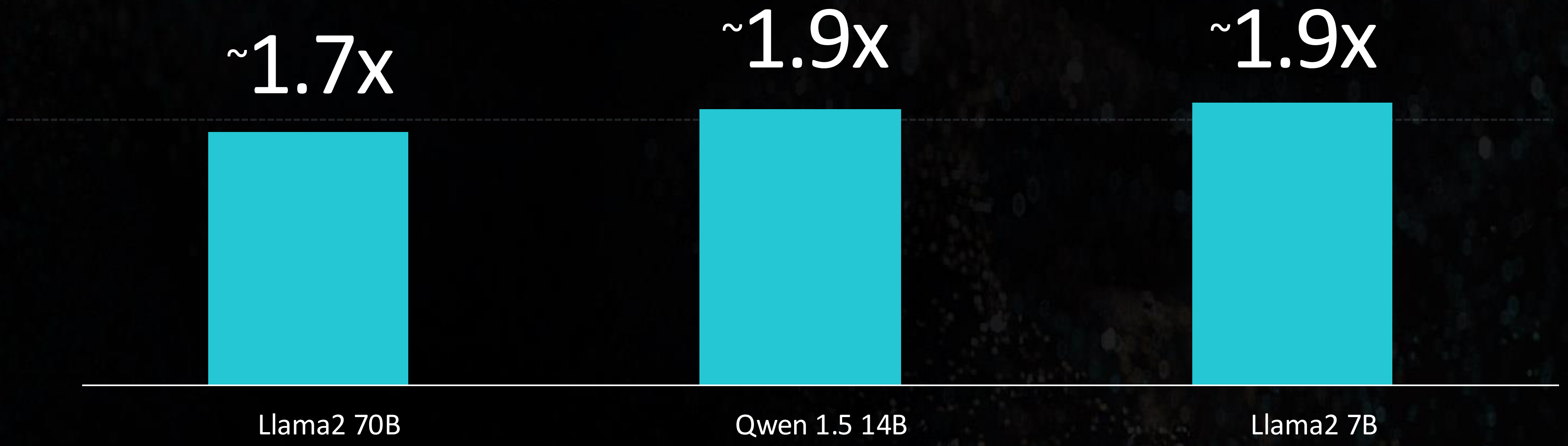
As of 9/30/2024. See endnote MI300-62.

Generational training improvement

ROCm™ 6.2 vs. ROCm 6.0

~1.8x

average performance improvement



Flash Attention 3 | Parallelization Strategy | Kernels | Scale out Efficiency

As of 9/30/2024. See endnote MI300-63.

AMD
SILO AI

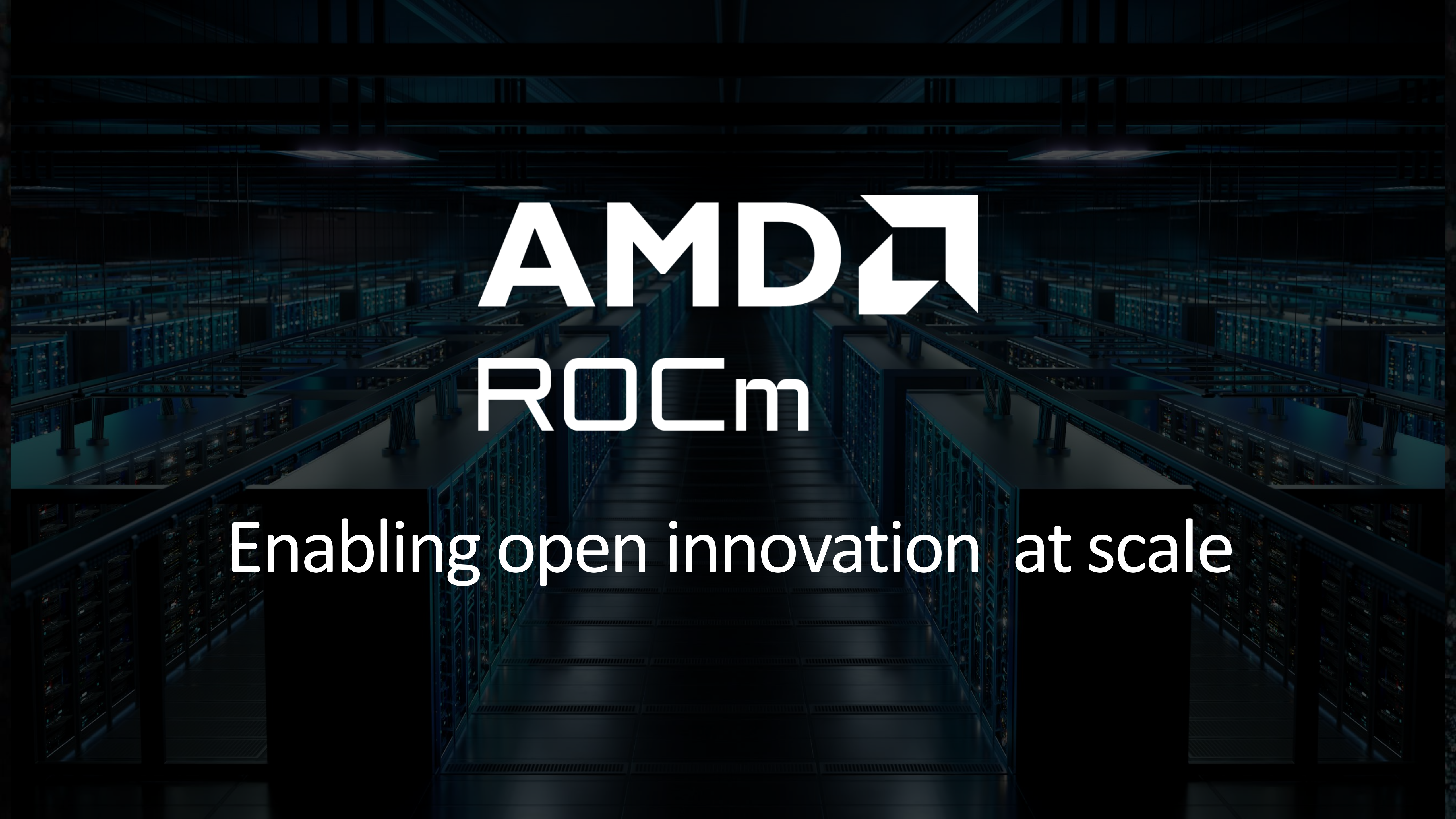
Solving the last mile of customer AI

200+ AI implementations

Helping clients succeed in building AI driven products and solutions

Open-source base models

European language LLMs trained on LUMI supercomputer with AMD Instinct™ accelerators



AMD

ROCm

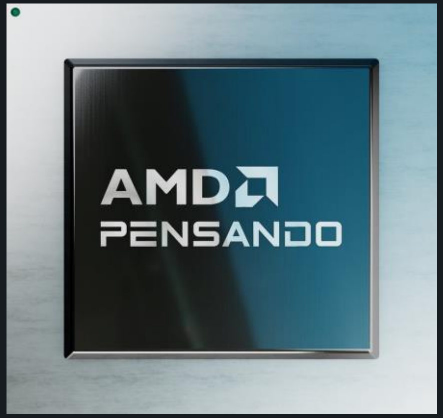
Enabling open innovation at scale

Advancing Data Center Solutions

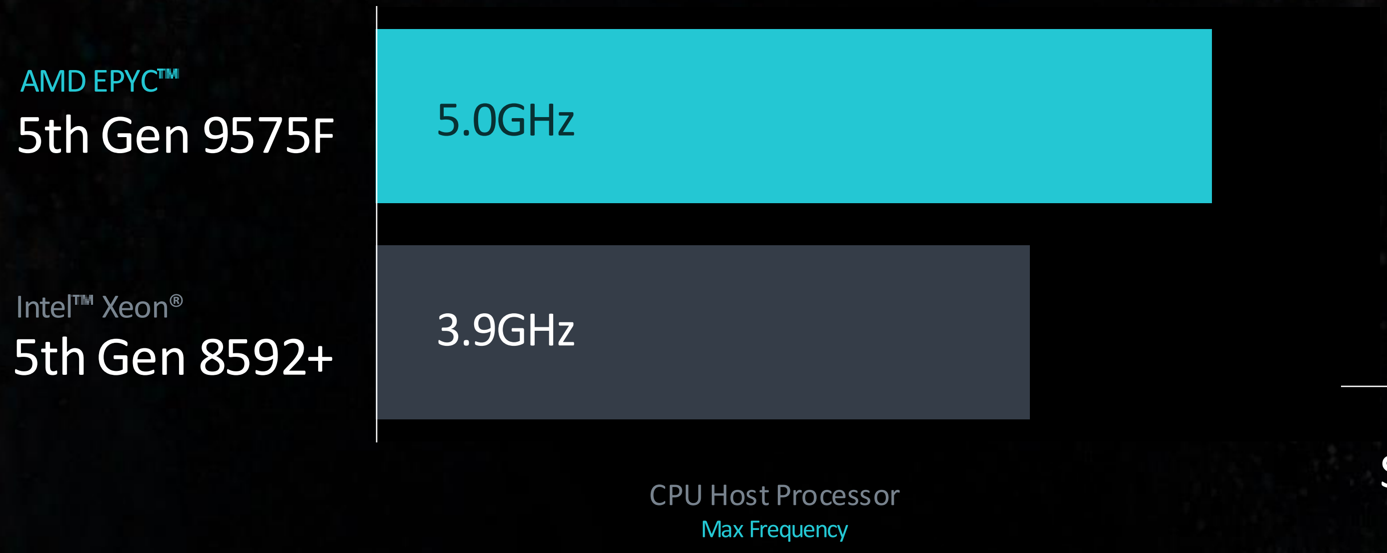
Data Center CPUs

Data Center GPUs

Networking



CPU Performance Enhances GPU Performance



28%

Faster processing for GPU orchestration tasks

Speeds up data prep, memory copies, kernel launch and task orchestration

AMD EPYC™ 9575F

Purpose built for GPU host nodes

~700,000

more inference tokens/s

on 1K node AI cluster running Llama3.1-70B

Up to 20%

faster training

with Stable Diffusion XL V2

Llama 3.1: 8.8% more perf on 1000 Node Cluster of Turin + 8xMI300X vs Emerald Rapids + 8xMI300X on Llama3.1-70B with 128 Input tokens, 2048 output tokens, batch size 1000
Stable Diffusion XL V2: 20% better training time on Turin + 8xMI300X vs Emerald Rapids + 8xMI300X
As of 10/4/2024. See endnote 9xx5-087, 9xx5-059a.

Programmable DPU

Evolving front-end network

Enables:

Faster Data Integration

Storage offload
and acceleration

Enables:

SDN and Security

Evolving network
infrastructure services

Secure multi-tenant access
Data privacy and integrity

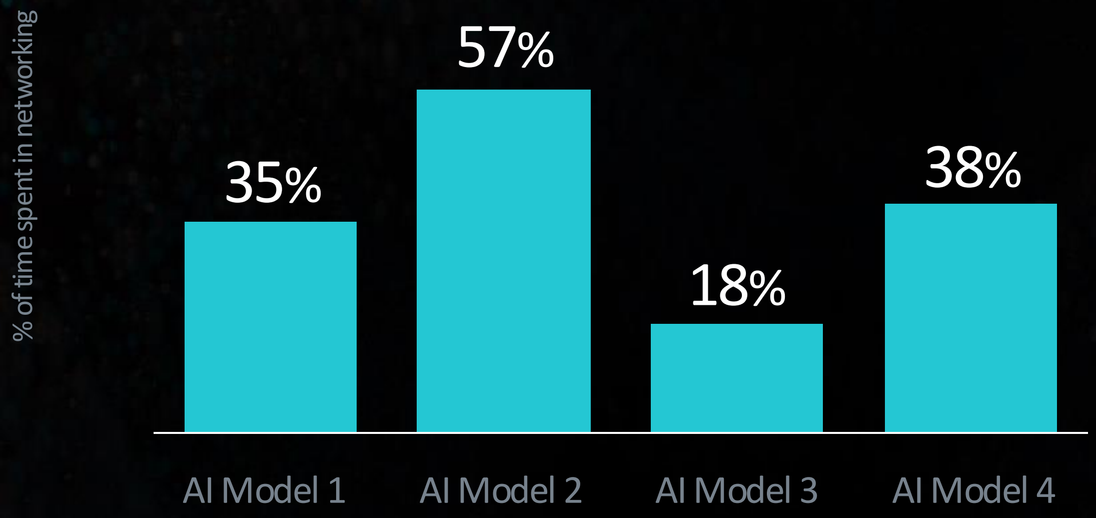
Enables:

Zero CPU Overhead

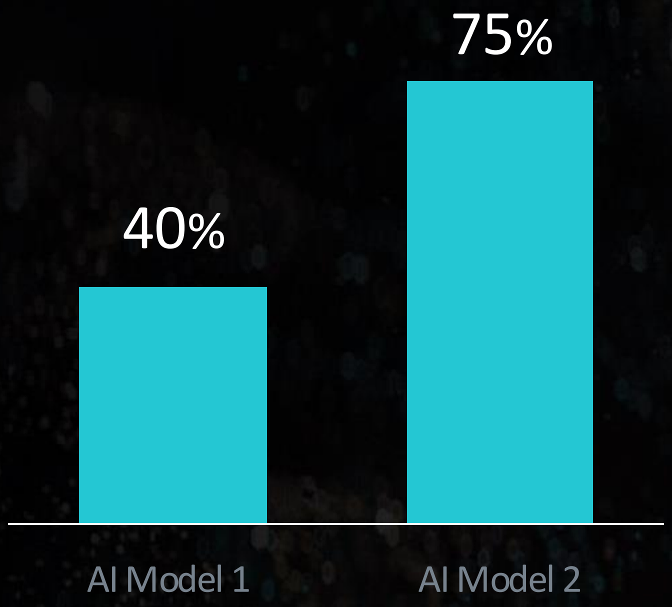
DPU accelerates infrastructure
services at line rate

Dedicated CPU for
AI workload processing

Back-end networks drive AI system performance



At an average 30% of training cycle time
is elapsed in waiting for networking



Communication accounts for 40%-75% of time
with Training and Distributed Inference Models[2]

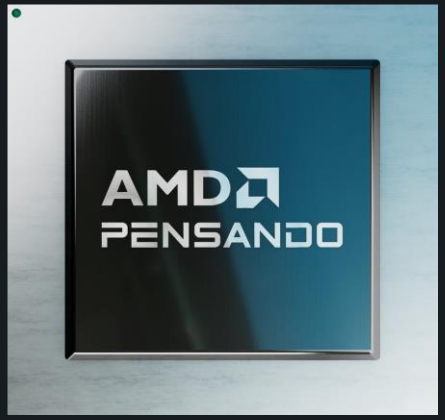
Source: 1) 2022 OCP Keynote by Alexis Bjorlin, VP at Meta, 2) Computation vs. Communication Scaling for Future Transformers on Future Hardware https://arxiv.org/pdf/2302.02825

Advancing Data Center Solutions

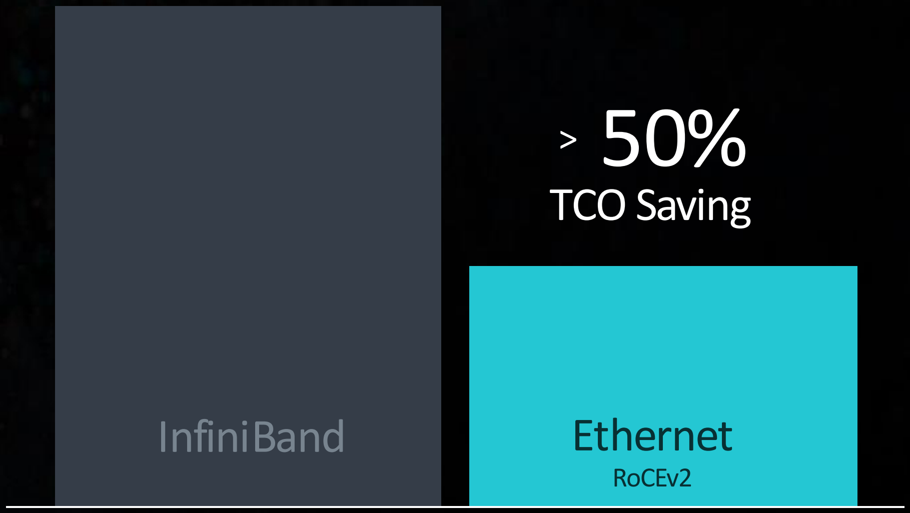
Data Center CPUs

Data Center GPUs

Networking

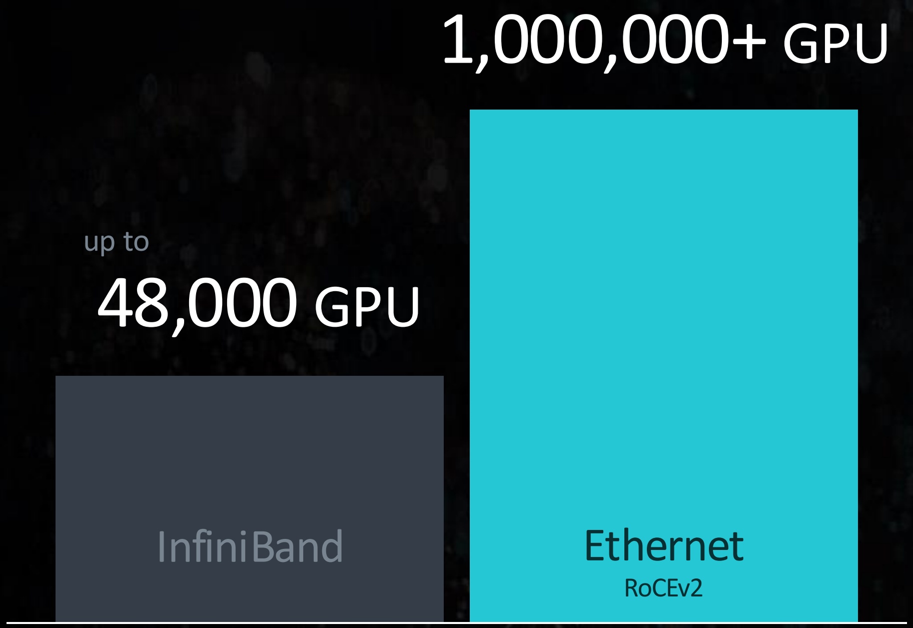


Ethernet is always the preferred choice



Total Cost of Ownership[1]

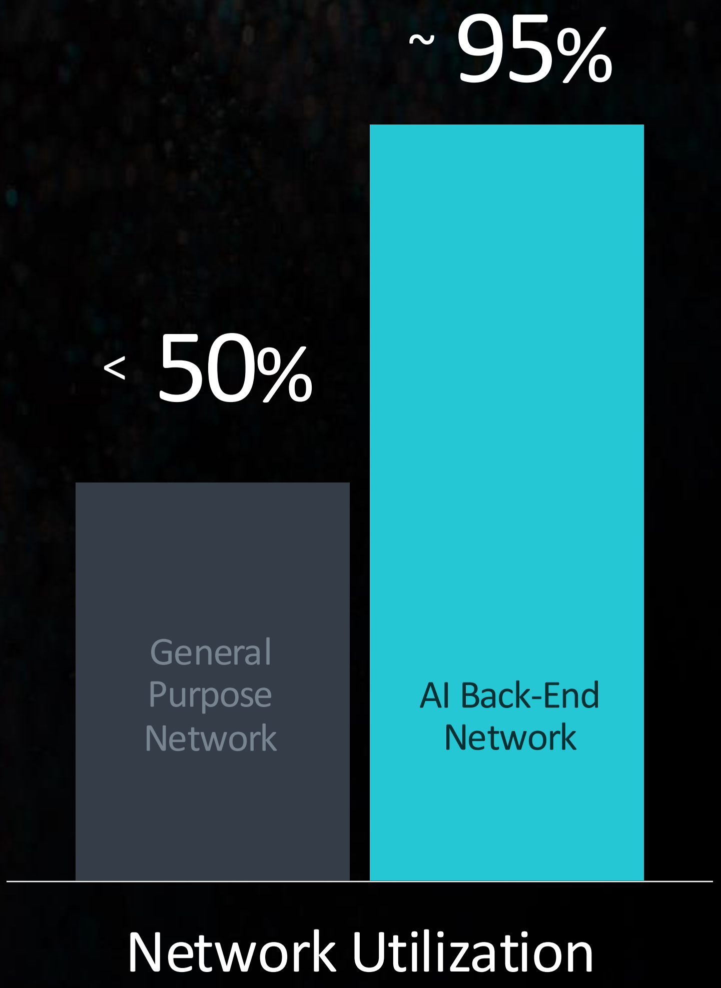
Lower is better



Scalability

Higher is better

Sources: 1) 650Group Datacenter AI Networking and Server SmartNIC Forecast Reports 2Q24 .



The challenge of high network utilization

AI backend networks drive sustained data transfers

Intelligent Load Balancing | Congestion Management | Fast Failover and Loss Recovery

Source: Alibaba Cloud Predictable network: Built for AI Infrastructure

Ultra Ethernet Consortium

Evolve ethernet as an open, interoperable, high performance, full-communications stack architecture to meet the growing network demands of AI and HPC at scale

UEC 1.0 Specification - Q1CY25

Performant | Scalable | Cost Effective

Ultra Ethernet
Consortium

Steering Members

AMD

ARISTA

BROADCOM

CISCO

EVIDEN
an atos business

Hewlett Packard
Enterprise

intel

Meta

Microsoft

ORACLE

Ultra Ethernet
Consortium

General Members

Alibaba Cloud | ARRCUS NETWORK DIFFERENT | Bai du 百度 | 世纪互联 VNET | ByteDance | cadence | CORNELIS NETWORKS | DELL Technologies | enfabrica

HUAWEI | IBM | JUNIPer NETWORKS | KEYSIGHT TECHNOLOGIES | Lawrence Livermore National Laboratory | Lenovo | MARVELL | H3C

NOKIA | NVIDIA | Preferred Networks | PURESTORAGE | Qualcomm | spirent Promise. Assured. | SYNOPSYS | ZTE

- Total 97 Members

RDMA outperforms RoCEv2

6x faster

message completion time

5x faster

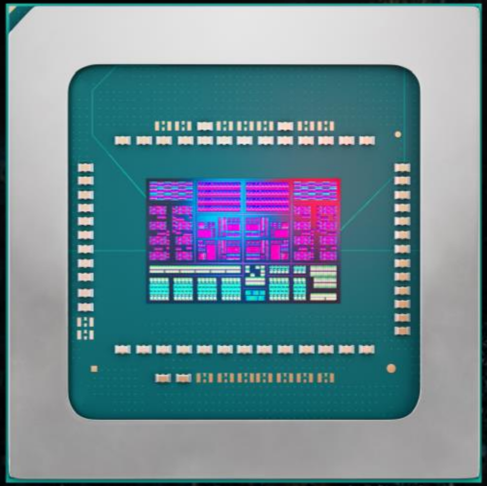
collective completion time

Intelligent packet spray and in-order message delivery

Path aware congestion avoidance

Selective retransmission and fast loss recovery

* Reference: STrack: A Reliable Multipath Transport for AI/ML Clusters, July 2024



3rd Gen AMD P4 Engines

Deliver network innovation at the speed of AI

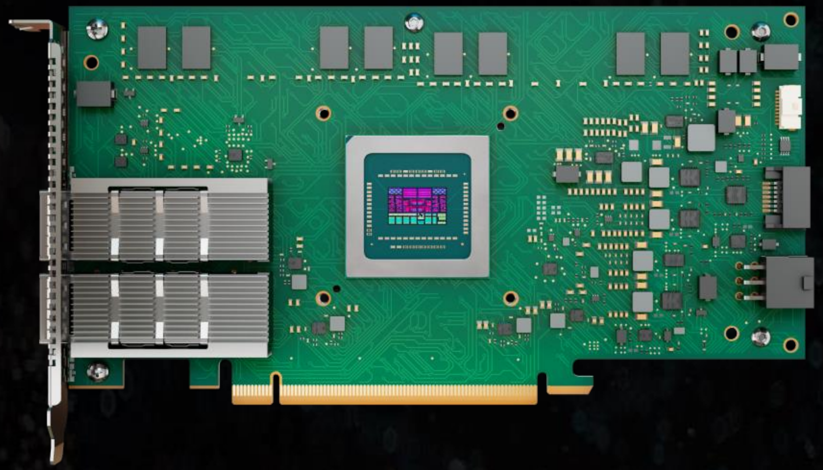
120M Packets/s	400Gb/s	5M Connections/s
Fully Programmable	Wire Rate	Concurrent Services (SDN, Security, Storage Acceleration)

Announcing Today: "Salina"

AMD Pensando™ Salina 400

Best DPU for evolving front-end networks

3rd Gen	400G	232 P4 MPU	2x DDR5	16 N1
Software Compatible	PCIe® Gen 5 2x400GE	Multi-Services	102GB/S Memory Bandwidth Up to 128 GB DDR	ARM Cores



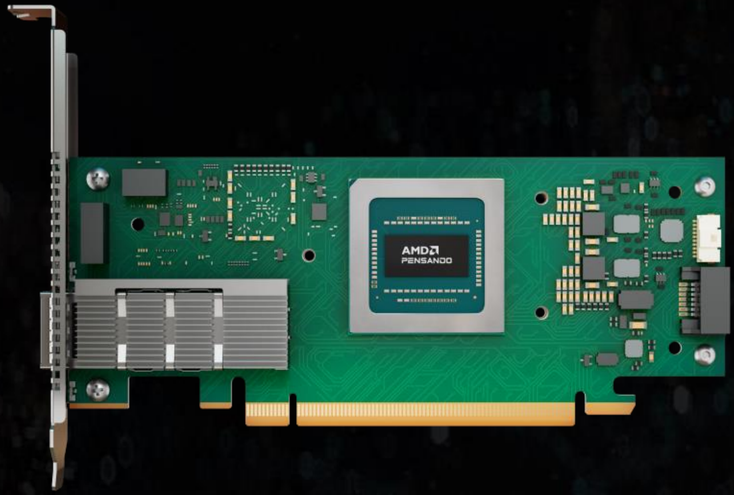
AMD Pensando™

DPU choice for hyperscalers

Announcing Today: “Pollara”

AMD Pensando™ Pollara 400

Industry’s first ultra ethernet consortium ready AI NIC



Ultra Ethernet Consortium

Programmable Hardware Pipeline

Up to 6x Performance Boost*

400 Gbps

- Open Ecosystem
- UEC Ready RDMA
- Reduction in Job Completion Times
- High Availability

Based on comparison of RDMA vs. RoCEv2. * Reference: STrack: A Reliable Multipath Transport for AI/ML Clusters , July 2024

Solutions you can rely on for your business

350+

server platforms

950+

cloud instances

Today at Advancing AI 2024

AMD end-to-end AI infrastructure leadership

Data Center CPUs

5th Generation
AMD EPYC™ "Turin"

Data Center GPUs

AMD Instinct™
MI325X and MI350 Series

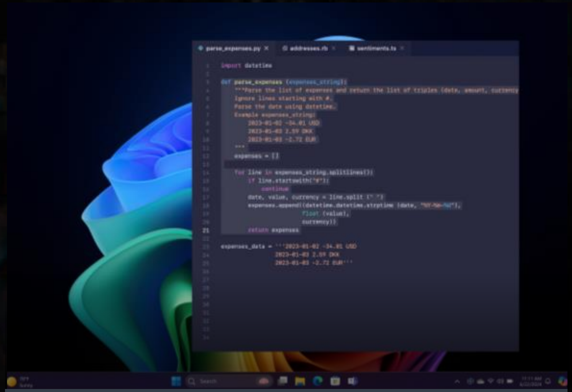
Networking

AMD Pensando™
Pollara 400/Salina 400

AI PCs

3rd Generation
AMD Ryzen™ AI PRO

Unprecedented transformational experiences with next-gen AI PCs



```
import datetime

def parse_expenses (expenses_string):
    """Parse the list of expenses and return the list of triples (date, amount, currency
    Ignore lines starting with #.
    Parse the date using datetime.
    Example expenses_string:
        2023-01-02 -34.01 USD
        2023-01-03 2.59 DKK
        2023-01-03 -2.72 EUR
    """
    expenses = []

    for line in expenses_string.splitlines():
        if line.startswith("#"):
            continue
        date, value, currency = line.split (" ")
        expenses.append((datetime.datetime.strptime (date, "%Y-%m-%d"),
                         float (value),
                         currency))
    return expenses

expenses_data = '''2023-01-02 -34.01 USD
                2023-01-03 2.59 DKK
                2023-01-03 -2.72 EUR'''
```

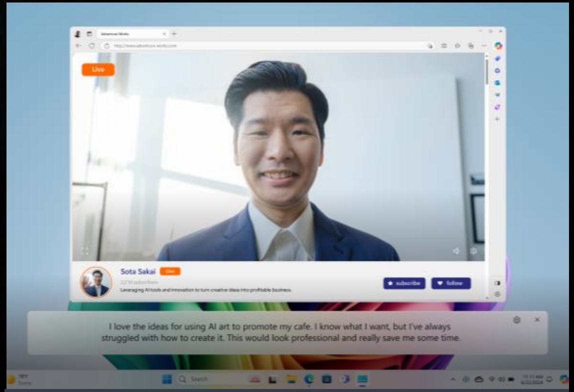
Enterprise Productivity

Business LLMs for data processing

Multilingual chatbots

Software coding assistants

Real-time threat detection



Leveraging AI tools and innovation to turn creative ideas into profitable business.

I love the ideas for using AI art to promote my cafe. I know what I want, but I've always struggled with how to create it. This would look professional and really save me some time.

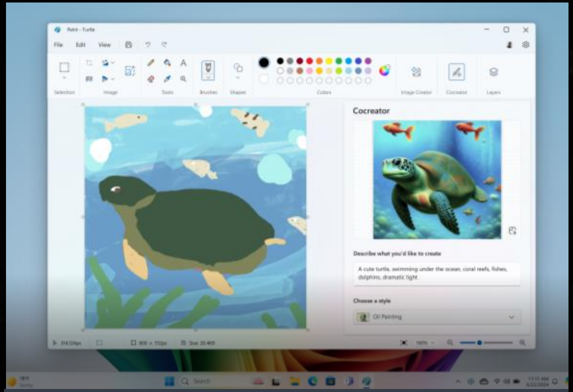
Immersive Collaboration

Live captions with translation

Speech recognition and transcription

Intelligent meeting assistance

Sentiment analysis



Cocreator

Describe what you'd like to create

A cute turtle, swimming under the ocean, coral reefs, fishes, dolphins, dramatic light

Choose a style

Oil Painting

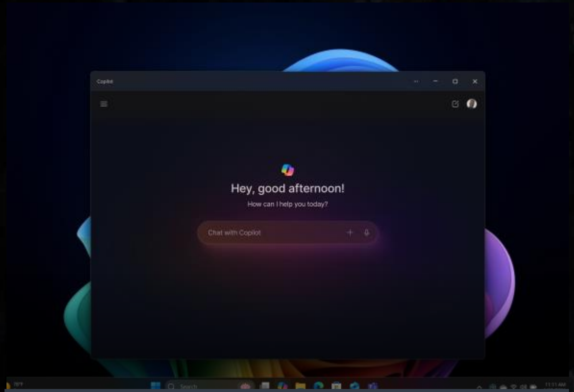
Revolutionary Creation and Editing

Automated content generation

Multimedia production

Art and design assist

Advanced video & audio effects



Hey, good afternoon!

How can I help you today?

Personal AI Assistance

Smart information retrieval

Document analysis

Calendar management

Travel planning

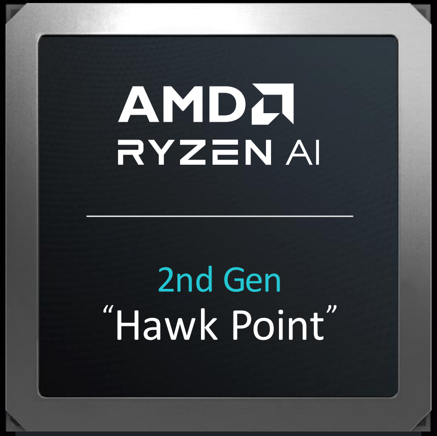
AMD Ryzen™ AI Leads the AI PC Era

Q1 2023

1st Gen
"Phoenix Point"

10
TOPS NPU

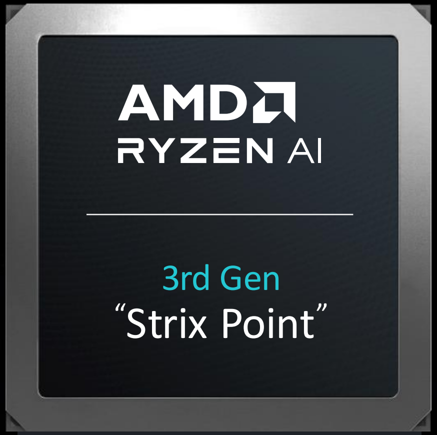
Q4 2023



2nd Gen
"Hawk Point"

16
TOPS NPU

Q2 2024



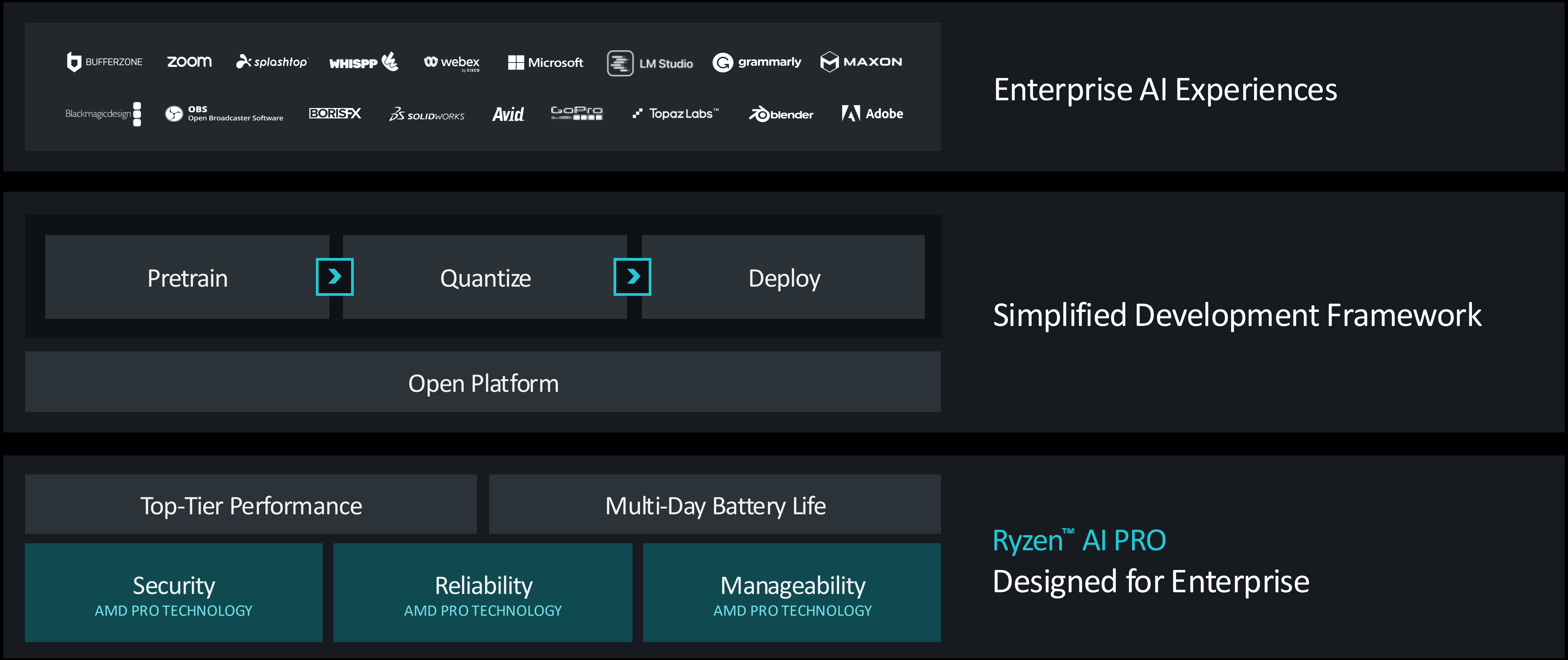
3rd Gen
"Strix Point"

50+
TOPS NPU

See endnote GD-243.

AMD RYZEN AI PRO

Best Processors for Copilot+ Enterprise PCs



BUFFERZONE · zoom · splashtop · WHISPP · webex by cisco · Microsoft · LM Studio · grammarly · MAXON

Blackmagicdesign · OBS Open Broadcaster Software · BORIS FX · SOLIDWORKS · Avid · GoPro · Topaz Labs™ · blender · Adobe

Enterprise AI Experiences

Pretrain › Quantize › Deploy

Open Platform

Simplified Development Framework

Top-Tier Performance

Multi-Day Battery Life

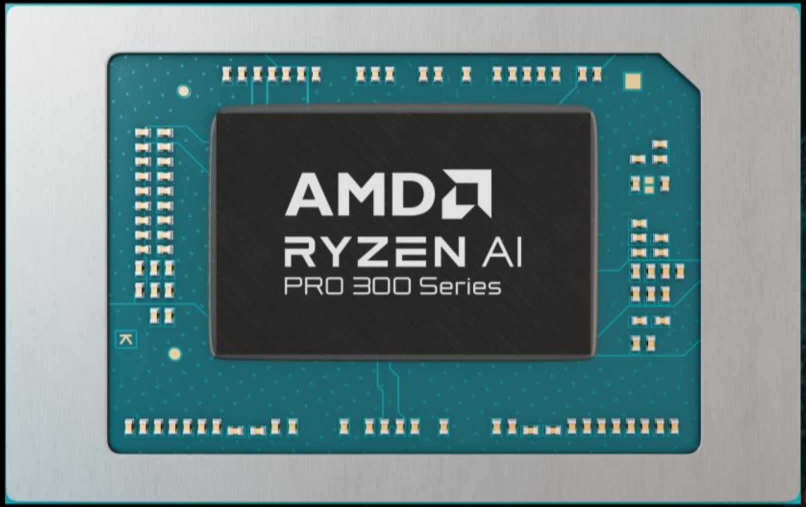
Security
AMD PRO TECHNOLOGY

Reliability
AMD PRO TECHNOLOGY

Manageability
AMD PRO TECHNOLOGY

Ryzen™ AI PRO
Designed for Enterprise

See endnote GD-173a, STXP-04.



Announcing today

AMD Ryzen™ AI PRO 300 Series

First Copilot+ laptops enabled for enterprise PCs

ZEN 5 | AMD XDNA 2 | AMD RDNA 3.5 | Copilot+PC

See endnote STXP-05

Up to

1.4x

multithreaded performance

Comparing AMD Ryzen™ AI 9 HX PRO 375 vs. Intel® Core® Ultra 7 165H vPro on Cinebench 2024 nT. See endnote STXP-12

Up to

23hrs

Multi day battery life

AMD Ryzen™ AI 9 HX PRO 375, VIdeo Playback. See endnote STXP-30

Up to

9hrs

battery life with Microsoft Teams

AMD Ryzen™ AI 9 HX PRO 375. See endnote STXP-32

Enterprise AI PC Application Ecosystem

Adobe

Microsoft

webex by CISCO

zoom

splashtop

WHISPP

SOLIDWORKS

BUFFERZONE

Camo

bitdefender secure your every bit

LM Studio

grammarly

Blackmagicdesign

nero

BORISFX

Avid

Rhinoceros

GoPro Be a HERO

Topaz Labs™

ARKRUNR

voicemy.ai

blender

RADiCAL AI-POWERED 3D ANIMATION

CyberLink

AFFINITY Photo 2

ACCA ACCA SOFTWARE

convai

OBS Open Broadcaster Software

MAXON

Announced today at Advancing AI 2024

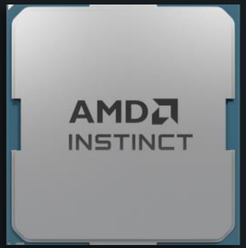
End-to-End AI Infrastructure Leadership

Data Center CPUs

5th Generation
AMD EPYC™ "Turin"

Shipping Now

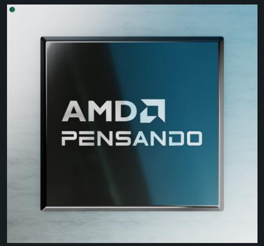
Data Center GPUs



AMD Instinct™
MI325X Series

Production Q4, Systems Q1

Solutions



AMD Pensando™
Pollara 400/Salina 400

Coming 1H25

AI PCs

3rd Generation
AMD Ryzen™ AI PRO

Shipping Now

AMD

Endnotes

9xx5-001: Based on AMD internal testing as of 9/10/2024, geomean performance improvement (IPC) at fixed-frequency. 5th Gen EPYC CPU Enterprise and Cloud Server Workloads generational IPC Uplift of 1.170x (geomean) using a select set of 36 workloads and is the geomean of estimated scores for total and all subsets of SPECrate®2017_int_base (geomean), estimated scores for total and all subsets of SPECrate®2017_fp_base (geomean), scores for Server Side Java multi instance max ops/sec, representative Cloud Server workloads (geomean), and representative Enterprise server workloads (geomean). “Genoa” Config (all NPS1): EPYC 9654 BIOS TQZ1005D 12c12t (1c1t/CCD in 12+1), FF 3GHz, 12x DDR5-4800 (2Rx4 64GB), 32Gbps xGMI; “Turin” config (all NPS1): EPYC 9V45 BIOS RVOT1000F 12c12t (1c1t/CCD in 12+1), FF 3GHz, 12x DDR5-6000 (2Rx4 64GB), 32Gbps xGMI Utilizing Performance Determinism and the Performance governor on Ubuntu® 22.04 w/ 6.8.0-40-generic kernel OS for all workloads. 5th Gen EPYC generational ML/HPC Server Workloads IPC Uplift of 1.369x (geomean) using a select set of 24 workloads and is the geomean of representative ML Server Workloads (geomean), and representative HPC Server Workloads (geomean). “Genoa Config (all NPS1) “Genoa” config: EPYC 9654 BIOS TQZ1005D 12c12t (1c1t/CCD in 12+1), FF 3GHz, 12x DDR5-4800 (2Rx4 64GB), 32Gbps xGMI; “Turin” config (all NPS1): EPYC 9V45 BIOS RVOT1000F 12c12t (1c1t/CCD in 12+1), FF 3GHz, 12x DDR5-6000 (2Rx4 64GB), 32Gbps xGMI. Utilizing Performance Determinism and the Performance governor on Ubuntu 22.04 w/ 6.8.0-40-generic kernel OS for all workloads except LAMMPS, HPCG, NAMD, OpenFOAM, Gromacs which utilize 24.04 w/ 6.8.0-40-generic kernel. SPEC® and SPECrate® are registered trademarks for Standard Performance Evaluation Corporation. Learn more at spec.org.

9xx5-002C: SPECrate®2017_int_base comparison based on published scores from www.spec.org as of 10/10/2024. 2P AMD EPYC 9965 (3000 SPECrate®2017_int_base, 384 Total Cores, 500W TDP, $14,813 CPU $), 6.060 SPECrate®2017_int_base/CPU W, 0.205 SPECrate®2017_int_base/CPU $, https://www.spec.org/cpu2017/results/res2024q3/cpu2017-20240923-44833.html). 2P AMD EPYC 9755 (2720 SPECrate®2017_int_base, 256 Total Cores, 500W TDP, $12,984 CPU $), 5.440 SPECrate®2017_int_base/CPU W, 0.209 SPECrate®2017_int_base/CPU $, https://www.spec.org/cpu2017/results/res2024q4/cpu2017-20240923-44837.pdf). 2P AMD EPYC 9754 (1950 SPECrate®2017_int_base, 256 Total Cores, 360W TDP, $11,900 CPU $), 5.417 SPECrate®2017_int_base/CPU W, 0.164 SPECrate®2017_int_base/CPU $, https://www.spec.org/cpu2017/results/res2023q2/cpu2017-20230522-36617.html). 2P AMD EPYC 9654 (1810 SPECrate®2017_int_base, 192 Total Cores, 360W TDP, $11,805 CPU $), 5.028 SPECrate®2017_int_base/CPU W, 0.153 SPECrate®2017_int_base/CPU $, https://www.spec.org/cpu2017/results/res2024q1/cpu2017-20240129-40896.html). 2P Intel Xeon Platinum 8592+ (1130 SPECrate®2017_int_base, 128 Total Cores, 350W TDP, $11,600 CPU $) 3.229 SPECrate®2017_int_base/CPU W, 0.097 SPECrate®2017_int_base/CPU $, http://spec.org/cpu2017/results/res2023q4/cpu2017-20231127-40064.html). 2P Intel Xeon 6780E (1410 SPECrate®2017_int_base, 288 Total Cores, 330W TDP) 4.273 SPECrate®2017_int_base/CPU W, 0.124 SPECrate®2017_int_base/CPU $, https://spec.org/cpu2017/results/res2024q3/ cpu2017-20240811-44406.html)SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. Intel CPU TDP at https://ark.intel.com/.

9xx5-005A: MySQL TPROC-C workload (SQL Server OLTP Brokerage) estimate based on internal AMD measurements as of 09/15/2024. The HammerDB TPROC-C workload is an open-source workload derived from TPC-BenchmarkTM Standard, and as such is not comparable to published TPC-C TM results, as the results do not comply with the TPC-C Benchmark Standard. Workload configs: MySQL 8.0.39, 8 core nodes (Multi-SUT), HammerDB-4.4, duration 5min, 32 v users, warehouses 128, aggregate New Orders Per Minute (NOPM). 2P AMD EPYC 9965 powered server (384 total cores), 2.35TB Memory, BIOS RVC100DB, OS VMWare ESXi 8.0.3 build 70965425, 1x1.6TB and 10x3.84TB storage. VM Configurations: 8 cores/VM, 48 VMs, 48GB memory, Ubuntu 22.04.4 LTS, Linux 5.15.0-119-generic, BOOT_IMAGE=/vmlinuz-5.15.0-119-generic root=/dev/mapper/ubuntu--vg-ubuntu--lv ro. 2P AMD EPYC 9755 powered server (256 total cores), 2.35TB Memory, BIOS RVOT1000C, OS VMWare ESXi 8.0.3 build 70965425, 1x1.6TB and 8x3.84TB storage. VM Configurations: 8 cores/VM, 32 VMs, 48GB memory, Ubuntu 22.04.4 LTS, Linux 5.15.0-119-generic, BOOT_IMAGE=/vmlinuz-5.15.0-119-generic root=/dev/mapper/ubuntu--vg-ubuntu--lv ro 2P AMD EPYC 9654 powered server (192 total cores), 1.5TB Memory, BIOS TVC100BD_2, OS VMWare ESXi 8.0.3 build 70965425, 1x1.6TB and 8x3.84TB storage. VM Configurations: 8 cores/VM, 24 VMs, 48GB memory, Ubuntu 22.04.4 LTS, Linux 5.15.0-119-generic, BOOT_IMAGE=/vmlinuz-5.15.0-119-generic root=/dev/mapper/ubuntu--vg-ubuntu--lv ro spec_rstack_overflow=off. 2P Intel Xeon 8592+ powered server (128 total cores), 1TB Memory, BIOS ESE124B, OS VMWare ESXi 8.0.3 build 24022510, 1x1.6TB and 8x3.84TB storage. VM Configurations: 8 cores/VM, 16 VMs, 48GB memory, Ubuntu 22.04.4 LTS, Linux 5.15.0-119-generic BOOT_IMAGE=/vmlinuz-5.15.0-119-generic root=/dev/mapper/ubuntu--vg-ubuntu--lv ro spec_rstack_overflow=off. CPU Score (TPM) Relative_8592+ Relative_9654. Intel 8592+ (64c) 9431248 1 0.523. AMD EPYC 9654 (96c) 18037794 1.913 1. AMD EPYC 9755 (128c) 32598005 3.456 1.807. AMD EPYC 9965 (192c) 36863796 3.909 2.043. Results may vary based on factors including but not limited to system configurations, software versions, and BIOS settings. TPC, TPC Benchmark, and TPC-C are trademarks of the Transaction Processing Performance Council.

.

Endnotes

9xx5-006: AMD internal testing as of 09/01/2024, on FFMPEG (Raw to VP9, 1080P, 302 Frames, 1 instance/thread, video source: https://media.xiph.org/video/derf/y4m/ducks_take_off_1080p50.y4m). System Configurations: 2P AMD EPYC™ 9965 reference system (2 x 192C) 1.5TB 24x64GB DDR5-6400 running at 6000MT/s, SAMSUNG MZWLO3T8HCLS-00A07, NPS=4, Ubuntu 22.04.3 LTS, Kernel Linux 5.15.0-119-generic, BIOS RVOT1000C (determinism enable=power), 10825484.25 Frames/Hour Median. 2P AMD EPYC™ 9654 production system (2 x 96C) 1.5TB 24x64GB DDR5-5600, , SAMSUNG MO003200KYDNC, NPS=4, Ubuntu 22.04.3 LTS, Kernel Linux 5.15.0-119-generic, BIOS 1.56 (determinism enable=power) , 5154133.333 Frames/Hour Median. 2P Intel Xeon Platinum 8592+ production system (2 x 64C) 1TB 16x64GB DDR5-5600, 3.2 TB NVME, Ubuntu 22.04.3 LTS, Kernel Linux 6.5.0-35-generic), BIOS ESE122V-3.10, 2712701.754 Frames/Hour Median. For 3.99x the performance with the AMD EPYC 9965 vs Intel Xeon Platinum 8592+ systems. For 1.90x the performance with the AMD EPYC 9654 vs Intel Xeon Platinum 8592+ systems. Results may vary based on factors including but not limited to BIOS and OS settings and versions, software versions and data used.

9xx5-012: TPCxAI @SF30 Multi-Instance 32C Instance Size throughput results based on AMD internal testing as of 09/05/2024 running multiple VM instances. The aggregate end-to-end AI throughput test is derived from the TPCx-AI benchmark and as such is not comparable to published TPCx-AI results, as the end-to-end AI throughput test results do not comply with the TPCx-AI Specification. 2P AMD EPYC 9965 (384 Total Cores), 12 32C instances, NPS1, 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT=off, Determinism=Power, Turbo Boost=Enabled). 2P AMD EPYC 9755 (256 Total Cores), 8 32C instances, NPS1, 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT0090F (SMT=off, Determinism=Power, Turbo Boost=Enabled) 2P AMD EPYC 9654 (192 Total cores) 6 32C instances, NPS1, 1.5TB 24x64GB DDR5-4800, 1DPC, 2 x 1.92 TB Samsung MZQL21T9HCJR-00A07 NVMe, Ubuntu 22.04.3 LTS, BIOS 1006C (SMT=off, Determinism=Power) Versus 2P Xeon Platinum 8592+ (128 Total Cores), 4 32C instances, AMX On, 1TB 16x64GB DDR5-5600, 1DPC, 1.0 Gbps NetXtreme BCM5719 Gigabit Ethernet PCIe, 3.84 TB KIOXIA KCMYXRUG3T84 NVMe, , Ubuntu 22.04.4 LTS, 6.5.0-35 generic (tuned-adm profile throughput-performance, ulimit -l 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT=off, Determinism=Power, Turbo Boost = Enabled) Results. CPU Median Relative Generational. Turin 192C, 12 Inst 6067.531 3.775 2.278. Turin 128C, 8 Inst 4091.85 2.546 1.536 Genoa 96C, 6 Inst 2663.14 1.657 1. EMR 64C, 4 Inst 1607.417 1 NA. Results may vary due to factors including system configurations, software versions and BIOS settings. TPC, TPC Benchmark and TPC-C are trademarks of the Transaction Processing Performance Council.

9xx5-022: Source: https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/performance-briefs/amd-epyc-9005-pb-gromacs.pdf

9xx5-038: AMD testing as of 09/18/2024. The detailed results show the average uplift of the performance metric (TFLOPS) of this benchmark for a 2P 96-Core AMD EPYC™ 9654 powered system compared to a 2P 64-Core Intel® Xeon® PLATINUM 8592+ powered system running select tests on Open-Source HPL v2.3. Uplifts for the performance metric normalized to the 64-Core Intel® Xeon® PLATINUM 8592+ follow for each benchmark: * hpl: ~1.12x. System Configurations: CPU: 2P 64-Core Intel® Xeon® PLATINUM 8592+ (128 total cores). Memory: 16x 64 GB DDR5-5600. Storage: KIOXIA KCMYXRUG3T84. Platform and BIOS: ThinkSystem SR650 V3 ESE122V-3.10. BIOS Options: SMT=Off. High Performance Mode. OS: rhel 9.4 5.14.0-427.16.1.el9_4.x86_64. Kernel Options: processor.max_cstate=1 intel_idle.max_cstate=0 iommu=pt mitigations=off. Runtime Options: cpupower frequency-set -g performance. echo 3 > /proc/sys/vm/drop_caches. echo 0 > /proc/sys/kernel/nmi_watchdog. echo 0 > /proc/sys/kernel/numa_balancing. echo 0 > /proc/sys/kernel/randomize_va_space. echo 'always' /sys/kernel/mm/transparent_hugepage/enabled. echo 'always' > /sys/kernel/mm/transparent_hugepage/defrag. CPU: 2P 96-Core AMD EPYC™ 9654 (192 total cores). Memory: 24x 64 GB DDR5-4800. Storage: SAMSUNG. MZQL21T9HCJR-00A07. Platform and BIOS: Titanite_4G RTI1009C. BIOS Options: SMT=Off. NPS=4. Power Determinism Mode. OS: rhel 9.4 5.14.0-427.16.1.el9_4.x86_64. Kernel Options: amd_iommu=on iommu=pt mitigations=off. Runtime Options: cpupower idle-set -d 2 cpupower frequency-set -g performance echo 3 > /proc/sys/vm/drop_caches echo 0 > /proc/sys/kernel/nmi_watchdog echo 0 > /proc/sys/kernel/numa_balancing echo 0 /proc/sys/kernel/randomize_va_space echo 'always' > /sys/kernel/mm/transparent_hugepage/enabled echo 'always' > /sys/kernel/mm/transparent_hugepage/defrag. Results may vary based on factors including but not limited to system configurations, software versions, and BIOS settings. 96C, NPS1 662.577 644.776 640.95 644.776 1.245 1 2P EMR 64C 517.986 421.053 553.846 517.986 1 NA Results may vary due to factors including system configurations, software versions and BIOS settings.

Endnotes

9xx5-039: AMD testing as of 09/18/2024.The detailed results show the average uplift of the performance metric (TFLOPS) of this benchmark for a 2P 192-Core AMD EPYC™ 9965 powered system compared to a 2P 64-Core Intel® Xeon® PLATINUM 8592+ powered system running select tests on Open-Source HPL v2.3. Uplifts for the performance metric normalized to the 64-Core Intel® Xeon® PLATINUM 8592+ follow for each benchmark: * hpl: ~3.10x

System Configurations: CPU: 2P 64-Core Intel® Xeon® PLATINUM 8592+ (128 total cores) Memory: 16x 64 GB DDR5-5600 Storage: KIOXIA KCMYXRUG3T84 Platform and BIOS: ThinkSystem SR650 V3 ESE122V-3.10 BIOS Options: SMT=Off High Performance Mode OS: rhel 9.4 5.14.0-427.16.1.el9_4.x86_64 Kernel Options: processor.max_cstate=1 intel_idle.max_cstate=0 iommu=pt mitigations=off Runtime Options: cpupower frequency-set -g performance echo 3 > /proc/sys/vm/drop_caches echo 0 > /proc/sys/kernel/nmi_watchdog echo 0 > /proc/sys/kernel/numa_balancing echo 0 > /proc/sys/kernel/randomize_va_space echo 'always' > /sys/kernel/mm/transparent_hugepage/enabled echo 'always' > /sys/kernel/mm/transparent_hugepage/defrag. CPU: 2P 192-Core AMD EPYC™ 9965 (384 total cores) Memory: 24x 64 GB DDR5-6000 Storage: SAMSUNG MZWLO3T8HCLS-00A07 Platform and BIOS: VOLCANO RVOT1000C BIOS Options: SMT=Off NPS=4Power Determinism Mode OS: rhel 9.4 5.14.0-427.16.1.el9_4.x86_64 Kernel Options: amd_iommu=on iommu=pt mitigations=off

Runtime Options: cpupower idle-set -d 2 cpupower frequency-set -g performance echo 3 > /proc/sys/vm/drop_caches echo 0 > /proc/sys/kernel/nmi_watchdog echo 0 > /proc/sys/kernel/numa_balancing echo 0 > /proc/sys/kernel/randomize_va_space echo 'always' > /sys/kernel/mm/transparent_hugepage/enabled echo 'always' > /sys/kernel/mm/transparent_hugepage/defrag. Results may vary based on factors including but not limited to system configurations, software versions, and

9xx5-040A : XGBoost (Runs/Hour) throughput results based on AMD internal testing as of 09/05/2024. XGBoost Configurations: v2.2.1, Higgs Data Set, 32 Core Instances, FP32 2P AMD EPYC 9965 (384 Total Cores), 12 x 32 core instances, 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS, 6.8.0-45-generic (tuned-adm profile throughput-performance, ulimit -l 198078840, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=1 2P AMD EPYC 9755 (256 Total Cores), 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198094956, ulimit -n 1024, ulimit -s 8192), BIOS RVOT0090F (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=1 2P AMD EPYC 9654 (192 Total cores), 1.5TB 24x64GB DDR5-4800, 1DPC, 2 x 1.92 TB Samsung MZQL21T9HCJR-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198120988, ulimit -n 1024, ulimit -s 8192), BIOS TTI100BA (SMT=off, Determinism=Power), NPS=1 Versus 2P Xeon Platinum 8592+ (128 Total Cores), AMX On, 1TB 16x64GB DDR5-5600, 1DPC, 1.0 Gbps NetXtreme BCM5719 Gigabit Ethernet PCIe, 3.84 TB KIOXIA KCMYXRUG3T84 NVMe®, Ubuntu 22.04.4 LTS, 6.5.0-35 generic (tuned-adm profile throughput-performance, ulimit -l 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT=off, Determinism=Power, Turbo Boost = Enabled) Results: CPU Run 1 Run 2 Run 3 Median Relative Throughput Generational 2P Turin 192C, NPS1 1565.217 1537.367 1553.957 1553.957 3 2.41 2P Turin 128C, NPS1 1103.448 1138.34 1111.969 1111.969 2.147 1.725 2P Genoa 96C, NPS1 662.577 644.776 640.95 644.776 1.245 1 2P EMR 64C 517.986 421.053 553.846 517.986 1 NA Results may vary due to factors including system configurations, software versions and BIOS settings.

9xx5-041: XGBoost (Runs/Hour) throughput results based on AMD internal testing as of 09/05/2024. XGBoost Configurations: v1.7.2, Airline Data Set, 32 Core Instances, FP32. 2P AMD EPYC 9965 (384 Total Cores), 12 x 32 core instances, 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=1. 2P AMD EPYC 9755 (256 Total Cores), 8 x 32 core instances, 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT0090F (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=4. 2P AMD EPYC 9654 (192 Total cores), 6 x 32 core instances, 1.5TB 24x64GB DDR5-4800, 1DPC, 2 x 1.92 TB Samsung MZQL21T9HCJR-00A07 NVMe®, Ubuntu 22.04.3 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS 1006C (SMT=off, Determinism=Power), NPS=1. Versus 2P Xeon Platinum 8592+ (128 Total Cores), 4 x 32 core instanc es, AMX On, 1TB 16x64GB DDR5-5600, 1DPC, 1.0 Gbps NetXtreme BCM5719 Gigabit Ethernet PCIe, 3.84 TB KIOXIA KCMYXRUG3T84 NVMe®, Ubuntu 22.04.4 LTS, 6.5.0-35 generic (tuned-adm profile throughput-performance, ulimit -l 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT=off, Determinism=Power, Turbo Boost = Enabled) Results: CPU Run 1 Run 2 Run 3 Median Average Throughput Relative Throughput Generational. 2P Turin 192C, 12x32C 28.836 28.461 28.729 28.729 1503.707 2.904 2.338. 2P Turin 128C, 8x32C 25.48 25.287 25.179 25.287 1138.925 2.2 1.771. 2P Genoa 96C, 6x32C 32.686 33.584 3 3.699 33.584 643.163 1.242 1. 2P EMR 64C, 4x32C 27.81 34.287 26.053 27.81 517.799 1 NA. Results may vary due to factors including system configurations, software versions and BIOS settings.

Endnotes

9xx5-048: AMD EPYC™ 9005 Series processors require OEM enablement and a BIOS update from your server or motherboard manufacturer if used with a motherboard designed for the SP5 socketed AMD EPYC™ 9004 Series processors. Contact your system manufacturer prior to purchase to determine compatibility.

9xx5-059A: Stable Diffusion XL v2 training results based on AMD internal testing as of 10/10/2024. SDXL configurations: DeepSpeed 0.14.0, TP8 Parallel, FP8, batch size 24, results in seconds 2P AMD EPYC 9575F (128 Total Cores) with 8x AMD Instinct MI300X-NPS1-SPX-192GB-750W, GPU Interconnectivity XGMI, ROCm™ 6.2.0-66, 2304GB 24x96GB DDR5-6000, BIOS 1.0 (power determinism = off), Ubuntu® 22.04.4 LTS, kernel 5.15.0-72-generic, 334.80 seconds. 2P Intel Xeon Platinum 8592+ (128 Total Cores) with 8x AMD Instinct MI300X-NPS1-SPX-192GB-750, GPU Interconnectivity XGMI, ROCm 6.2.0-66, 2048GB 32x64GB DDR5-4400, BIOS 2.0.4, (power determinism= off), Ubuntu 22.04.4 LTS, kernel 5.15.0-72-generic, 400.43 seconds. For 19.600% training performance increase.Results may vary due to factors including system configurations, software versions and BIOS settings.

9xx5-069A: SPECrate®2017_int_base comparison based on published scores from www.spec.org as of 10/10/2024. Generational scores are based on highest published scores from www.spec.org from respective launch years. 2P AMD EPYC 9965 (3000 SPECrate®2017_int_base, 384 Total Cores, https://www.spec.org/cpu2017/results/res2024q4/cpu2017-20240923-44837.pdf) 2P AMD EPYC 9654 (1790 SPECrate®2017_int_base, 192 Total Cores, https://www.spec.org/cpu2017/results/res2022q4/cpu2017-20221024-32607.html) 2P AMD EPYC 7763 (861 SPECrate®2017_int_base, 128 Total Cores, https://www.spec.org/cpu2017/results/res2021q4/cpu2017-20211121-30148.html) 2P AMD EPYC 7742 (701 SPECrate®2017_int_base, 128 Total Cores, https://www.spec.org/cpu2017/results/res2019q4/cpu2017-20191125-20001.html) 2P AMD EPYC 7601 (275 SPECrate®2017_int_base, 64 Total Cores, https://www.spec.org/cpu2017/results/res2017q4/cpu2017-20171211-01594.html) SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. Intel CPU TDP at https://ark.intel.com/. SPEC - Standard Performance Evaluation Corporation

9xx5-071: VMmark® 4.0.1 host/node FC SAN comparison based on "independently published" results as of 10/10/2024. Configurations: 2 node, 2P AMD EPYC 9575F (128 total cores) powered server running VMware ESXi8.0 U3, 3.31 @ 4 tiles, https://www.infobellit.com/BlueBookSeries/VMmark4-FDR-1003. 2 node, 2P AMD EPYC 9554 (128 total cores) powered server running VMware ESXi 8.0 U3, 2.64 @ 3 tiles, https://www.infobellit.com/BlueBookSeries/VMmark4-FDR-1002. 2 node, 2P Intel Xeon Platinum 8592+ (128 total cores) powered server running VMware ESXi 8.0 U3, 2.06 @ 2.4 Tiles, https://www.infobellit.com/BlueBookSeries/VMmark4-FDR-1001. VMmark is a registered trademark of VMware in the US or other. countries.

9xx5-083::5th Gen EPYC processors support DDR5-6400 MT/s for targeted customers and configurations. 5th Gen production SKUs support up to DDR5-6000 MT/s to enable a broad set of DIMMs across all OEM platforms and maintain SP5 platform compatibility.

9xx5-087: As of 10/10/2024; this scenario contains several assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. Referencing 9XX5-056A: "2P AMD EPYC 9575F powered server and 8x AMD Instinct MI300X GPUs running Llama3.1-70B select inference workloads at FP8 precision vs 2P Intel Xeon Platinum 8592+ powered server and 8x AMD Instinct MI300X GPUs has ~8% overall throughput increase across select inference use cases" and 8763.52 tokens/s (9575F) versus 8,048.48 tokens/s (8592+) at 128 input / 2048 output tokens, 500 prompts for 1.089x the tokens/s or 715.04 more tokens/s. 1 Node = 2 CPUs and 8 GPUs. Assuming a 1000 node cluster, 1000 * 715.04 = 715,040 tokens/s. For ~700,000 more tokens/s. Results may vary due to factors including system configurations, software versions and BIOS settings.

Endnotes

99xx5TCO-002A: This scenario contains many assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. The AMD Server & Greenhouse Gas Emissions TCO (total cost of ownership) Estimator Tool - version 1.12, compares the selected AMD EPYC™ and Intel® Xeon® CPU based server solutions required to deliver a TOTAL_PERFORMANCE of 391000 units of SPECrate2017_int_base performance as of October 10, 2024. This estimation compares a legacy 2P Intel Xeon 28 core Platinum_8280 based server with a score of 391 versus 2P EPYC 9965 (192C) powered server with a score of 3000 (https://www.spec.org/cpu2017/results/res2024q4/cpu2017-20240923-44837.pdf) along with a comparison upgrade to a 2P Intel Xeon Platinum 8592+ (64C) based server with a score of 1130 (https://spec.org/cpu2017/results/res2024q3/cpu2017-20240701-43948.pdf). Actual SPECrate®2017_int_base score for 2P EPYC 9965 will vary based on OEM publications. Environmental impact estimates made leveraging this data, using the Country / Region specific electricity factors from the 2024 International Country Specific Electricity Factors 10 – July 2024, and the United States Environmental Protection Agency 'Greenhouse Gas Equivalencies Calculator'. For additional details, see https://www.amd.com/en/legal/claims/epyc.html#q=epyc4#SP9xxTCO-002A.

EPYC-029C: Comparison based on thread density, performance, features, process technology and built-in security features of currently shipping servers as of 10/10/2024. EPYC 9005 series CPUs offer the highest thread density [EPYC-025B], leads the industry with 500+ performance world records [EPYC-023F] with performance world record enterprise leadership Java® ops/sec performance [EPYCWR-20241010-260], top HPC leadership with floating-point throughput performance [EPYCWR-2024-1010-381], AI end-to-end performance with TPCx-AI performance [EPYCWR-2024-1010-525] and highest energy efficiency scores [EPYCWR-20241010-326]. The 5th Gen EPYC series also has 50% more DDR5 memory channels [EPYC-033C] with 70% more memory bandwidth [EPYC-032C] and supports 70% more PCIe® Gen5 lanes for I/O throughput [EPYC-035C], has up to 5x the L3 cache/core [EPYC-043C] for faster data access, uses advanced 3-4nm technology, and offers Secure Memory Encryption + Secure Encrypted Virtualization (SEV) + SEV Encrypted State + SEV-Secure Nested Paging security features. See the AMD EPYC Architecture White Paper (https://library.amd.com/l/3f4587d147382e2/) for more information.

MI300-53: Testing completed on 05/28/2024 by AMD performance lab attempting text generated throughput measured using Mistral-7B model comparison. Tests were performed using batch size 1 and 2048 input tokens and 2048 output tokens for Mistral-7B **Configurations: **2P AMD EPYC 9534 64-Core Processor based production server with 8x AMD InstinctTM MI300X (192GB, 750W) GPU, Ubuntu® 22.04.1, and ROCm™ 6.1.1 Vs. 2P Intel Xeon Platinum 8468 48-Core Processor based production server with 8x NVIDIA Hopper H100 (80GB, 700W) GPU, Ubuntu 22.04.3, and CUDA® 12.2. 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.

MI300-54: Testing completed on 05/28/2024 by AMD performance lab attempting text generated Llama3-70B using batch size 1 and 2048 input tokens and 128 output tokens for each system. **Configurations:** 2P AMD EPYC 9534 64-Core Processor based production server with 8x AMD InstinctTM MI300X (192GB, 750W) GPU, Ubuntu® 22.04.1, and ROCm™ 6.1.1 Vs. 2P Intel Xeon Platinum 8468 48-Core Processor based production server with 8x NVIDIA Hopper H100 (80GB, 700W) GPU, Ubuntu 22.04.3, and CUDA® 12.2 **8 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.

MI300-62: Testing conducted by internal AMD Performance Labs as of September 29, 2024 inference performance comparison between ROCm 6.2 software and ROCm 6.0 software on the systems with 8 AMD Instinct™ MI300X GPUs coupled with Llama 3.1-8B, Llama 3.1-70B, Mixtral-8x7B, Mixtral-8x22B, and Qwen 72B models. ROCm 6.2 with vLLM 0.5.5 performance was measured against the performance with ROCm 6.0 with vLLM 0.3.3, and tests were performed across batch sizes of 1 to 256 and sequence lengths of 128 to 2048. Configurations: 1P AMD EPYC™ 9534 CPU server with 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, Supermicro AS-8125GS-TNMR2, NPS1 (1 NUMA per socket), 1.5 TiB (24 DIMMs, 4800 mts memory, 64 GiB/DIMM), 4x 3.49TB Micron 7450 storage, BIOS version: 1.8, , ROCm 6.2.0-00, vLLM 0.5.5, PyTorch 2.4.0, Ubuntu® 22.04 LTS with Linux kernel 5.15.0-119-generic. Vs. 1P AMD EPYC 9534 CPU server with 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, Supermicro AS-8125GS-TNMR2, NPS1 (1 NUMA per socket), 1.5TiB 24 DIMMS, 4800 mts memory, 64 GiB/DIMM), 4x 3.49TB Micron 7450 storage, BIOS version: 1.8, ROCm 6.0.0-00, vLLM 0.3.3, PyTorch 2.1.1, Ubuntu 22.04 LTS with Linux kernel 5.15.0-119-generic. Server manufacturers may vary configurations, yielding different results. Performance may vary based on factors including but not limited to different versions of configurations, vLLM, and drivers.

Endnotes

MI300-63: Testing conducted by internal AMD Performance Labs as of September 29, 2024 training performance comparison between ROCm 6.2 software with compared to ROCm 6.0 software both with Megatron-LM on systems with 8 AMD Instinct™ MI300X GPUs running Llama 2-7B, Llama 2-70B (4K), Qwen1.5-14B models using custom docker container for each system. ROCm 6.2 with megatron-LM TFLOPs was measured against the TFLOPs with ROCm 6.0 with megatron-LM. Configurations:CPU: 1P AMD EPYC 9454 48-core processor,Host memory: 2x3.5 T GB GPU: AMD Instinct MI300X. 1P AMD EPYC™ 9454 CPU, 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, American Megatrends International LLC BIOS version: 1.8, ROCm 6.2 internal release, Megatron-LM code branches hanl/disable_te_llama2 for Llama 2-7B, guihong_dev for LLama 2-70B, renwuli/disable_te_qwen1.5 for Qwen1.5-14B, PyTorch 2.4, Ubuntu 22.04 LTS with Linux kernel 5.15.0-117-generic. Vs. 1P AMD EPYC 9454 CPU 48-core processor, 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, American Megatrends International LLC BIOS version: 1.8, ROCm 6.0.0, Megatron-LM code branches hanl/disable_te_llama2 for Llama 2-7B, guihong_dev for LLama 2-70B, renwuli/disable_te_qwen1.5 for Qwen1.5-14B, PyTorch 2.2, Ubuntu 22.04 LTS with Linux kernel 5.15.0-72-generic. Server manufacturers may vary configurations, yielding different results. Performance may vary based on factors including but not limited to different versions of configurations, megatron-LM, and drivers. Results: MI300X with ROCm 6.2 delivers average 1.83X the (83% higher) training throughput than ROCm 6.0.

MI300-64: Based on testing completed on 10/09/2024 by AMD performance lab measuring overall throughput for text generated using LLaMA 3.1-405B model using FP8 datatype. Test was performed using various input token length and an output token length for the following configurations of AMD Instinct™ MI325X 8xGPU platform and NVIDIA H100 platform. Configurations: AMD Instinct™ MI300X platform: Supermicro AS - 8125GS-TNMR2 server with 2x AMD EPYC 9654 Processors, 8x AMD MI300X (192GB, 750W) GPUs, Ubuntu 22.04).ROCm 6.2 NVIDIA H100 HGX platform: Supermicro AS - 8125GS-TNHR server with 2x AMD EPYC 9654 Processors, 8x Nvidia H100 (80GB, 700W) GPUs, Ubuntu 22.04) Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. MI300-64

MI325-001A: Calculations conducted by AMD Performance Labs as of September 26th, 2024, based on current specifications and /or estimation . The AMD Instinct™ MI325X OAM accelerator will have 256GB HBM3E memory capacity and 6 TB/s GPU peak theoretical memory bandwidth performance. Actual results based on production silicon mayvary. The highest published results on the NVidia Hopper H200 (141GB) SXM GPU accelerator resulted in 141GB HBM3E memory capacity and 4.8 TB/s GPU memory bandwidth performance. https://nvdam.widen.net/s/nb5zzzsjdf/hpc-datasheet-sc23-h200-datasheet-3002446. The highest published results on the NVidia Blackwell HGX B100 (192GB) 700W GPU accelerator resulted in 192GB HBM3E memory capacity and 8 TB/s GPUmemory bandwidth performance. The highest published results on the NVidia Blackwell HGX B200 (192GB) GPU accelerator resulted in 192GB HBM3E memory capacity and 8 TB/s GPU memory bandwidth performance. Nvidia Blackwell specifications at https://resources.nvidia.com/en-us-blackwell-architecture?_gl=1*1r4pme7*_gcl_aw*R0NMLjE3MTM5NjQ3NTAuQ2p3S0NBancyNkt4QmhCREVpd0F1NktYdDlweXY1dlUtaHNKNmhPdHM4UVdPSlM3dFdQaE40WkI4THZBaWFVajFyTGhYd3hLQmlZQ3pCb0NsVElRQXZEX0J3RQ..*_gcl_au*MTIwNjg4NjU0Ny4xNzExMDM1NTQ3

Endnotes

MI325-02: Calculations conducted by AMD Performance Labs as of May 28th, 2024 for the AMD Instinct™ MI325X GPU resulted in 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. Actual performance will vary based on final specifications and system configuration. Published results on Nvidia H200 SXM (141GB) GPU: 989.4 TFLOPS peak theoretical half precision tensor (FP16 Tensor), 989.4 TFLOPS peak theoretical Bfloat16 tensor format precision (BF16 Tensor), 1,978.9 TFLOPS peak theoretical 8-bit precision (FP8), 1,978.9 TOPs peak theoretical INT8 floating-point performance. BFLOAT16 Tensor Core, FP16 Tensor Core, FP8 Tensor Core and INT8 Tensor Core performance were published by Nvidia using sparsity; for the purposes of comparison, AMD converted these numbers to non-sparsity/dense by dividing by 2, and these numbers appear above. Nvidia H200 source: https://nvdam.widen.net/s/nb5zzzsjdf/hpc-datasheet-sc23-h200-datasheet-3002446 and https://www.anandtech.com/show/21136/nvidia-at-sc23-h200-accelerator-with-hbm3e-and-jupiter-supercomputer-for-2024 Note: Nvidia H200 GPUs have the same published FLOPs performance as H100 products https://resources.nvidia.com/en-us-tensor-core/.

MI325-004: Based on testing completed on 9/28/2024 by AMD performance lab measuring text generated throughput for Mixtral-8x7B model using FP16 datatype. Test was performed using input length of 128 tokens and an output length of 4096 tokens for the following configurations of AMD Instinct™ MI325X GPU accelerator and NVIDIA H200 SXM GPU accelerator. 1x MI325X at 1000W with vLLM performance Vs. 1x H200 at 700W with TensorRT-LLM v0.13 Configurations: AMD Instinct™ MI325X reference platform:
1x AMD Ryzen™ 9 7950X CPU, 1x AMD Instinct MI325X (256GiB, 1000W) GPU, Ubuntu® 22.04, and ROCm™ 6.3 pre-release Vs NVIDIA H200 HGX platform: Supermicro SuperServer with 2x Intel Xeon® Platinum 8468 Processors, 8x Nvidia H200 (140GB, 700W) GPUs [only 1 GPU was used in this test], Ubuntu 22.04) CUDA® 12.6. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI325-005:: Based on testing completed on 9/28/2024 by AMD performance lab measuring overall latency forLLaMA 3.1-70B model using FP8 datatype. Test was performed using input length of 2048 tokens and an output length of 2048 tokens for the following configurations of AMD Instinct™ MI325X GPU accelerator and NVIDIA H200 SXM GPU accelerator. MI325X at 1000W with vLLM performance: 48.025 sec (latency in seconds) Vs. 1x H200 at 700W with TensorRT-LLM v 0.13: 56.310 sec (latency in seconds) Configurations: AMD Instinct™ MI325X reference platform: 1x AMD Ryzen™ 9 7950X 16-Core Processor CPU, 1x AMD Instinct MI325X (256GiB, 1000W) GPU, Ubuntu® 22.04, and ROCm™ 6.3 pre-release Vs NVIDIA H200 HGX platform: Supermicro SuperServer with 2x Intel Xeon® Platinum 8468 Processors, 8x Nvidia H200 (140GB, 700W) GPUs, Ubuntu 22.04), CUDA 12.6. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI325-006: Based on testing completed on 9/28/2024 by AMD performance lab measuring overall latency for LLaMA 3.1-70B model using FP8 datatype. Test was performed using input length of 2048 tokens and an output length of 2048 tokens for the following configurations of AMD Instinct™ MI325X GPU accelerator and NVIDIA H200 SXM GPU accelerator. MI325X at 1000W with vLLM performance: 48.025 sec (latency in seconds) Vs. 1x H200 at 700W with TensorRT-LLM v 0.13: 56.310 sec (latency in seconds) Configurations: AMD Instinct™ MI325X reference platform: 1x AMD Ryzen™ 9 7950X 16-Core Processor CPU, 1x AMD Instinct MI325X (256GiB, 1000W) GPU, Ubuntu® 22.04, and ROCm™ 6.3 pre-release Vs NVIDIA H200 HGX platform: Supermicro SuperServer with 2x Intel Xeon® Platinum 8468 Processors, 8x Nvidia H200 (140GB, 700W) GPUs, Ubuntu 22.04), CUDA 12.6. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

Endnotes

MI355-004: Calculations conducted by AMD Performance Labs as of September 26th, 2024 for the AMD Instinct™ MI300X GPU platform and AMD Instinct™ MI300X GPU platform performance comparing FP16, FP8 and FP4 datatypes. MI355X 8xGPU Platform Peak theoretical Half Precision (FP16) Performance - 18.5 PFLOPs. Peak theoretical Eight-bit Precision (FP8) Performance - 37 PFLOPs. Peak theoretical Four-bit Precision (FP4) Performance - 74 PFLOPs . MI325X 8xGPU Platform: Peak theoretical Half Precision (FP16) Performance - 10.4 PFLOPs Peak theoretical Eight-bit Precision (FP8) Performance - 20.88 PFLOPs. MI300X 8xGPU Platform: Peak theoretical Half Precision (FP16) Performance - 10.4 PFLOPs. Actual performance will vary based on final specifications and system configuration.

MI355-005: Calculations conducted by AMD Performance Labs as of October 2nd, 2024 for the AMD Instinct™ MI300X GPU accelerator, AMD Instinct™ MI325X GPU accelerator and AMD Instinct™ MI350X GPU accelerator performance comparing FP16, FP8 and FP4 datatypes. MI300X GPU Accelerator Peak theoretical Half Precision (FP16) Performance - 1.3 PFLOPs Peak theoretical Eight-bit Precision (FP8) Performance - 2.61 PFLOPs. MI325X GPU Accelerator. Peak theoretical Half Precision (FP16) Performance - 1.3 PFLOPs. Peak theoretical Eight-bit Precision (FP8) Performance - 2.61 PFLOPs. MI355X GPU Accelerator: Peak theoretical Half Precision (FP16) Performance - 2.3 PFLOPs. Peak theoretical Eight-bit Precision (FP8) Performance - 4.614 PFLOPs. Peak theoretical Six-bit Precision (FP6) Performance – 9.227 PFLOPs.Peak theoretical Four-bit Precision (FP4) Performance - 9.227 PFLOPs. Actual performance will vary based on final specifications and system configuration

GD-173a: AMD defines “All Day Battery Life” as at least 8 hours of continuous battery life and "Multi-Day battery Life" as continuous runtime above 8 hours. All battery life scores are approximate. Actual battery life will vary based on several factors, including, but not limited to: system configuration and software, settings, product use and age, and operating conditions.

GD-243: Trillions of Operations per Second (TOPS) for an AMD Ryzen processor is the maximum number of operations per second that can be executed in an optimal scenario and may not be typical. TOPS may vary based on several factors, including the specific system configuration, AI model, and software version.

STXP-04: Based on product specifications and competitive products announced as of Oct 2024 and testing as of Sept 2024 by AMD performance labs using the following systems: HP EliteBook X G1a with AMD Ryzen AI 9 HX PRO 375 processor @23W, Radeon 880M graphics, 32GB of RAM, 512GB SSD, VBS=ON, Windows 11 PRO; Dell Latitude 7450 with Intel Core Ultra 7 165U processor @15W (vPro enabled), Intel Iris Xe Graphics, VBS=ON, 32GB RAM, 512GB NVMe SSD, Microsoft Windows 11 Professional; Dell Latitude 7450 with Intel Core Ultra 7 165H processor @28W (vPro enabled), Intel Iris Xe Graphics, VBS=ON, 16GB RAM, 512GB NVMe SSD, Microsoft Windows 11 Pro. All systems were tested in Best Performance Mode. AI PC is defined as a laptop PC with a processor that includes a neural processing unit (NPU).

STXP-05: Based on Microsoft Copilot+ requirements of minimum 40 TOPS using AMD product specifications and competitive products announced as of Oct 2024. Microsoft requirements found here - https://support.microsoft.com/en-us/topic/copilot-pc-hardware-requirements-35782169-6eab-4d63-a5c5-c498c3037364.

STXP-12: Testing as of Sept 2024 by AMD performance labs on an HP EliteBook X G1a (14in) (40W) with AMD Ryzen AI 9 HX PRO 375 processor, Radeon™ 890M graphics, 32GB of RAM, 512GB SSD, VBS=ON, Windows 11 Pro vs. a Dell Latitude 7450 with an Intel Core Ultra 7 165H processor (vPro enabled), Intel Arc Graphics, VBS=ON, 16GB RAM, 512GB NVMe SSD, Microsoft Windows 11 Pro in the application(s) (Best Performance Mode): Cinebench R24 nT. Laptop manufactures may vary configurations yielding different results. STXP-12.

Endnotes

STXP-30: Video Playback Battery Life vs Intel Core Ultra 165H: Intel Core Ultra 7 165H: 13.1 or 13 hours and 6 min (100.0% Baseline Score) - Score: AMD Ryzen™ AI 9 HX PRO 375: 23.6 or 23 hours and 36 minutes (23.6 / 13.1 = ~180% or 80%). Video Playback Battery Life vs Apple M3 Pro 12-Core: 14.4 or 14 hours and 24 min (100.0% Baseline Score) - Score: AMD Ryzen™ AI 9 HX PRO 375: 23.6 or 23 hours and 36 minutes (23.6 / 14.4 = ~163.88% or 64%). PC manufacturers may vary configurations yielding different results. All scores are an average of 3 runs with the same settings. Performance may vary based on use of latest drivers.

STXP-32: Video Playback Battery Life vs Intel Core Ultra 165H: Intel Core Ultra 7 165H: 13.1 or 13 hours and 6 min (100.0% Baseline Score) - Score: AMD Ryzen™ AI 9 HX PRO 375: 23.6 or 23 hours and 36 minutes (23.6 / 13.1 = ~180% or 80%) Video Playback Battery Life vs Apple M3 Pro 12-Core: 14.4 or 14 hours and 24 min (100.0% Baseline Score) - Score: AMD Ryzen™ AI 9 HX PRO 375: 23.6 or 23 hours and 36 minutes (23.6 / 14.4 = ~163.88% or 64%) PC manufacturers may vary configurations yielding different results. All scores are an average of 3 runs with the same settings. Performance may vary based on use of latest drivers.

DISCLAIMER AND ATTRIBUTIONS

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'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