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.

2024 AMD AI SOLUTIONS DAY
AI 無限進化 . AMD 驅動未來

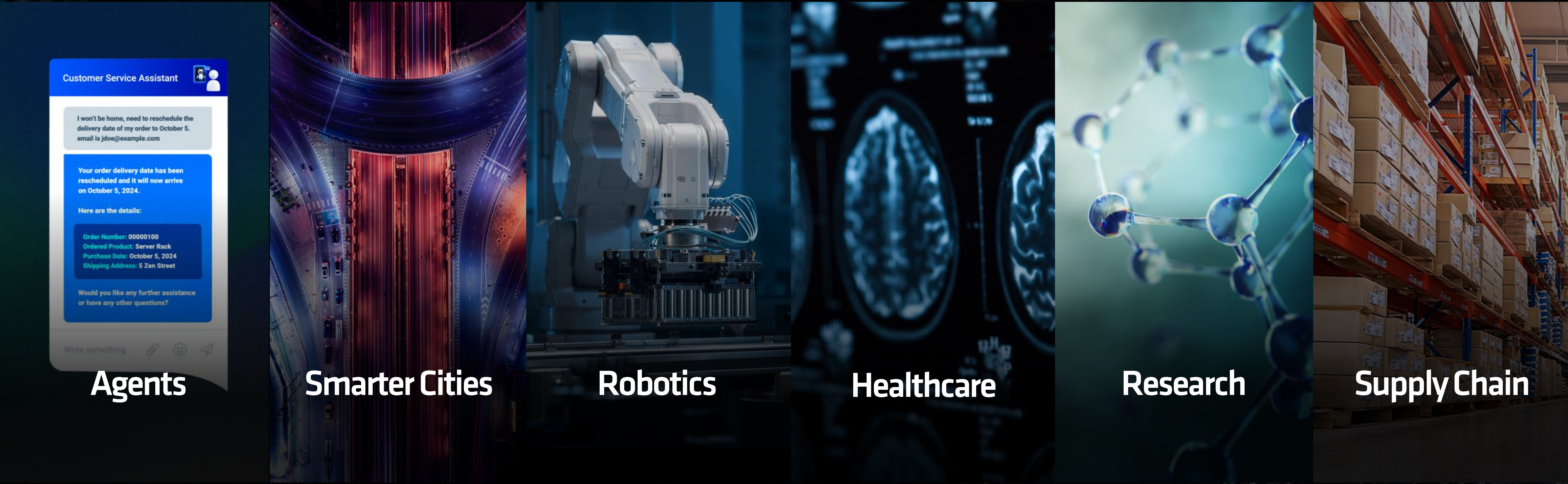
AMD 台灣區商用市場業務處
資深業務副總經理

林建誠 Ken Lin

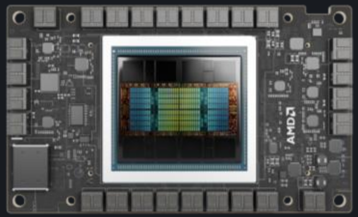
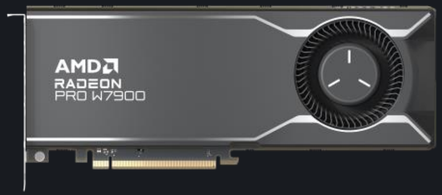
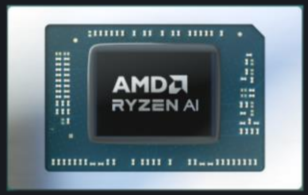
AMD
together we advance_

AI

Most transformational technology in 50 years



Broad portfolio to address diverse spectrum of requirements

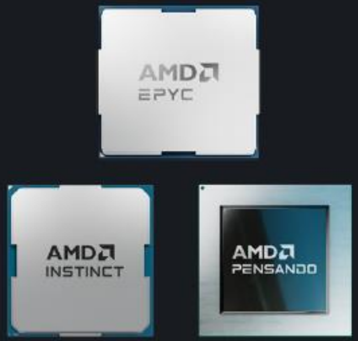
AMD EPYC™ Processors	AMD Instinct™ Accelerators	AMD Radeon™ Graphics	AMD Ryzen™ Mobile Processors	AMD Versal™ Adaptive SoCs
				

From cloud to endpoint to embedded

Advancing end-to-end AI infrastructure

Cloud	Endpoint	Edge

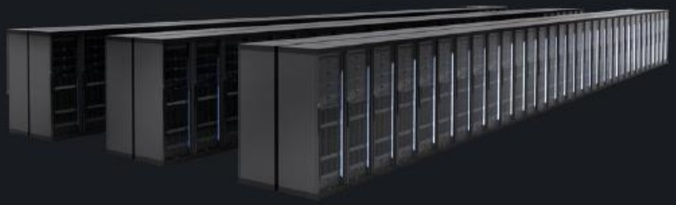
Evolving Complexity: 250KW Racks, 100K node clusters



Silicon

Server

Rack



Data Center

3x chip power

15x rack power/density

32x GPU POD size

100x cluster size

10-100x datacenter power

16x networking bandwidth

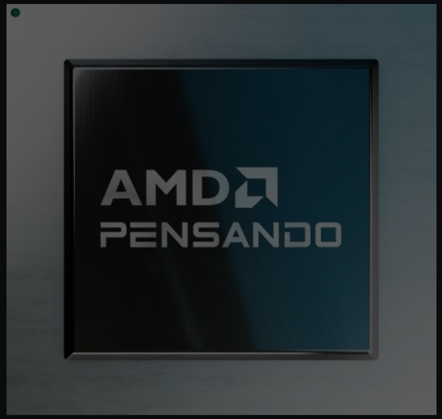
AMD internal estimates

Advancing Data Center Solutions

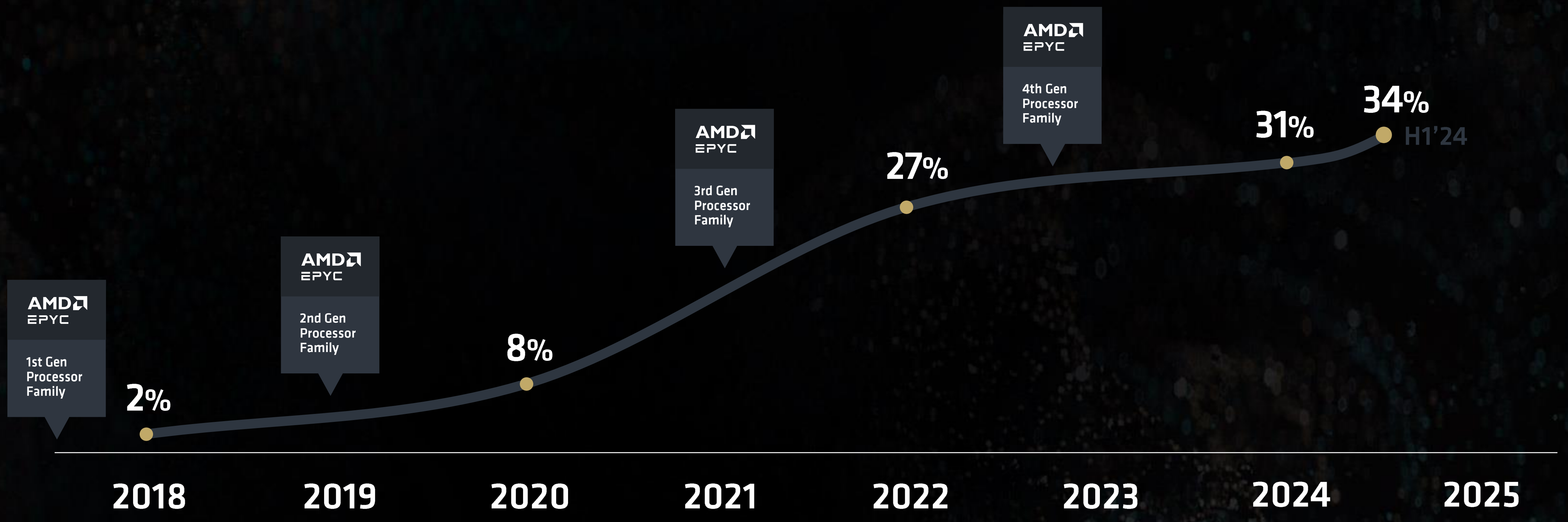
Data Center CPUs

Data Center GPUs

Networking



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

Hyperscale leaders power internal workloads with AMD, serving billions worldwide

Trusted by industry leaders on-prem

5th Gen AMD EPYC™

World's best CPU for Cloud, Enterprise and AI

ZEN 5

3nm
4nm

150 billion transistors

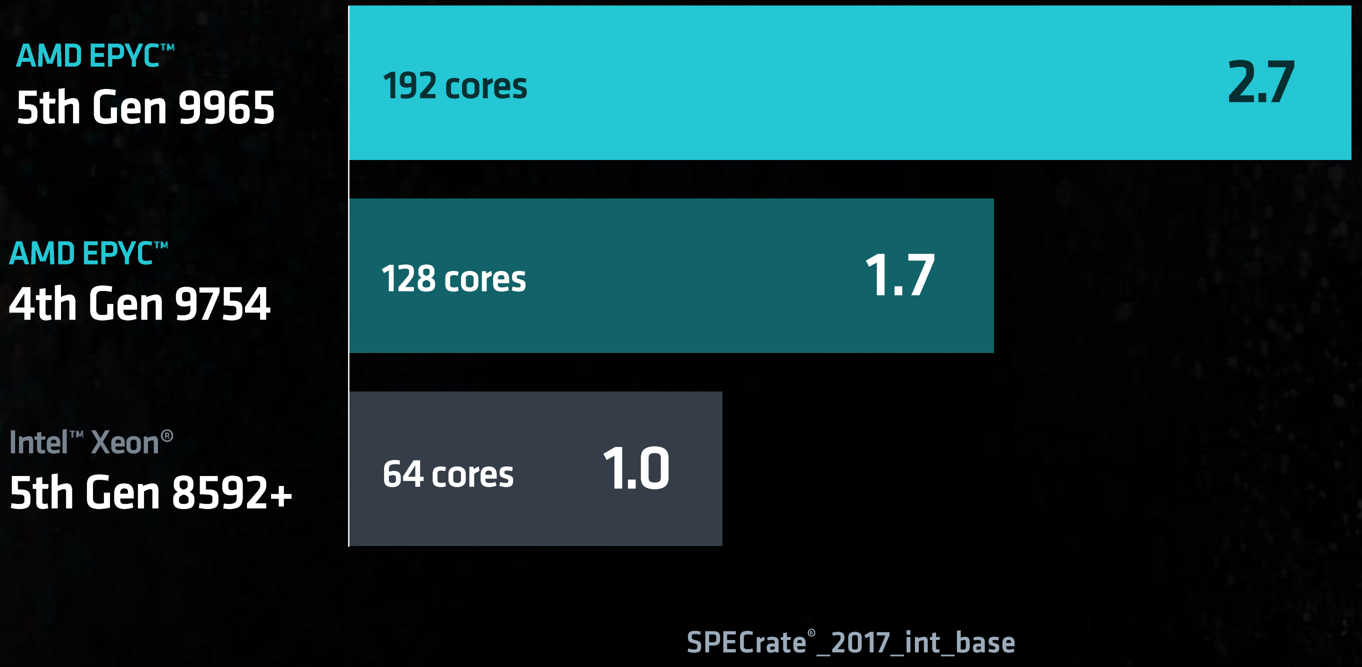
Up to 192 cores
384 threads

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

Industry's Highest Performing Server CPU

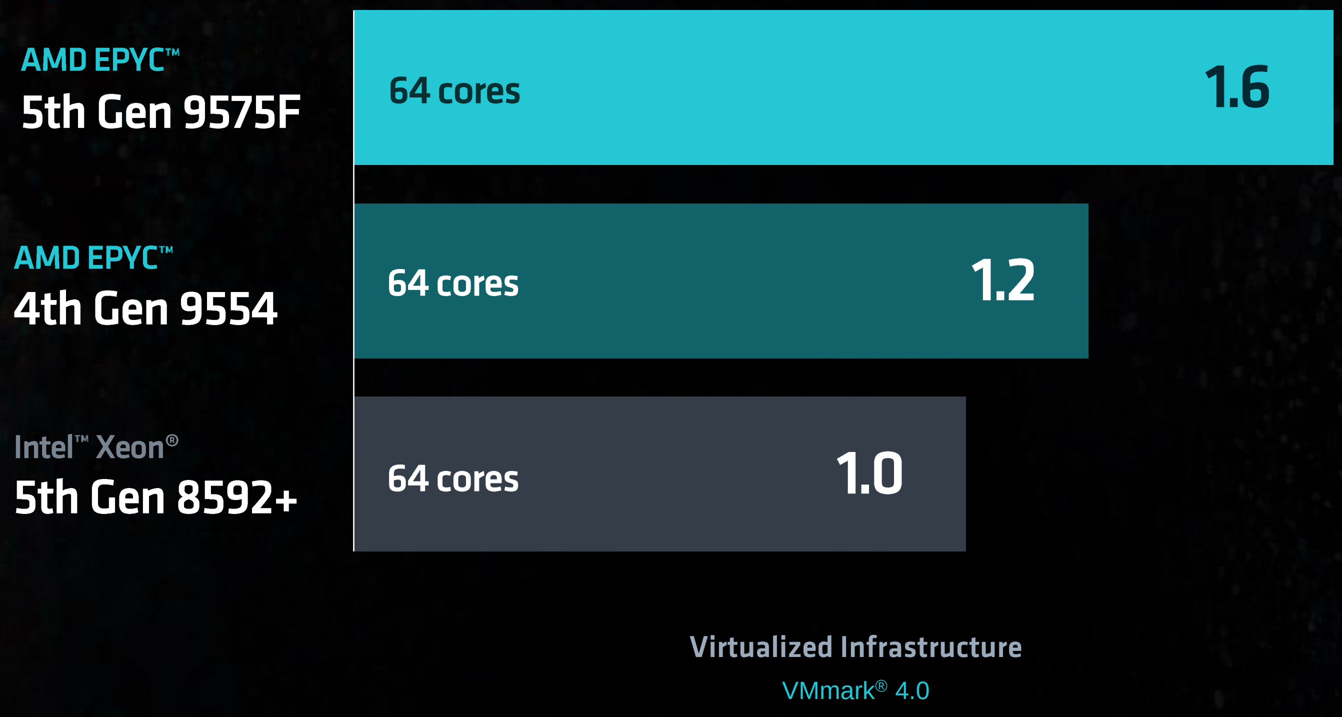


2.7x

vs. top-of-stack "Emerald Rapids"

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

60% More Performance at the Same Licensing Cost

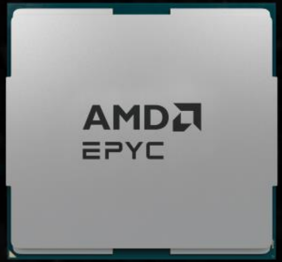


up to **1.6x**

Performance per core in virtualized infrastructure

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

Leadership Engines for Enterprise AI Workloads



x86 CPU

Data Input › Data Cleaning › Pre-Processing › Model Training › Deployment

GPU

Data from x86-Based Systems

‹ **Agentic AI** ›

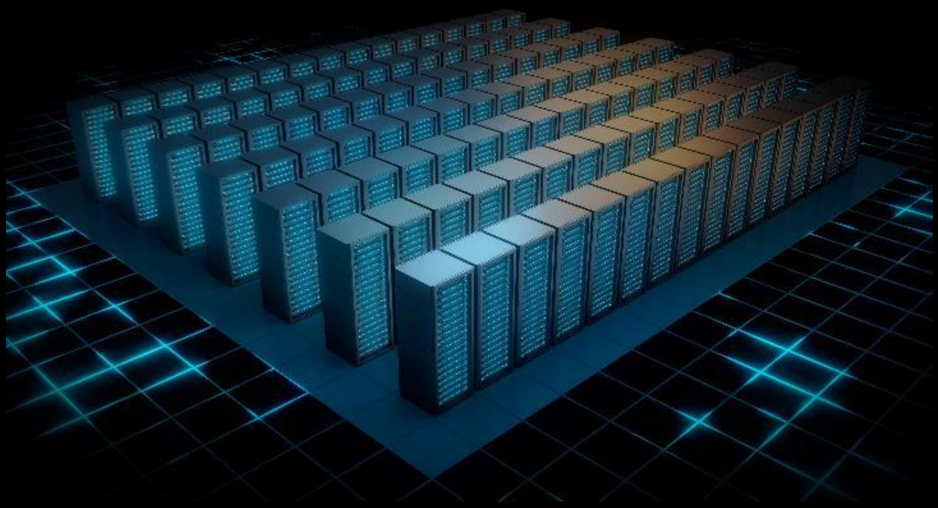
LLM Training and Inference

From **analytics** to **generative AI** to **agentic AI**

Easily Upgrade to 5th Gen AMD EPYC™ CPUs

Modernize your data center – Add more capacity for your compute needs

1000 Old Servers
2P Intel® Xeon® Platinum 8280 servers



7 to 1

Easy to migrate to AMD
- X86 architecture
- Mature ecosystem
- Robust tools

131 Modern Servers
2P AMD EPYC™ 9965 servers

Up to **68%**
Less power

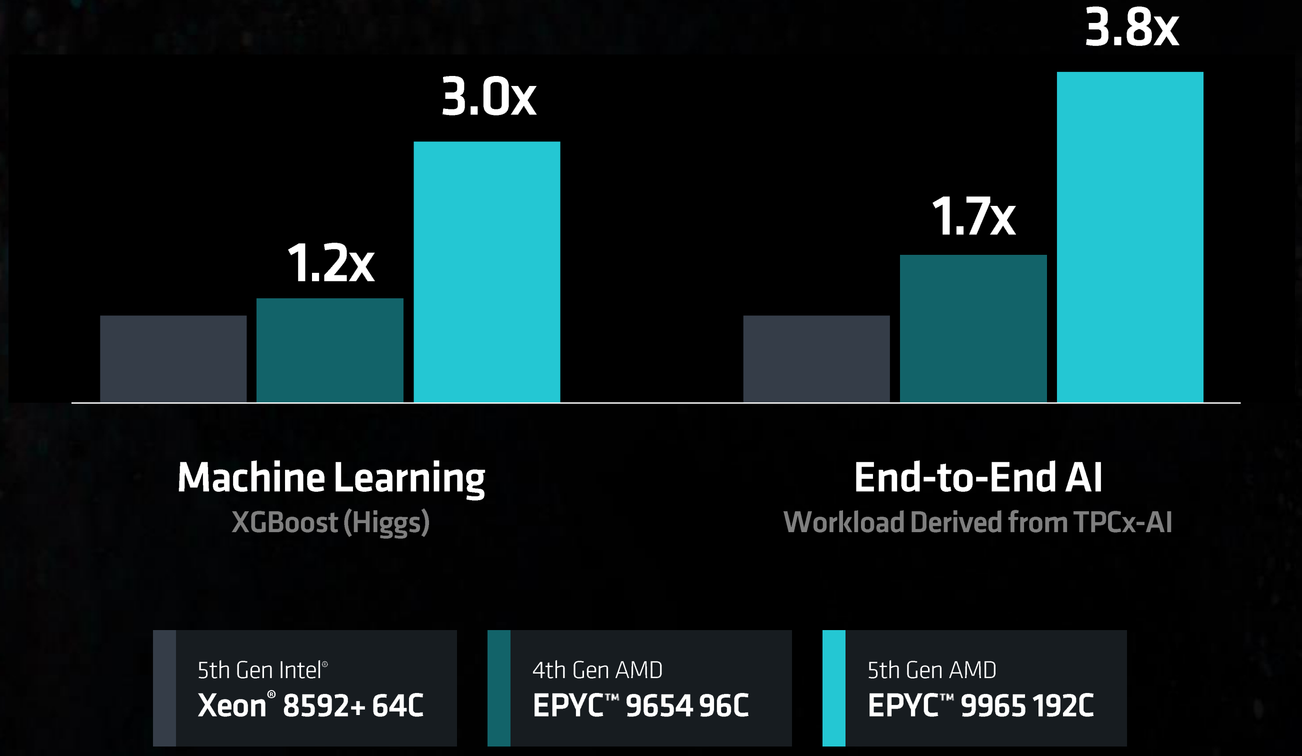
Up to **87%**
Fewer Servers

Up to **67%**
Lower 3-yr TCO

Servers required to achieve a total of 391,000 SPECrate®2017_ int_base performance score

See endnotes 9xx5TCO-002A

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

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

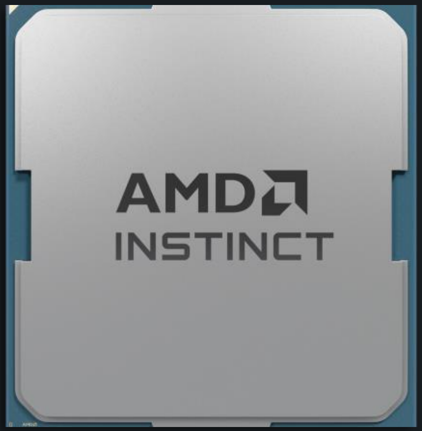
with Stable Diffusion XL V2

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.

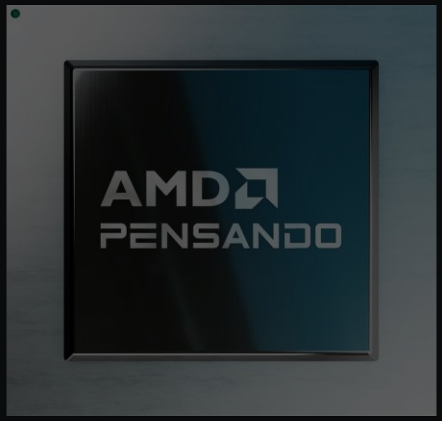
Advancing Data Center Solutions

Data Center CPUs

Data Center GPUs



Networking



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 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

AMD INSTINCT

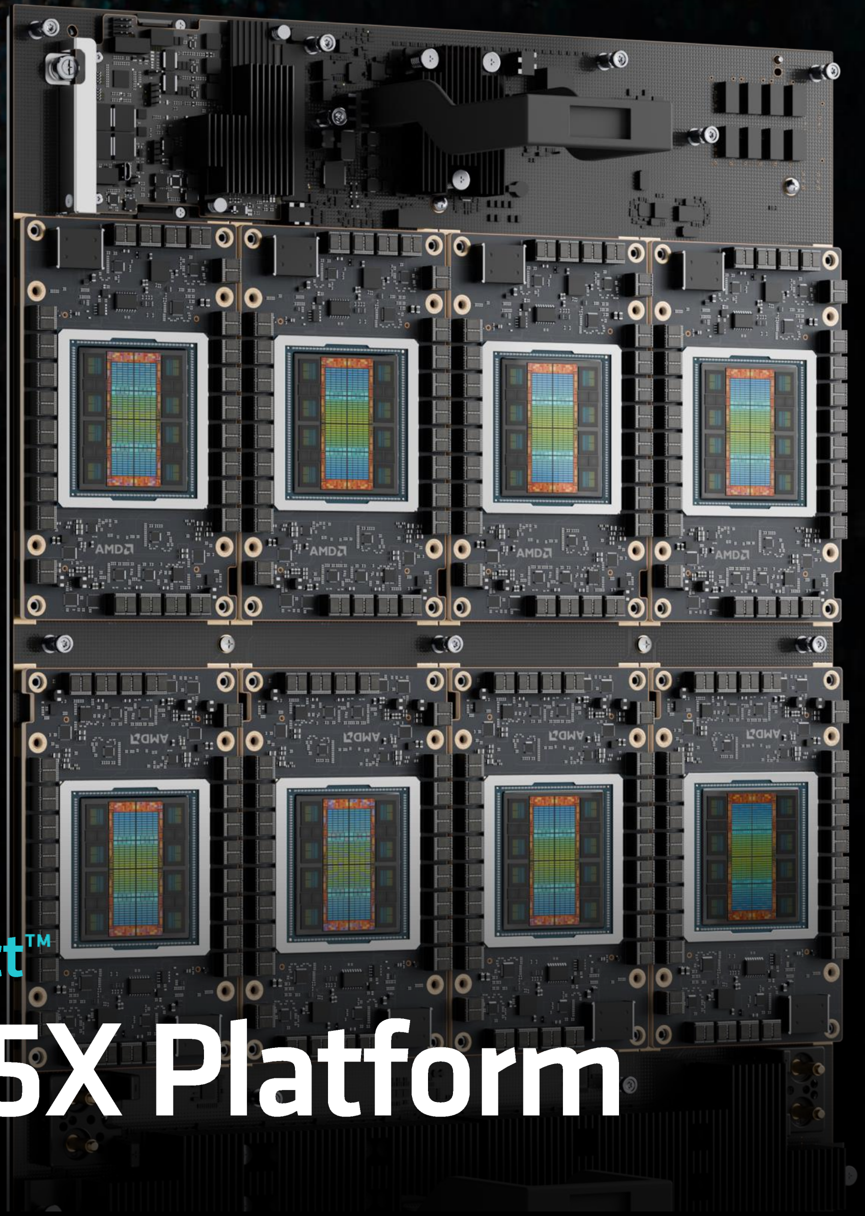
Solutions from leading OEMs and cloud

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



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**

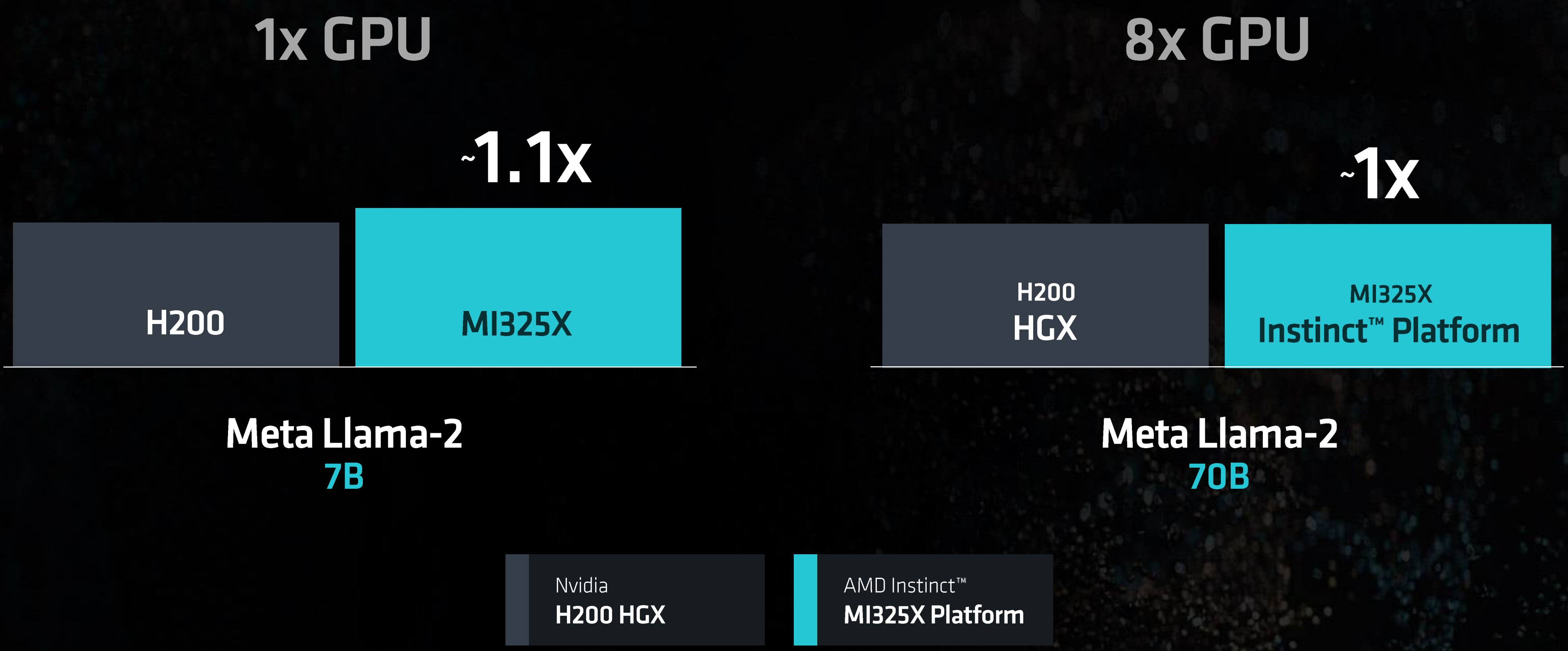
1.3x compute flops vs. H200 HGX

See endnotes MI325-001A, MI325-002

*Dense flops

World-Class Training Performance

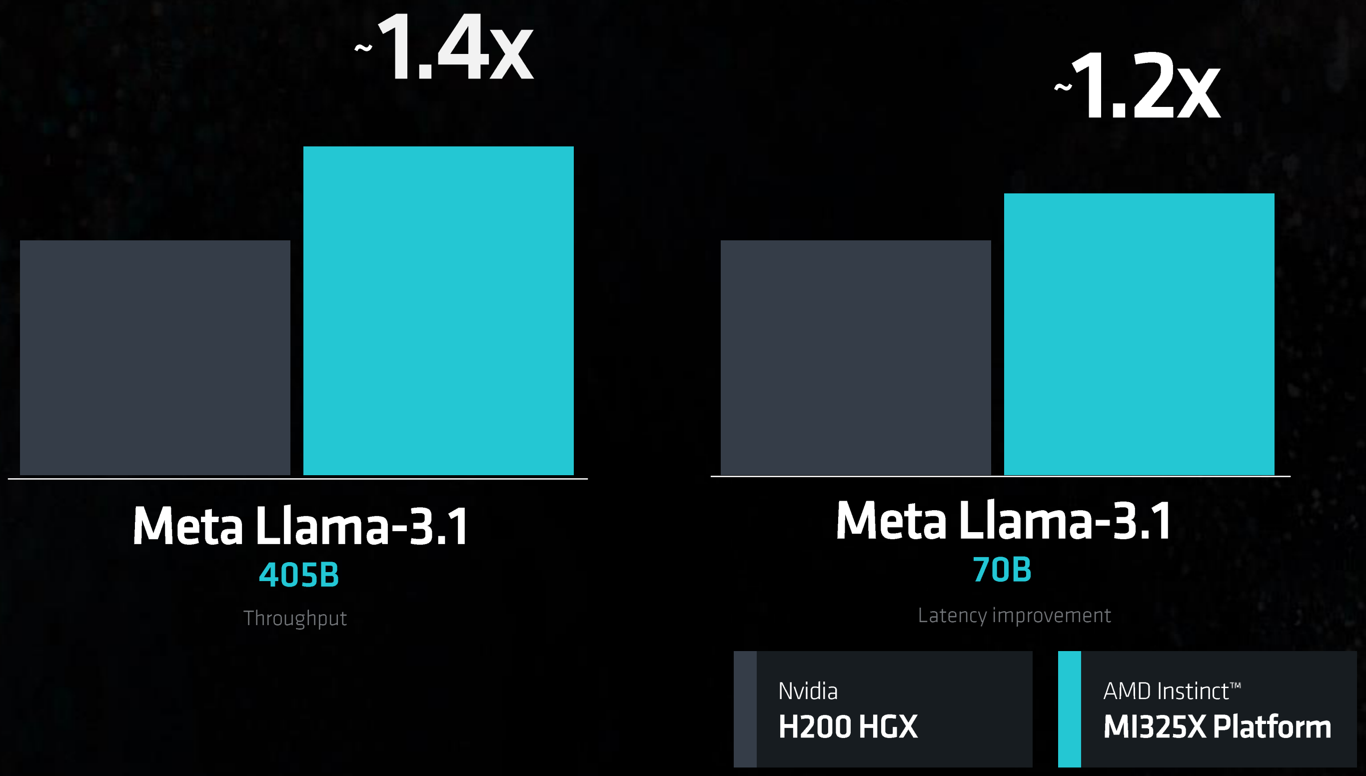
Single GPU and 8 GPU Training



See endnotes MI325-013, MI325-012.

AMD Instinct™ MI325X Platform

Leadership Inference performance using 8x MI325X



up to **1.4x**

Inference performance

See endnotes MI325-015, MI325-014

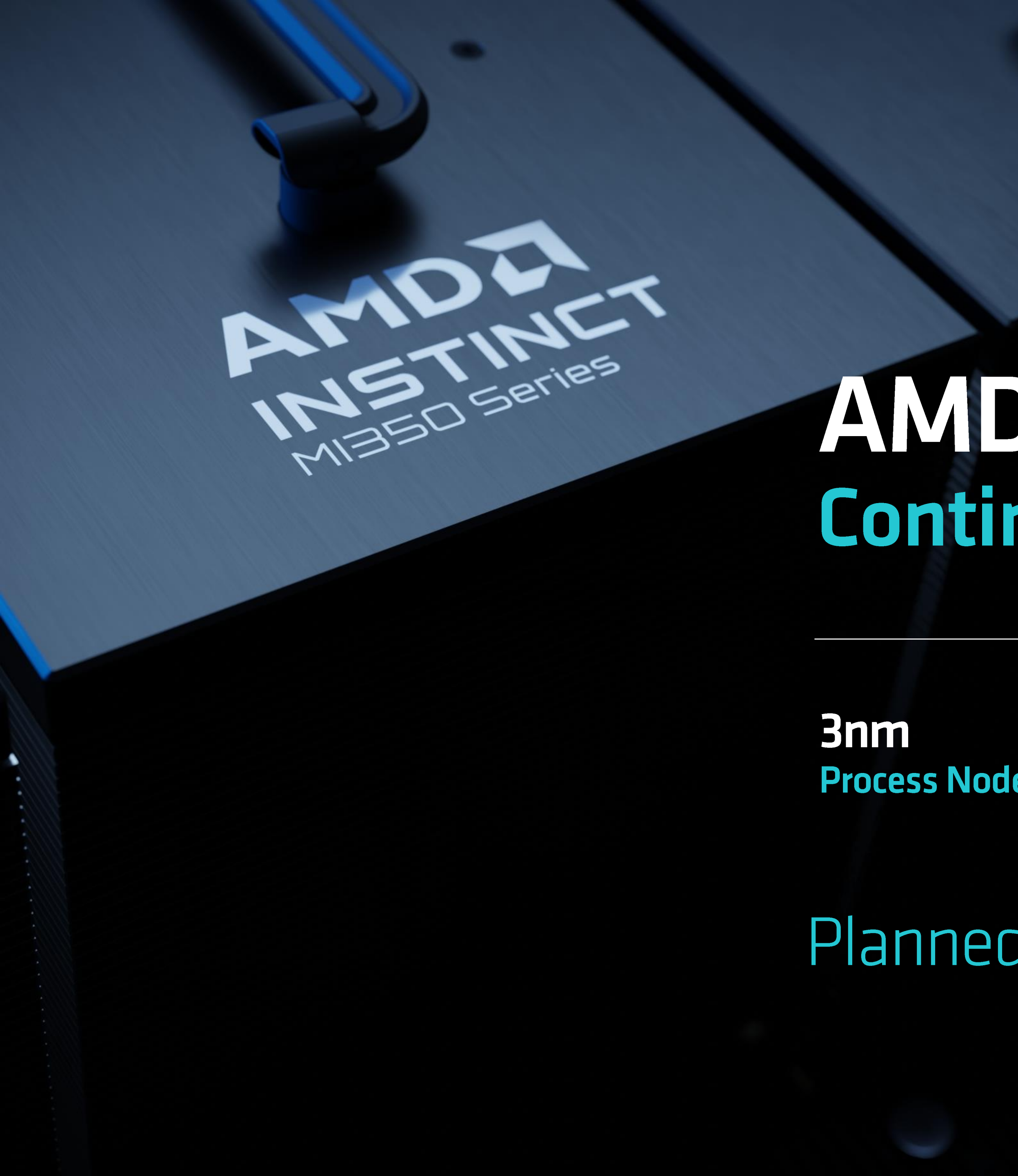
AMD Instinct™ MI325X GPU

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™ MI350 Series

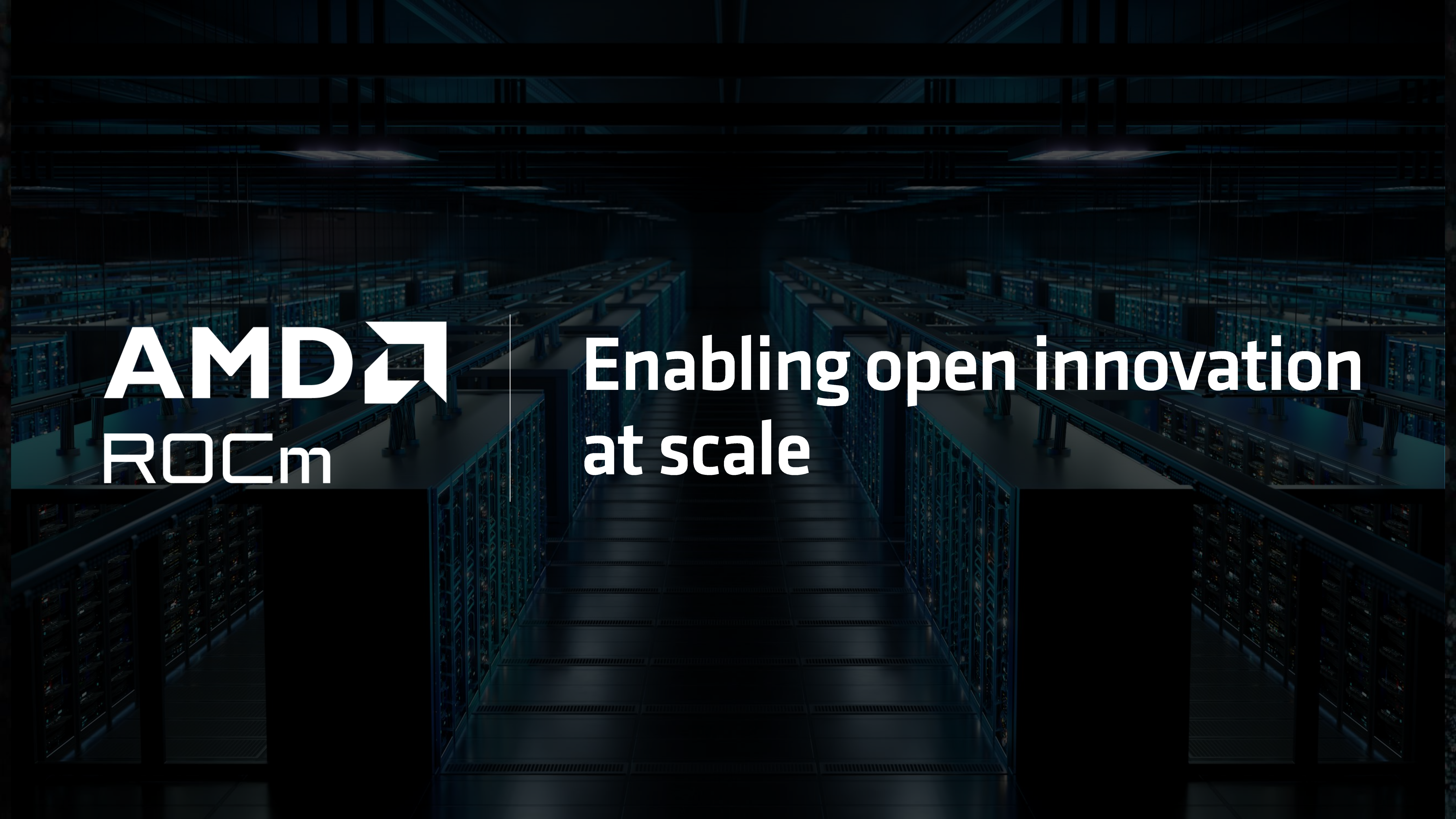
Continued Gen AI Leadership

3nm Process Node	Up to **288GB** HBM3E	**FP4 / FP6** Datatype Support	AMD CDNA 4

Planned availability **2H 2025**

AMD ROCm

Enabling open innovation at scale



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

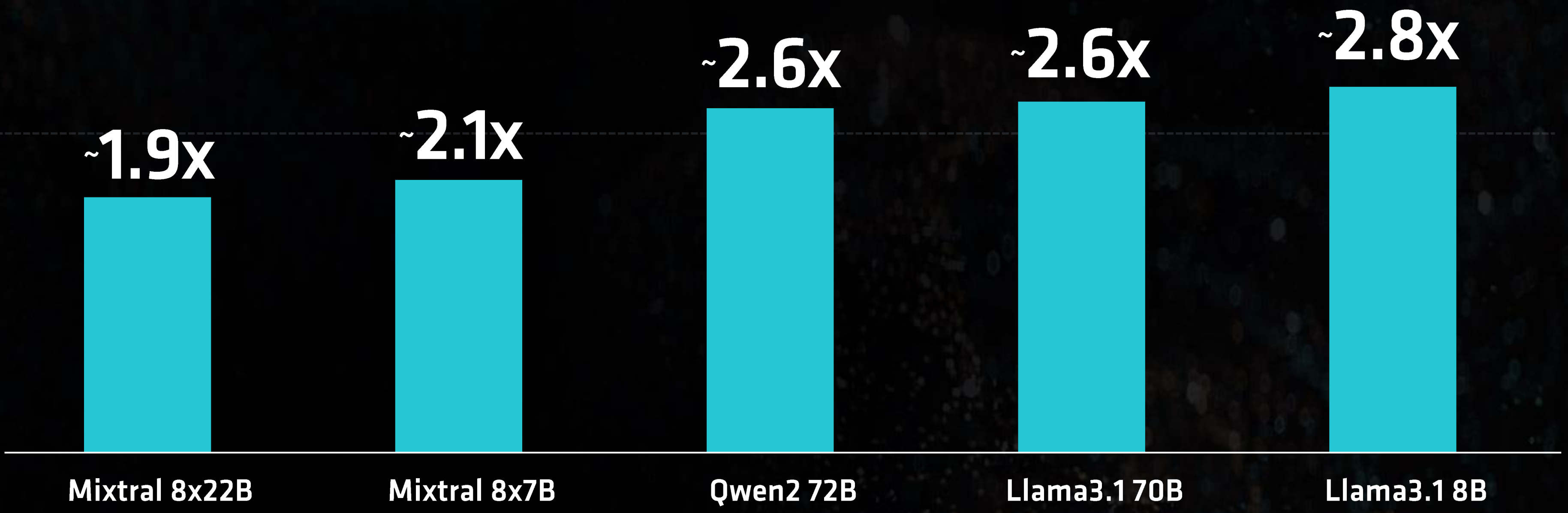
MI300X vs. H100

Leadership performance on popular models

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.

Advancing Data Center Solutions

Data Center CPUs

Data Center GPUs

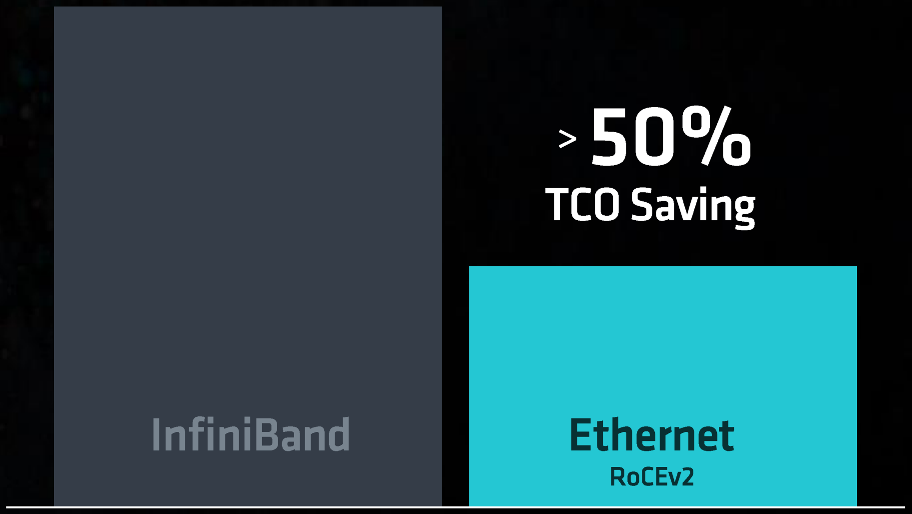
Networking

Ultra Ethernet Consortium

Steering Members

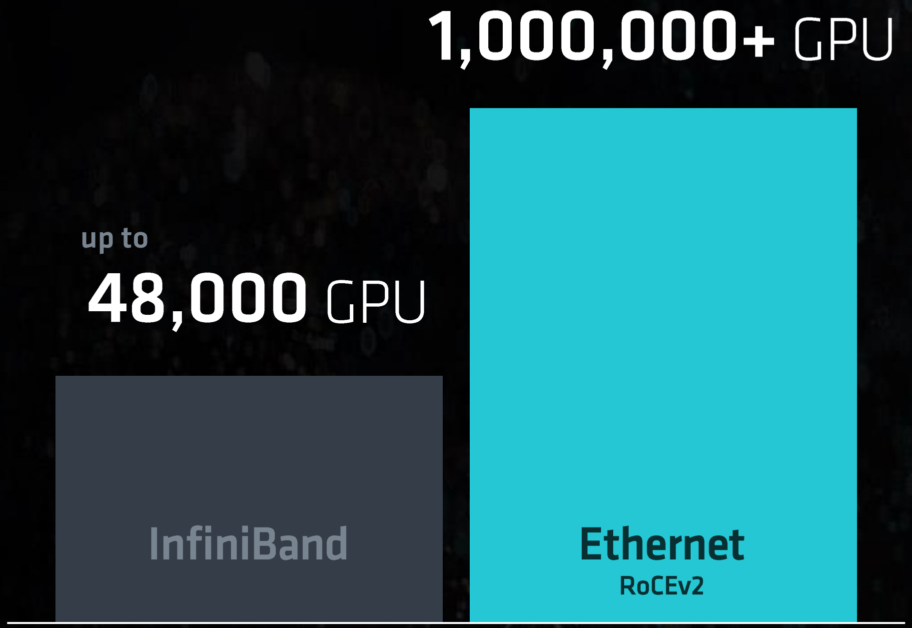
AMD
ARISTA
BROADCOM
CISCO
EVIDEN an atos business
Hewlett Packard Enterprise
intel
Meta
Microsoft
ORACLE

Ethernet is always the preferred choice



Total Cost of Ownership[1]

Lower is better



Scalability

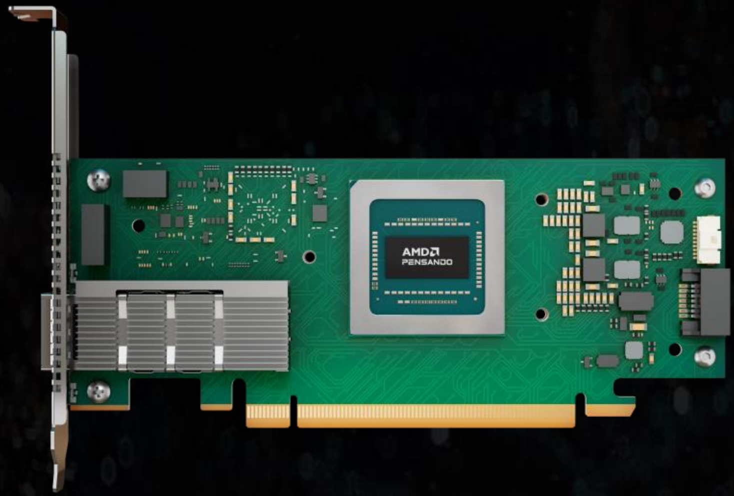
Higher is better

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

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

AMD | Advancing the AI Data Center

CPUs	GPUs	Networking	Software Solutions	Cluster Level Systems Design
AMD EPYC™	AMD Instinct™	DPUs, UALink + Ultra Ethernet	Open Software Stack	

AMD Pensando · Ultra Ethernet Consortium · UA Link | Ultra Accelerator Link · AMD ROCm · ZT Systems

Announced acquisition expected to close 1H 2025

AMD

Advancing end-to-end AI infrastructure

Cloud | Endpoint | Edge

AMD Ryzen™ AI Leads the AI PC Era

Q1 2023

AMD RYZEN AI

1st Gen
"Phoenix Point"

10
TOPS NPU

Q4 2023

AMD RYZEN AI

2nd Gen
"Hawk Point"

16
TOPS NPU

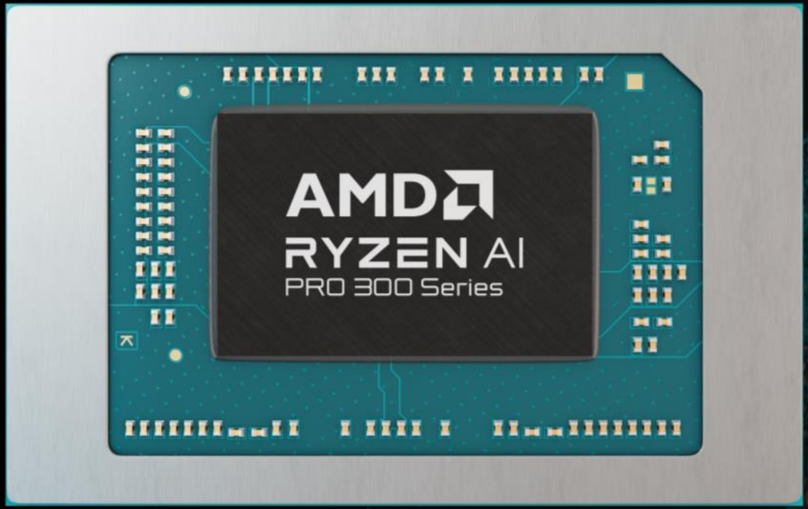
Q2 2024

AMD RYZEN AI

3rd Gen
"Strix Point"

50+
TOPS NPU

See endnote GD-243.



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

Enterprise AI PC Application Ecosystem

Adobe | Microsoft | webex by CISCO | zoom

splashtop | WHISPP | SOLIDWORKS | BUFFERZONE | Camo | bitdefender | LM Studio | grammarly

Blackmagicdesign | nero | BORISFX | Avid | Rhinoceros | GoPro | Topaz Labs™ | ARKRUNR | voicemy.ai

blender | RADiCAL | CyberLink | AFFINITY Photo 2 | ACCA SOFTWARE | convai | OBS Open Broadcaster Software | MAXON

Advancing end-to-end AI infrastructure

Cloud	Endpoint	Edge

General Disclaimer and Attribution Statement 2024

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/.

9xx5-012: 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.

.

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 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 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 33.699 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

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 may vary. 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 GPU memory 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

MI325-012: Overall GPU-normalized Training Throughput (processed tokens per second) for text generation using the Llama2-7b chat model running Megatron-LM v0.12 (BF16) when using a maximum sequence length of 4096 tokens comparison based on AMD internal testing as of 10/4/2024. Batch size according to largest micro-batch that fits in GPU memory for each system. AMD Instinct batch size 8, Nvidia batch size 2. Configurations: AMD Development system: 1P AMD Ryzen 9 7950X (16-core), 1x AMD Instinct™ MI325X (256GB, 1000W) GPU, 128 GiB memory, ROCm 6.3.0 (pre-release), Ubuntu 22.04.2 LTS with Linux kernel 5.15.0-72-generic, PyTorch 2.4.0.Vs. An Nvidia DGX H200 with 2x Intel Xeon Platinum 8468 Processors, 1x Nvidia H200 (141GB, 700W) GPUs, 2 TiB (32 DIMMs, 64 GiB/DIMM), CUDA 12.6.37-1, 560.35.03, Ubuntu 22.04.5, PyTorch 2.5.0a0+872d972e41.nv24.8. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. MI325-012

MI325-013: Overall GPU-normalized Training Throughput (processed tokens per second) for text generation using the Llama2-70b chat model running Megatron-LM v0.12 (BF16) when using a maximum sequence length of 4096 tokens comparison based on AMD internal testing as of 10/7/2024. Batch size according to largest micro-batch that fits in GPU memory for each system. AMD Instinct global batch size 32, Nvidia global batch size 49. Configurations:2P Intel Xeon Platinum 8480+ CPU server with 8x AMD Instinct™ MI325X (256GB, 1000W) GPUs, 4 TiB memory, ROCm® 6.3 (Pre-release), Ubuntu® 22.04.2, PyTorch 2.4.0. An Nvidia DGX H200 with 2x Intel Xeon Platinum 8468 Processors, 8x Nvidia H200 (141GB, 700W) GPUs, 2 TiB memory, CUDA 12.6.37-1, 560.35.03, Ubuntu 22.04.5, PyTorch 2.5.0a0+872d972e41.nv24.8.. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. MI325-013

Endnotes

MI325-002: 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 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.

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

- MI355-005: 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. MI355-005

- MI325-005: Based on testing completed on 9/28/2024 by AMD performance lab measuring overall latency for Mistral-7B model using FP16 datatype. Test was performed using input length of 128 tokens and an output length of 128 tokens for the AMD Instinct™ MI325X GPU accelerator and NVIDIA H200 SXM GPU accelerator. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. MI325-005

- 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. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. MI325-006

- MI325-014: Based on testing completed on 10/08/2024 by AMD performance lab measuring text generated throughput for LLaMA 3.1-405B model using FP8 datatype. Test was performed using input length of 128 tokens and an output length of 2048 tokens for the following configurations of AMD Instinct™ MI325X 8xGPU platform and NVIDIA H200 HGX GPU platform. 8xGPU MI325X platform with vLLM performance Vs. NVIDIA published results Configurations: MI325X 8xGPU Platform Configuration Dell PowerEdge XE9680 with 2x Intel Xeon Platinum 8480+ Processors, 8x AMD Instinct MI325X (256GiB, 1000W) GPUs, Ubuntu 22.04, and a pre-release build of ROCm 6.3 vs Nvidia published results for TensorRT-LLM v0.13 were captured from: https://github.com/NVIDIA/TensorRT-LLM/blob/v0.13.0/docs/source/performance/perf-overview.md - 3039.7 output tokens/s. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. MI325-014

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. MI300-62

- 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: 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 public 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 2.2, 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. MI300-63

- 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. MI300-53

ENDNOTES

- MI325-002 -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/. MI325-002
- MI300-55: Inference performance projections as of May 31, 2024 using engineering estimates based on the design of a future AMD CDNA 4-based Instinct MI350 Series accelerator as proxy for projected AMD CDNA™ 4 performance. A 1.8T GPT MoE model was evaluated assuming a token-to-token latency = 70ms real time, first token latency = 5s, input sequence length = 8k, output sequence length = 256, assuming a 4x 8-mode MI350 series proxy (CDNA4) vs. 8x MI300X per GPU performance comparison. Actual performance will vary based on factors including but not limited to final specifications of production silicon, system configuration and inference model and size used.
- MI325-001A: Calculations conducted by AMD Performance Labs as of September 26th, 2024, based on current specifications and /or estimation. The AMD Instinct™ MI325X OAMaccelerator will have 256GB HBM3e memory capacity and 6 TB/s GPU peak theoretical memory bandwidth performance. Actual results based on production silicon may vary. 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 GPU memory 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 . MI325-001A
- 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 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. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. MI325-004
- MI355-003. Calculations conducted by internal AMD Performance Labs as of September 26, 2024 on current specifications and/or internal engineering calculations. Comparison of the AMD Instinct™ MI355X platform (2.3 TB HBM3e) vs AMD Instinct™ MI300X platform (1.5 TB HBM3) on Large Language Model (LLM) to determine the maximum model size supported Max Model Size Supported on a Platform

  FP16 Datatype
  MI300X Platform: 1.5 TB HBM3 (in FP16 datatype)
  [(Max Model Size in Billion Parameters) * 2.1 = 1.5 TB*1000
  Max Model Size = ~715 billion Parameters
  FP4 Datatype
  MI355X Platform: 2.3 TB HBM3 (in FP4 datatype)
  [(Max Model Size in Billion Parameters) * 0.55 = 2.3 TB*1000
  Max Model Size = ~4181 billion parameters/4.2 trillion parameters
  Assumptions:
  Batch size 1
  Memory needs for model = 2 Bytes per Parameter
  Memory size needs for activations and others = +10%

Actual maximum LLM parameter size that can run on each platform may vary upon performance testing with physical servers. Calculations rely on published and sometimes preliminary model memory sizes. Model size results estimated on MI355X, MI325X and MI300X platforms due to system/part availability. Actual performance will vary based on final specifications and system configuration. MI355-003

- MI355-004: Calculations conducted by AMD Performance Labs as of September 26th, 2024 for the AMD Instinct™ MI300X GPU platform and accelerator and AMD Instinct™ MI300X GPU platform and accelerator performance comparing FP16 and FP4 datatypes.

  MI355X 8xGPU Platform
  Peak theoretical Four-bit Precision (FP4) Performance - 74 PFLOPs

  MI300X 8xGPU Platform
  Peak theoretical Half Precision (FP16) Performance - 10.4 PFLOPs

Actual performance will vary based on final specifications and system configuration. MI355-004

MI325-015: Based on testing completed on 10/08/2024 by AMD performance lab measuring overall latency for text generated using LLaMA 3.1-405B model using FP8 datatype. Test was performed using input length of 2048 tokens and an output length of 2048 tokens with a batch size of 32 for the following configurations of AMD Instinct™ MI325X 8xGPU platform and NVIDIA H200 HGX GPU platform. Configurations: AMD Instinct™ MI325X platform: Dell PowerEdge XE9680 with 2x Intel Xeon Platinum 8480+ Processors, 8x AMD Instinct MI325X (256GiB, 1000W) GPUs, Ubuntu 22.04, and a pre-release build of ROCm 6.3 NVIDIA H200 HGX platform: Supermicro SuperServer with 2x Intel Xeon Platinum 8468 Processors, 8x Nvidia H200 (140GB, 700W) GPUs, Ubuntu 22.04) 8x MI325X platform with vLLM performance: 23.033 seconds (latency in seconds) Vs. 8x H200 DGX platform with TensorRT-LLM: 27.743 seconds (latency in seconds) Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. MI325-015

AMD
together we advance_