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 including the AMD InstinctTM MI350 Series, AMD ROCm 7 and the associated products with AMD “Helios” and Next-Gen AI Racks; projected data center AI accelerator TAM in 2028; AMD’s AI strategy; and AMD’s ability to accelerate momentum across its AI capabilities, 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.

驅動開源運算革命
重塑AI運算版圖

AMD 台灣區商用市場業務處
資深業務副總經理

林建誠 Ken Lin

AI Innovation is Accelerating

Training is Evolving

+

Inference Scaling Accelerates

+

Explosion of Models

+

Reasoning & Agents Surge

AMD

AI Strategy

Leadership Compute Engines

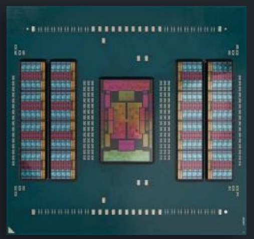
Open Ecosystem

Full Stack Solutions

Best End-to-End AI Compute Portfolio in the Industry

AMD EPYC™

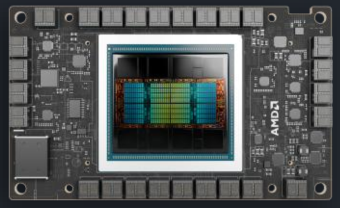
Processors



Leading server CPU

AMD Instinct™

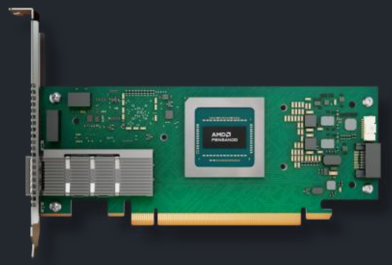
Accelerators



World's best GPU accelerator

AMD Pensando™

Networking



Premier programmable DPUs & AI NICs

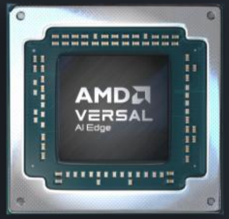
AMD Ryzen™ AI AMD Radeon™ AI

Processors

Most powerful client AI processors

AMD Versal™

Adaptive SOCs



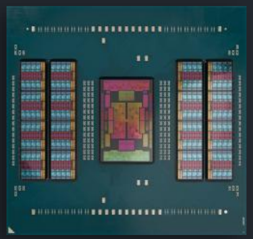
Leadership AI Processing at the edge

See endnote: SHO-000

Best End-to-End AI Compute Portfolio in the Industry

AMD EPYC™

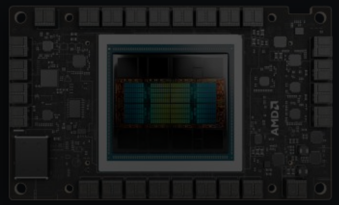
Processors



Leading server CPU

AMD Instinct™

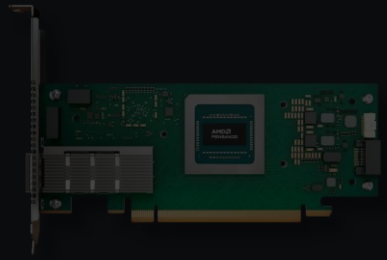
Accelerators



World's best GPU accelerator

AMD Pensando™

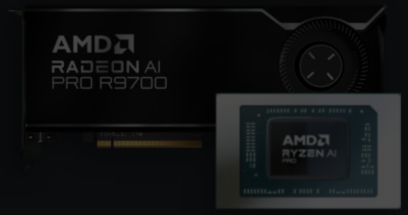
Networking



Premier programmable
DPUs & AI NICs

AMD Ryzen™ AI
AMD Radeon™ AI

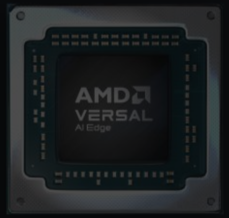
Processors



Most powerful client
AI processors

AMD Versal™

Adaptive SOCs

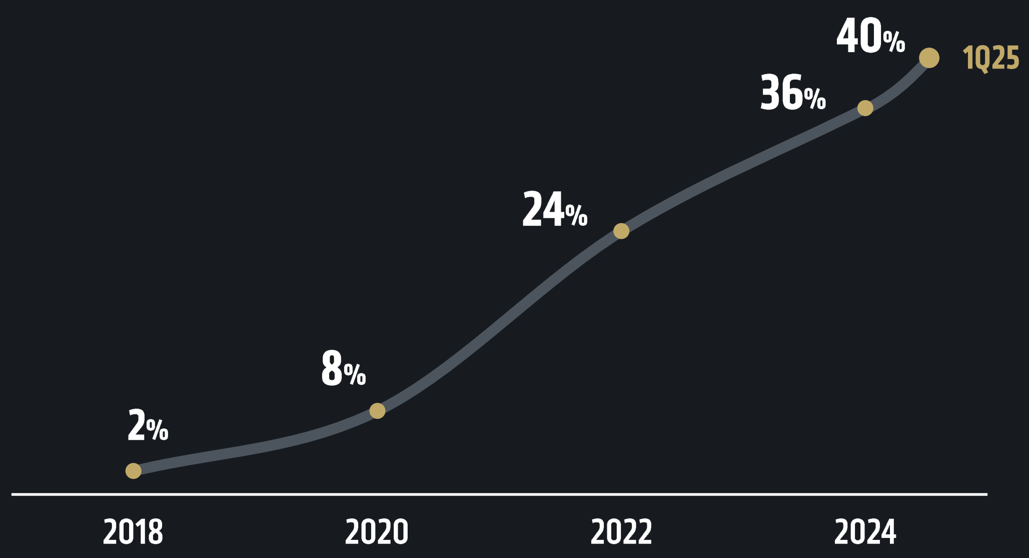


Leadership AI
Processing at the edge

See endnote: SHO-000

EPYC Momentum Accelerates…

>18x Server CPU Market Share Growth

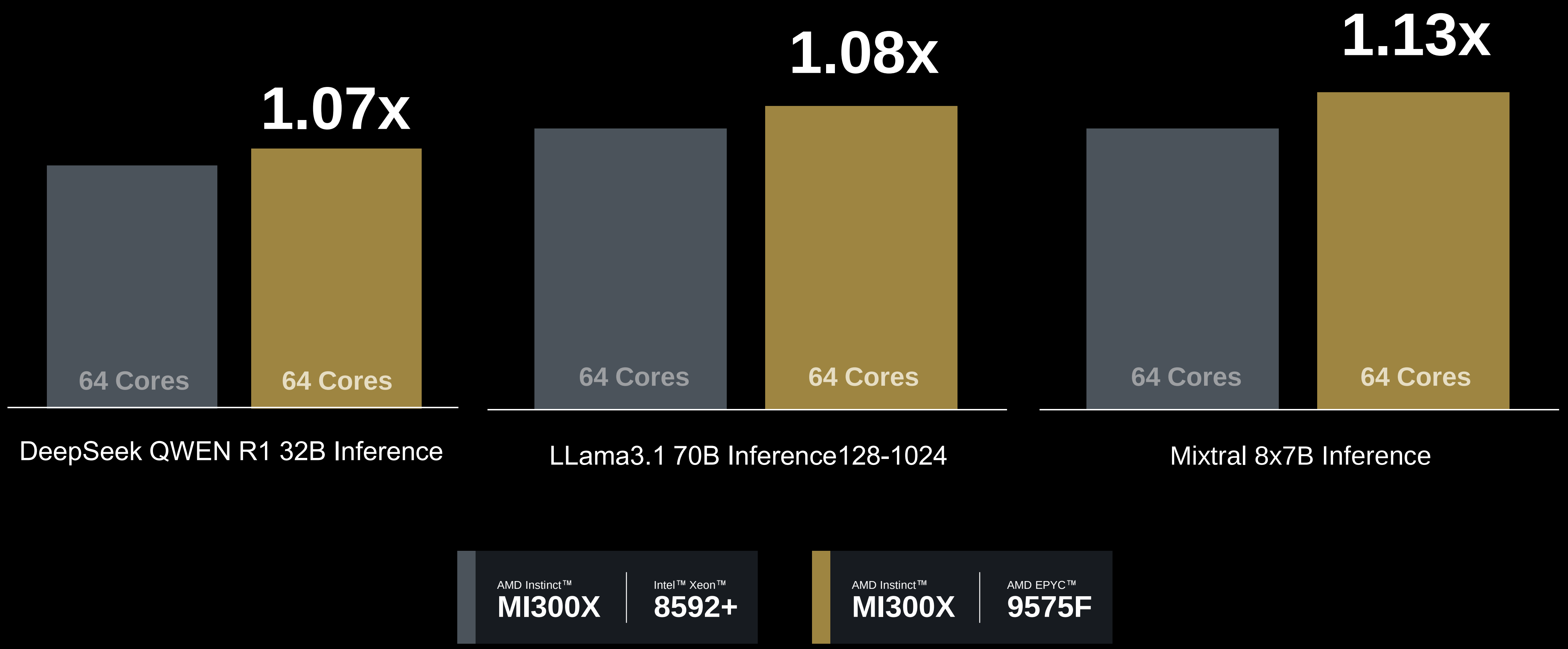


Industry Leaders Run on EPYC™

Cloud	aws	Microsoft	Google	ORACLE	
Digital	NETFLIX	Uber	Meta	zoom	
Enterprise	BEST BUY	IBM	Emirates NBD	NISSAN	
OEM	DELL Technologies	Hewlett Packard Enterprise	Lenovo	SUPERMICRO	cisco

Source: Mercury

AMD EPYC™ CPU Advantage for End-to-End System Performance

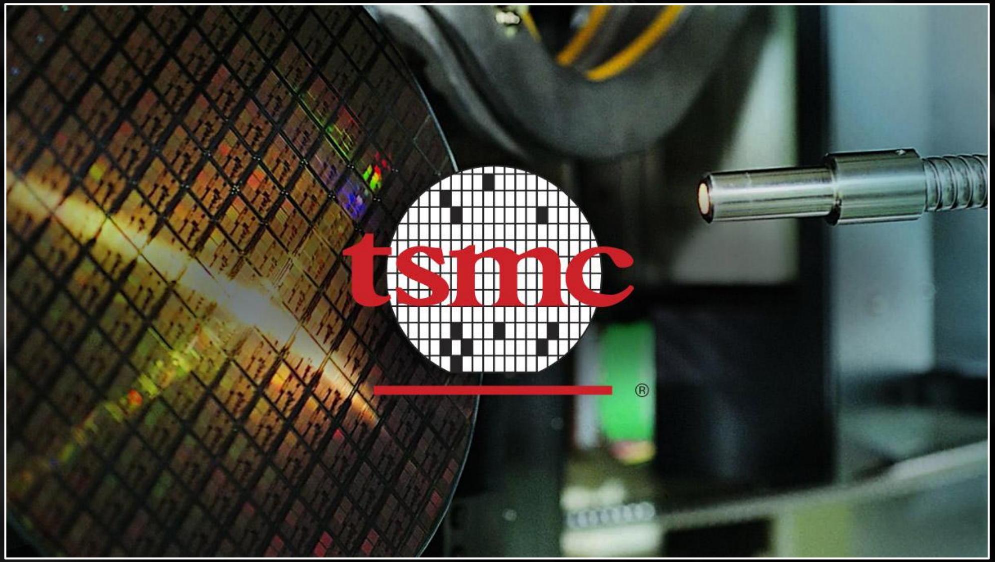


See Endnote 9xx5-168; Source: http://www.amd.com/content/dam/amd/en/documents/epyc-business-docs/white-papers/maximize-ai-gpu-efficiency-with-amd-epyc-processors.pdf

TSMC ENABLES SEMICONDUCTOR FABRICATION EXPANSION WITH AMD

“By deploying 4th Gen AMD EPYC CPUs, we could buy fewer servers while increasing the computing performance by 30 to 40 percent."

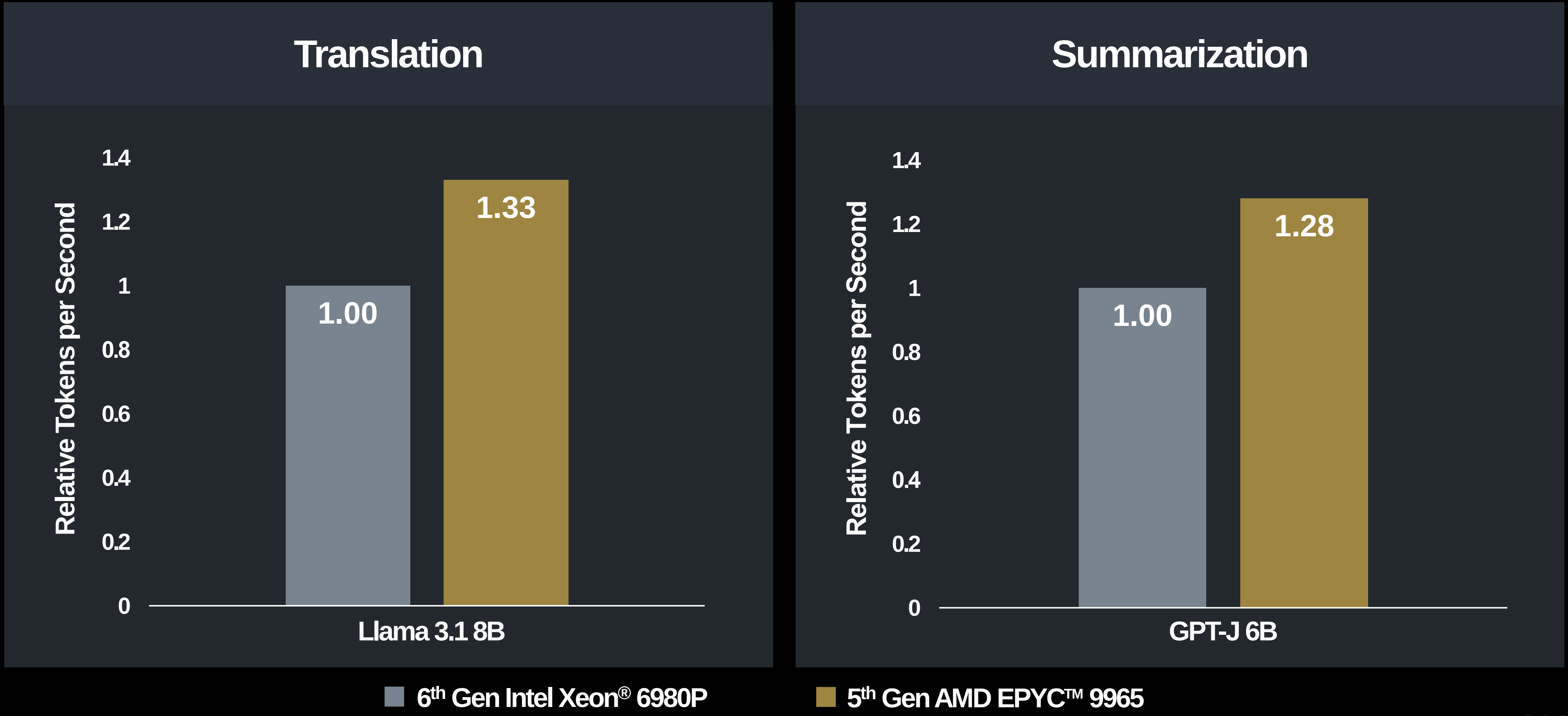
"With the most advanced TSMC fab, more than 90 percent of the workload now runs on 4th Gen AMD EPYC CPUs. We now have close to 20,000 servers across our three workload areas, with 6,600 already powered by 4th Gen AMD EPYC CPUs.”



Simon Wang, Director of Infrastructure and Communication Services Division at TSMC

Read more at: https://www.amd.com/en/resources/case-studies/tsmc.html

AMD EPYC™ CPU Advantage for Generative Workloads



As of 04/08/2025. 32C instances, BS 32, BF16. Results may vary. See endnotes 9xx5-156, 9xx5-158

AMD
together we advance_

CHT ITG DELIVERS SUSTAINABLE PERFORMANCE CLOUD WITH AMD

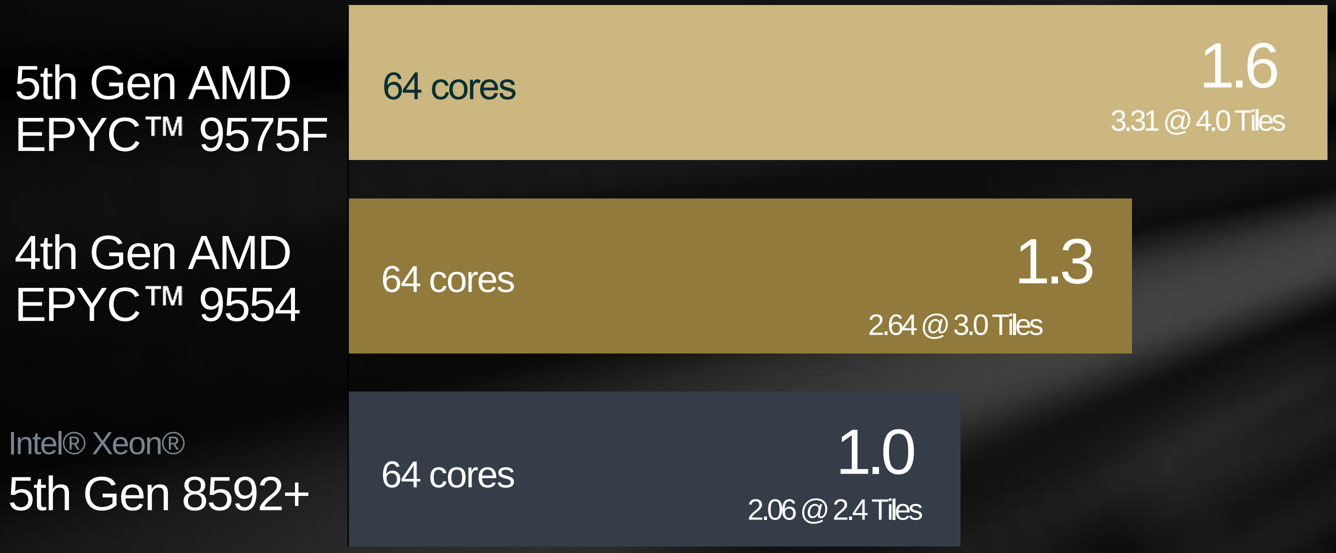
"The high performance of AMD EPYC™ processors allows the company to use fewer servers to provide higher computing capability, significantly optimizing hardware costs."

"We'll continue to work with AMD because the high-core count CPUs will be an ideal solution for training and inference for future AI applications."

Chung-Shuo Lin, Advisor, CHT ITG

Read more at: https://www.amd.com/en/resources/case-studies/chunghwa.html

HIGH FREQUENCY: 60% MORE PERFORMANCE THAN XEON AT THE SAME LICENSING COST



Virtualized Infrastructure

VMmark® 4.0

up to 1.6x

Performance per core in virtualized infrastructure

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

PHISON ELECTRONICS ACCELERATES EDA AND SSD VALIDATION WITH AMD EPYC™ PROCESSORS

"The message from AMD was compelling: the performance of **one AMD9004 Genoa CPU equals that of two CPUs from other brands.**"

Louis Liao, Product Manager, Phison

"Our testing showed that we could complete critical tasks **21% faster on Cadence and 18% faster on Synopsys** thanks to the AMD EPYC™ processors."

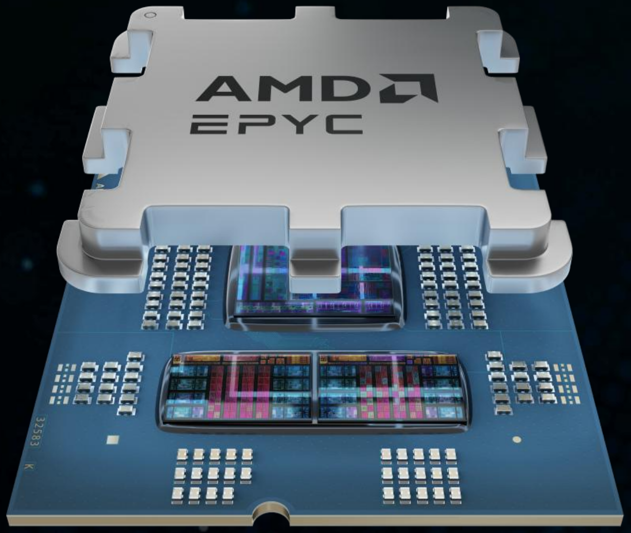
Mobe Chang, IT Manager, Phison

Read more at: https://www.amd.com/en/resources/case-studies/phison-electronics-corporation.html

5th Gen AMD EPYC™ Processors

AMD EPYC™ 4005 Series

A trusted choice for small business and hosted services embracing the AI era



"Zen 5" cores

up to 16 cores

2 ch DDR5 ECC

Up to 28 lanes of PCIe® Gen 5*

Advanced security features with AMD Infinity Guard

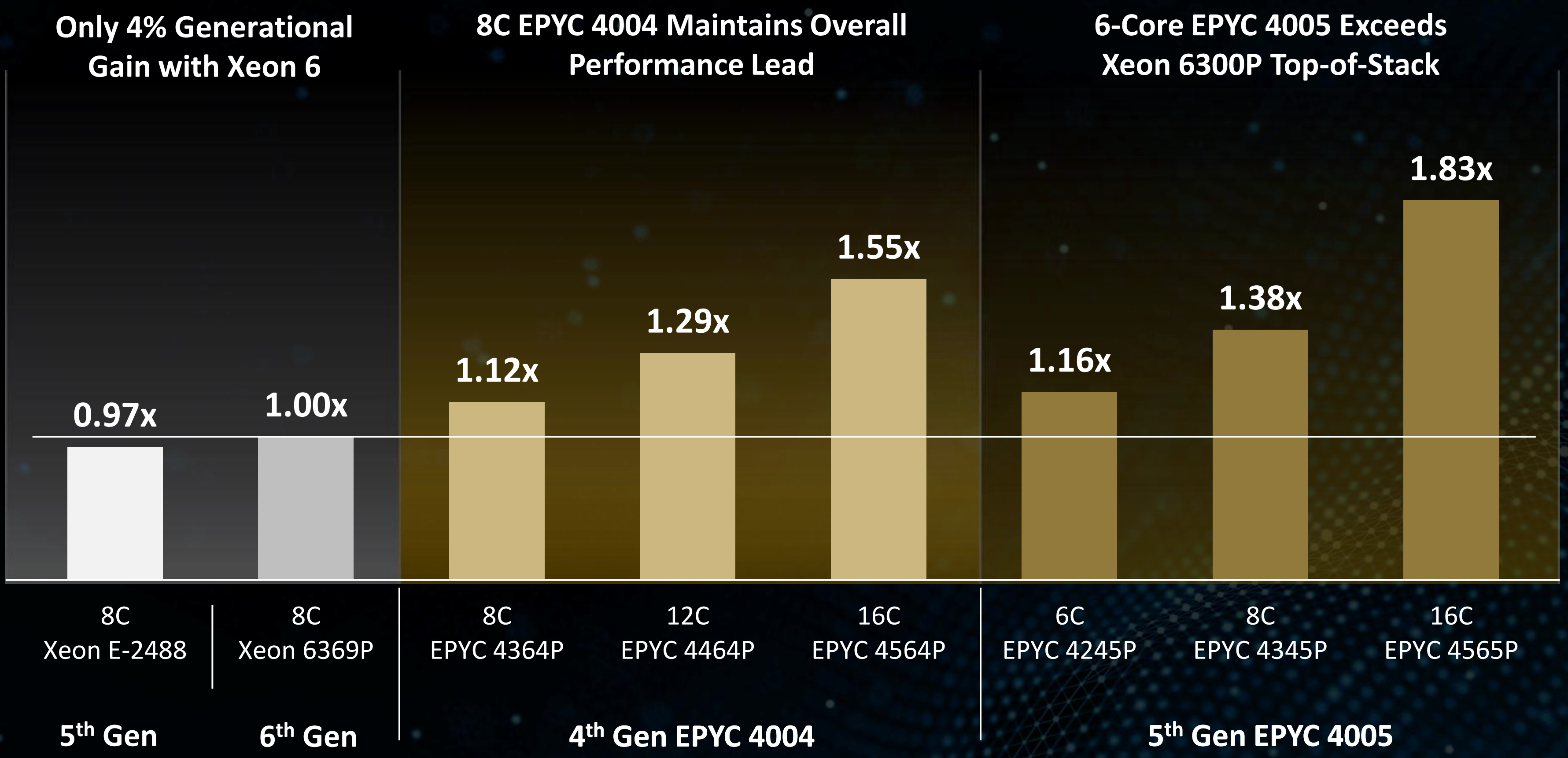
* Additional I/O available via optional chipset.

AMD Infinity Guard features vary by EPYC™ Processor series and/or generations. Infinity Guard security features must be enabled by server OEMs and/or Cloud Service Providers to operate. Check with your OEM or provider to confirm support of these features. Learn more about Infinity Guard at https://www.amd.com/en/technologies/infinity-guard. GD-183A

Leadership Performance

Phoronix Test Suite

(Geometric Mean of All Tests)

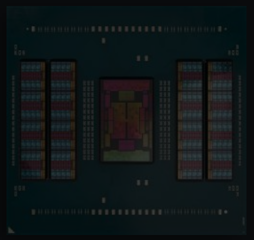


Results may vary based on system configurations, software versions and BIOS settings.
Based on Phoronix Test Suite paid testing as of 04/01/2025. Results not independently verified by AMD. See endnote E4K-021.

Best End-to-End AI Compute Portfolio in the Industry

AMD EPYC™

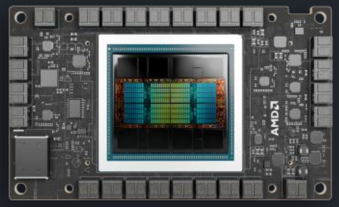
Processors



Leading server CPU

AMD Instinct™

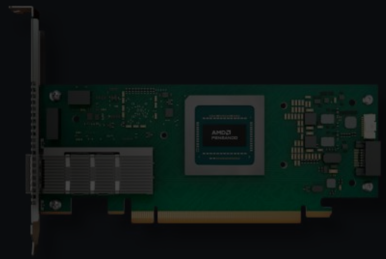
Accelerators



World's best GPU accelerator

AMD Pensando™

Networking



Premier programmable DPUs & AI NICs

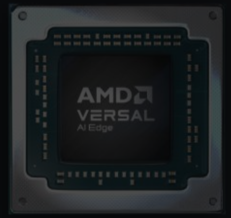
AMD Ryzen™ AI AMD Radeon™ AI

Processors

Most powerful client AI processors

AMD Versal™

Adaptive SOCs



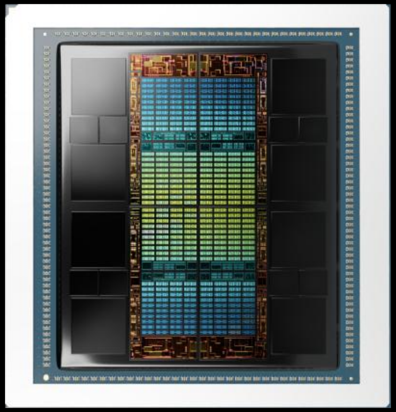
Leadership AI Processing at the edge

See endnote: SHO-000

AMD INSTINCT

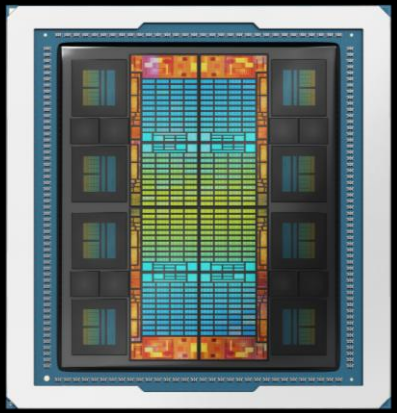
Delivering on Annual Roadmap Commitment

AMD Instinct™
MI300A/X



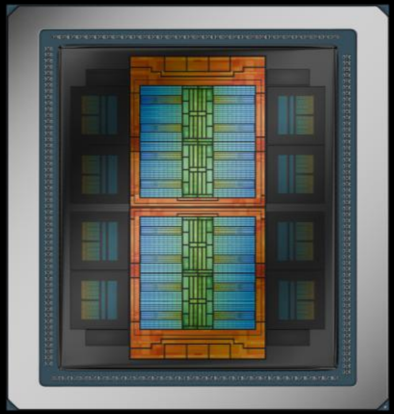
2023

AMD Instinct™
MI325X



2024

AMD Instinct™
MI350 SERIES



2025

AMD Instinct™
MI400 SERIES

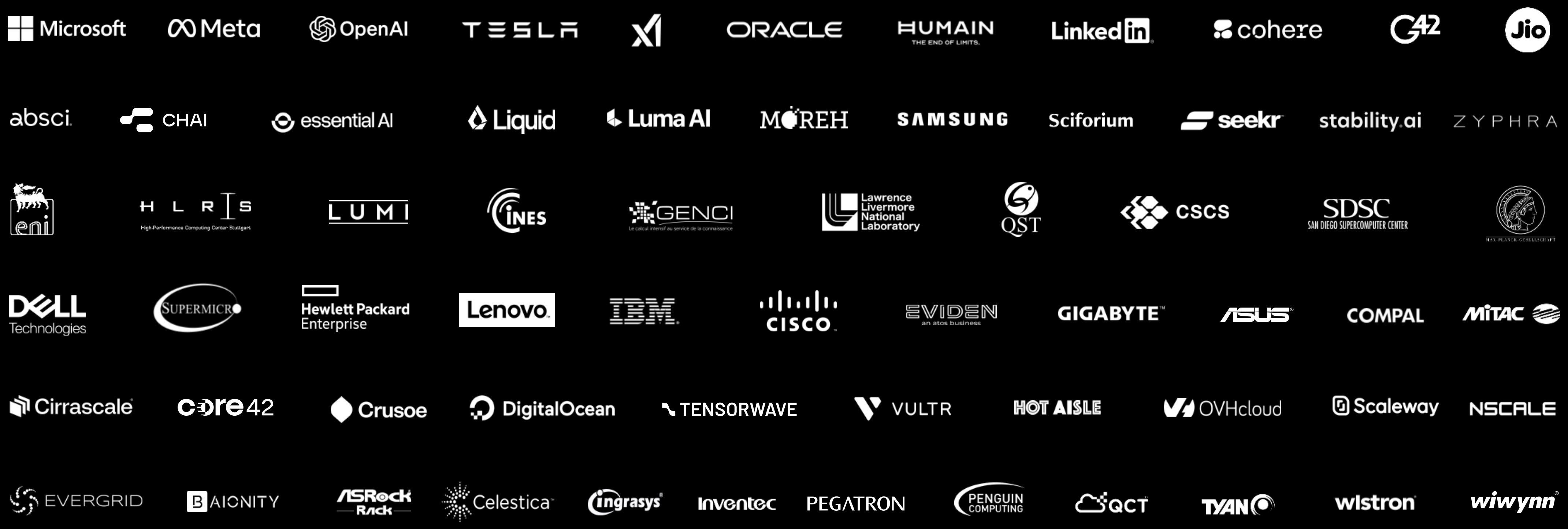
2026

Roadmap subject to change

AMD INSTINCT

Growing Industry Adoption

7 of 10 Largest AI Companies Use AMD Instinct



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.

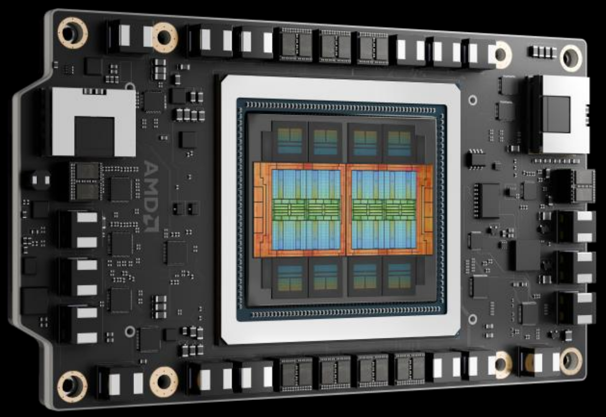
In addition to Llama Inference, Meta's AI Recommendation Inference & Training models run on AMD MI300X GPUs

AMD × Meta

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™ MI350 Series GPUs

AMD Instinct
MI350X GPU



AMD Instinct
MI355X GPU

20 PF | 10 PF
FP4 | FP8
Flops

288 GB
HBM3E Capacity

8 TB /s
Memory Bandwidth

UBB8 Design
in Air Cooled or Liquid Cooled

Leadership Performance | Cost Efficient | Fully Open-Source

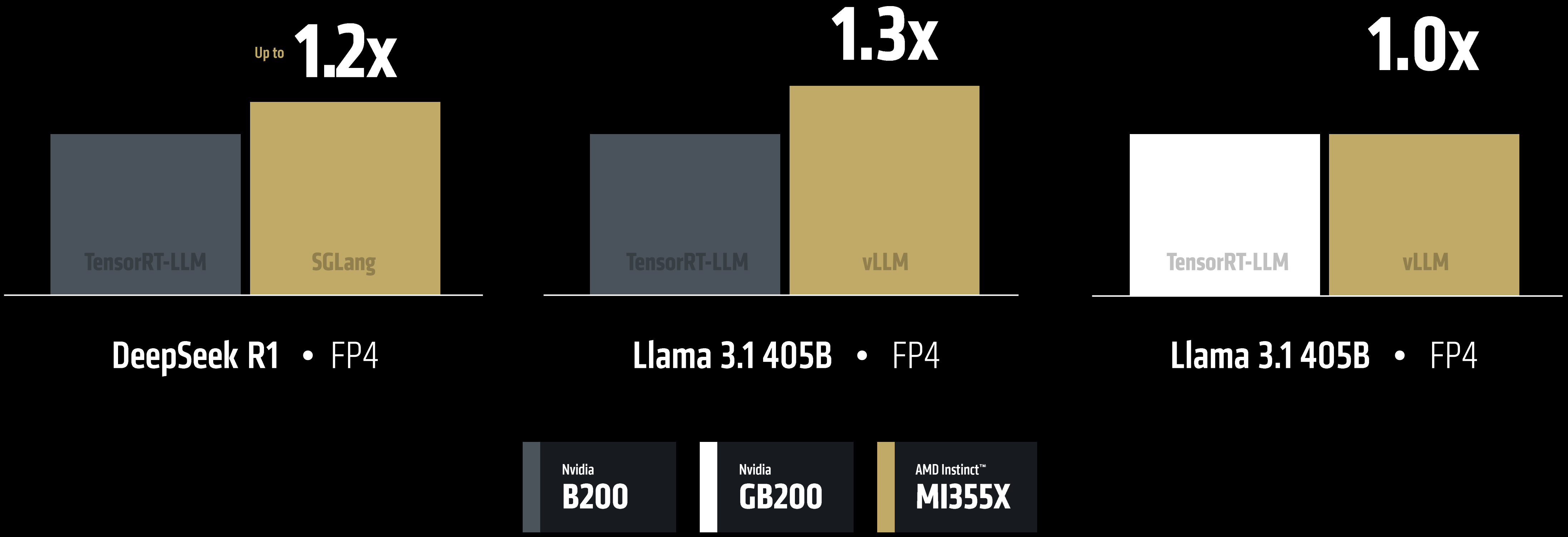
Instinct™ MI350 Series Advantage

	vs. GB200	vs. B200
MEMORY	1.6x	1.6x
MEMORY BANDWIDTH	1.0x	1.0x
FP64	2.0x	2.1x
FP16	1.0x	1.1x
FP8	1.0x	1.1x
FP6	2.0x	2.2x
FP4	1.0x	1.1x

See endnote: MI350-008, 009

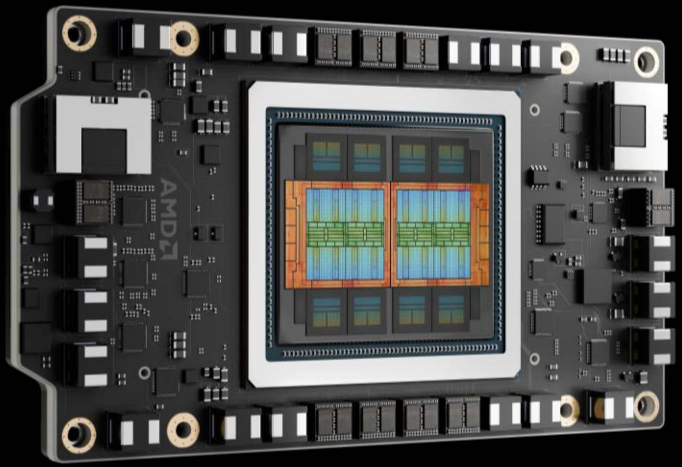
MI355X Delivers the Highest Inference Throughput

For Large Models



Inference performance, throughput

See endnote: MI350-038, 039, 040



Up to 40% More Tokens / $

Using AMD Instinct™ MI355X vs. B200

See endnote: MI350-049

AMD ROCm

Deepening Ecosystem Collaboration

Pytorch

Day 0 support
daily performance CI

Triton
v3.3

Performance focus

Hugging Face
1.8 million models

Nightly CI/CD,
finetuning support

vLLM v1
SGL
llm-d

Serving leadership
Distributed
inference

LLaMA 4 Meta
Gemma 3
deepseek
QwQ-32B
Command R+
Grok
MISTRAL AI_

Support for
SOTA models

ONNX
deepspeed
TensorFlow
OpenXLA
MLIR

Expanding open-source footprint

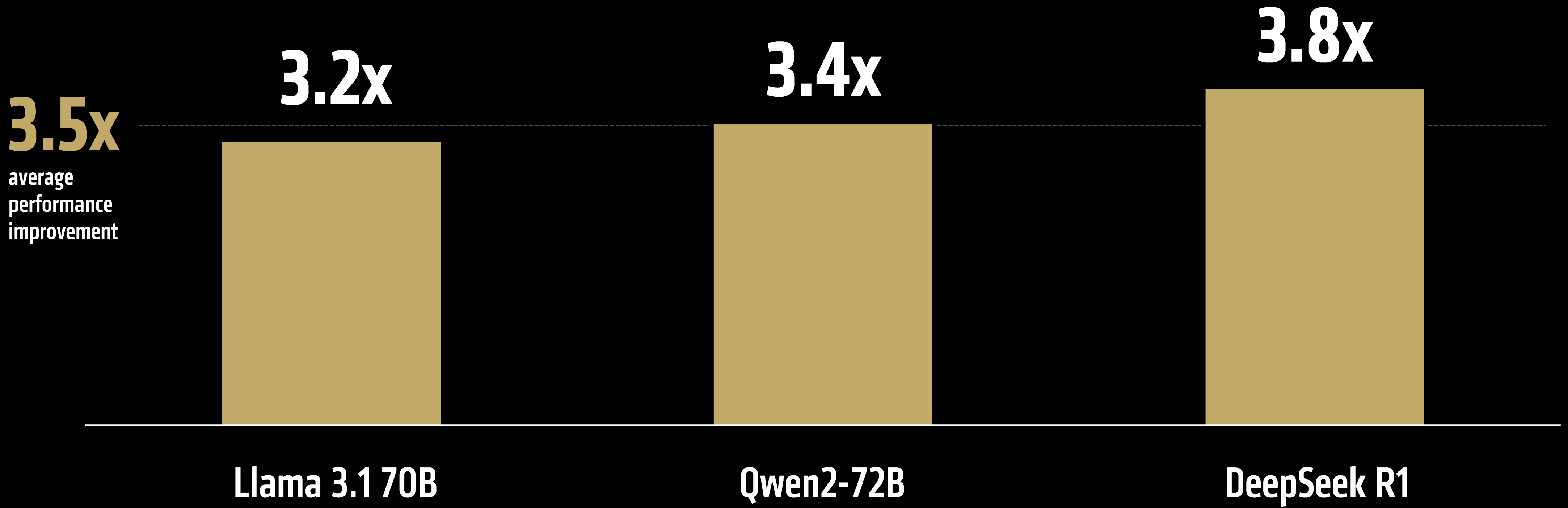
Use or mention of third-party marks, logos, products, services, or solutions herein is for informational purposes only and no endorsement by AMD is intended or implied. GD-83.

Introducing AMD ROCm™ 7

Accelerating AI Innovation & Developer Productivity

Latest Algorithms & Models	Advanced Features for Scaling AI	MI350 Series Support	Cluster Management	Enterprise Capabilities

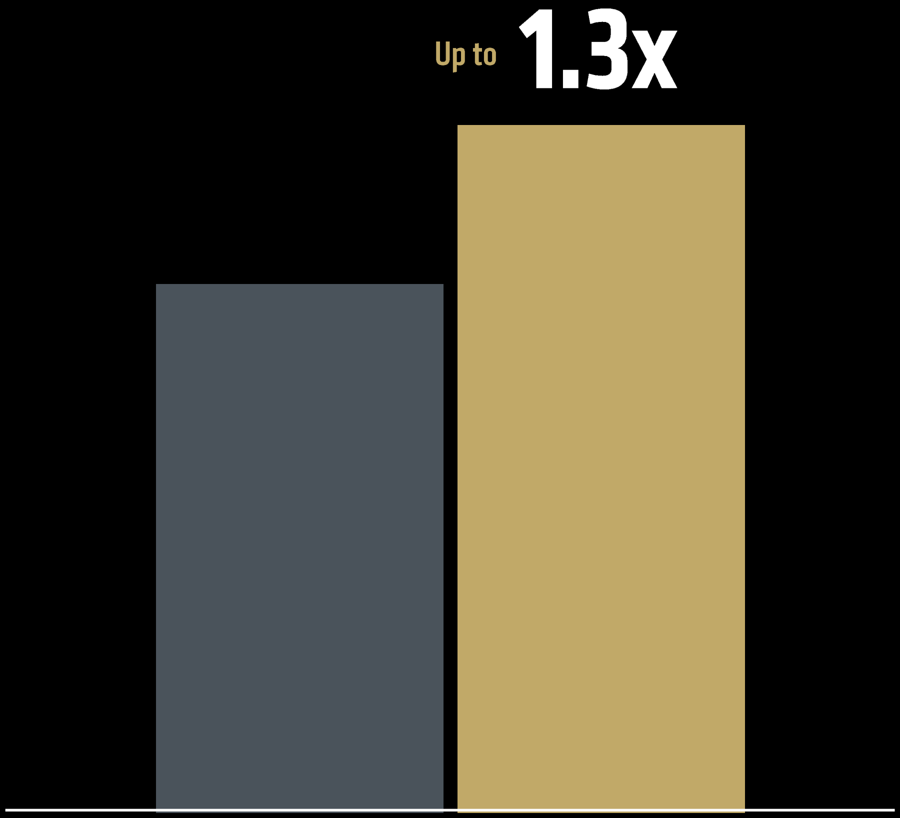
Accelerating Inference Performance



ROCm 7 vs. ROCm 6

See endnote: MI300-080

Open Source: Feature Velocity & Leadership Performance



DeepSeek R1

FP8 Throughput

FP8 Model Support					
vLLM	✓	SGL	✓	TRT-LLM	☒

Nvidia
B200

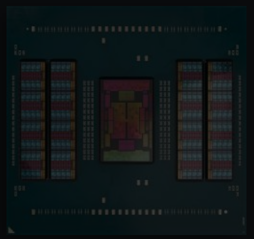
AMD Instinct
MI355X

See endnote: MI350-025

AMD

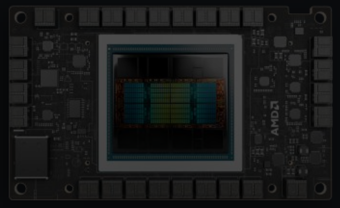
Best End-to-End AI Compute Portfolio in the Industry

AMD EPYC™
Processors



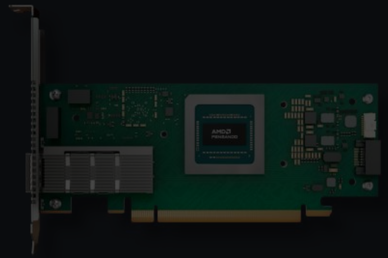
Leading server CPU

AMD Instinct™
Accelerators



World's best GPU accelerator

AMD Pensando™
Networking

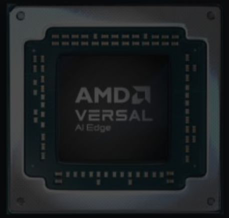


Premier programmable
DPUs & AI NICs

AMD Ryzen™ AI
AMD Radeon™ AI
Processors

Most powerful client
AI processors

AMD Versal™
Adaptive SOCs



Leadership AI
Processing at the edge

See endnote: SHO-000

The Ultimate Client AI Solutions for Every Need

AMD Ryzen™ AI 300

Up to **24B parameters**

AMD Ryzen™ AI Max

Up to **70B parameters**

AMD Threadripper™ + Radeon™ AI

Up to **128B parameters**

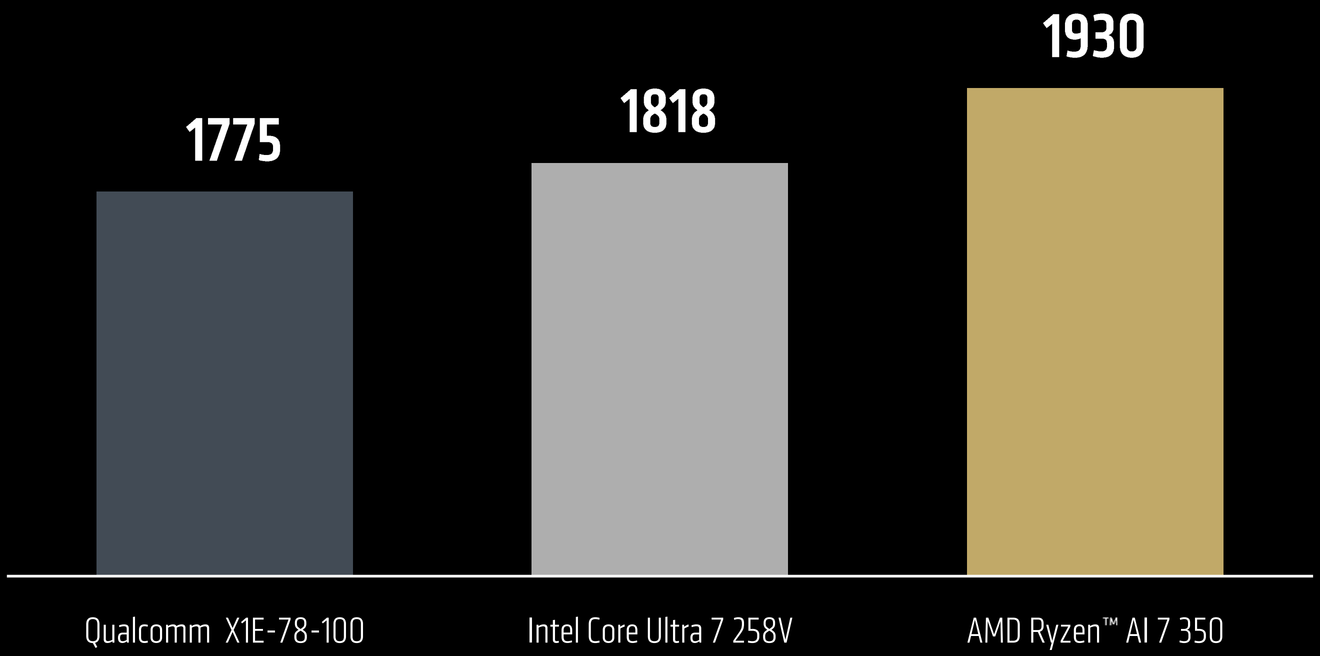
Leadership AI Performance

DeepSeek Performance:
Tokens Per Second

Up to	**Up to**
60%	**80%**
Faster	Faster
DeepSeek R1 Distill Qwen 7b	DeepSeek R1 Distill Qwen 1.5b

AMD Ryzen™ AI 9 HX 370 Processor vs similar system
w/ Intel Core Ultra 7 258v processor w/vPro

NPU AI Performance (Procyon)

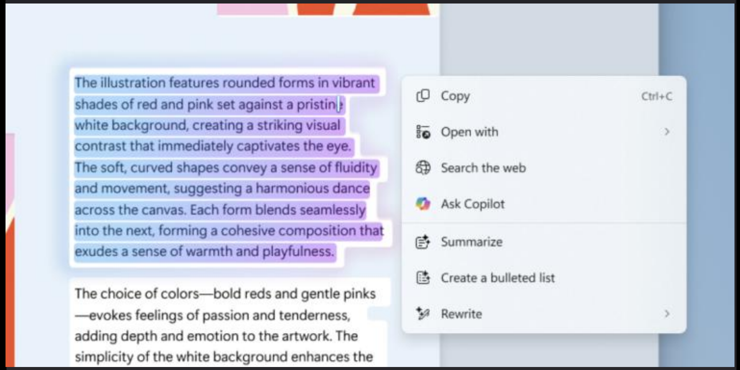


See endnote: STX-112, STX-113

AMD
together we advance_

Copilot+PC

Click to Do

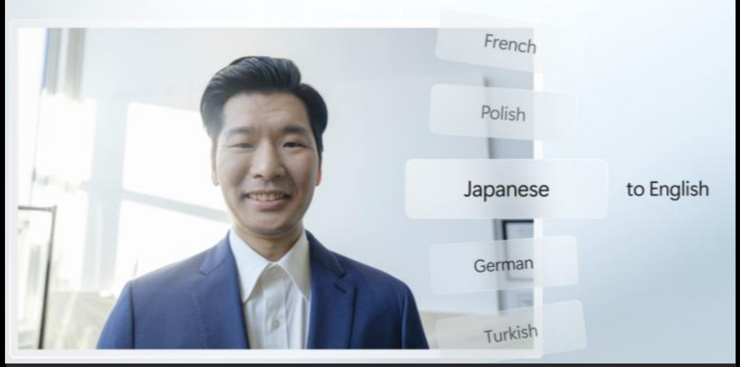


The illustration features rounded forms in vibrant shades of red and pink set against a pristine white background, creating a striking visual contrast that immediately captivates the eye. The soft, curved shapes convey a sense of fluidity and movement, suggesting a harmonious dance across the canvas. Each form blends seamlessly into the next, forming a cohesive composition that exudes a sense of warmth and playfulness.

The choice of colors—bold reds and gentle pinks—evokes feelings of passion and tenderness, adding depth and emotion to the artwork. The simplicity of the white background enhances the

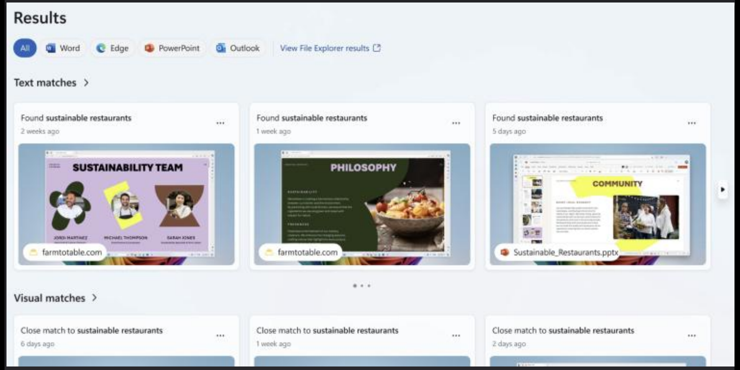
Seamless integration of AI capabilities within existing Windows applications

Live Captions



Reducing language barriers across worldwide organizations

Enhanced Search



Results
All
Word
Edge
PowerPoint
Outlook
View File Explorer results
Text matches
Found sustainable restaurants
2 weeks ago
SUSTAINABILITY TEAM
farmtotable.com
Found sustainable restaurants
1 week ago
PHILOSOPHY
farmtotable.com
Found sustainable restaurants
5 days ago
COMMUNITY
Sustainable_Restaurants.pptx
Visual matches
Close match to sustainable restaurants
6 days ago
Close match to sustainable restaurants
1 week ago
Close match to sustainable restaurants
2 days ago

Get to what you need faster

NPU-Enabled Workflow Improvement

AI PC Momentum With Enterprise ISVs

- BUFFERZONE
- SOLIDWORKS
- zoom
- Microsoft
- Adobe
- MAXON
- blender
- splashtop
- WHISPP
- webex by CISCO
- Rhinoceros
- Camo
- bitdefender secure your every bit
- LM Studio
- grammarly
- Blackmagicdesign
- nero
- BORISFX
- Avid
- GoPro Be a HERO.
- Topaz Labs™
- ARKRUNR
- voicemy.ai
- RADiCAL AI-POWERED 3D ANIMATION
- CyberLink
- AFFINITY Photo 2
- ACCA ACCA SOFTWARE
- convai
- OBS Open Broadcaster Software

AMD
together we advance_

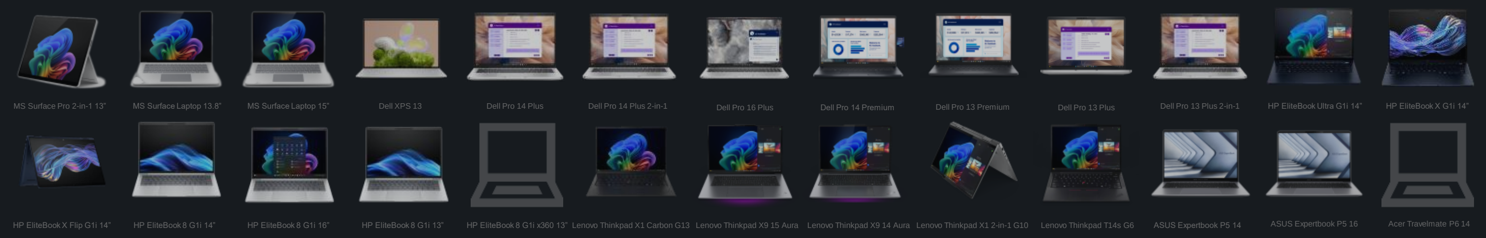
Commercial Copilot+

Platform Availability

AMD

Dell PRO Max 16
Dell PRO Max 14
Dell PRO 13 Plus
Dell PRO 13 Plus 2-in-1
Dell PRO 14 Plus
Dell PRO 14 Plus 2-in-1
Dell PRO 16 Plus
Dell PRO 14
Dell PRO 16
ASUS Expertbook 14
ASUS Expertbook 16
Lenovo Thinkbook 14 G7
Lenovo Thinkbook 16 G7
Lenovo Thinkpad 14s G6
Lenovo Thinkpad 14 G6
Lenovo Thinkpad L14 G6
Lenovo Thinkpad L16 G2
Lenovo Thinkpad T16
Lenovo Thinkpad P14s G6
Lenovo Thinkpad P16s G4
Lenovo Thinkpad X13
HP Elitebook G1a 13”
HP Zbook Ultra G1a
HP Elitebook 8 G1a 13”
HP Elitebook 8 G1a 14”
HP Elitebook 8 G1a 16”
HP Elitebook 6 G1a 14”
HP Elitebook 6 G1a 16”
HP Zbook 8 G1a

INTEL



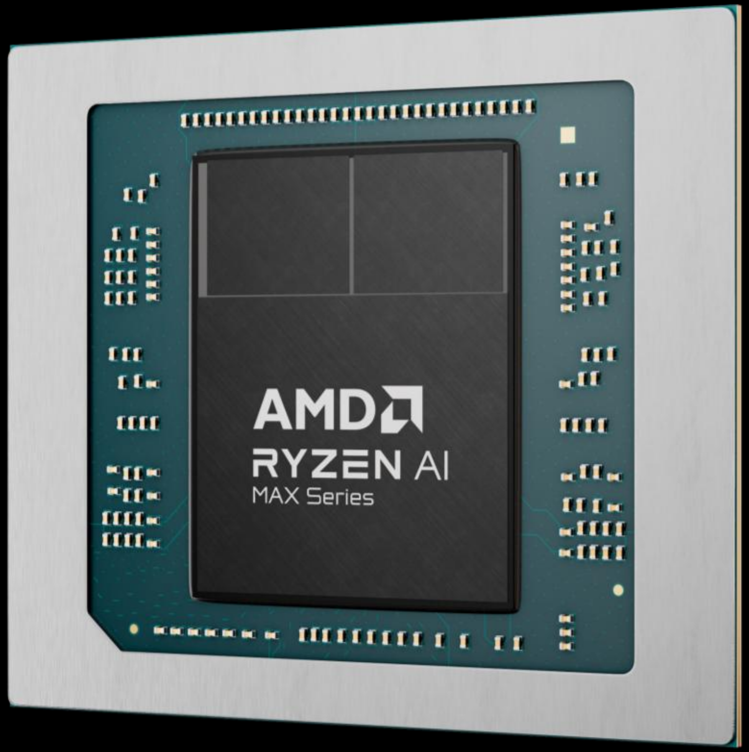
MS Surface Pro 2-in-1 13”
MS Surface Laptop 13.8”
MS Surface Laptop 15”
Dell XPS 13
Dell Pro 14 Plus
Dell Pro 14 Plus 2-in-1
Dell Pro 16 Plus
Dell Pro 14 Premium
Dell Pro 13 Premium
Dell Pro 13 Plus
Dell Pro 13 Plus 2-in-1
HP EliteBook Ultra G1i 14”
HP EliteBook X G1i 14”
HP EliteBook X Flip G1i 14”
HP EliteBook 8 G1i 14”
HP EliteBook 8 G1i 16”
HP EliteBook 8 G1i 13”
HP EliteBook 8 G1i x360 13”
Lenovo Thinkpad X1 Carbon G13
Lenovo Thinkpad X9 15 Aura
Lenovo Thinkpad X9 14 Aura
Lenovo Thinkpad X1 2-in-1 G10
Lenovo Thinkpad T14s G6
ASUS Expertbook P5 14
ASUS Expertbook P5 16
Acer Travelmate P6 14

QUALCOMM

Lenovo ThinkPad T14s G6
Lenovo ThinkBook 16 G7
Dell XPS 13
Dell Latitude 5455
Dell Latitude 7455
HP Elitebook Ultra G1q
HP Probook 4 G1q
HP Elitebook 6 G1q
MS Surface Pro 2-in-1 12”
MS Surface Pro 2-in-1 13”
MS Surface Laptop 13”
MS Surface Laptop 13.8”
MS Surface Laptop 15”

AMD Ryzen™ AI Max Series Processors

Featuring The World's Most Powerful Processor for Next Gen AI PCs



ZEN 5 | AMD RDNA 3.5 | AMD RYZEN AI | Up to 96GB Graphics Memory | Copilot+PC

See endnote SHO-06.

AMD together we advance_

The world's first Copilot+ PC processor to run 70B LLM

Up to **2.2x**

faster AI Performance*

AMD Ryzen™ AI Max+ 395

vs. NVIDIA GeForce RTX 4090 24GB

AMD RYZEN AI MAX Series

LM Studio

*tokens / second

Up to **87%**

lower TDP

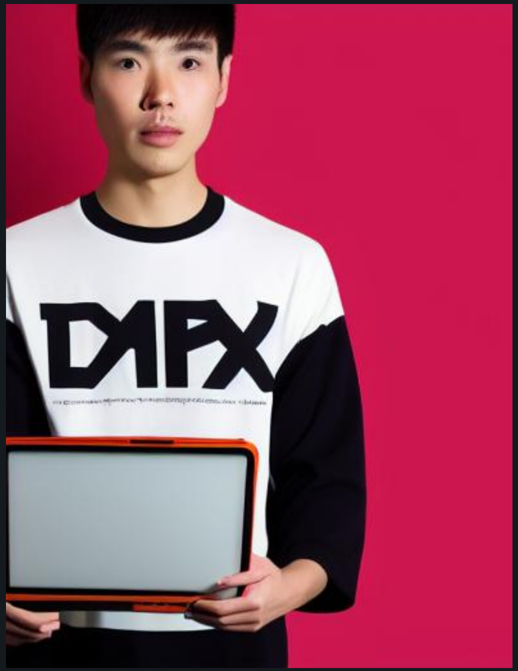
AMD Ryzen™ AI Max+ 395

vs. NVIDIA GeForce RTX 4090 24GB

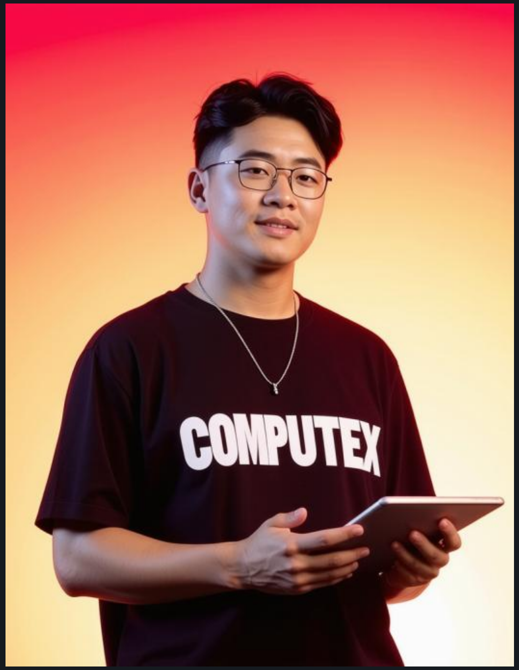
All results are up to. Total TDP of Ryzen AI MAX+ 395 is 55W and tGP of GeForce GTX 4090 is 450W. See endnote: SHO-14
Running Llama 70b in Q4 quants requires 43 GB of graphics memory (shared and/or dedicated)

Unlocking Higher Quality AI Models at the Edge

Text to Image



1B • FP16



8B • FP16

Prompt Wide and low angle shot of Taiwanese male wearing a shirt that says “Computex” while holding a laptop. Background is a gradient of red, pink, and orange.

Reasoning

4/20 CORRECT

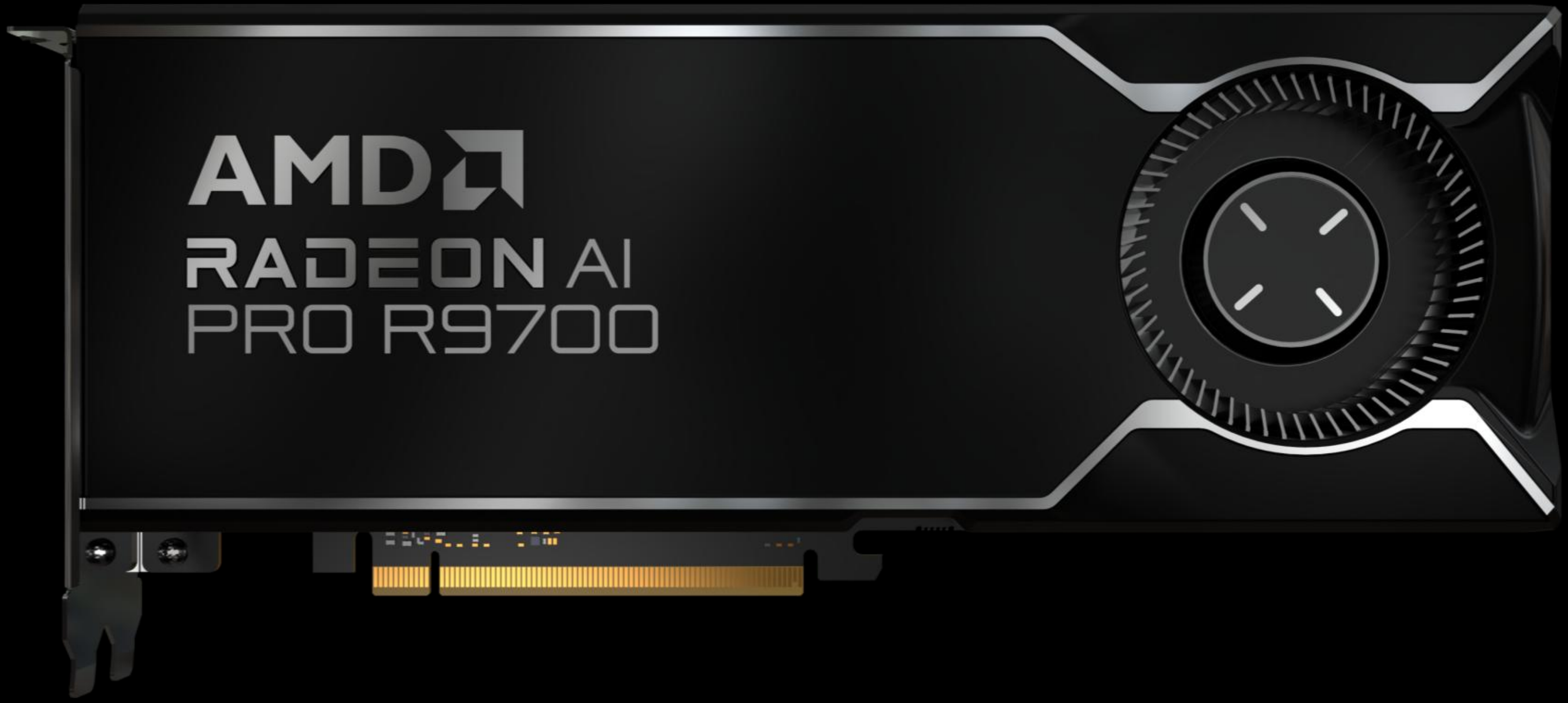
20/20 CORRECT

8B • 6-bit

32B • 6-bit

Prompt How much water can a pickleball hold?

AMD Radeon™ AI PRO R9700 Graphics



128 AI Accelerators

32GB GDDR6

Up to
96 TFLOPS
Peak Half-Precision

Up to
1531 TOPS
INT4 Sparse

300W TDP

See endnote GD-243

Next-Gen Scalability

Multi-GPU PCIe® 5 platform

32GB

4x
Radeon AI PRO R9700

Mistral Large Instruct
123B, GPTQ4
116GB

DeepSeek R1 Distill
Llama 70B, FP8
112GB

*Estimated, incl. typical KV Cache allocation

Investing in Full-Stack Solutions

Acquisitions Span Entire AI Value Chain

zt Systems

nod

Mipsology

brium

SILO AI

LAMINI

ENOSEMI

PENSANDO

XILINX

Over 25 AI Acquisitions & Investments in the Last Year Alone

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.

Open Development Drives Value & Innovation

Open Hardware + Open Software + Open Ecosystem

OPEN Compute Project · ULTRA ACCELERATOR LINK · UltraEthernet Consortium · AMD ROCm · Hugging Face · PyTorch · Triton · vLLM · SGL

Choice | Flexibility | Rapid Co-Innovation | Portability | Proven

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

AI Innovation is a Global, Collective Effort

Endnotes

SHO-06: Testing as of Dec 2024 using the following benchmark scores compared to Intel Core Ultra 9 288V and Qualcomm Snapdragon X Elite X1E-84-100. Cinebench 2024 nT, 3Dmark Wildlife Extreme, and Blender.. Next gen AI PC defined as a Windows PC with a processor that includes a NPU with at least 40 TOPS. Configuration for AMD Ryzen™ AI Max+ 395 processor: AMD reference board, Radeon™ 8060S graphics, 32GB RAM, 1TB SSD, VBS=ON, Windows 11. Configuration for Qualcomm Snapdragon X Elite X1E-84-100 processor: Samsung Galaxybook, Adreno Graphics, 16GB RAM, Microsoft Windows 11. Configuration for Intel Core Ultra 9 288V: ASUS Zenbook X 14, Intel Arc Graphics, 32GB RAM, 1TBSSD, Microsoft Windows 11 Home. Laptop manufacturers may vary configurations yielding different results.

MI300-080: Testing by AMD Performance Labs as of May 15, 2025, measuring the inference performance in tokens per second (TPS) of AMD ROCm 6.x software, vLLM 0.3.3 vs. AMD ROCm 7.0 preview version SW, vLLM 0.8.5 on a system with (8) AMD Instinct MI300X GPUs running Llama 3.1-70B (TP2), Qwen 72B (TP2), and Deepseek-R1 (FP16) models with batch sizes of 1-256 and sequence lengths of 128-204. Stated performance uplift is expressed as the average TPS over the (3) LLMs tested. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of the latest drivers and optimizations.

MI300-081: AMD Instinct MI300X platform (8x GPUs) and AMD ROCm 7.0 preview version software running Llama2-70B, Qwen1.5-14B, Llama3.1-8B, Megatron-LM using the FP16 and FP8 datatypes,shows a combined average of 3.04x or average of 304%) better training performance (TFLOPS) vs. AMD Instinct MI300X platform (8x GPUs) with ROCm 6.0 SW.

MI350-004: Based on calculations by AMD Performance Labs in May 2025, to determine the peak theoretical precision performance of eight (8) AMD Instinct™ MI355X and MI350X GPUs (Platform) and eight (8) AMD Instinct MI325X, MI300X, MI250X and MI100 GPUs (Platform) using the FP16, FP8, FP6 and FP4 datatypes with Matrix. Server manufacturers may vary configurations, yielding different results. Results may vary based on use of the latest drivers and optimizations.

MI350-008: Based on measurements taken by AMD Performance Labs in May 2025, of the peak theoretical precision performance of an AMD Instinct™ MI355X GPU with FP64 datatype with Matrix vs. Nvidia Grace Blackwell GB200 accelerator with FP64 datatype with Tensor; MI355X: FP32 with Matrix vs. GB200: FP32 datatype with Vector; and MI355X: FP6 datatype with Sparsity vs. GB200: FP6 datatype with Sparsity. Results may vary based on configuration, datatype. **MI350-008**

MI350-009: Based on calculations by AMD Performance Labs in May 2025, to determine the peak theoretical precision performance for the AMD Instinct™ MI350X / MI355X GPUs, when comparing FP64, FP32, TF32, FP16, FP8, FP6 and FP4, INT8, and bfloat16 datatypes with Vector, Matrix, Sparsity or Tensor with Sparsity as applicable, vs. NVIDIA Blackwell B200 accelerator. Server manufacturers may vary configurations, yielding different results.

MI350-025: Testing by AMD Performance Labs as of May 25, 2025, measuring the inference performance in tokens per second (TPS) of the AMD Instinct MI355X platform with ROCm 7.0 pre-release build 16047, running DeepSeek R1 LLM on SGLang versus NVIDIA Blackwell B200 platform with CUDA version 12.8. Server manufacturers may vary configurations, yielding different results. Performance may vary based on hardware configuration, software version, and the use of the latest drivers and optimizations.

MI350-030: Based on calculations by AMD internal testing as of 6/4/2025. Using 8 GPU AMD Instinct™ MI355X Platform for overall GPU-normalized Training Throughput (processed tokens per second) for text generation using the Llama3-70B chat model running Torchtitan (FP8) when using a maximum sequence length of 8192 tokens compared to published 64 GPU Nvidia B200 Platform performance running NeMo (FP8) when using a maximum sequence length of 8192 tokens. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

Endnotes

MI350-031: Based on calculations by AMD internal testing as of 6/4/2025. Using 8 GPU AMD Instinct™ MI355X Platform for overall GPU-normalized Training Throughput (processed tokens per second) for text generation using both LLaMA3-70B and LLaMA3-8B chat models running Torchtitan (BF16) or Megatron-LM (BF16) where applicable when using a maximum sequence length of 8192 tokens compared to 8 GPU Nvidia B200 Platform performance running NeMo (BF16) when using a maximum sequence length of 8192 tokens. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-032: Based on calculations by AMD internal testing as of 6/4/2025. Using 8 GPU AMD Instinct™ MI355X Platform for overall GPU-normalized Training Throughput (processed tokens per second) for text generation using both LLaMA3-70B and LLaMA3-8B chat models running Torchtitan (BF16) or Megatron-LM (BF16) where applicable when using a maximum sequence length of 8192 tokens compared to 8 GPU Nvidia B200 Platform performance running NeMo (BF16) when using a maximum sequence length of 8192 tokens. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-033: Based on calculations by AMD internal testing as of 6/5/2025. Using 8 GPU AMD Instinct™ MI355X Platform for overall GPU-normalized Training Throughput (time to complete) for fine-tuning using the Llama2-70B LoRA chat model (FP8) compared to published 8 GPU Nvidia B200 and 8 GPU Nvidia GB200 Platform performance (FP8). Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-034: Based on AMD internal testing as of 6/4/2025, using an (8) GPU AMD Instinct™ MI355X Platform for overall GPU-normalized Training Throughput (processed tokens per second) for text generation using the LLaMA3-70B and LLaMA3-8B chat models running Torchtitan or Megatron-LM (FP8 and BF16) as applicable, using a maximum sequence length of 8192 tokens, compared to an (8) GPU AMD Instinct™ MI300X Platform using Megatron-LM (FP8 and BF16). Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI350-035: Based on AMD internal testing as of 6/5/2025. Using 8 GPU AMD Instinct™ MI355X Platform for overall GPU-normalized Training Throughput (time to complete) for fine-tuning using the Llama2-70B LoRA chat model (FP8) compared 8 GPU AMD Instinct™ MI300X Platform performance with (FP8). Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI350-038: Based on testing by AMD internal labs as of 6/6/2025 measuring text generated throughput for LLaMA 3.1-405B model using FP4 datatype. Test was performed using input length of 128 tokens and an output length of 2048 tokens for AMD Instinct™ MI355X 8xGPU platform compared to NVIDIA B200 HGX 8xGPU platform published results. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-039: Based on Lucid automation framework testing by AMD internal labs as of 6/6/2025 measuring text generated throughput for LLaMA 3.1-405B model using FP4 datatype. Test was performed using 4 different combinations (128/2048) of input/output lengths to achieve a mean score of tokens per second for AMD Instinct™ MI355X 4xGPU platform compared to NVIDIA DGX GB200 4xGPU platform. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI350-040: Based on testing (tokens per second) by AMD internal labs as of 6/6/2025 measuring text generated online serving throughput for DeepSeek-R1 chat model using FP4 datatype. Test was performed using input length of 3200 tokens and an output length of 800 tokens with concurrency up to 64 looks, serviceable with 30ms ITL threshold for AMD Instinct™ MI355X 8xGPU platform median total tokens compared to NVIDIA B200 HGX 8xGPU platform results. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-041: Based on AMD internal testing as of 6/5/2025. Using 8 GPU AMD Instinct™ MI355X Platform measuring text generated offline inference throughput for Llama4 Maverick chat model (FP4) compared 8 GPU AMD Instinct™ MI300X Platform performance with (FP8). MI355X ran 8xTP1 (8 copies of model on 1 GPU) compared to MI300X running 2xTP4 (2 copies of model on 4 GPUs). Tests were conducted using a synthetic dataset with different combinations of 128 and 2048 input tokens, and 128 and 2048 output tokens. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

Endnotes

MI350-042: Based on AMD internal testing as of 6/5/2025. Using 8 GPU AMD Instinct™ MI355X Platform measuring text generated offline inference throughput for Llama 3.1-405B chat model (FP4) compared 8 GPU AMD Instinct™ MI300X Platform performance with (FP8). MI355X ran 8xTP1 (8 copies of model on 1 GPU) compared to MI300X running 2xTP4 (2 copies of model on 4 GPUs). Tests were conducted using a synthetic dataset with different combinations of 128 and 2048 input tokens, and 128 and 2048 output tokens. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-043: Based on AMD internal testing as of 6/5/2025. Using 8 GPU AMD Instinct™ MI355X Platform measuring text generated online serving inference throughput for DeepSeek-R1 chat model (FP4) compared 8 GPU AMD Instinct™ MI300X Platform performance with (FP8). Test was performed using input length of 3200 tokens and an output length of 800 tokens with concurrency set to maximize the throughput on each platform, 128 for MI300X and 2048 for MI355X platforms. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI350-044: Based on AMD internal testing as of 6/9/2025. Using 8 GPU AMD Instinct™ MI355X Platform measuring text generated online serving inference throughput for Llama 3.1-405B chat model (FP4) compared 8 GPU AMD Instinct™ MI300X Platform performance with (FP8). Test was performed using input length of 32768 tokens and an output length of 1024 tokens with concurrency set to best available throughput to achieve 60ms on each platform, 1 for MI300X (35.3ms) and 64ms for MI355X platforms (50.6ms). Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.Based on AMD internal testing as of 6/9/2025. Using 8 GPU AMD Instinct™ MI355X Platform measuring text generated online serving inference throughput for Llama 3.1-405B chat model (FP4) compared 8 GPU AMD Instinct™ MI300X Platform performance with (FP8). Test was performed using input length of 32768 tokens and an output length of 1024 tokens with concurrency set to best available throughput to achieve 60ms on each platform, 1 for MI300X (35.3ms) and 64ms for MI355X platforms (50.6ms). Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI350-047: Based on engineering projections by AMD Performance Labs in June 2025, to estimate the peak theoretical precision performance of seventy-two (72) AMD Instinct™ MI400X GPUs (Rack) vs. an 8xGPU AMD Instinct MI355X platform using the FP6 Matrix datatype. Results subject to change when products are released in market.

MI350-048: Based on AMD internal testing as of 6/9/2025. Using 8 GPU AMD Instinct™ MI355X Platform measuring text generated offline inference throughput for Llama 3.3-70B chat model (FP4) compared 8 GPU AMD Instinct™ MI300X Platform performance with (FP8). MI355X ran 8xTP1 (8 copies of model, one per GPU) compared to MI300X running 8xTP1 (8 copies of model, one per GPU). Tests were conducted using a synthetic dataset with different combinations of 128 and 2048 input tokens, and 128 output tokens. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-049: Based on performance testing by AMD Labs as of 6/6/2025, measuring the text generated inference throughput on the LLaMA 3.1-405B model using the FP4 datatype with input length of 128 tokens and an output length of 2048 tokens on the AMD Instinct™ MI355X 8x GPU, and published results for the NVIDIA B200 HGX 8xGPU. Performance per dollar calculated with current pricing for NVIDIA B200 and Instinct MI355X based cloud instances. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. Current customer pricing as of June 10, 2025, and subject to change

MI400-001: Performance projection as of 06/05/2025 using engineering estimates based on the design of a future AMD Instinct MI400 Series GPU compared to the Instinct MI355x, with 2K and 16K prefill with TP8, EP8 and projected inference performance, and using a GenAI training model evaluated with GEMM and Attention algorithms for the Instinct MI400 Series .Results may vary when products are released in market.

Endnotes

MI350-54: Based on calculations by AMD internal testing as of 6/10/2025. Using 8 GPU AMD Instinct™ MI355X Platform for overall GPU-normalized Training Throughput (time to complete) for fine-tuning using the Llama2-70B LoRA chat model (FP8) compared to published 8 GPU Nvidia H200 Platform performance (FP8). Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-055: Based on engineering projections by AMD Performance Labs in June 2025, to estimate the peak theoretical precision performance of seventy-two (72) AMD Instinct™ MI400X GPUs (Rack) using the FP4 Matrix datatype vs. an 8xGPU AMD Instinct MI00 platform using the FP16 Matrix datatype. Results subject to change when products are released in market.

MI350-056: Based on calculations by AMD Performance Labs in June 2025, to determine the peak theoretical precision performance of 8x GPU AMD Instinct MI355X platform with FP6 Matix datatype vs. an 8x GPU AMD Instinct MI325X/MI300X platforms with FP8 Matrix datatype. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of the latest drivers and optimizations.

MI350-057: *Based on calculations by AMD Performance Labs in June 2025, to determine the peak theoretical precision performance of 8x GPU AMD Instinct MI325X/MI300X platform with FP8 Matix datatype vs. an 8x GPU AMD Instinct MI250X platform with FP16 Matrix datatype. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of the latest drivers and optimizations.

MI350-058: *Based on calculations by AMD Performance Labs in June 2025, to determine the peak theoretical precision performance of 8x GPU AMD Instinct MI325X/MI300X platform with FP8 Matix datatype vs. an 8x GPU AMD Instinct MI250X platform with FP16 Matrix datatype. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of the latest drivers and optimizations.

VEN-003: PCIe Gen comparison based on PCI-SIG published statements, https://pcisig.com/pci-express-6.0-specification.   2P 6th Gen EPYC CPU with 128 lanes of PCIe Gen 6 and 5th Gen EPYC with 128 lanes of PCIe Gen 5 as of 6/3/2025. PCIe is a registered trademark of PCI-SIG Corporation

End Notes

MI350-025: AMD Instinct MI355X platform with AMD ROCm 7.0 (pre-release build 16047), running Deepseek-R1 on SGLang shows up to 1.3x TPS inference performance advantage with the FP8 datatype (Up to 1.3x faster or 30% faster) vs. NVIDIA Blackwell B200 platform, with CUDA version 12.8.
Testing by AMD Performance Labs as of May 25, 2025, measuring the inference performance in tokens per second (TPS) of the AMD Instinct MI355X platform with ROCm 7.0 pre-release build 16047, running DeepSeek R1 LLM on SGLang versus NVIDIA Blackwell B200 platform with CUDA version 12.8. Server manufacturers may vary configurations, yielding different results. Performance may vary based on hardware configuration, software version, and the use of the latest drivers and optimizations.

Additional Hardware Configuration(s)
1P AMD EPYC™ 9575F CPU server with 8x AMD Instinct™ MI355X (288GB, 1400W) GPUs, Supermicro AS-4126GS-NMR0LCC, 3 TiB (24 D IMMs, 6400 mts memory, 128 GiB/DIMM), 2x 3.49TB Micron 7450 storage, BIOS version: 1.4a.
1P Intel Xeon 692P CPU server with 8x NVIDIA B200 (180GB, 1000W) GPUs, Supermicro SYS-A22GA-NBRT, 2.95 TiB (24 DIMMs, 4800 mts memory, 128 GiB/DIMM), 2x 3.5 TB Micron 7450 storage, BIOS version: 1.8.
Additional Software Configuration(s)
Ubuntu 22.04 LTS with Linux kernel 6.8.0-59-generic, ROCm 7.0.0 (pre-release build 16047) + amdgpu 6.14.5 (build 2168543)
Pre-release Docker: rocm/aigmodels-private:experimental_950_5_26 (cache off, --chunked prefill size 131072, torch compile), TP8+DP8
vs.
Ubuntu 22.04.5 LTS with Linux kernel 5.15.0-72-generic, Driver Version: 570.133.20 CUDA Version: 12.8
Public Docker: lmsysorg/sglang:Blackwell.

MI350-038: Based on testing by AMD internal labs as of 6/6/2025 measuring text generated throughput for LLaMA 3.1-405B model using FP4 datatype. Test was performed using input length of 128 tokens and an output length of 2048 tokens for AMD Instinct™ MI355X 8xGPU platform compared to NVIDIA published B200 HGX 8xGPU platform results. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-039: Based on Lucid automation framework testing by AMD internal labs as of 6/6/2025 measuring text generated throughput for LLaMA 3.1-405B model using FP4 datatype. Test was performed using 4 different combinations (128/2048) of input/output lengths to achieve a mean score of tokens per second for AMD Instinct™ MI355X 4xGPU platform compared to NVIDIA DGX GB200 4xGPU platform. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI350-040:Based on testing (tokens per second) by AMD internal labs as of 6/6/2025 measuring text generated online serving throughput for DeepSeek-R1 chat model using FP4 datatype. Test was performed using input length of 3200 tokens and an output length of 800 tokens with concurrency up to 64 looks, serviceable with 30ms ITL threshold for AMD Instinct™ MI355X 8xGPU platform median total tokens compared to NVIDIA B200 HGX 8xGPU platform results. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. MI350-040

GENERAL DISLCAIMER

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

Endnotes

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

GD-243: TOPS :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. GD-243.

SHO-06: Testing as of Dec 2024 using the following benchmark scores compared to Intel Core Ultra 9 288V and Qualcomm Snapdragon X Elite X1E-84-100. Cinebench 2024 nT, 3Dmark Wildlife Extreme, and Blender. Next gen AI PC defined as a Windows PC with a processor that includes a NPU with at least 40 TOPS. Configuration for AMD Ryzen™ AI Max+ 395 processor: AMD reference board, Radeon™ 8060S graphics, 32GB RAM, 1TB SSD, VBS=ON, Windows 11. Configuration for Qualcomm Snapdragon X Elite X1E-84-100 processor: Samsung Galaxybook, Adreno Graphics, 16GB RAM, Microsoft Windows 11. Configuration for Intel Core Ultra 9 288V: ASUS Zenbook X 14, Intel Arc Graphics, 32GB RAM, 1TBSSD, Microsoft Windows 11 Home. Laptop manufacturers may vary configurations yielding different results. SHO-06

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

Endnotes

PEN-012: Measurements conducted by AMD Performance Labs as of Aug 27, 2024 on the current specification for the AMD Pensando™ Salina DPU accelerator designed with AMD Pensando™ 5nm process technology, projected to result in delivering 400Gb/s line-rate estimated performance.

Estimated delivered results calculated for AMD Pensando™ Elba DPU designed with AMD Pensando 7nm process technology resulted in 200Gb/s line-rate performance.
Actual results based on production silicon may vary.

Salina projected performance:
Bandwidth: 400Gbps
Connections per second: 10M
Packets per Second: 100MPPS
Encryption Offloads: 400 Gbps
Storage IOPS: 4 Million

Actual results and specifications may vary based on production silicon.

PEN-013: Testing was conducted by AMD Performance Labs as of 3/30/2025 on the Pollara 400 vs Broadcom Thor2 on a test system using identical GPU cluster configurations (16 SuperMicro server nodes, 128 AMD Instinct™ MI300 GPU, 128 Pollara or Thor2 NICs, rail based network topology using 64-port x 400G Broadcom Tomahawk5   based Ethernet switching). CPU on all are: 2P Intel® Xeon® Platinum 8468.Memory: 2048G DDDR5 4800mhz 64GB dual rank dims, OS: Ubuntu® 22.0.4, Kernel 6.5.0-45-generic LTS, 2 rear m.2 NVME, Bios version: 2.3.5., ROCm™ 6.3.0-39. Results may vary due to factors including but not limited to software versions, network speeds and system configurations.

PEN-014: AMD comparison and pricing as of 4/23/2025 of a generic big-buffer Jericho3 based DEF Leaf-Spine Network versus a Naddod Tomahawk5 leaf-spine Network. The Tomahawk system would employ an AMD Pensando™ Pollara NIC in the AI network; the generic system would employ a competitive NIC, costs of the NICs are assumed to be comparable.

Naddod Tomahawk5 800G Leaf-Spine Network Cost: $26.65M
188 Leaf & Spine units (Naddod N9600-64OD) at $26,999 each = $5,075,812 https://www.naddod.com/products/102322.html?srsltid=AfmBOor58tjyTO2mhb8lSaa3A13INrL_RSiS4BOHYTsCQN6X99fJ-tPZ
8K AOC 10m (QDD-400-AOC10M) cables at $1,059 each = $8,472,000
https://www.fs.com/products/146389.html
16K SR4 (QDD-SR4-400G) optics at $819 each = $13,104,000
https://www.fs.com/products/226577.html?attribute=94273&id=3632994

Comparison Large Buffer, Fully Scheduled 800G Leaf-Spine (Jericho3-AI/Ramon3) Network Cost: $31.84M
Network comprised of 222 units of Leaf-Spine Network Cost: $31.84M;
8K AOC 10m cables at $1059 each = $8,472,000
17,778 SR4 Opticss @ $819 each = 14,560,182.

Cost Savings (%): 16%

Prices subject to change. Comparison for specific network configurations only, and may not be representative of all possible network configurations and comparisons.

Endnotes

9xx5-014A: Llama3.1-70B inference throughput results based on AMD internal testing as of 09/01/2024. Llama3.1-70B configurations: TensorRT-LLM 0.9.0, nvidia/cuda 12.5.0-devel-ubuntu22.04 , FP8, Input/Output token configurations (use cases): [BS=1024 I/O=128/128, BS=1024 I/O=128/2048, BS=96 I/O=2048/128, BS=64 I/O=2048/2048]. Results in tokens/second. 2P AMD EPYC 9575F (128 Total Cores) with 8x NVIDIA H100 80GB HBM3, 1.5TB 24x64GB DDR5-6000, 1.0 Gbps 3TB Micron_9300_MTFDHAL3T8TDP NVMe®, BIOS T20240805173113 (Determinism=Power,SR-IOV=On), Ubuntu 22.04.3 LTS, kernel=5.15.0-117-generic (mitigations=off, cpupower frequency-set -g performance, cpupower idle-set -d 2, echo 3> /proc/syss/vm/drop_caches) , 2P Intel Xeon Platinum 8592+ (128 Total Cores) with 8x NVIDIA H100 80GB HBM3, 1TB 16x64GB DDR5-5600, 3.2TB Dell Ent NVMe® PM1735a MU, Ubuntu 22.04.3 LTS, kernel-5.15.0-118-generic, (processor.max_cstate=1, intel_idle.max_cstate=0 mitigations=off, cpupower frequency-set -g performance), BIOS 2.1, (Maximum performance, SR-IOV=On), I/O Tokens Batch Size EMR Turin Relative Difference 128/128 1024 814.678 1101.966 1.353 287.288 128/2048 1024 2120.664 2331.776 1.1 211.112 2048/128 96 114.954 146.187 1.272 31.233 2048/2048 64 333.325 354.208 1.063 20.833 For average throughput increase of 1.197x. When scaling to a 1000 node cluster (1 node = 2 CPUs and 8 GPUs) comparing the AMD EPYC 9575F system and Intel Xeon 8592+ system: 128/128 achieves 287,288 more tokens/s 128/2048 achieves 211,112 more tokens/s 2048/128 achieves 31,233 more tokens/s 2048/2028 achieves 20,833 morere tokens/s Results may vary due to factors including system configurations, software versions and BIOS settings.

9xx5-156: Llama3.1-8B throughput results based on AMD internal testing as of 04/08/2025. Llama3.1-8B configurations: BF16, batch size 32, 32C Instances, Use Case Input/Output token configurations: [Summary = 1024/128, Chatbot = 128/128, Translate = 1024/1024, Essay = 128/1024]. 2P AMD EPYC 9965 (384 Total Cores), 1.5TB 24x64GB DDR5-6400, 1.0 Gbps NIC, 3.84 TB Samsung MZWLO3T8HCLS-00A07, Ubuntu® 22.04.5 LTS, Linux 6.9.0-060900-generic, BIOS RVOT1004A, (SMT=off, mitigations=on, Determinism=Power), NPS=1, ZenDNN 5.0.1 2P AMD EPYC 9755 (256 Total Cores), 1.5TB 24x64GB DDR5-6400, 1.0 Gbps NIC, 3.84 TB Samsung MZWLO3T8HCLS-00A07, Ubuntu® 22.04.4 LTS, Linux 6.8.0-52-generic, BIOS RVOT1004A, (SMT=off, mitigations=on, Determinism=Power), NPS=1, ZenDNN 5.0.1 2P Xeon 6980P (256 Total Cores), AMX On, 1.5TB 24x64GB DDR5-8800 MRDIMM, 1.0 Gbps Ethernet Controller X710 for 10GBASE-T, Micron_7450_MTFDKBG1T9TFR 2TB, Ubuntu 22.04.1 LTS Linux 6.8.0-52-generic, BIOS 1.0 (SMT=off, mitigations=on Performance Bias), IPEX 2.6.0 Results: CPU 6980P 9755 9965 Summary 1 n/a1.093 Translate 1 1.062 1.334 Essay 1 n/a 1.14 Results may vary due to factors including system configurations, software versions, and BIOS settings.

9xx5-158: GPT-J-6B throughput results based on AMD internal testing as of 04/08/2025. GPT-J-6B configurations: BF16, batch size 32, 32C Instances, Use Case Input/Output token configurations: [Summary = 1024/128, Chatbot = 128/128, Translate = 1024/1024, Essay = 128/1024]. 2P AMD EPYC 9965 (384 Total Cores), 1.5TB 24x64GB DDR5-6400, 1.0 Gbps NIC, 3.84 TB Samsung MZWLO3T8HCLS-00A07, Ubuntu® 22.04.5 LTS, Linux 6.9.0-060900-generic, BIOS RVOT1004A, (SMT=off, mitigations=on, Determinism=Power), NPS=1, ZenDNN 5.0.1, Python 3.10.12 2P AMD EPYC 9755 (256 Total Cores), 1.5TB 24x64GB DDR5-6400, 1.0 Gbps NIC, 3.84 TB Samsung MZWLO3T8HCLS-00A07, Ubuntu® 22.04.4 LTS, Linux 6.8.0-52-generic, BIOS RVOT1004A, (SMT=off, mitigations=on, Determinism=Power), NPS=1, ZenDNN 5.0.1, Python 3.10.12 2P Xeon 6980P (256 Total Cores), AMX On, 1.5TB 24x64GB DDR5-8800 MRDIMM, 1.0 Gbps Ethernet Controller X710 for 10GBASE-T, Micron_7450_MTFDKBG1T9TFR 2TB, Ubuntu 22.04.1 LTS Linux 6.8.0-52-generic, BIOS 1.0 (SMT=off, mitigations=on, Performance Bias), IPEX 2.6.0, Python 3.12.3 Results: CPU 6980P 9755 9965 Summary 1 1.034 1.279 Chatbot 1 0.975 1.163 Translate 1 1.021 0.93 Essay 1 0.978 1.108 Caption 1 0.913 1.12 Overall 1 0.983 1.114 Results may vary due to factors including system configurations, software versions, and BIOS settings.

9xx5-162: XGBoost (Runs/Hour) throughput results based on AMD internal testing as of 04/08/2025. XGBoost Configurations: v1.7.2, Higgs Data Set, 32 Core Instances, FP32 2P AMD EPYC 9965 (384 Total Cores), 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1.0 Gbps NIC, 3.84 TB Samsung MZWLO3T8HCLS-00A07, Ubuntu® 22.04.5 LTS, Linux 5.15 kernel, BIOS RVOT1004A, (SMT=off, mitigations=on, Determinism=Power), NPS=1 2P AMD EPYC 9755 (256 Total Cores), 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1.0 Gbps NIC, 3.84 TB Samsung MZWLO3T8HCLS-00A07, Ubuntu® 22.04.4 LTS, Linux 5.15 kernel, BIOS RVOT1004A, (SMT=off, mitigations=on, Determinism=Power), NPS=1 2P Xeon 6980P (256 Total Cores), 1.5TB 24x64GB DDR5-8800 MRDIMM, 1.0 Gbps Ethernet Controller X710 for 10GBASE-T, Micron_7450_MTFDKBG1T9TFR 2TB, Ubuntu 22.04.1 LTS Linux 6.8.0-52-generic, BIOS 1.0 (SMT=off, mitigations=on, Performance Bias) Results: CPU Throughput Relative 2P 6980P 400 1 2P 9755 436 1.090 2P 9965 771 1.928 Results may vary due to factors including system configurations, software versions and BIOS settings.

9xx5-164: FAISS (Runs/Hour) throughput results based on AMD internal testing as of 04/08/2025. FAISS Configurations: v1.8.0, sift1m Data Set, 32 Core Instances, FP32 2P AMD EPYC 9965 (384 Total Cores), 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1.0 Gbps NIC, 3.84 TB Samsung MZWLO3T8HCLS-00A07, Ubuntu® 22.04.5 LTS, Linux 5.15 kernel, BIOS RVOT1004A, (SMT=off, mitigations=on, Determinism=Power), NPS=1 2P AMD EPYC 9755 (256 Total Cores), 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1.0 Gbps NIC, 3.84 TB Samsung MZWLO3T8HCLS-00A07, Ubuntu® 22.04.4 LTS, Linux 5.15 kernel, BIOS RVOT1004A, (SMT=off, mitigations=on, Determinism=Power), NPS=1 2P Xeon 6980P (256 Total Cores), 1.5TB 24x64GB DDR5-8800 MRDIMM, 1.0 Gbps Ethernet Controller X710 for 10GBASE-T, Micron_7450_MTFDKBG1T9TFR 2TB, Ubuntu 22.04.1 LTS Linux 6.8.0-52-generic, BIOS 1.0 (SMT=off, mitigations=on, Performance Bias) Results: Throughput Relative 2P 6980P 36.63 1 2P 9755 46.86 1.279 2P 9965 58.6 1.600 Results may vary due to factors including system configurations, software versions and BIOS settings.

9xx5-166: Llama3.2-1B throughput results based on AMD internal testing as of 04/08/2025. Llama3.3-1B configurations: BF16, batch size 32, 32C Instances, Use Case Input/Output token configurations: [Summary = 1024/128, Chatbot = 128/128, Translate = 1024/1024, Essay = 128/1024]. 2P AMD EPYC 9965 (384 Total Cores), 1.5TB 24x64GB DDR5-6400, 1.0 Gbps NIC, 3.84 TB Samsung MZWLO3T8HCLS-00A07, Ubuntu® 22.04.5 LTS, Linux 6.9.0-060900-generic, BIOS RVOT1004A, (SMT=off, mitigations=on, Determinism=Power), NPS=1, ZenDNN 5.0.1, Python 3.10.2 2P Xeon 6980P (256 Total Cores), AMX On, 1.5TB 24x64GB DDR5-8800 MRDIMM, 1.0 Gbps Ethernet Controller X710 for 10GBASE-T, Micron_7450_MTFDKBG1T9TFR 2TB, Ubuntu 22.04.1 LTS Linux 6.8.0-52-generic, BIOS 1.0 (SMT=off, mitigations=on, Performance Bias), IPEX 2.6.0, Python 3.12.3 Results: CPU 6980P 9965 Summary 1 1.213 Translation 1 1.364 Essay 1 1.271 Results may vary due to factors including system configurations, software versions, and BIOS settings.

9xx5-168: DeepSeek-Qwen-R1 (32B) inference using SGLang and EchoSwift inference throughput results based on AMD internal testing as of 05/14/2025. DeepSeek-Qwen-R1 (32B) configurations: SGLang 0.4.6, TP8 Parallel, results in tokens/second. System: Supermicro AS -8125GS-TNMR2, 2P AMD EPYC 9575F (128 Total Cores) with 8x AMD Instinct MI300X, GPU Interconnectivity XGMI, ROCm™ 6.4.0, 1536GB 24x64GB DDR5-6000, BIOS 3.3, Ubuntu® 24.04.1 LTS, kernel 6.8.0-59-generic. System: Supermicro SYS-821GE-TNMR2, 2P Intel Xeon Platinum 8592+ (128 Total Cores) with 8x AMD Instinct MI300X, GPU Interconnectivity XGMI, ROCm 6.2.0-66, 2048GB 32x64GB DDR5-4400, BIOS 2.4, Ubuntu 24.04.1 LTS, kernel 6.8.0-52-generic. Results may vary due to factors including system configurations, software versions and BIOS settings.

Endnotes

Performance

1. STX-04a: Based on AMD product specifications and competitive products announced as of October 2024. AMD Ryzen™ AI 300 Series processors' NPU offer up to 50+ peak TOPS. AI PC is defined as a laptop PC with a processor that includes a neural processing unit (NPU). STX-04a.
2. STXP-06a: Based on AMD product specifications and competitive products announced as of March 2025. AMD Ryzen™ AI PRO 300 Series processors' NPU offers up to 55 peak TOPS. This is the most TOPS offered on any system found in enterprise today. AI PC is defined as a laptop PC with a processor that includes a neural processing unit (NPU). STXP-06a.
3. KRKP-31: Testing as of March 2025 by AMD using Procyon Office Productivity, Teams + Procyon Office Productivity Word, Teams + Procyon Office Productivity Excel, Teams + Procyon Office Productivity Power Point, Teams + Procyon Office Productivity Outlook benchmarks tested in Balanced mode on a Dell Pro 14 Plus system with AMD Ryzen AI 7 PRO 350 @28W, Radeon 860M graphics driver 32.0.13022.4002, 64GB @5600MHz, 1TB SSD, Win 11 Pro 26100 vs. a similarly configured system with Intel Core Ultra 7 268V@17W, Intel Arc 140V, Graphics driver 32.0.101.6556, 32GB @ 8533MHz, 512GB SSD, Win11 Pro 26100. Laptop manufacturers may vary configurations yielding different results. KRKP-31
4. KRKP-26: Testing as of 2/4/25 by AMD performance labs on a Dell Pro 14 with AMD Ryzen™ AI 7 PRO 350 processor (28W), Radeon™ 860M graphics, 64GB of RAM, 1TB NVMe SSD, VBS=ON, Windows 11 Pro vs. a Dell Pro 14 Plus with an Intel Core Ultra 7 268v processor (17W) (vPro enabled), Intel Arc Graphics, VBS=ON, 32GB RAM, 512GB NVMe SSD, Microsoft Windows 11 Pro running, on DC power (unplugged from wall power) in "Balanced Mode", a Teams video conference call while simultaneously running the following benchmarks: Procyon Office Productivity, Procyon Office Productivity Excel, Procyon Office Productivity Outlook, Procyon Office Productivity Power Point, Procyon Office Productivity Word. Laptop manufactures may vary configurations yielding different results. KRKP-26.
5. KRKP-41: Testing as of Jan 2025 by AMD performance labs on a ASUS Vivobook S 14 powered by Intel Core Ultra 7 258V processor @17W, Intel Arc 140V GPU, 16GB RAM, 1TB SSD, VBS=ON, compared to an ASUS Vivobook S 14 powered by AMD Ryzen AI 7 350 processor @28W, AMD Radeon™ 860M Graphics VBS=On. (Balanced Mode): Running Cinebench nT (2024) and 1T (2024) benchmarks tests across AC and DC power. Laptop manufactures may vary configurations yielding different results. KRKP-41.
6. STX-112: Testing as of March 2024 by AMD using Deepseek-R1 distill Qwen 1.5b and 7b models running Ryzen AI Software 1.3 (AMD NPU + igpu) and OpenVino GenAI (Intel NPU). Configuration for AMD Ryzen™ AI 9 HX 375 processor: HP OmniBook Ultra 14, 32GB RAM, Windows 11 Pro. Configuration for Intel Core Ultra 7 258V: ASUS Zenbook S 14, 32GB RAM, Windows 11 Pro. Both tested in Best Performance mode with VBS ON. Laptop manufactures may vary configurations yielding different results. STX-112
7. STX-113: Testing as of 3/4/2025 by AMD. All tests conducted on LM Studio 0.3.11. Vulkan Llama.cpp runtime 1.18. Performance may vary.Tokens/s: Sustained performance average of multiple runs with specimen prompt "How long would it take for a ball dropped from 10 meter height to hit the ground?". Models tested: DeepSeek R1 Distill Qwen 1.5b Q4 K M, DeepSeek R1 Distill Qwen 7b Q4 K M, DeepSeek R1 Distill Qwen 8b Q4 K M, DeepSeek R1 Distill Qwen 14b Q4 K M, Phi 4 Mini Instruct 3.8b, Phi 4 Q4 K M, Llama 3.2 3b Instruct. TTFT: Sustained average of multiple runs in DeepSeek R1 Distill Qwen 1.5b Q4 K M. Short prompt = same as Token/s. Long prompt = "Summarize the following in exactly 5 lines: [Scene 1, Act 1 of Romeo and Juliet by Shakespeare]. AMD Ryzen™ AI HX 370 processor on an HP Zenbook S16 with 32GB 7500 MT/s memory, Windows 11 Pro 24H2 and Adrenalin 25.1.1 Optional. VGM = High (16GB) Intel Core Ultra 7 258V on an HP Zenbook S14 with 32GB 8533 MT/s memory, Windows 11 Pro 24H2 and Intel Graphics Driver 32.0.101.6559. STX-113.
8. STXP-35: Testing as of Dec 2024 by Signal65 (3rd party), on a Lenovo ThinkPad T14s Gen 6 with an AMD Ryzen™ AI 7 PRO 360 processor @22W, Radeon™ 880M graphics, 32GB RAM, 512GB SSD, VBS=ON, Windows 11 Pro vs. a 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 Enterprise and an IT image(s) on both system. Calculation of total cost savings include comparing the following for an example organization with 25k employees: initial system acquisition cost (per employee) and time value savings per employee (using multitasking performance on typical office workloads). STXP-35
9. PXD-19: Testing as of July 2025 by AMD performance labs on a Lenovo Qitian M550 desktop system with AMD Ryzen™ 7 8700G desktop processor (65W), Radeon™ 780M graphics, 32GB RAM, 1TB PCIe 4.0 SSD vs. a Dell Optiplex 7020 Intel Core i7 processor 14700, Intel UHD Graphics 770, 32GB RAM, 1TB PCIe 4.0 SSD, in the following benchmarks and application(s): PCMark10, Passmark, Procyon AI Computer Vision, LM Studio Deepseek. Laptop manufactures may vary configurations yielding different results. PDX-19
10. PXD-20: Testing as of July 2025 by AMD performance labs on a Lenovo ThinkCentre M75s AMD Ryzen 7 PRO 8700G desktop processor (65W), Radeon™ RX 6400 graphics, 16GB RAM, 1TB PCIe 4.0 SSD, Camera 1080p, 4K vs. a DELL 09M47G Intel Core i7 processor 14700, Intel UHD Graphics 770, 16GB RAM, 1TB PCIe 4.0 SSD, Camera 1080p, 4K, measuring wall power in best performance mode across the following test(s): Teams + Procyon Office. Laptop manufactures may vary configurations yielding different results. PXD-20.

Legal Disclaimers

- SHO-06: Testing as of Dec 2024 using the following benchmark scores compared to Intel Core Ultra 9 288V and Qualcomm Snapdragon X Elite X1E-84-100. Cinebench 2024 nT, 3Dmark Wildlife Extreme, and Blender.. Next gen AI PC defined as a Windows PC with a processor that includes a NPU with at least 40 TOPS. Configuration for AMD Ryzen™ AI Max+ 395 processor: AMD reference board, Radeon™ 8060S graphics, 32GB RAM, 1TB SSD, VBS=ON, Windows 11. Configuration for Qualcomm Snapdragon X Elite X1E-84-100 processor: Samsung Galaxybook, Adreno Graphics, 16GB RAM, Microsoft Windows 11. Configuration for Intel Core Ultra 9 288V: ASUS Zenbook X 14, Intel Arc Graphics, 32GB RAM, 1TBSSD, Microsoft Windows 11 Home. Laptop manufacturers manufactures may vary configurations yielding different results. SHO-06
- SHO-08: Based on AMD product specifications and competitive products announced as of December 2024. AMD Ryzen™ AI Max Series processors' NPU offer up to 50 peak TOPS. AI PC is defined as a laptop PC with a processor that includes a neural processing unit (NPU). SHO-08
- SHO-13: Testing as of Dec 2024 using Llama 70b 3.1 Nemotron Q4 K M quantization running through llama.cpp and LM Studio. Input prompt length 100 token prompt. Next Gen AI PC defined as a PC with a minimum 40 TOPS NPU. System configuration for Ryzen AI Max+ 395: AMD reference board, 55W TDP, Radeon™ 8060S graphics, 128GB unified memory (96GB graphics , 32GB rest of system), 1TB SSD, using Llama 3.1. Configuration for Nvidia RTX 4070: ASUS ProArt P16, Ryzen AI 9 HX 370 processor, 64GB RAM, 2 TB SSD, Windows 11. (https://blogs.nvidia.com/blog/ai-decoded-lm-studio/). Manufactures may vary configurations yielding different results. SHO-13
- SHO-14: Testing as of Dec 2024 using Llama 70b 3.1 Nemotron Q4 K M quantization running through llama.cpp and LM Studio. Input prompt length 100 token prompt. System configuration for Ryzen AI Max+ 395: AMD reference board, 55W TDP, Radeon™ 8060S graphics, 128GB RAM, 1TB SSD, using Llama 3.1. Configuration for Nvidia RTX 4090: ASUS ProArt X670E-CREATOR WIFI motherboard, AMD Ryzen 9 7900X processor, 32GB system RAM, 40GB GPU memory, 1TB SSD, Windows 11. (https://blogs.nvidia.com/blog/ai-decoded-lm-studio/). Manufactures may vary configurations yielding different results. SHO-14
- SHO-16: Based on manufacturer provided specifications of AMD Strix Halo platform Thermal Design Power (TDP) (55W) compared to NVIDIA RTX 4090 Total Graphics Power (TGP) (450W) as of December 2024. SHO-16.
- SHO-22: Testing by AMD as of February 2025 using the following benchmark scores compared to an Intel Core Ultra 9 288V: Cinebench 2024, Blender, Vray, and Corona. Configuration for AMD Ryzen™ AI Max+ 395 processor: Asus ROG Flow Z13, Radeon™ 8060S graphics, 32GB RAM, 1TB SSD, VBS=ON, Windows 11. Configuration for Intel Core Ultra 9 288V: ASUS Zenbook X 14, Intel Arc Graphics, 32GB RAM, 1TB SSD, Microsoft Windows 11 Home. Laptop manufacturers manufactures may vary configurations yielding different results. SHO-22
- SHO-23: Testing by AMD as of February 2025 using 3DMark scores compared to Intel Core Ultra 9 288V. Configuration for AMD Ryzen™ AI Max+ 395 processor: Asus ROG Flow Z13, Radeon™ 8060S graphics, 32GB RAM, 1TB SSD, VBS=ON, Windows 11. Configuration for Intel Core Ultra 9 288V: ASUS Zenbook X 14, Intel Arc Graphics, 32GB RAM, 1TB SSD, Microsoft Windows 11 Home. Laptop manufacturers manufactures may vary configurations yielding different results. SHO-23
- SHO-24: Testing by AMD as of February 2025 using the following benchmarks compared to Apple M4 Pro (12 core and 14 core CPU models): Blender, Corona, and Vray. Configuration for AMD Ryzen™ AI Max+ 395 processor: Asus ROG Flow Z13, Radeon™ 8060S graphics, 32GB RAM, 1TB SSD, VBS=ON, Windows 11. Configuration for Apple M4 Pro (14''/12 core CPU and 16''/14 core CPU): Apple Macbook Pro 2024, 16/20 core GPU, 48GB RAM, macOS Sequoia (x64) Build 15.1.1. Laptop manufacturers manufactures may vary configurations yielding different results. SHO-24
- SHO-25: Testing by AMD as of February 2025 across selected game title FPS scores at 1080p resolution and high settings compared to Intel Core Ultra 9 288V. Configuration for AMD Ryzen™ AI Max+ 395 processor: Asus ROG Flow Z13, Radeon™ 8060S graphics, 32GB RAM, 1TB SSD, VBS=ON, Windows 11. Configuration for Nvidia compare: ASUS ROG Flow Z13, Intel Core i9-13900H processor, Nvidia RTX 4070 graphics, 32GB RAM, 1TB SSD, Microsoft Windows 11 Home. Laptop manufacturers manufactures may vary configurations yielding different results. SHO-25
- GD-150: Max boost for AMD Ryzen processors is the maximum frequency achievable by a single core on the processor running a bursty single-threaded workload. Max boost will vary based on several factors, including, but not limited to: thermal paste; system cooling; motherboard design and BIOS; the latest AMD chipset driver; and the latest OS updates . GD-150
- GD-220d: Ryzen™ AI is defined as the combination of a dedicated AI engine, AMD Radeon™ graphics engine, and Ryzen processor cores that enable AI capabilities. OEM and ISV enablement is required, and certain AI features may not yet be optimized for Ryzen AI processors. Ryzen AI is compatible with: (a) AMD Ryzen 7040 and 8040 Series processors except Ryzen 5 7540U, Ryzen 5 8540U, Ryzen 3 7440U, and Ryzen 3 8440U processors; (b) AMD Ryzen AI 300 Series processors and AMD Ryzen AI 300 Series PRO processors; and (c) all AMD Ryzen 8000G Series desktop processors except the Ryzen 5 8500G/GE and Ryzen 3 8300G/GE. Please check with your system manufacturer for feature availability prior to purchase. GD-220d.
- 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. GD-243.

Legal Disclaimers

- SHO-06: Testing as of Dec 2024 using the following benchmark scores compared to Intel Core Ultra 9 288V and Qualcomm Snapdragon X Elite X1E-84-100. Cinebench 2024 nT, 3Dmark Wildlife Extreme, and Blender.. Next gen AI PC defined as a Windows PC with a processor that includes a NPU with at least 40 TOPS. Configuration for AMD Ryzen™ AI Max+ 395 processor: AMD reference board, Radeon™ 8060S graphics, 32GB RAM, 1TB SSD, VBS=ON, Windows 11. Configuration for Qualcomm Snapdragon X Elite X1E-84-100 processor: Samsung Galaxybook, Adreno Graphics, 16GB RAM, Microsoft Windows 11. Configuration for Intel Core Ultra 9 288V: ASUS Zenbook X 14, Intel Arc Graphics, 32GB RAM, 1TBSSD, Microsoft Windows 11 Home. Laptop manufacturers manufactures may vary configurations yielding different results. SHO-06
- SHO-08: Based on AMD product specifications and competitive products announced as of December 2024. AMD Ryzen™ AI Max Series processors' NPU offer up to 50 peak TOPS. AI PC is defined as a laptop PC with a processor that includes a neural processing unit (NPU). SHO-08
- SHO-13: Testing as of Dec 2024 using Llama 70b 3.1 Nemotron Q4 K M quantization running through llama.cpp and LM Studio. Input prompt length 100 token prompt. Next Gen AI PC defined as a PC with a minimum 40 TOPS NPU. System configuration for Ryzen AI Max+ 395: AMD reference board, 55W TDP, Radeon™ 8060S graphics, 128GB unified memory (96GB graphics , 32GB rest of system), 1TB SSD, using Llama 3.1. Configuration for Nvidia RTX 4070: ASUS ProArt P16, Ryzen AI 9 HX 370 processor, 64GB RAM, 2 TB SSD, Windows 11. (https://blogs.nvidia.com/blog/ai-decoded-lm-studio/). Manufactures may vary configurations yielding different results. SHO-13
- SHO-14: Testing as of Dec 2024 using Llama 70b 3.1 Nemotron Q4 K M quantization running through llama.cpp and LM Studio. Input prompt length 100 token prompt. System configuration for Ryzen AI Max+ 395: AMD reference board, 55W TDP, Radeon™ 8060S graphics, 128GB RAM, 1TB SSD, using Llama 3.1. Configuration for Nvidia RTX 4090: ASUS ProArt X670E-CREATOR WIFI motherboard, AMD Ryzen 9 7900X processor, 32GB system RAM, 40GB GPU memory, 1TB SSD, Windows 11. (https://blogs.nvidia.com/blog/ai-decoded-lm-studio/). Manufactures may vary configurations yielding different results. SHO-14
- SHO-16: Based on manufacturer provided specifications of AMD Strix Halo platform Thermal Design Power (TDP) (55W) compared to NVIDIA RTX 4090 Total Graphics Power (TGP) (450W) as of December 2024. SHO-16.
- SHO-22: Testing by AMD as of February 2025 using the following benchmark scores compared to an Intel Core Ultra 9 288V: Cinebench 2024, Blender, Vray, and Corona. Configuration for AMD Ryzen™ AI Max+ 395 processor: Asus ROG Flow Z13, Radeon™ 8060S graphics, 32GB RAM, 1TB SSD, VBS=ON, Windows 11. Configuration for Intel Core Ultra 9 288V: ASUS Zenbook X 14, Intel Arc Graphics, 32GB RAM, 1TB SSD, Microsoft Windows 11 Home. Laptop manufacturers manufactures may vary configurations yielding different results. SHO-22
- SHO-23: Testing by AMD as of February 2025 using 3DMark scores compared to Intel Core Ultra 9 288V. Configuration for AMD Ryzen™ AI Max+ 395 processor: Asus ROG Flow Z13, Radeon™ 8060S graphics, 32GB RAM, 1TB SSD, VBS=ON, Windows 11. Configuration for Intel Core Ultra 9 288V: ASUS Zenbook X 14, Intel Arc Graphics, 32GB RAM, 1TB SSD, Microsoft Windows 11 Home. Laptop manufacturers manufactures may vary configurations yielding different results. SHO-23
- SHO-24: Testing by AMD as of February 2025 using the following benchmarks compared to Apple M4 Pro (12 core and 14 core CPU models): Blender, Corona, and Vray. Configuration for AMD Ryzen™ AI Max+ 395 processor: Asus ROG Flow Z13, Radeon™ 8060S graphics, 32GB RAM, 1TB SSD, VBS=ON, Windows 11. Configuration for Apple M4 Pro (14"/12 core CPU and 16"/14 core CPU): Apple Macbook Pro 2024, 16/20 core GPU, 48GB RAM, macOS Sequoia (x64) Build 15.1.1. Laptop manufacturers manufactures may vary configurations yielding different results. SHO-24
- SHO-25: Testing by AMD as of February 2025 across selected game title FPS scores at 1080p resolution and high settings compared to Intel Core Ultra 9 288V. Configuration for AMD Ryzen™ AI Max+ 395 processor: Asus ROG Flow Z13, Radeon™ 8060S graphics, 32GB RAM, 1TB SSD, VBS=ON, Windows 11. Configuration for Nvidia compare: ASUS ROG Flow Z13, Intel Core i9-13900H processor, Nvidia RTX 4070 graphics, 32GB RAM, 1TB SSD, Microsoft Windows 11 Home. Laptop manufacturers manufactures may vary configurations yielding different results. SHO-25
- GD-150: Max boost for AMD Ryzen processors is the maximum frequency achievable by a single core on the processor running a bursty single-threaded workload. Max boost will vary based on several factors, including, but not limited to: thermal paste; system cooling; motherboard design and BIOS; the latest AMD chipset driver; and the latest OS updates . GD-150
- GD-220d: Ryzen™ AI is defined as the combination of a dedicated AI engine, AMD Radeon™ graphics engine, and Ryzen processor cores that enable AI capabilities. OEM and ISV enablement is required, and certain AI features may not yet be optimized for Ryzen AI processors. Ryzen AI is compatible with: (a) AMD Ryzen 7040 and 8040 Series processors except Ryzen 5 7540U, Ryzen 5 8540U, Ryzen 3 7440U, and Ryzen 3 8440U processors; (b) AMD Ryzen AI 300 Series processors and AMD Ryzen AI 300 Series PRO processors; and (c) all AMD Ryzen 8000G Series desktop processors except the Ryzen 5 8500G/GE and Ryzen 3 8300G/GE. Please check with your system manufacturer for feature availability prior to purchase. GD-220d.
- 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. GD-243.