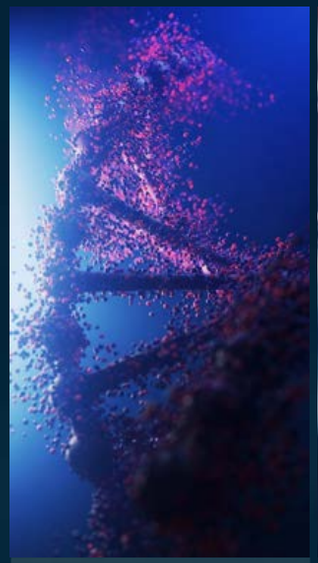
CAUTIONARY STATEMENT

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, product roadmaps, partnerships and collaborations; AI demand; AMD's total addressable markets; and AMD's strategy, 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.

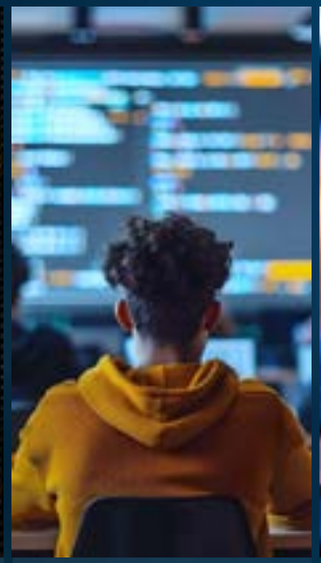
AI Is Transforming Every Industry



HEALTHCARE

MANUFACTURING

FINANCE



EDUCATION

SCIENCE

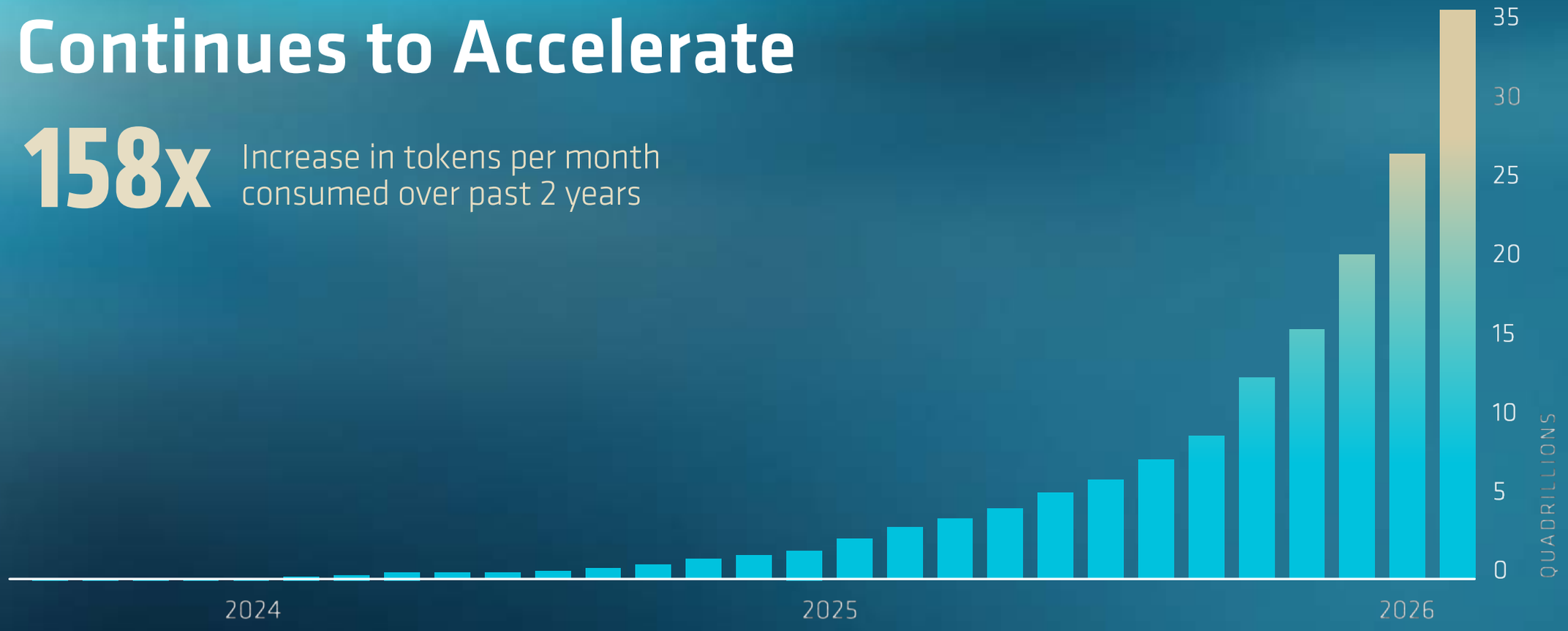
SOFTWARE

AGRICULTURE

MEDIA & ENTERTAINMENT

Demand for AI Continues to Accelerate

158x Increase in tokens per month consumed over past 2 years



Source: Exponential View, State of the AI Economy, June 25, 2026

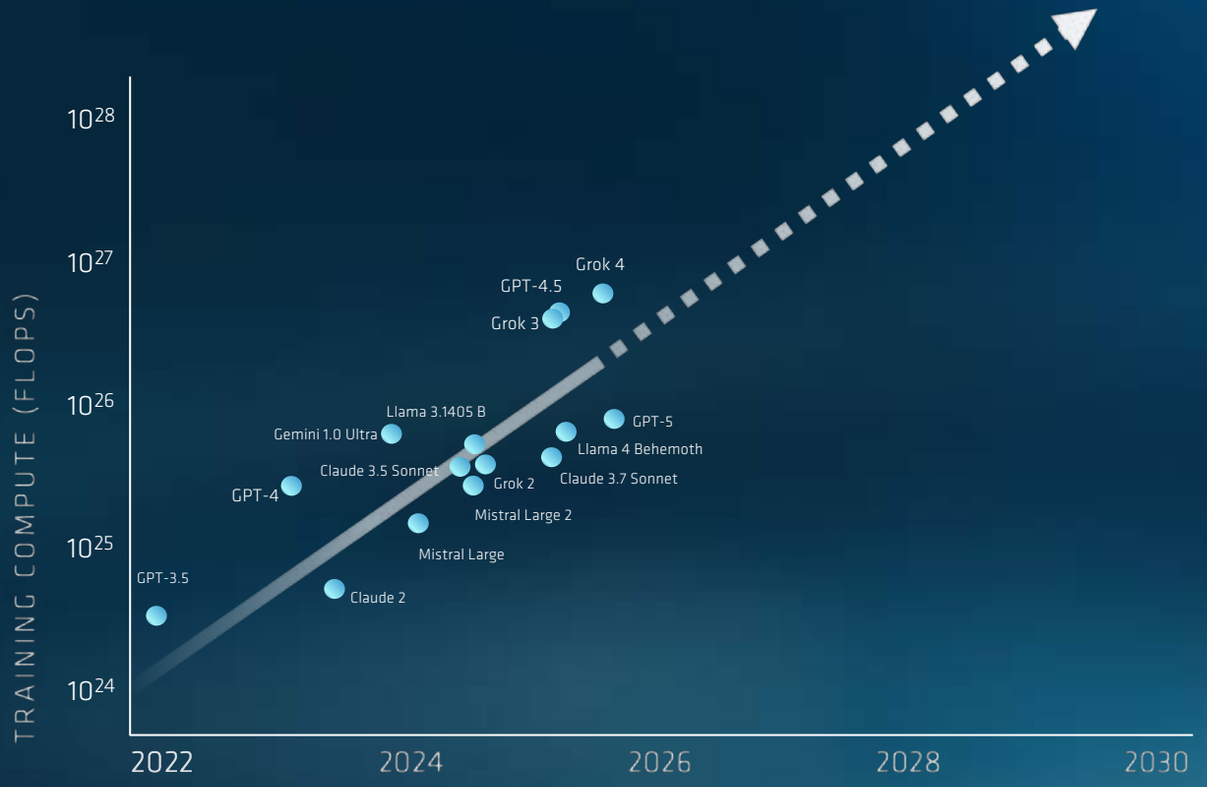
Training Continues to Scale

TRAINING FLOPS INCREASING 5X EVERY YEAR SINCE 2020

Source: Epoch AI, Trends in Artificial Intelligence, February 5, 2026

Training Continues to Scale

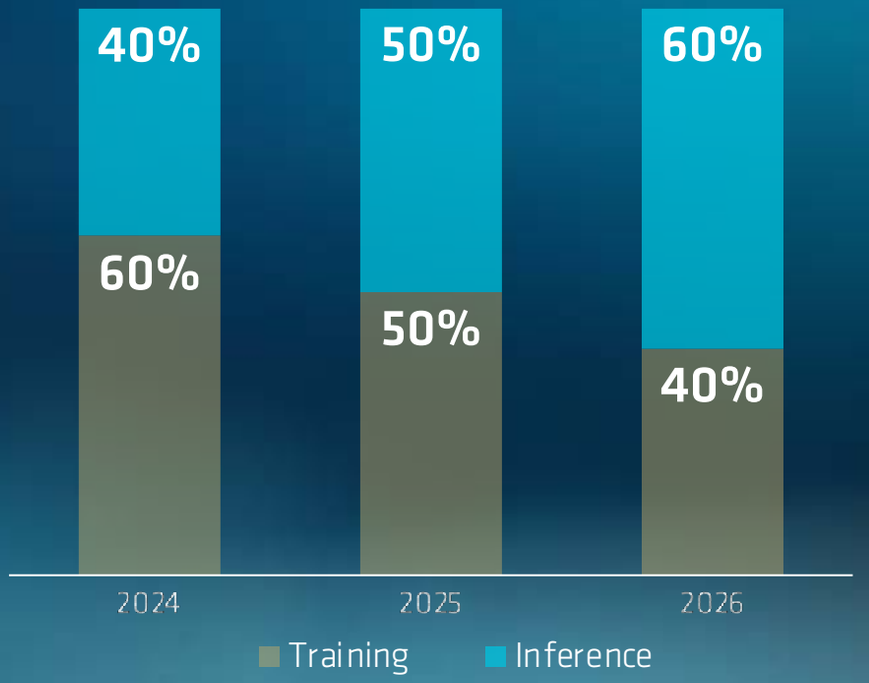
TRAINING FLOPS INCREASING 5X EVERY YEAR SINCE 2020



Source: Epoch AI, Trends in Artificial Intelligence, February 5, 2026

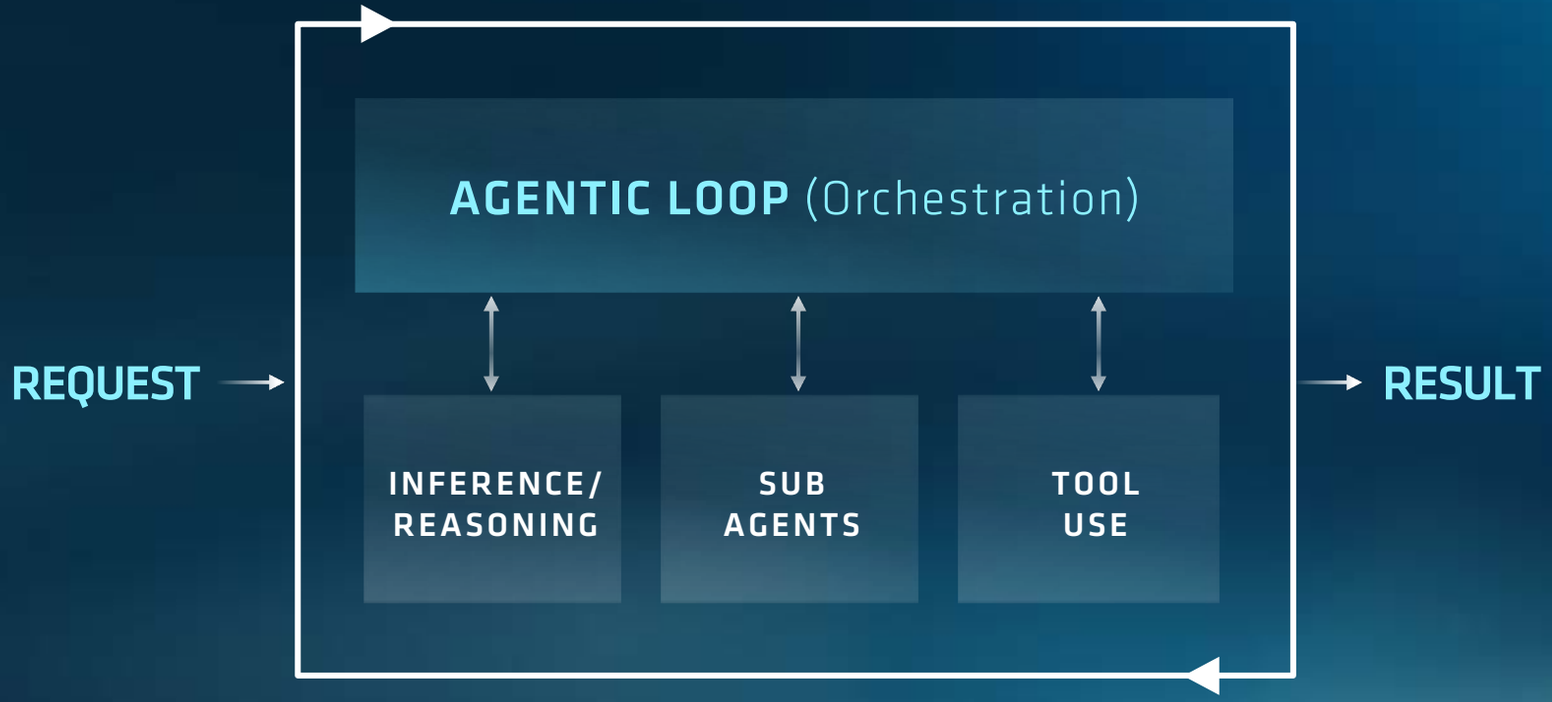
But Inference is Now the #1 AI Workload

TRAINING VS INFERENCE SPLIT



Source: AMD Internal Projections

Agentic AI Further Accelerates Demand for Compute



More Inference

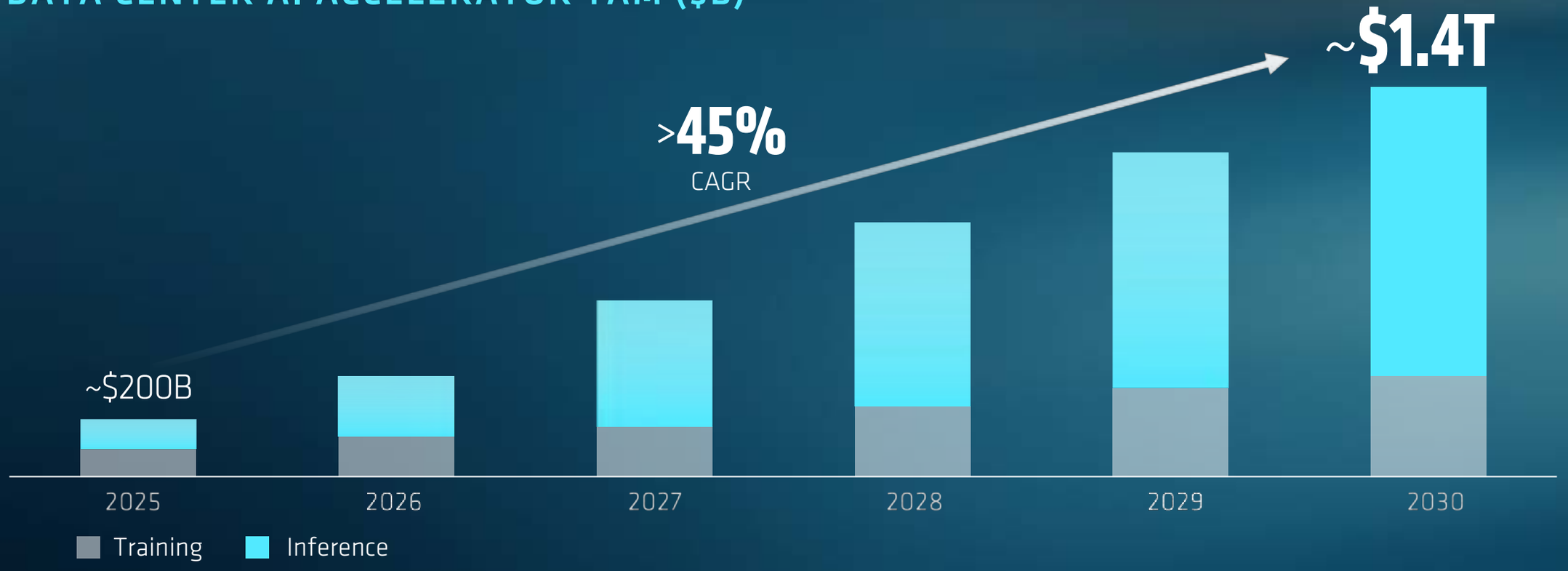
More Agents

More Tools

More Compute

AI Demand
Driving Acceleration in TAM

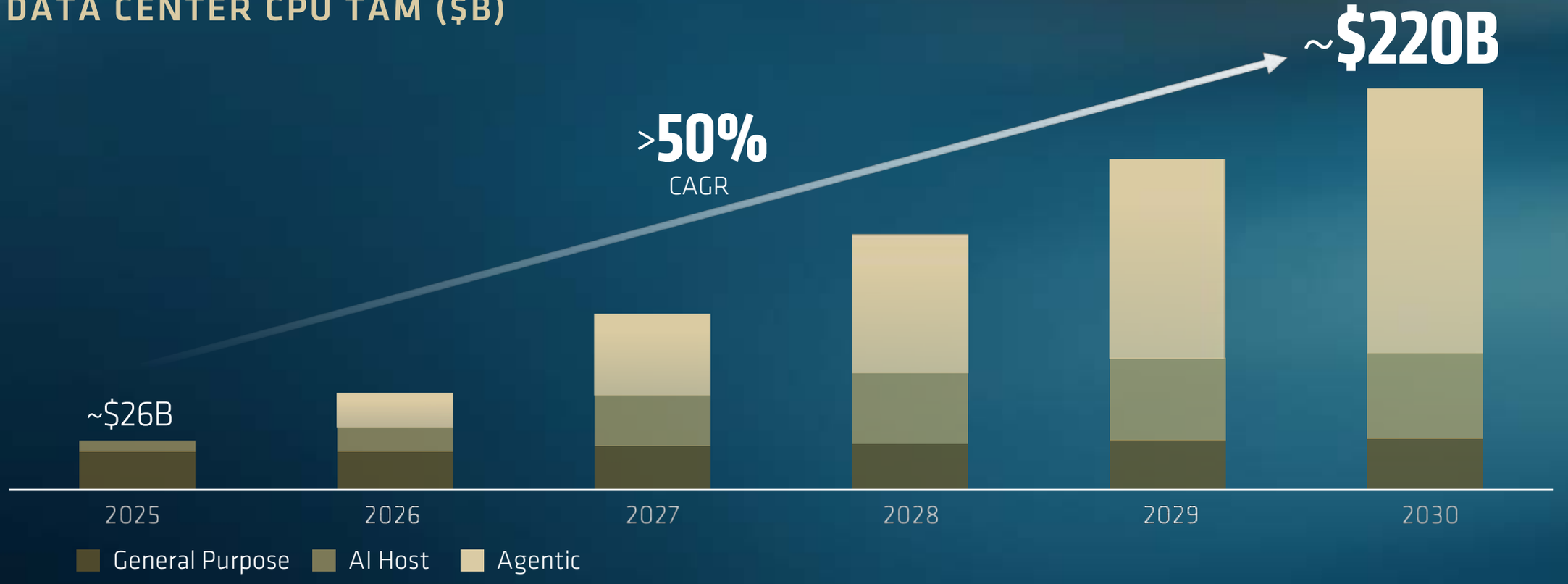
DATA CENTER AI ACCELERATOR TAM ($B)



Source: AMD Internal Projections

AI Demand
Driving Acceleration in TAM

DATA CENTER CPU TAM ($B)



Source: AMD Internal Projections

AI Will Run Wherever the Work Happens

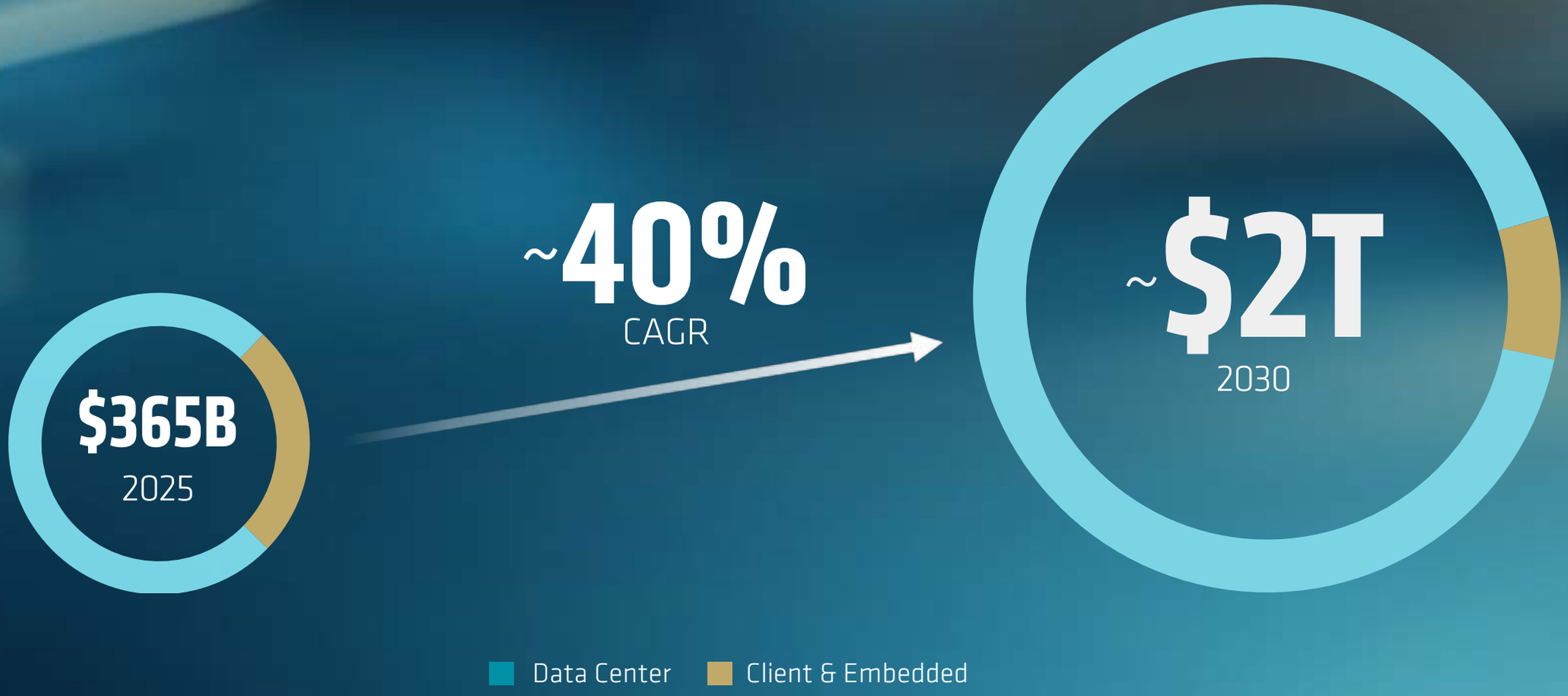
CLOUD

DATA CENTER

CLIENT

EDGE

AMD Total Compute TAM



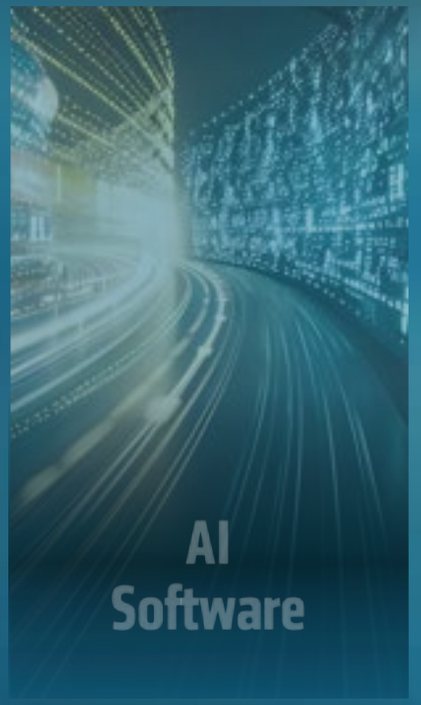
Source: AMD Internal Projections

AMD

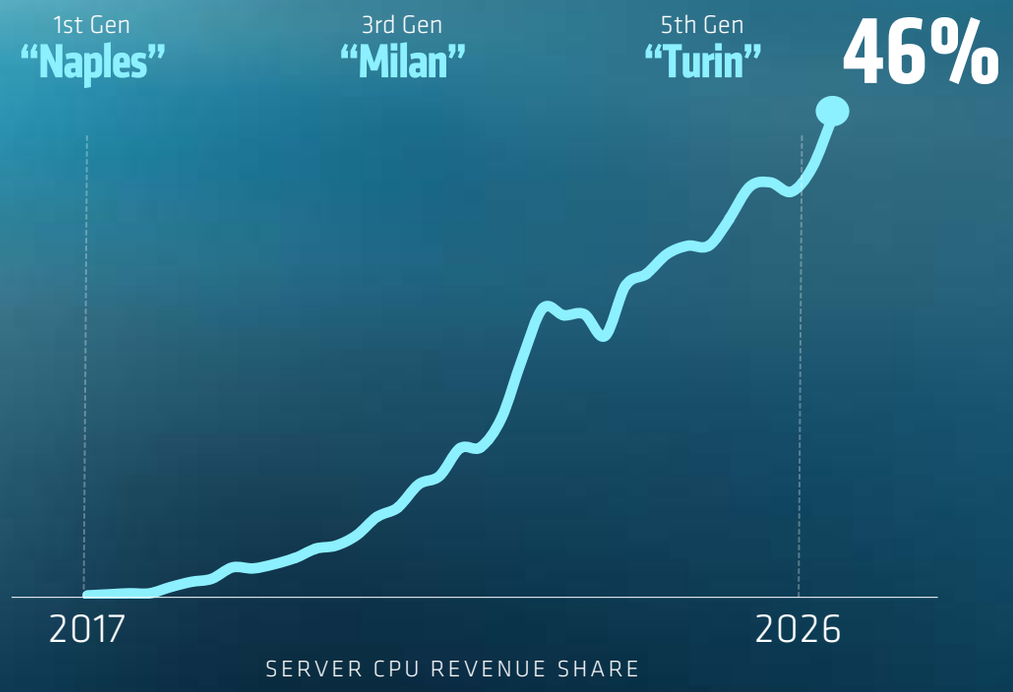
Strategy

AMD | Advancing AI 2026

AMD | Advancing AI 2026



EPYC Momentum Continues to Accelerate



The Industry Standard For Server CPUs

AI	OpenAI	Meta	xAI	GENCI	
Cloud	aws	Microsoft	Google	ORACLE	
Digital	zoom	Uber		NETFLIX	
Enterprise	BOEING	Walmart	AT&T		
OEM	DELL Technologies	SUPERMICRO	HPE	Lenovo	CISCO

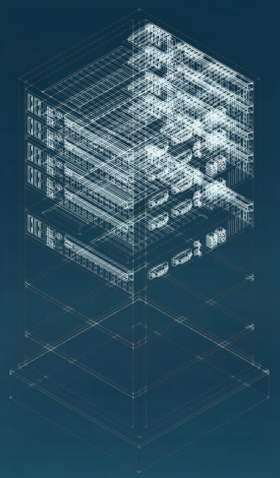
See Endnote: EPYC-055D

AI Leaders Trust AMD Instinct At Scale

Frontier AI Requires Integrated Rack Architectures

Leadership Compute Capability

+



Open Rack Architecture

+

AMD PENSANDO

Ultra Ethernet Consortium

ESUN

Open Fabric to Scale AI

+

Rack Scale Efficiency & Serviceability

+

Turnkey Solutions

LAUNCHING TODAY AT **ADVANCING AI 2026**

AMD HELIOS

Rackscale Performance For At-Scale AI

AMD Helios™ Delivers Leadership Across All Dimensions

#1 COMPUTE
CPU Core Performance
GPU Compute

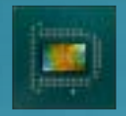
#1 MEMORY
HBM4 Memory Capacity
HBM4 Memory Bandwidth

#1 NETWORKING
Scale Out Bandwidth
Scale Up Bandwidth

MI455X GPU
5th Generation Instinct

“Venice” CPU
6th Generation EPYC

Salina DPU
AMD Pensando



Vulcano AI NIC
AMD Pensando

See Endnote: MI400-005, MI400-007, MI400-019

AMD Helios Compute Architecture Delivers The Highest Performing GPU With Instinct MI455X

40 PF | 20 PF FP4 & FP8 Flops	**432 GB** HBM4 Capacity	**23.3 TB/s** Memory Bandwidth
CDNA 5 GPU Architecture	**2nm | 3nm** Advanced Process	**320B** Transistors

Dense FP4 Flops

AMD Helios Networking Powers Rackscale Performance

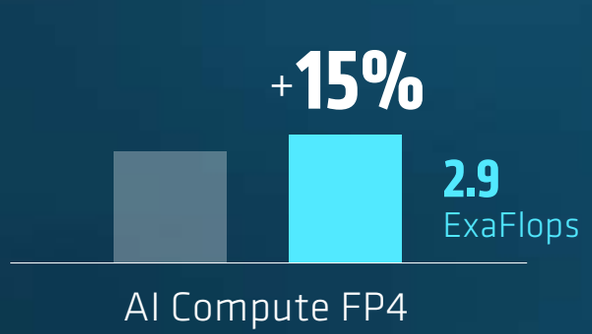
The Only Fully Programmable Networking Solutions Built for AI

Front-End Salina DPU	**Scale-Out** 800G AI NIC	**Scale-Up** UALoE
Secures Network Traffic	AMD Pensando Vulcano	End-to-end fabric reliability
Accelerates SDN and Storage	Broadcom Thor Ultra	Industry Leading Switching

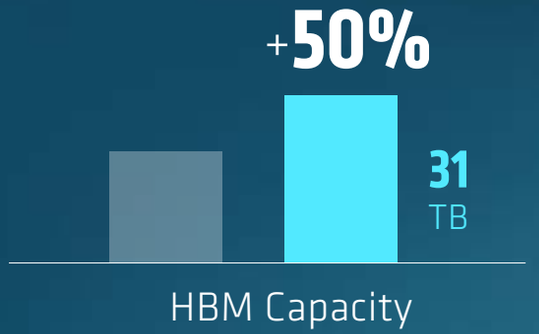
Ultra Ethernet Consortium

ESUN

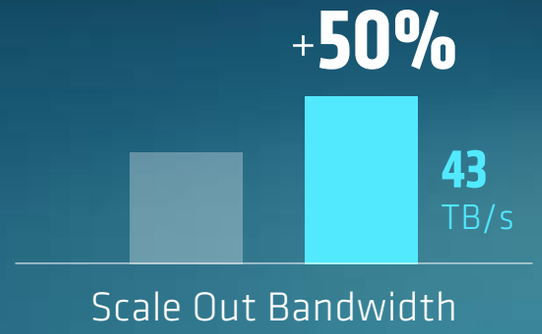
AMD Helios Delivers
The Most Powerful Rackscale AI Infrastructure



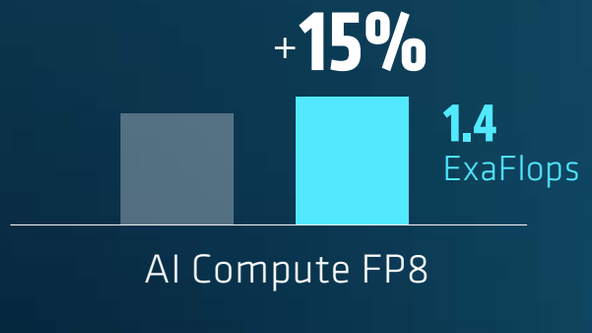
AI Compute FP4



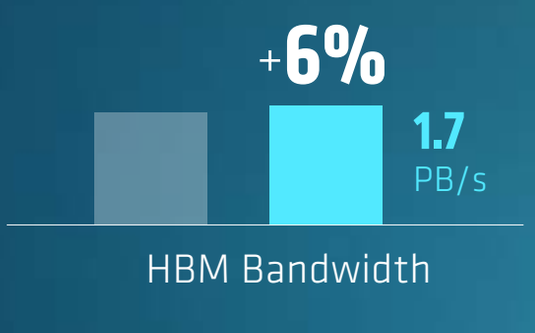
HBM Capacity



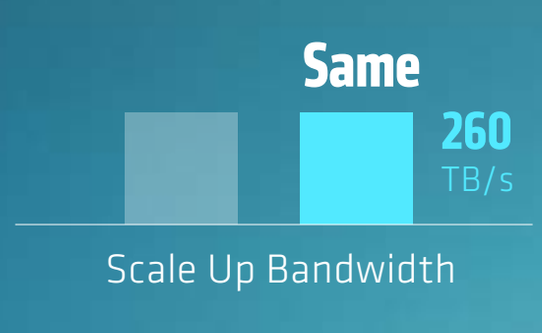
Scale Out Bandwidth



AI Compute FP8



HBM Bandwidth



Scale Up Bandwidth

Vera Rubin AMD Helios

Dense FP4 Flops

See Endnote: MI400-005, MI400-007, MI400-019

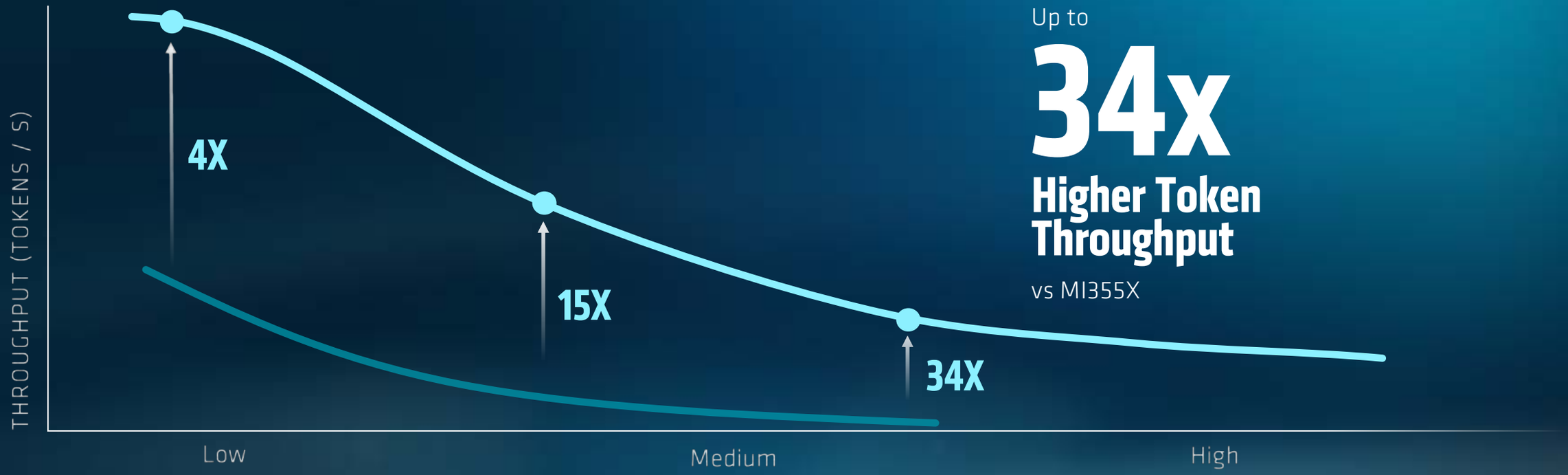
The Most Powerful Rackscale AI Infrastructure

COMPUTE	PERFORMANCE	MEMORY	NETWORKING
4,600 CPU Cores **18,000** GPU Compute Units	**2.9 ExaFlops** FP4 Compute	**31 TB** HBM Capacity	**260 TB/s** Scale Up Bandwidth


AMD HELIOS
AMD EPYC
AMD INSTINCT
AMD PENSANDO
AMD ROCm
IN PRODUCTION TODAY


AMD
×
ANTHROP\C

Instinct MI455X Delivers
Generational Increase In Inference Throughput



Up to

34x

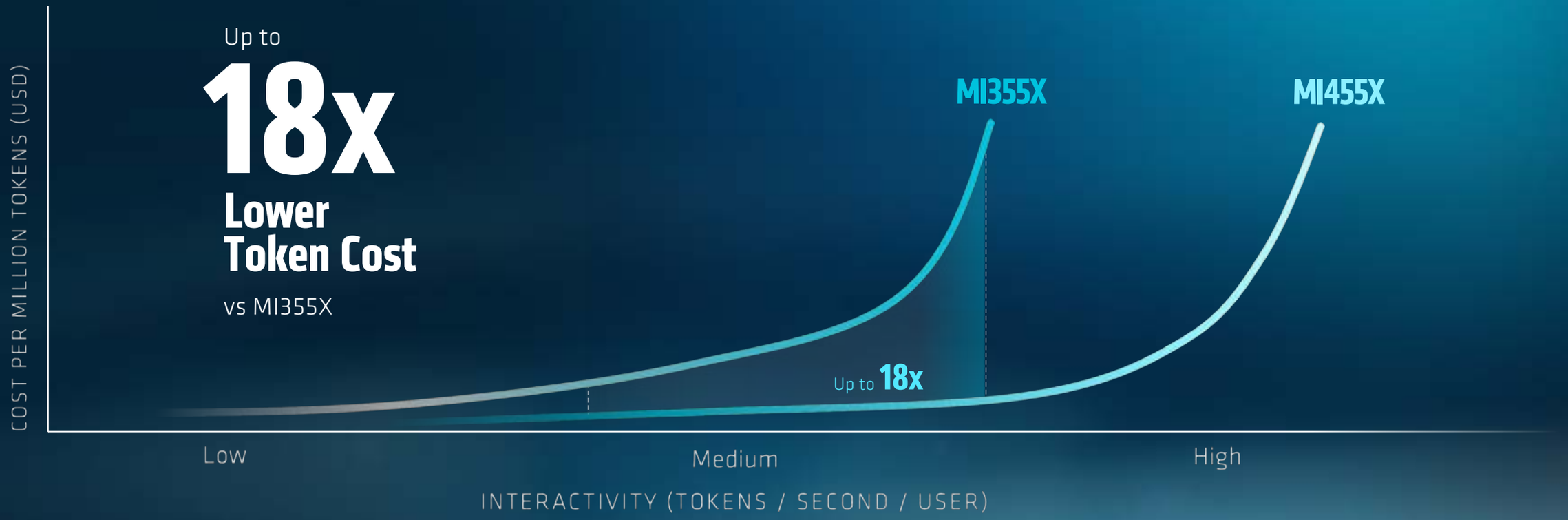
Higher Token Throughput

vs MI355X

DeepSeek V4 Flash

See Endnote: MI400-020

Instinct MI455X Delivers
Generational Improvement In Inference Cost

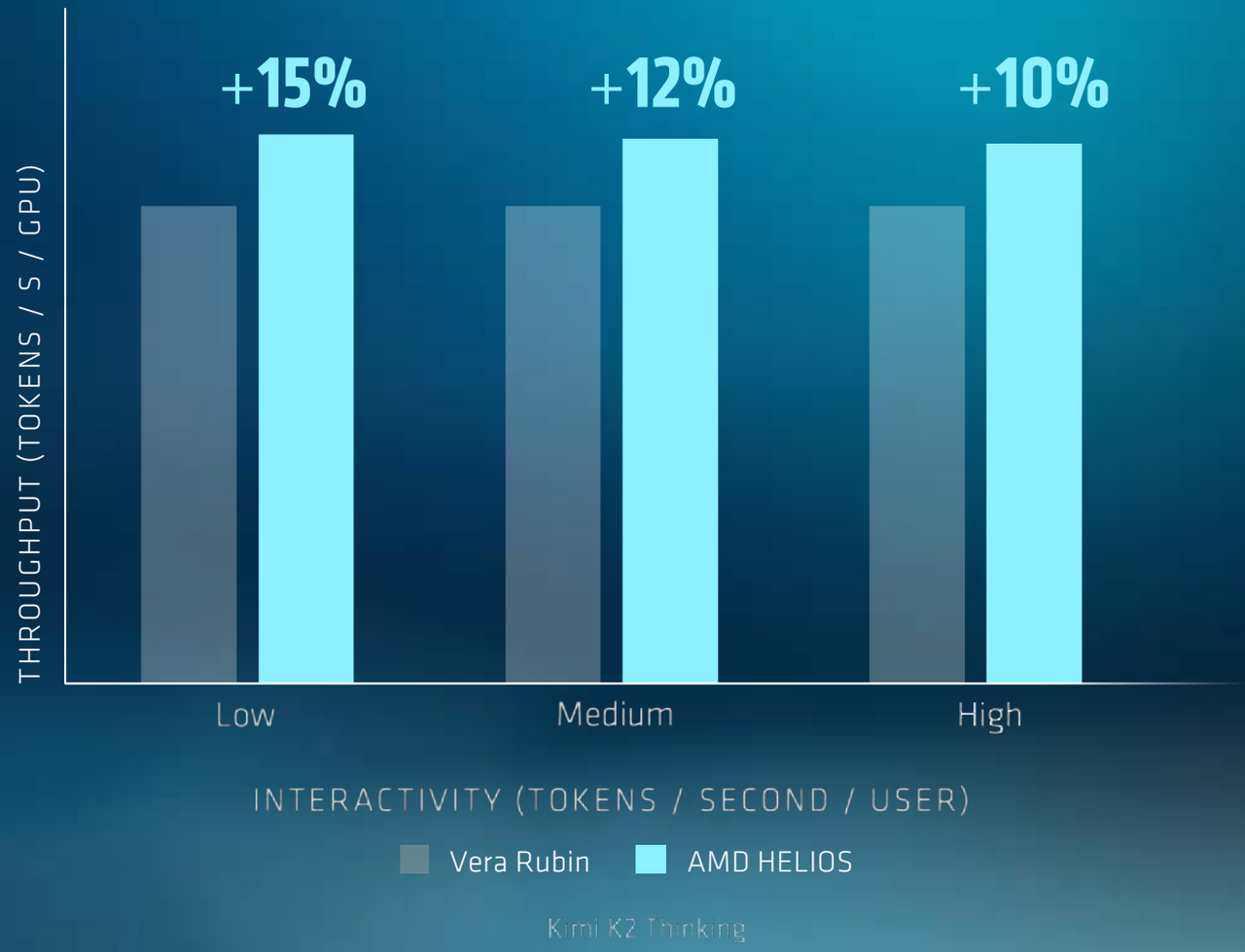


DeepSeek V4 Flash

See Endnote: MI400-022

AMD Helios Delivers

Industry Leading Rack Scale Performance



See Endnote MI400-023

Up To 30% More Tokens / $

Using AMD Helios vs Vera Rubin NVL72

See Endnote: MI400-025

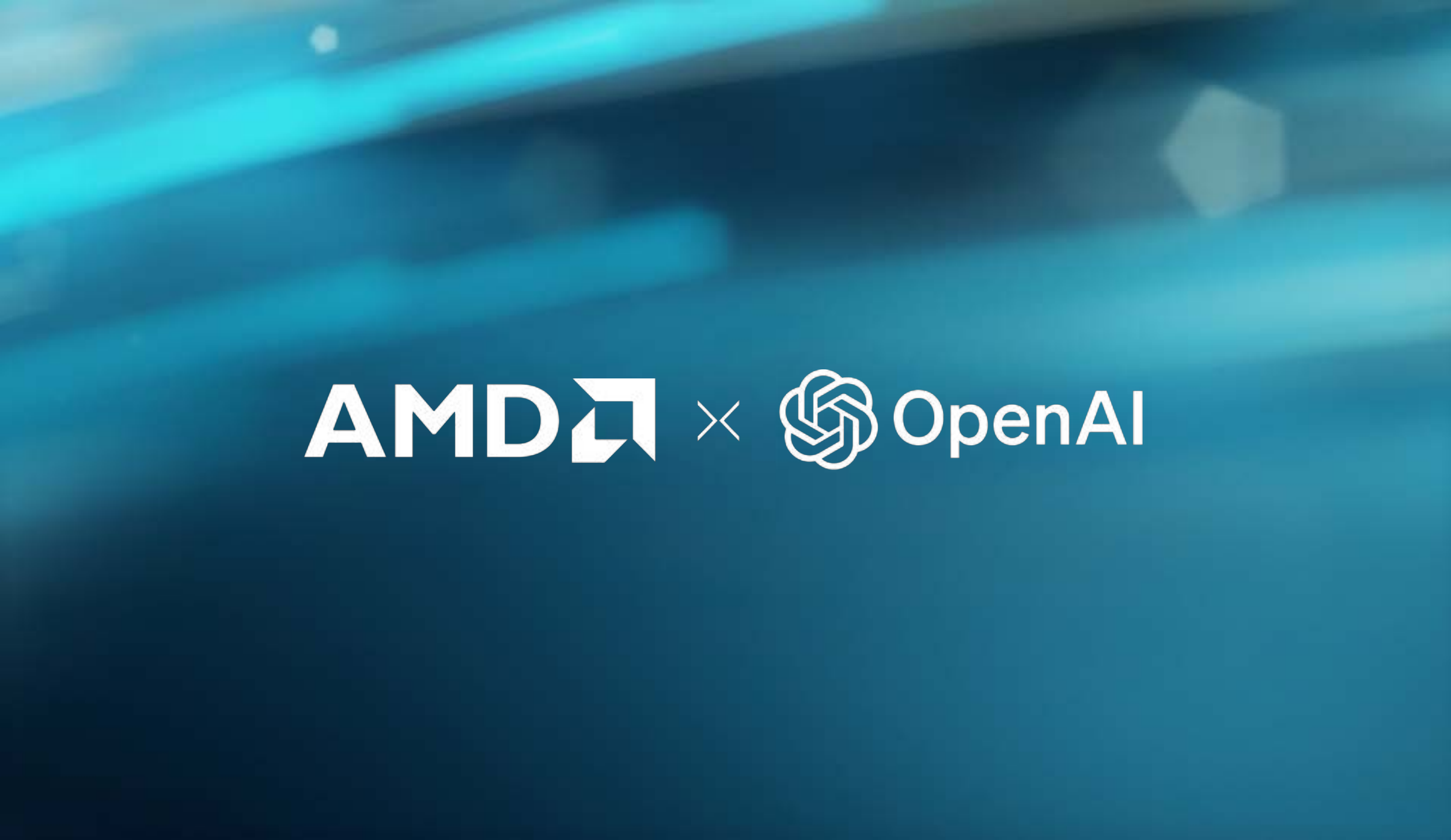
AMD Helios
Adopted by Leading AI Companies

AT SCALE AI PARTNERS

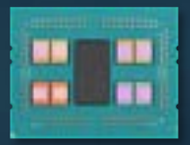
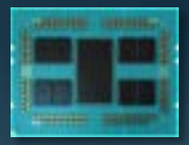
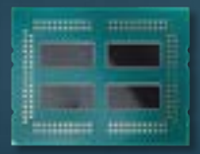
OpenAI
Meta
ANTHROPIC
Microsoft
ORACLE

INFRASTRUCTURE PARTNERS

HPE
Lenovo
SUPERMICRO
Bull
SANMINA
wiwynn
HUMAIN
IBM
TENSORWAVE
VULTR
Crusoe
Cirrascale
BROADCOM
cisco
AsteraLabs
Celestica
Aligned
OneQode


AMD
×
OpenAI

EPYC: 5 Generations of The World's Most Powerful Server CPUs

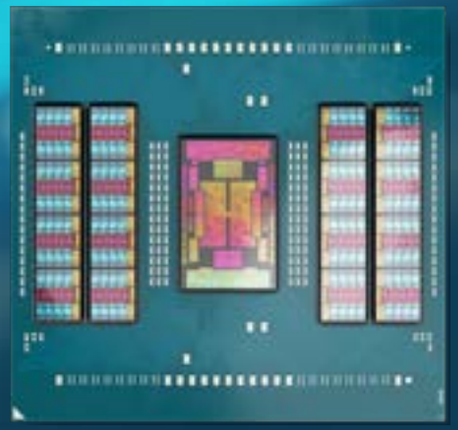
5th Generation **"Turin"**		AI OPTIMIZATION
4th Generation **"Genoa"**		PURPOSE-BUILT DESIGN POINTS
3rd Generation **"Milan"**		CORE PERFORMANCE LEADERSHIP
2nd Generation **"Rome"**		>2X THREAD DENSITY
1st Generation **"Naples"**		1ST "ZEN" GENERATION

EPYC Delivers

#1

- Core Performance
- Socket Performance
- Thread Count
- Performance/Watt
- Performance/Dollar

See Endnote: 9xx6-004a, 9xx6-047, EPYC-025D



EPYC “Turin” The Best Server CPU in Production Today

PERFORMANCE LEADERSHIP	DENSITY LEADERSHIP	I/O BANDWIDTH LEADERSHIP
#1 Performance	Up to **384 Threads** 192 Cores	PCIe® GEN 5 @ **32Gbps**

See Endnote: 9xx5-266

Three Tiers of Agentic AI Infrastructure Driving CPU Demand

GPU SERVERS
"HOST NODE"

High frequency cores with leadership I/O

AGENT CPU SERVERS
"SANDBOXES"

Maximum # of performant cores per watt

GENERAL PURPOSE
CPU SERVERS

Diverse balance of performance, core counts, power, and cost

INTRODUCING

6th Gen AMD EPYC “Venice”

World’s Best Server CPUs for the Agentic Era

See Endnote: EPYC-067

6th Gen AMD EPYC “Venice”

World’s Best Server CPUs for the Agentic Era

PERFORMANCE LEADERSHIP	DENSITY LEADERSHIP	MEMORY BANDWIDTH LEADERSHIP	TECHNOLOGY LEADERSHIP
#1	**512 Threads**	**1.6TB/s**	**“Zen 6”**
			CPU Architecture
1.8x	**1.3x**	**2.6x**	**2nm | 6nm**
Est. SPECint® 2026 vs Turin	vs Turin	vs Turin	Advanced Process

See Endnote: 9xx6-001a, 9xx6-005a, 9xx6-006a

EPYC “Venice”
The Best Server CPU Portfolio

GPU SERVERS
“HOST NODE”

“Venice” HF
RDIMM / MRDIMM

“Verano”
LPDDR

High Frequency
CPU to GPU Bandwidth

AGENT CPU SERVERS
“SANDBOXES”

“Venice” 256c

High Core Count
Low Power

GENERAL PURPOSE
CPU SERVERS

“Venice” SP7

High Performance
High Density

“Venice” SP8

Balanced
Performance & Power

“Venice-X”

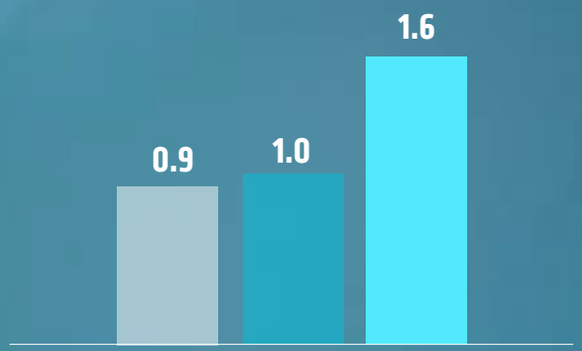
Technical
Compute & HPC

See Endnote: EPYC-068

AMD EPYC “Venice”
Delivers More Tokens, Agents & Performance vs Competition

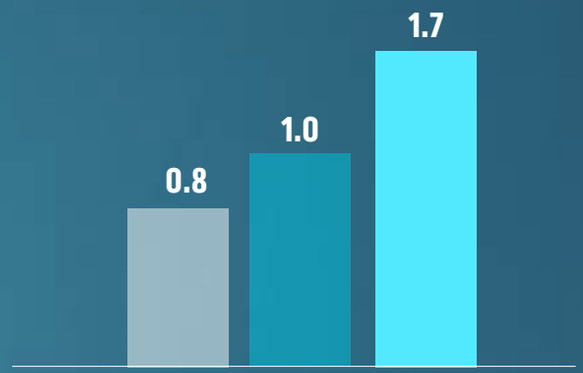
GPU SERVERS
“HOST NODE”

Up to **1.8x**
Tokens/second for Frontier Models



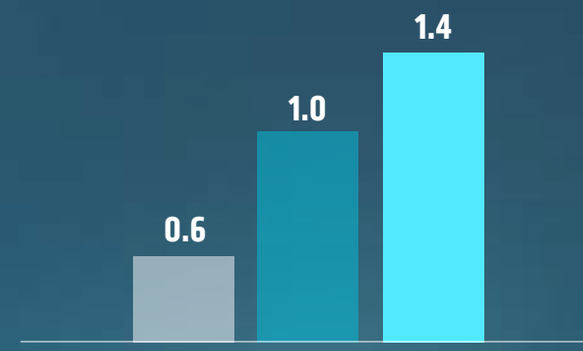
AGENTIC CPU SERVERS
“SANDBOXES”

Up to **2.1x**
Agents per Watt



GENERAL PURPOSE CPU SERVERS

Up to **2.3x**
Performance per Watt



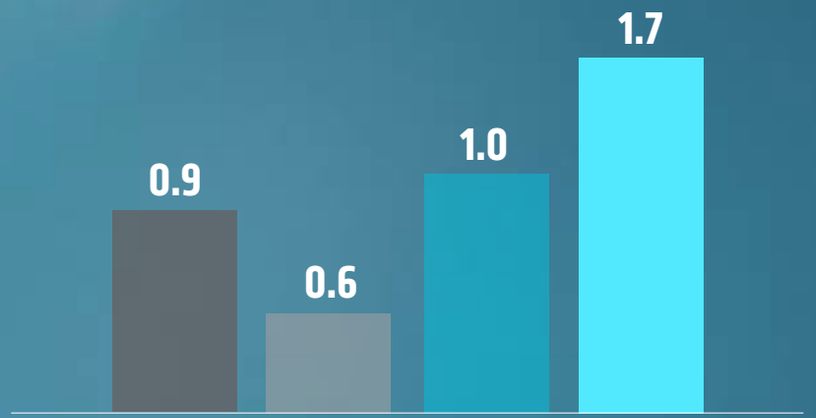
6th Gen Intel 5th Gen EPYC 6th Gen EPYC

See Endnote: 9xx6-022, 9xx6-004a, 9xx6-014

AMD EPYC “Venice”
Delivers More Agents & Performance vs ARM-based CPUs

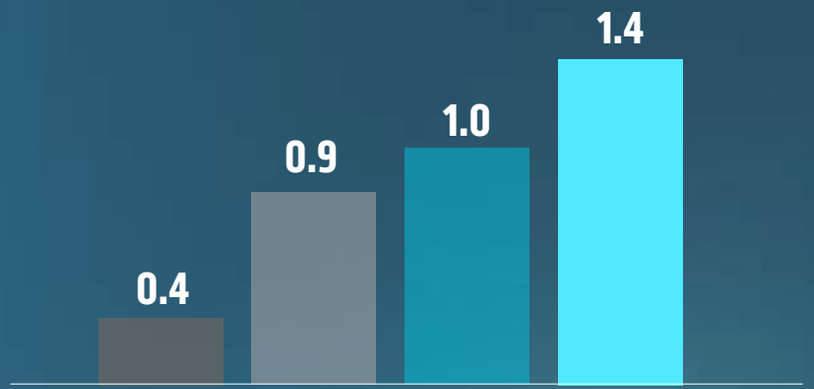
AGENTIC CPU SERVERS
“SANDBOXES”

Up to **2.8x**
Agents per Watt



GENERAL PURPOSE
CPU SERVERS

Up to **3.3x**
Performance per Watt



NVIDIA Vera ARM AGI 5th Gen EPYC 6th Gen EPYC

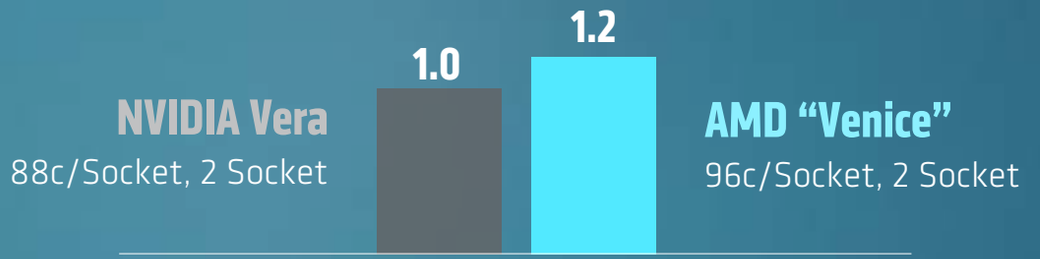
See Endnote: 9xx6-021, 9xx6-014

AMD EPYC “Venice” vs Nvidia Vera

Generational Lead Single Core Performance & Throughput

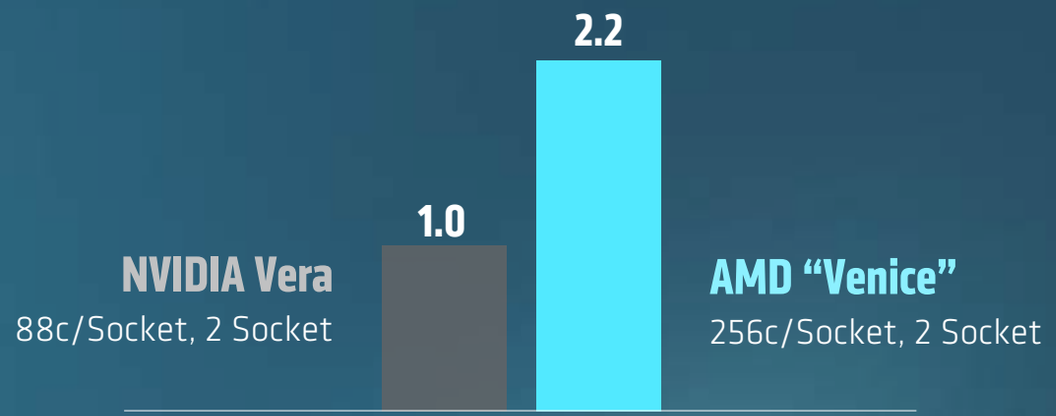
SINGLE CORE PERFORMANCE

Up to **1.2x**
vs Vera



THROUGHPUT PERFORMANCE

Up to **2.2x**
vs Vera

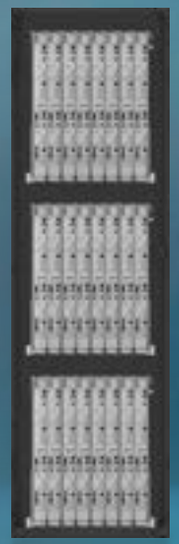


NVIDIA Vera 6th Gen EPYC

See Endnote: 9xx6-049, 9xx6-050

6th Gen AMD EPYC “Venice”
World’s Best Agentic CPU Racks

Lenovo



24,576
Cores

DELL Technologies

36,864
Cores

SUPERMICRO

43,008
Cores

GIGABYTE™

49,152
Cores

HPE

49,152
Cores

“Venice” 256c OEM

22,528
Cores

Vera 88c MGX

Vera Rack Image Source: NVIDIA Corporation | Media Assets (used for identification purposes only)

AMD EPYC “Venice”
In Production Today

HYPERSCALERS

Microsoft Meta aws Google ORACLE

OEM PROVIDERS

DELL Technologies SUPERMICRO HPE Lenovo

CLOUD INSTANCES & OEM PLATFORMS AVAILABLE STARTING IN Q4


AMD
Meta

Inference Market Is Segmenting

High Batch, Cost Optimized

- Offline Tasks
- Elastic Workloads
- Deep Research

Balanced Throughput & Latency

- Chatbots
- Copilots
- AI Assistants

Ultra-Low Latency

- Code Completion
- Real-Time Copilots
- Agentic Workflows

Disaggregated Inference Is Emerging
As A Powerful Way To Deliver Ultra Low Latency

INFERENCE WORKFLOW

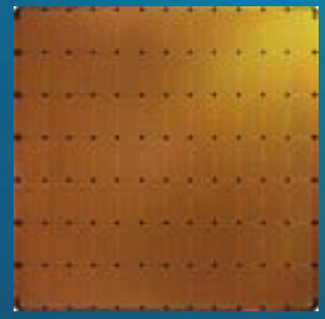

AMD
×
cerebras

The Most Powerful Solution For Ultra Low Latency Inference

High-Throughput Scalable Engine

432 GB HBM4 Capacity
23.3 TB/s Memory Bandwidth
72 GPU Domain

+

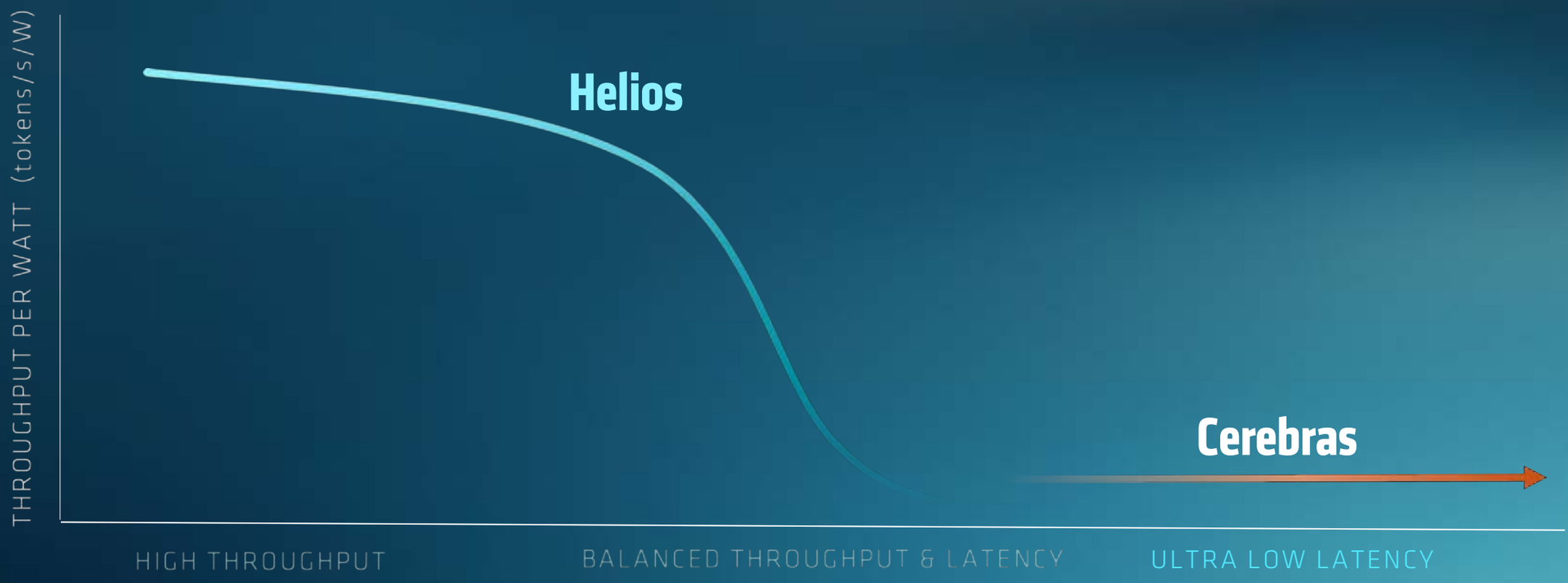


Wafer-scale, Ultra Low Latency Engine

44 GB on-chip SRAM
21,000 TB/s Memory Bandwidth
900,000 AI Cores

AMD × cerebras

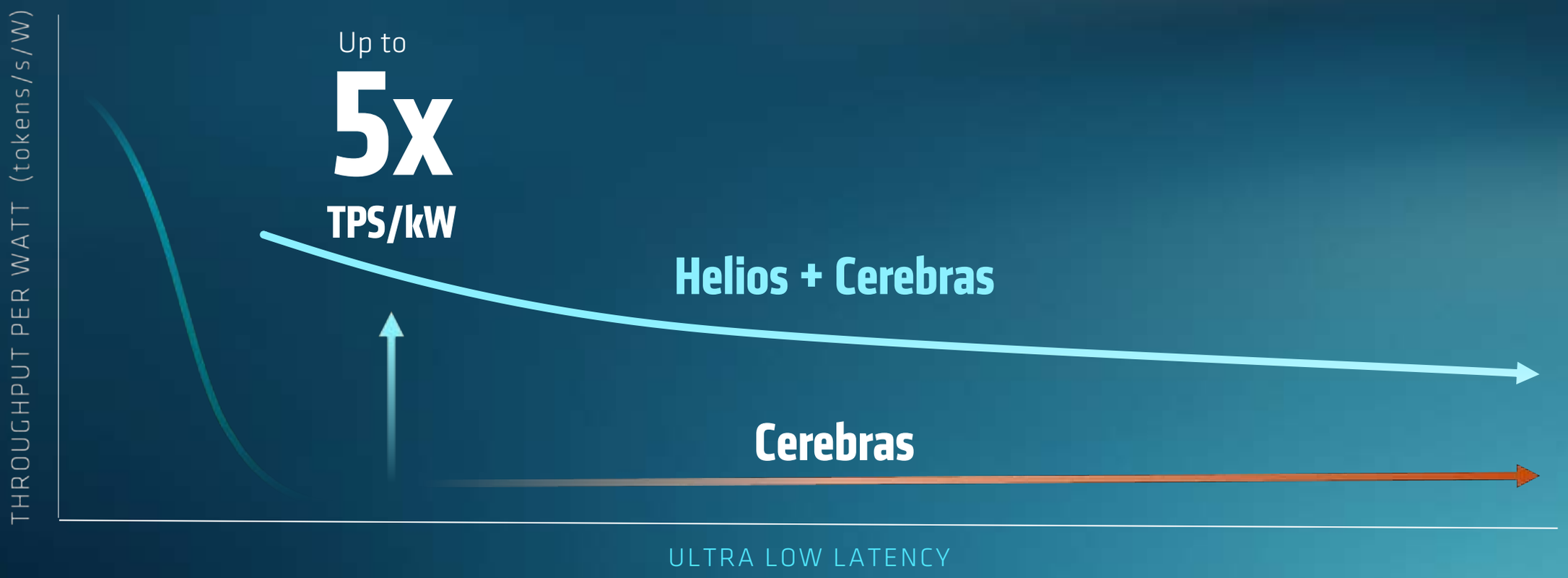
Ultra-Fast Inference, Higher TPS/W on Leading AI Models



See Endnote:MI400-021

AMD × cerebras

Ultra-Fast Inference, Higher TPS/W on Leading AI Models



See Endnote:MI400-021

Unlock Higher Throughput Ultra Low Latency Inference

Fastest Tokens In The World
For Leading AI Models

AMD HELIOS AI RACK

+

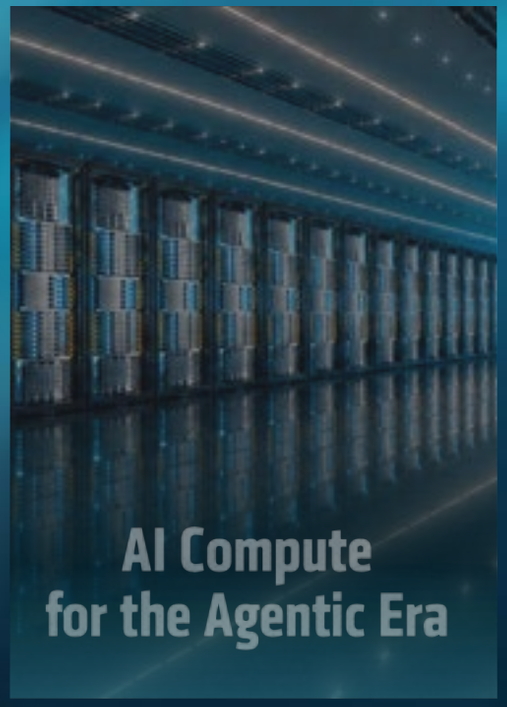
CEREBRAS

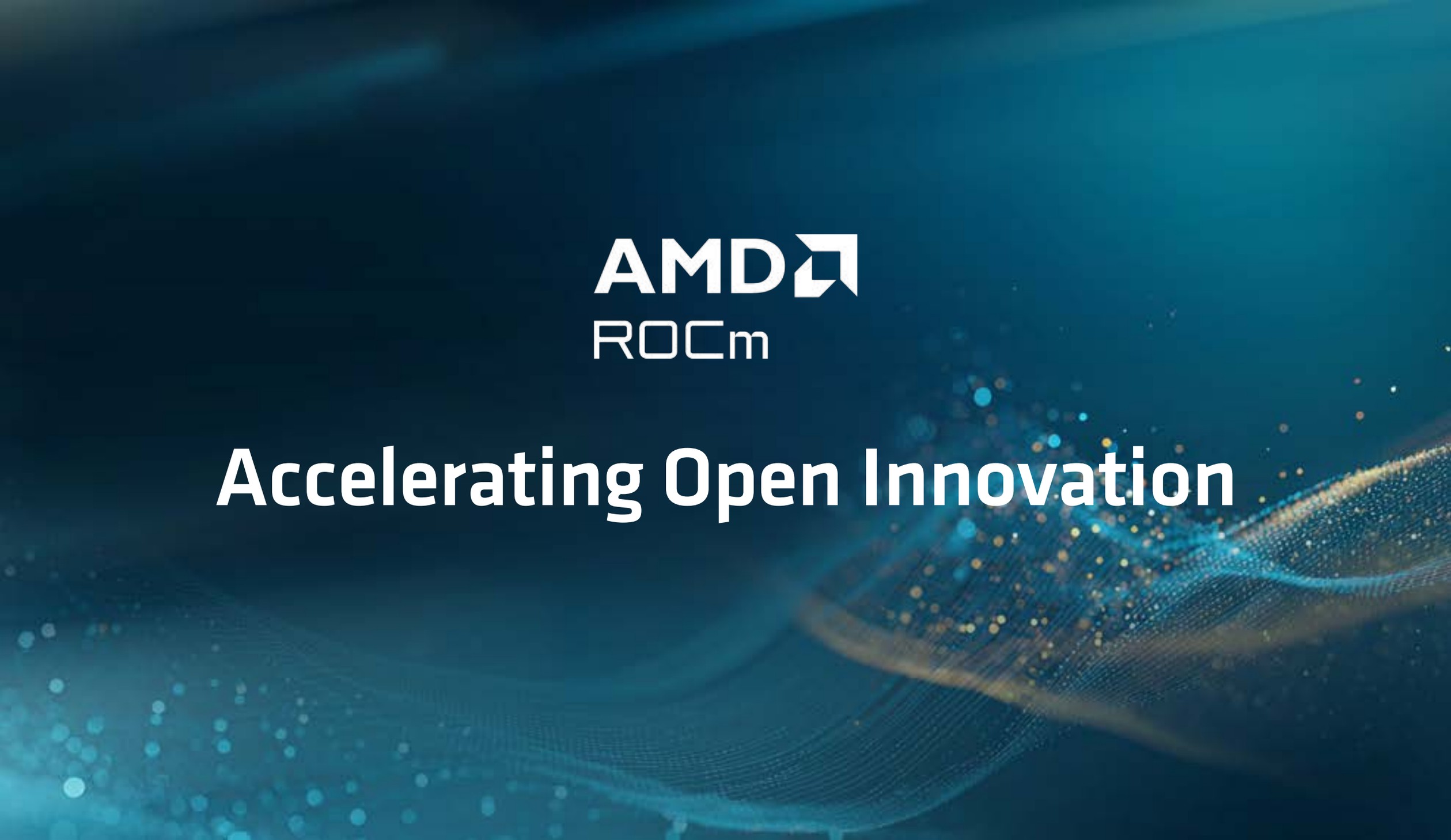
AVAILABLE IN 2H 2026
STARTING IN CEREBRAS INFERENCE CLOUD

See Endnote:MI400-023


AMD
×
cerebras

AMD | Advancing AI 2026




AMD
ROCm
Accelerating Open Innovation

Relentless Focus on Developers

Feature Velocity

Performance Optimization

Deep Open Ecosystem Partnerships

Richer Out-of-the-box Experience

3M+ Models on Hugging Face Running out-of-the-box

TOP 10 AI Open-Source Projects Natively Supported on AMD

>10x Increase in OSS Contributions

ROCm Strategy and Evolution

Open Source

Driving Rapid Innovation

Abstraction

Enabling Higher Productivity

Open Source Strategy Accelerating Progress

OPEN SOURCE COMMUNITY ENABLING AMD

PyTorch · Hugging Face · vLLM · SGL

TensorFlow · unsloth

LLaMA.cpp · llm-d · LMCache

Comfy · Mooncake · ONNX · VERL

WORLD'S MOST IMPORTANT WORKLOADS RUN ON AMD

GPT-OSS · MINIMAX · deepseek

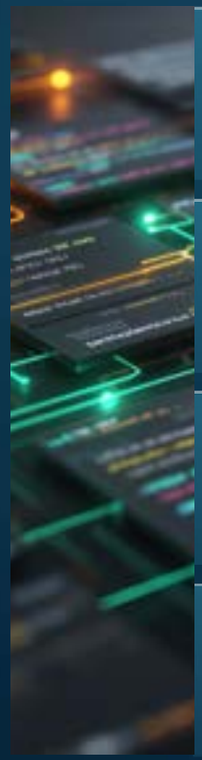
KIMI · LLaMA (Meta) · Gemma · GLM

Xiaomi MIMO · Tencent Hy · Qwen

Mistral · PHI-4 · Granite

Abstraction Makes Deployment Easy on AMD

Program at Every Level. Leveraging Rich Ecosystem



FRAMEWORKS	Rapid innovation at the highest level	PyTorch · JAX · vLLM · SGLang · TensorFlow · ATOM
PYTHONIC DSLS	Express intent in Python	OpenAI Triton · Gluon · TileLang · FlyDSL
LIBRARIES	Performance with productivity	AITER · hipBLASLt · HipKittens · RCCL · MoRI
LOW-LEVEL PROGRAMMING	Maximum control and performance	HIP C++ · AMDGPU Assembly

The Next Evolution:
“AI Assisted” GPU Programming

AI Assist

Automating GPU Programming

Open Source

Driving Rapid Innovation

Abstraction

Enabling Higher Productivity

Introducing AMD ROCm.AI
AI-Driven Platform for Developers

AMD
ROCm.AI

AI-Driven Development Platform

Accelerating development velocity and performance optimization

AI SKILLS Popular agents become ROCm superusers	**Skills**	Claude · Codex · Cursor · Gemini
AI ASSISTED OPTIMIZATION AI that optimizes workloads	**Hyperloom**	End-to-end AI Workload Optimization
ROCm CORE Compilers, tools, libraries, frameworks	**Framework Integration**	vLLM · SGL
	Core SDK	Libraries Compilers and Tools Runtime

Accelerating Inference with ROCm.AI

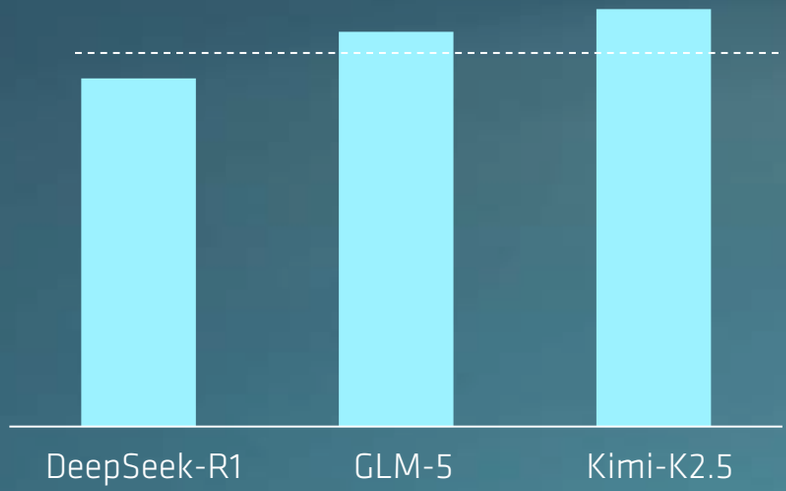
Parallelization & Scheduling
Expert parallelism, DP-attention,
Compute communications overlap

Memory Management
Paged attention, KV cache quantization,
Unified attention (prefill + decode)

Optimized Kernels
MHA prefill, MLA decode, Block-scale fused MoE

RELEASE-TO-RELEASE IMPROVEMENT
INFERENCE

3.3x
average performance improvement



ROCm AI vs ROCm 7

See Endnote: MI350-081

Accelerating Training with ROCm.AI

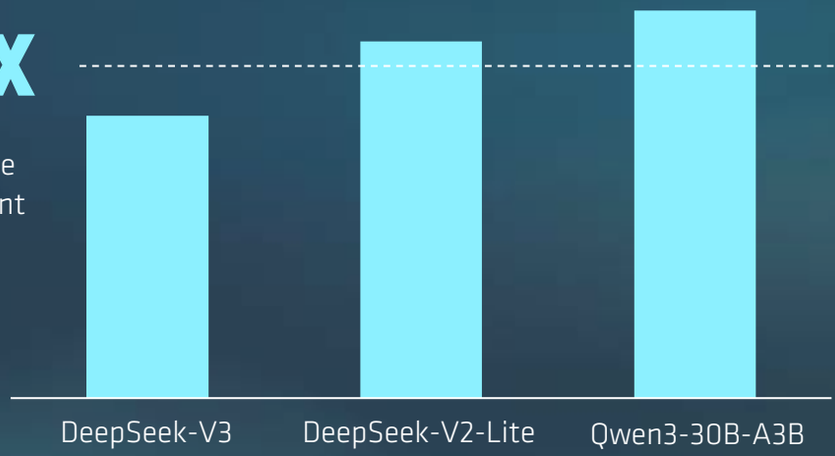
Parallelization & Scheduling
Pipeline parallelism, Sharded data parallel,
Gradient AllReduce overlap

Memory Management
Activation checkpointing, FP8 mixed precision,
Fused gradient accumulation

Optimized Kernels
Fused FlashAttention (fwd + bwd),
Grouped GEMM, Fused optimizer step

RELEASE-TO-RELEASE IMPROVEMENT
TRAINING

2.4x
average performance improvement



ROCm AI vs ROCm 7

See Endnote: MI350-082

ROCm.AI Delivers Leadership Performance on MI455X

MEMORY	COMPUTE	SCALE UP NETWORKING	SCALE OUT NETWORKING
20 TB/s	**20 PF**	**3.2 TB/s**	**190 GB/s**
Measured MLA (Decode, FP8)	Measured FP4 Compute	Measured Bandwidth	Measured Bandwidth
3.8x	**3.3x**	**3.5x**	**2x**
More Performance	More AI Compute	More Scale Up Bandwidth	More Scale Out Bandwidth

MI455X VS MI355X

See Endnote: MI400-027, MI400-028, MI400-029, MI400-030

ROCm.AI Ready for
MI455X Day 0 Support

PyTorch

"PyTorch Core tests run successfully on MI455X, and full support is on track."

- Jana van Greunen, Engineering Director at Meta, PyTorch

Hugging Face

"We tested the MI455X and love what we're seeing. Models run out of the box, and 432GB handles 3x more concurrent users."

- Clem Delangue, CEO, Hugging Face

vLLM

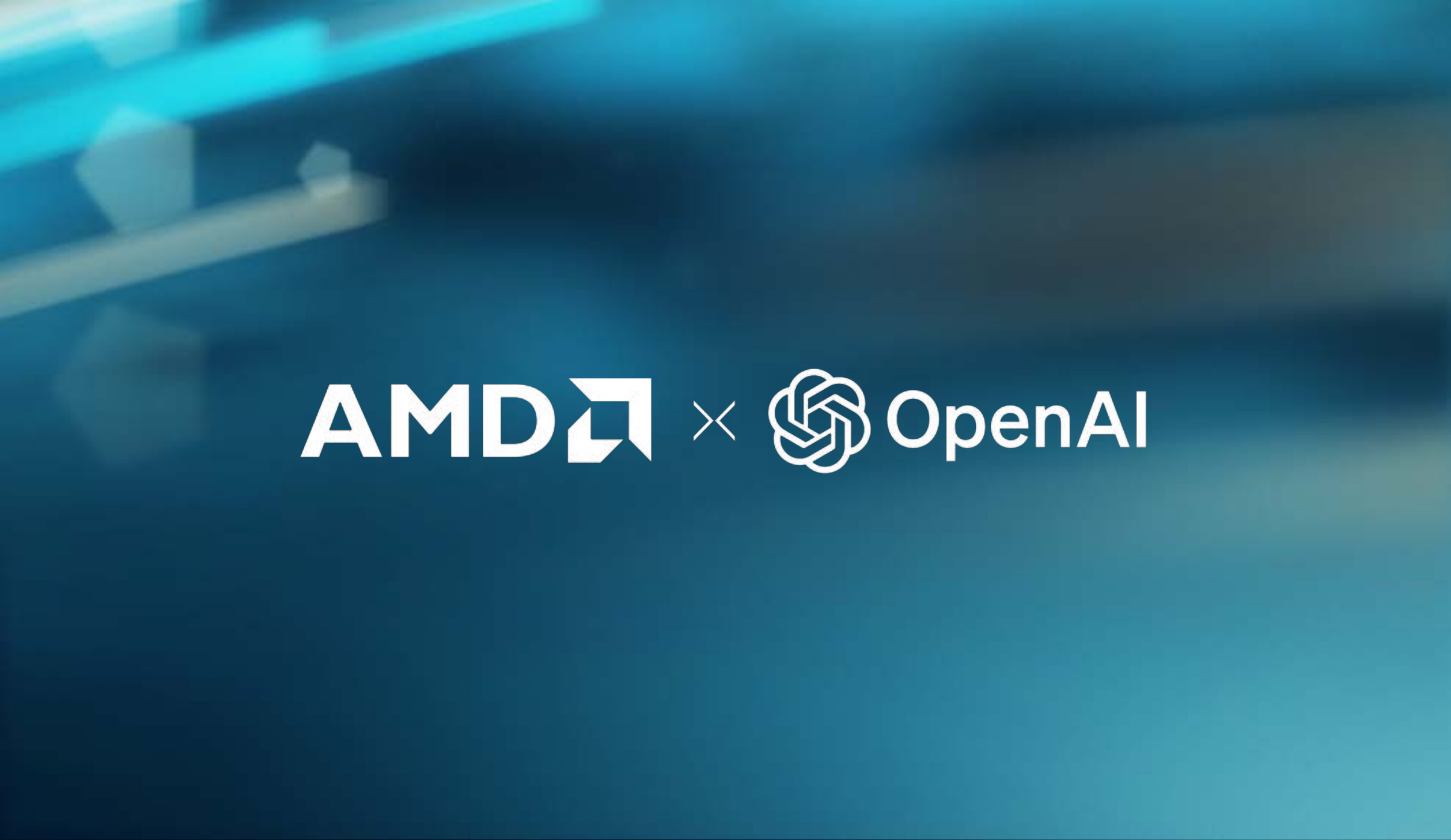
"vLLM is working well on Helios and we are excited to see what our community can build on top of it!"

- Simon Mo, CEO, Inferact

SGL

"We are thrilled to have SGLang enabled on day zero on MI455."

- Ying Sheng, CEO, RadixArk


AMD
×
OpenAI

AI

Training, Inference, Intelligence

SCIENCE/HPC

Simulations, Discoveries, Impact

AMD

ROCm.AI

Open Software Stack Built for Every Workload

Leadership Platforms Powering AI & HPC
Require Breadth of Accelerated Compute

AT SCALE AI TRAINING & INFERENCE

AMD Instinct MI455X

#1
- AI COMPUTE
- HBM4 MEMORY
- NETWORKING PERFORMANCE

SOVEREIGN AI & HPC

AMD Instinct MI430X

#1
- HARDWARE BASED FP64
- HBM4 MEMORY CAPACITY
- HBM4 MEMORY BANDWIDTH

See Endnote: MI400-005
Source: www.amd.com/en/blogs/2026/agentic-ai-needs-rack-scale-cpu-performance-amd-epyc.html

SHIPPING IN 1H 2027

AMD Instinct™ MI430X

Most Advanced HPC & Sovereign AI Accelerator

HPC FP64 COMPUTE	HBM4 CAPACITY	HBM4 BANDWIDTH
288TF	**432 GB**	**23.3 TB/s**
8.7x vs Vera Rubin	**+50%** vs Vera Rubin	**+6%** vs Vera Rubin

See Endnote: MI400-001b, MI400-004

US and Europe's First Exascale Sovereign AI Factories Are Deploying MI430X

OAK RIDGE National Laboratory

Discovery

CINES GENCI

Alice Recoque

AMD
ROCm.AI

The Future of GPU Development

Open Ecosystem

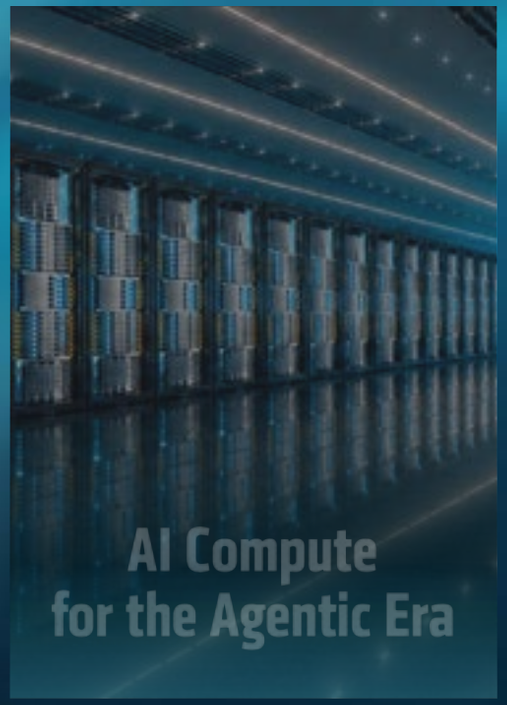
Higher Level Abstractions

AI-Assisted Development

ROCm Everywhere

ROCm.AI AVAILABLE AUGUST 2026

AMD | Advancing AI 2026



Powering AI Everywhere

ENTERPRISE AI

Enterprise
Data Center

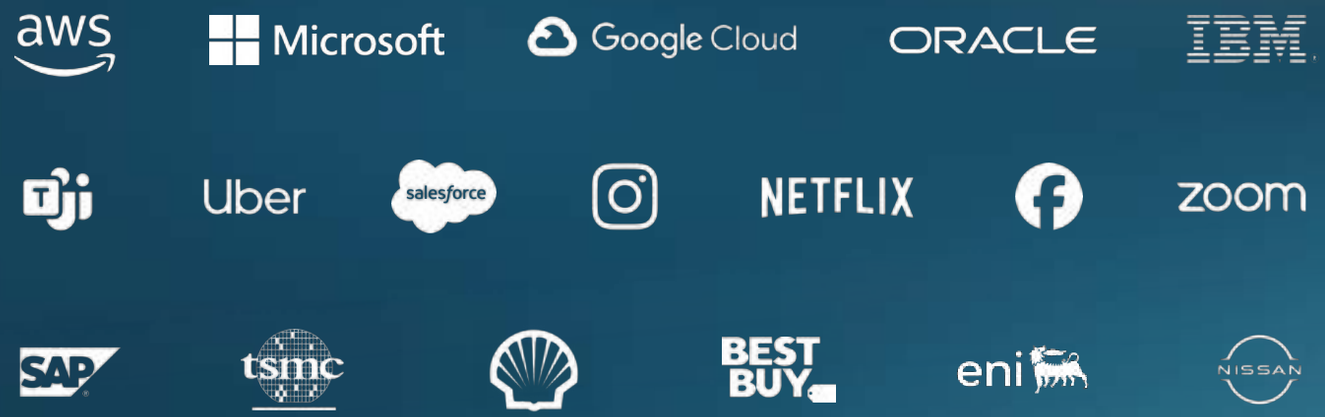
PERSONAL AI

Individual Productivity
& Development

PHYSICAL AI

Robotics, Autonomous Vehicles,
Industrial AI, Healthcare

EPYC is the Established Leader for Enterprises



60% of Fortune 100 Companies Trust Their Data Center to AMD EPYC™ CPUs

76% Cloud VM Adoption CAGR

8/10 Top Financial Companies Use EPYC

“AMD Is the Company to Beat for Enterprise AI Server CPUs”

Gartner®

Source: Gartner, AI Vendor Race: AMD Is the Company to Beat for Enterprise AI Server CPUs [Accessible to Gartner Subscribers], by Warren Peng & Gaurav Gupta, June 10, 2025

Enterprise Workloads are Diverse

MEDIA PROCESSING	REAL TIME SERVICES	WEB APPLICATIONS	BUSINESS APPLICATIONS	DATA ENCRYPTION
				
Thread Count	Memory Bandwidth	I/O Performance	Cache Efficiency	Core Performance

“Venice” Extends Enterprise Leadership

MEDIA PROCESSING	REAL TIME SERVICES	WEB APPLICATIONS	BUSINESS APPLICATIONS	DATA ENCRYPTION
Up to **2.7x**	Up to **2.9x**	Up to **3.7x**	Up to **2.6x**	Up to **2.5x**
FFmpeg	Redis	NGINX	MySQL	OpenSSL

6th Gen Intel 6th Gen EPYC

See Endnote: 9xx6-036, 9xx6-037, 9xx6-038, 9xx6-039, 9xx6-040

Agentic AI is the Force Multiplier for Enterprise But it Comes with New Challenges for CIOs

95%

of enterprises surveyed
expect to use agentic AI by 2027

CIO Considerations

- Cost
- Data Governance
- Integration

Source: Deloitte Insights

Agentic AI Deployment Will Be Distributed to Address These Challenges

FRONTIER	CLOUD	ON PREM	CLIENT
			

Agentic AI Deployment Will Be Distributed to Address These Challenges

ON PREM

CONSTRAINTS

- Power & Cooling
- Cost
- Ease of Deployment

AMD Instinct MI350P

The Easiest Way to Add AI to Your Data Center

PERF/$ LEADERSHIP	LARGE MODEL CAPABLE	STANDARD SERVER COMPATIBLE
Up to **4.2x**	Up to **260 Billion**	**Air Cooled**
Tokens/second/$ vs RTX Pro 6000	Model parameters @ FP4	450-600W TBP

See Endnote: MI350P-007

AMD Instinct MI350P
Delivering Higher Productivity

Tokens/Second vs RTX Pro 6000

WORKFLOW AUTOMATION Llama 3.3 70B Instruct	**5.1x**
CUSTOMER SUPPORT Mistral Small 3.2 24B	**2.5x**
CONTRACT ANALYSIS GPT-OSS 120B	**2.1x**

See Endnote: MI350P-003, MI350P-006, MI350P-001

Case Study
AMD IT Optimizes Costs by Routing to the Right Model

24/7 Autonomous Threat Detection

Personal Agents (OpenClaw)

TOKEN ROUTER

Frontier

On Prem EPYC+MI350P
Qwen 72B

On Prem EPYC+MI350P
Gemma 31B

43%
Cost Savings
vs 100% Frontier tokens

2.9x
Faster Responses
vs Frontier

See Endnote: MI350P-005

A Complete Ecosystem for Enterprise AI


AMD
AT&T

AMD × AT&T

Delivering AI at Scale

>45B	#1	122
Tokens Per Day	Leadership on 7 Key AI Benchmarks	Production RAG+FT Pipelines
40%	#1	5x
Specialized Models on AMD	Announcing OTel 2.0 Trained on AMD	In-year Free Cash Flow Impacting ROI

See Endnote: GD-181a

AMD Delivers
End-to-End Solutions for Enterprise AI

CLOUD

Cloud Instances

ON PREM SERVER

Full Racks

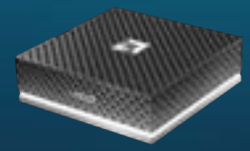
Instinct PCIe Cards

OEM Servers

CLIENT

Ryzen AI Laptops

Agent Computers



Radeon Pro PCIe Cards

AI Agents Are the Next Great Productivity Multiplier

PERSONAL AGENT

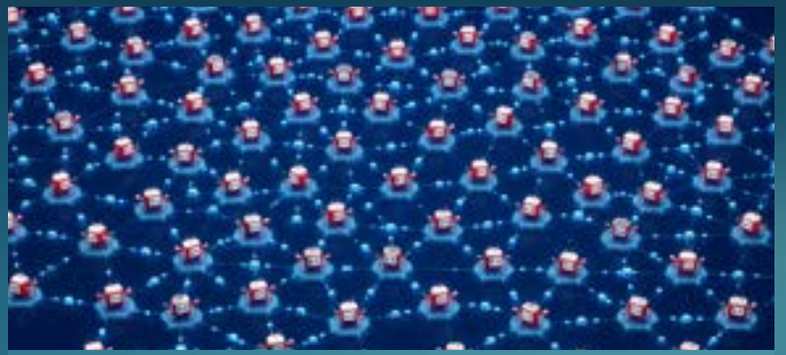
2x

Force Multiplier

AGENT TEAM

10x

Force Multiplier



MULTI-CAPABILITY AGENTS

100x

Force Multiplier

Directional figures for illustrative purposes only

Why Local AI is Taking Off

SECURITY & PRIVACY

Full control over data and no third-party handling

“FREE TOKEN” INFERENCE

Optimize recurring inference costs

RELIABLE ANYWHERE

Continue working without connectivity

Small Models are Becoming Incredibly Efficient

GPT-OSS

Number of Parameters **120B**

GPQA Score*
Graduate-Level Google-Proof Q&A **80.1**

August 2025

Qwen 3.5

Number of Parameters **9B**

GPQA Score*
Graduate-Level Google-Proof Q&A **81.7**

March 2026

Source: LLM Stats, May 20, 2026 (https://llm-stats.com/models/compare/qwen3.5-9b-vs-gpt-oss-120b)

Smaller Models are Closing The Frontier Gap

Smaller open models can now outperform yesterday's frontier models while using a fraction of the cost.

Source: LLM Stats, May 20, 2026 (https://llm-stats.com/models/compare/claude-opus-4-5-20251101-vs-qwen3.6-27b)

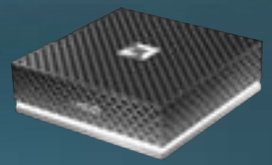
Efficient Models Unlock Personal AI

Small Efficient Models Can Fit Easily Across AMD Agentic PCs

Ryzen™ AI 400

9B
Model Size

24B
Model Size Limit



Ryzen™ AI MAX

27B
Model Size

200B
Model Size Limit

RYZEN AI HALO

DEVELOPER PLATFORM

128GB Unified Memory

200B Models Natively

+ Speed to Productivity
+ Simplicity of Setup
+ Efficiency of Cost
+ Continuous Developer Support

One Software Stack. Every AMD Device.
Write Once. Run Anywhere.

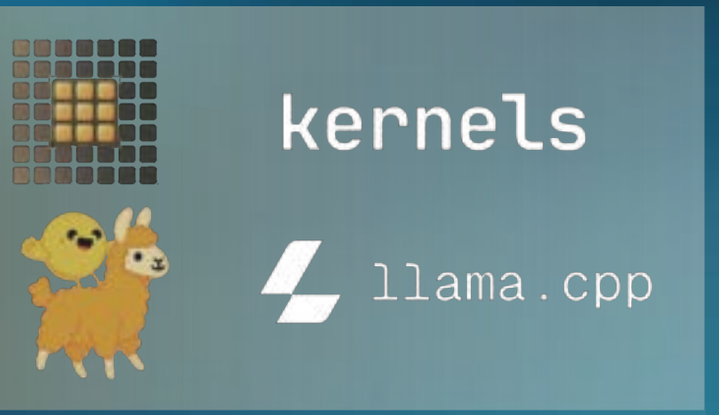
AMD ROCm

Agent Frameworks	Lemonade · OpenClaw · LM Studio · Ollama · CrewAI
AI Frameworks	PyTorch™ · JAX · vLLM · ONNX
ROCm Libraries	rocBLAS · MIOpen · hipBLAS · Triton ROCm™
Runtime & Compiler	HIP · HIPCC · ROCm-CPU
AMD Silicon	Ryzen™ AI · Radeon™ AI PRO · Instinct™ MI · EPYC™

AMD × Hugging Face

Deepening our Partnership with the World's Premier Open-Source AI Community

Native Halo support in Hugging Face with optimized models and workflows



Co-engineered Halo Local AI experiences + Toolkits to Accelerate Development

Every Halo Box Ships with 1 Year Hugging Face PRO

Pushing inference further with

“GORGON HALO”

Largest unified memory

from 128GB to 192GB

Extending AI models

from 200B to 300B

Run cloud-scale AI

Locally on Your Desktop


AMD
CISCO

Agents Operate at Machine Speed

Bursty – demand spikes, then idles

INFRASTRUCTURE DEMAND

CHATBOTS

Steady state – always on, at machine speed

INFRASTRUCTURE DEMAND

AGENTIC & PHYSICAL AI


AMD
CISCO

AMD × CISCO

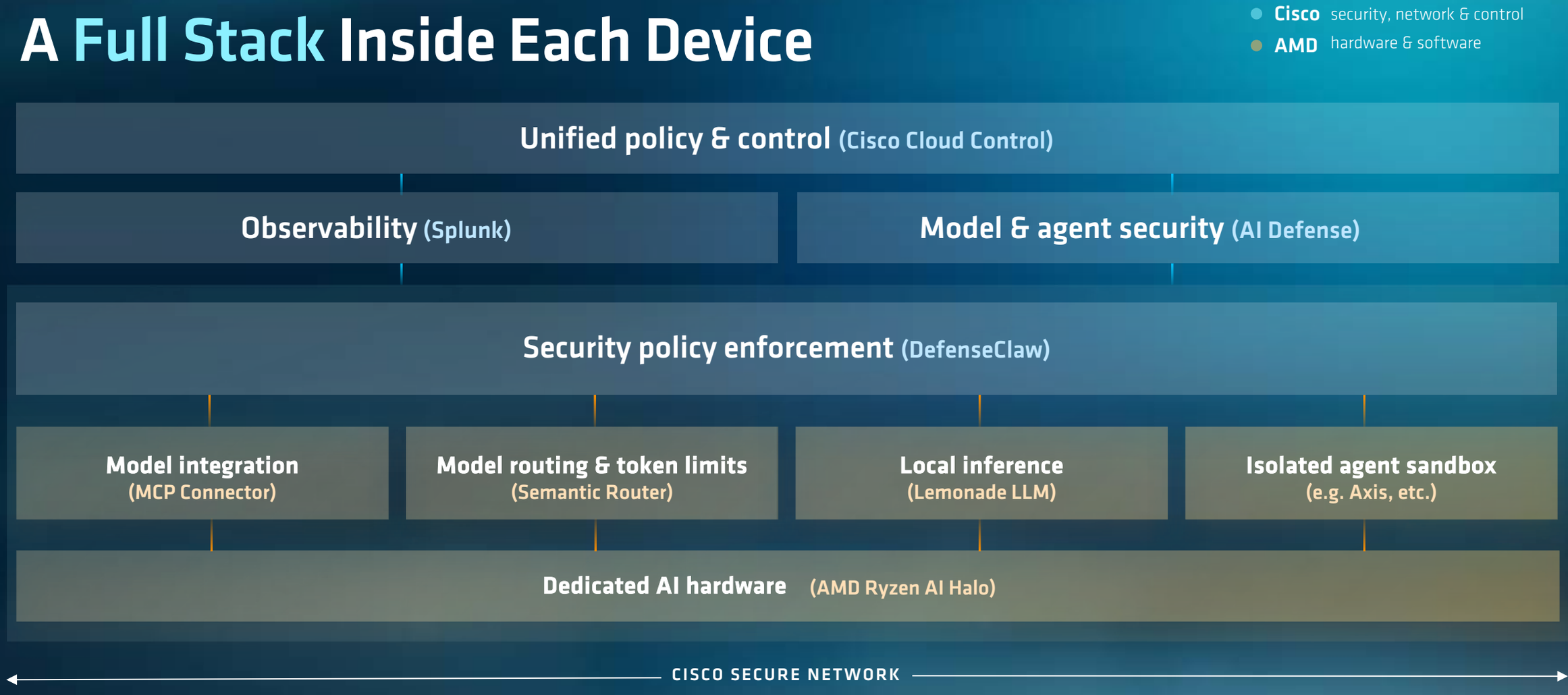
NETWORK INFRASTRUCTURE

TOKENOMICS

AGENT BEHAVIOR

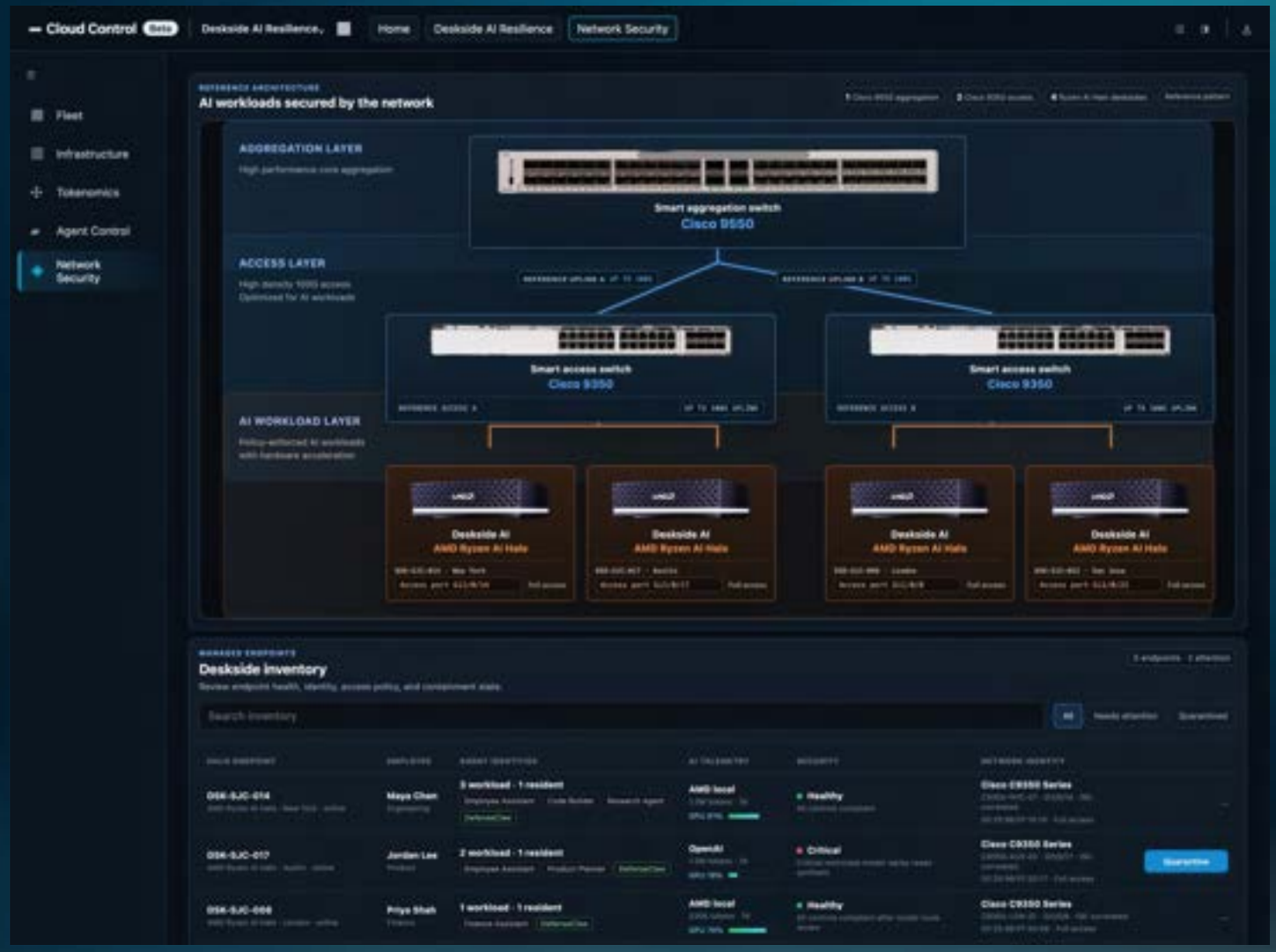
SECURITY

A Full Stack Inside Each Device



Security & Governance for Agentic AI

- Limit agent access to approved resources
- Apply consistent security policies
- Identify and isolate activity in real time



Intelligence
for Enterprise Inference

- Gain visibility into AI costs and savings
- Improve model and token efficiency
- Identify adoption and usage trends


AMD
CISCO

Powering AI Everywhere

ENTERPRISE AI

Enterprise
Data Center

PERSONAL AI

Individual Productivity & Development

PHYSICAL AI

Robotics, Autonomous Vehicles,
Industrial AI, Healthcare

Physical AI

The Ultimate Application of Agentic AI

Physical AI that Amplifies Human Potential

From today's deployment to tomorrow's possibilities

ROBOTIC SURGERY

SMART MANUFACTURING

INTELLIGENT AGRICULTURE

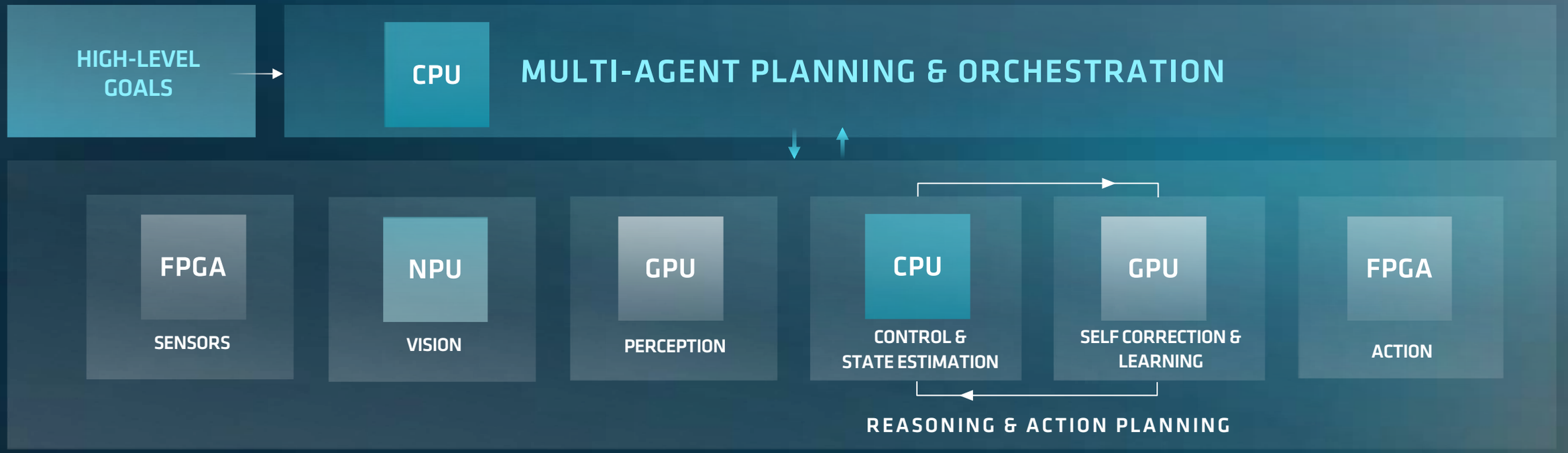
WAREHOUSING

FACTORY AUTOMATION

MULTI-PURPOSE


AMD is a Trusted Leader in Robotics
Enabling sensing, safety, and control systems for over 20 years
ABB
CASTEC INTERNATIONAL
DEXORY
FANUC
FOUNDATION
GENERATIVE BIONICS
INTUITIVE
Kawasaki
Muks Robotics
NUWA ROBOTICS
OMRON
REV ROBOTICS
Rexroth Bosch Group
Robotnik
SICK Sensor Intelligence
UNIVERSAL ROBOTS

The Next Era of Robotics is Autonomous

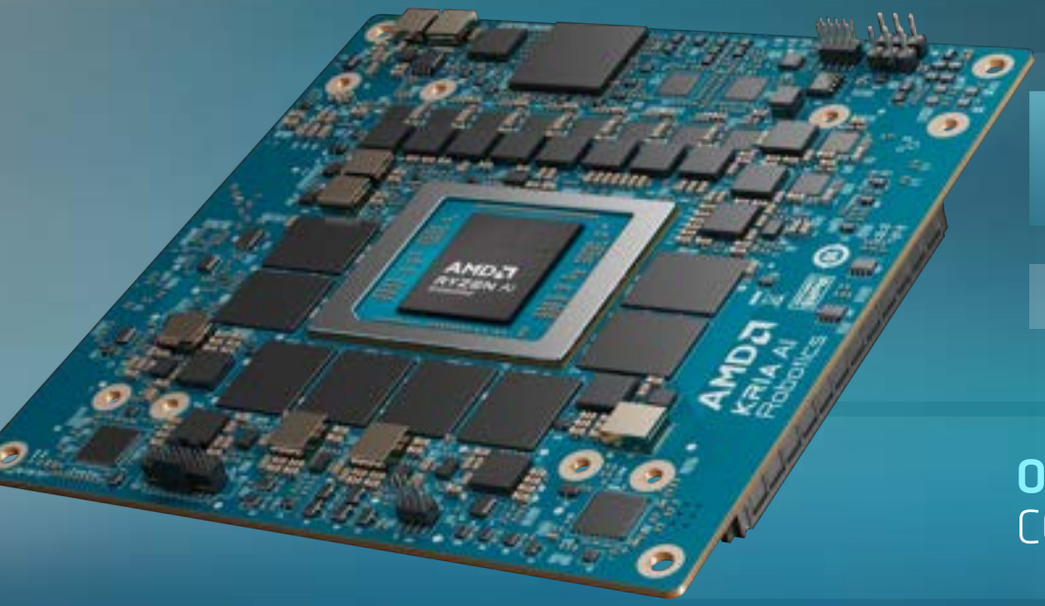


Plans, reasons, learns and adapts to achieve desired outcomes

INTRODUCING

AMD KRIA™ AI System on Module

Powered by Ryzen™ AI Embedded X100 Series



INDUSTRY-LEADING SYSTEM PERFORMANCE

CPU	**125 µs** Deterministic Control	
GPU	**92 ms** AI Reasoning	
NPU	**<0.4 ms** Vision Classification	

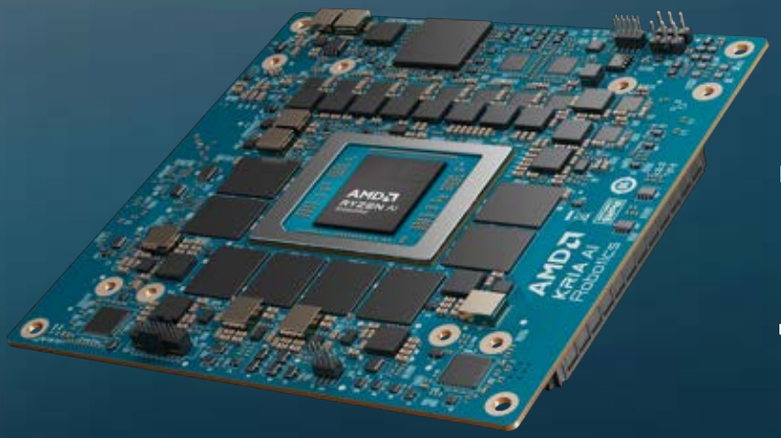
Unified Memory

OPEN STANDARDS
COM-HPC Form Factor

DEPLOY FASTER
Pre-engineered Platform

ctrlX OS | rexroth A Bosch Company | embedl EFFICIENT DEEP LEARNING

Results based on 3rd party and AMD internal testing. 8,000 Decisions per second (125 µs) obtained by benchmark testing by Bosch using a Bosch Rexroth controller on AMD X100 Series GPU to measure CPU real-time control loop latency. Robot reasoning in <100 msbenchmarking by embedl targeting GPU VLA inference latency (based on Pi0.5 VLA). Vision classification results based on AMD internal testing using mobilenetv3. Results provided by 3rd party organizations have not been independently verified by AMD. Performance and cost benefits are impacted by a variety of variables. Results herein are specific to such 3rd party organization and may not be typical



KRIA™ AI System On Module

The New Standard
for the Robotics Brain

REACTION TIME	AGENTIC AI CAPACITY	ROBOT CAPABILITY
3.4x	**2.3x**	**1.6x**
Better real-time results	More Concurrent agents	More CPU Capacity

VS NVIDIA JETSON THOR™

OPEN NAVIGATION mimik

Based on 3rd party benchmarks; REX-015, REX-018, REX-019

INTRODUCING

AMD KRIA™ AI Robotics Developer Platform

Industry's First Turnkey Open, Fully Integrated Platform

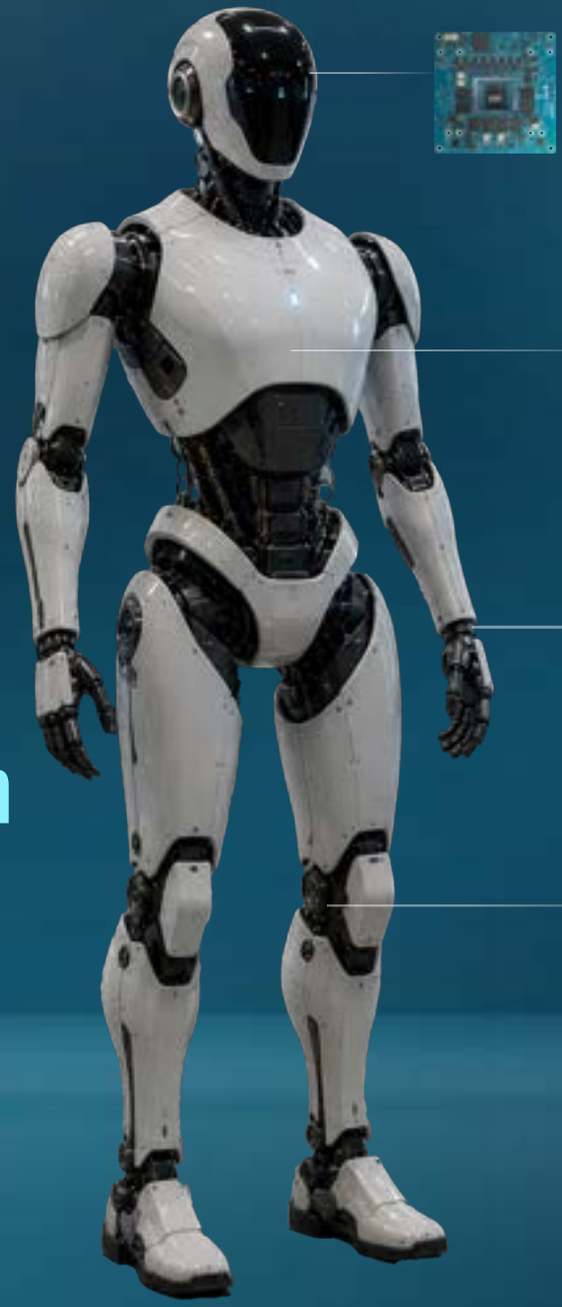
From concept to prototype in days

Enabled with Robotics SW Suite
built on AMD ROCm™ and ROS 2

Faster deployment with open
baseboard schematic & FPGA design

Only AMD Delivers the Complete Autonomous Robotics Solution

From Body To Brain



AMD KRIA AI Robotics

AMD VERSAL AI Edge Gen 2

AMD ZYNQ UltraScale+

AMD SPARTAN UltraScale+

BRAIN

High-level decision making, communication and human-machine interface

Optimized for inference throughput

SPINE

Time-critical communication coordination between brain and joints

Optimized for reliability

JOINTS

Real-time multi-axis actuation

Optimized for low-latency

SENSORS

Any sensor connectivity

Optimized for real-time sensor ingest and fusion

Advancing the Open Robotics Ecosystem

AI MODELS & SOFTWARE

Archetype AI
embedl
Liquid
mimik
Nota AI
OPEN NAVIGATION
open robotics
OpenCV
QNX
TRENER
ultralytics
Xaba

ROBOTIC TECHNOLOGIES

ANALOG DEVICES
AHEAD OF WHAT'S POSSIBLE™
machineering
MAKARENA LABS
multicoreware
ROBOTEC AI
rexroth
A Bosch Company
SICK
Sensor Intelligence.
Voyant
VVDN TECHNOLOGIES
World Wide Technology

HARDWARE PARTNERS

ARBOR
congatec
iBASE
iEi
kontron
SAPPHIRE
SEAVO 信步科技 SEAVO TECHNOLOGY

INDUSTRY GROUPS

digital twin CONSORTIUM
An EDM Association Community
mass robotics

Announced today at Advancing AI 2026

“Helios”

“Venice”

MI430X

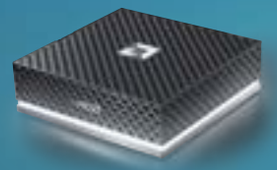
MI350P

ROCm.ai

AMD
ROCm.AI

“Gorgon Halo”

DEVELOPER PLATFORM



Kria AI Robotics

SOM & DEVELOPER PLATFORM

AMD EPYC
Leadership Server CPU Roadmap

"Turin"
5th Generation EPYC

"Zen 5"
"Zen 5c"

- Wider/Deeper Compute
- Full 512 Bit AI Vectors
- Re-optimized Cache Hierarchy

2024

"Venice"
6th Generation EPYC

"Zen 6"
"Zen 6c"

- First to 2nm & 512 Threads
- First to JEDEC MRDIMM
- First x86 to PCIe Gen 6

2026

"Florence"
7th Generation EPYC

"Zen 7"
"Zen 7c"

- Next Gen Process Node
- ACE: AI Compute Extensions
- Next Gen Memory: MRDIMM & LPDDR

2028

"Ravenna"
8th Generation EPYC

"Zen 8"
Family

In Development

2030

All roadmaps subject to change

COMING IN 2028

EPYC “Florence” Extends EPYC Leadership For the Agentic Era

“Florence”

SP7

PERFORMANCE
Leadership Core Count and Performance

“Florence”

SP8

ENTERPRISE
Optimized for Performance per System $

“Ferrara”

AI HOST NODE
Optimized for high-performance AI Host Node

“Fidenza”

AGENTIC “SANDBOX”
Optimized for Performance/System Watt

AMD Instinct
Leadership GPU Roadmap

MI350 Series

- Increased HBM3e Capacity & Bandwidth
- Block Scaled AI Formats
- Optimized Platform Solutions

2025

MI400 Series

- Increased HBM4 Capacity & Bandwidth
- Expanded AI Formats with Higher Throughput
- Standard-Based Rack-Scale Networking
  (UALoE, UEC)

2026

MI500 Series

- Next Gen HBM4e Memory
- Leadership GPU Scale Up Domain
- Copper And Optical Interconnect

2027

MI600 Series

- In Development

2028

*All roadmaps subject to change

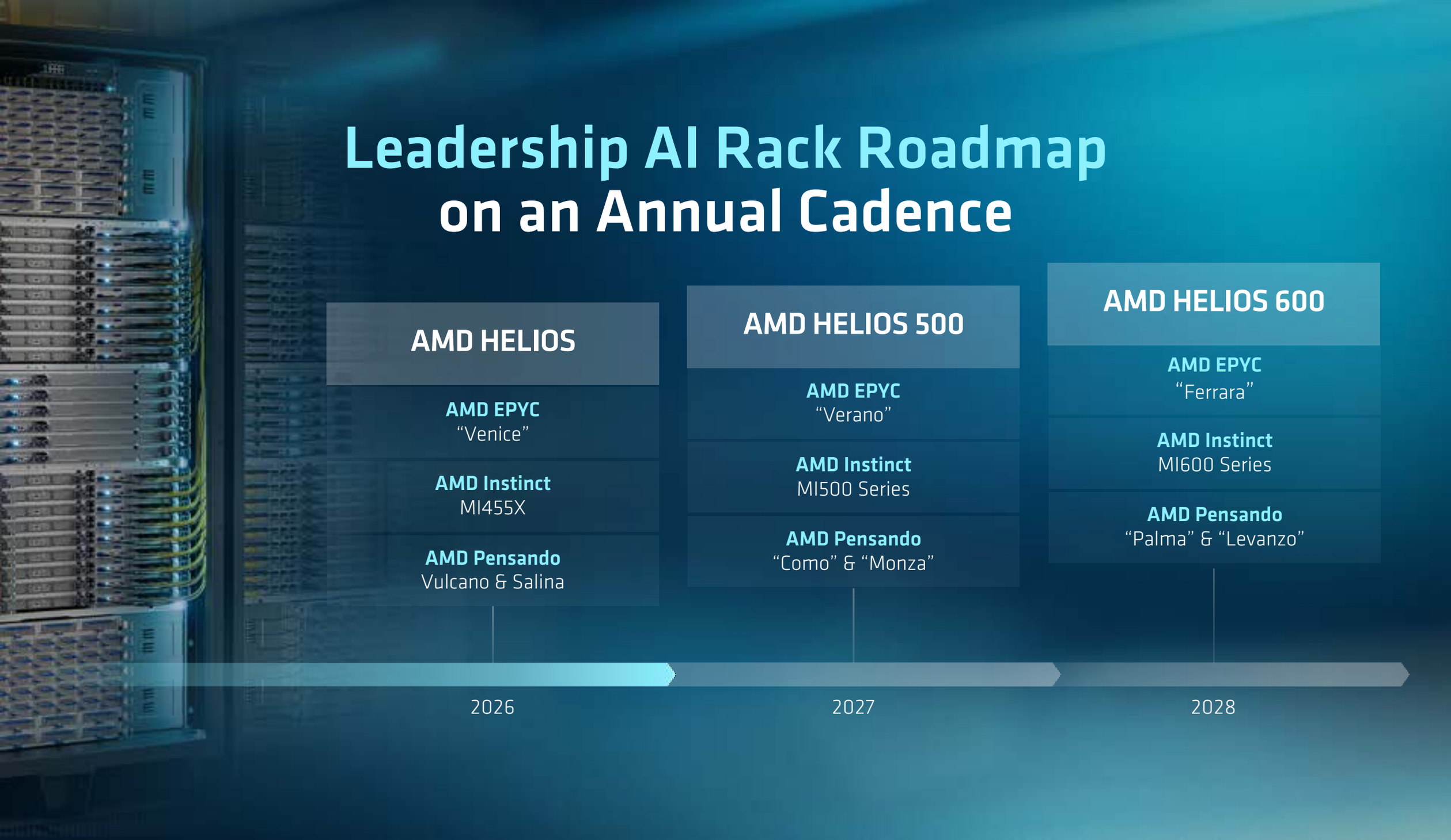
Instinct MI500 The Next Leap In AI Performance

INFERENCE THROUGHPUT

>2,000X

Increase in 4 Years

See Endnote: MI500-001


Leadership AI Rack Roadmap
on an Annual Cadence
AMD HELIOS
AMD EPYC
“Venice”
AMD Instinct
MI455X
AMD Pensando
Vulcano & Salina
AMD HELIOS 500
AMD EPYC
“Verano”
AMD Instinct
MI500 Series
AMD Pensando
“Como” & “Monza”
AMD HELIOS 600
AMD EPYC
“Ferrara”
AMD Instinct
MI600 Series
AMD Pensando
“Palma” & “Levanzo”
2026
2027
2028

Announced today at Advancing AI 2026

“Helios”

In Production Now

“Venice”

In Production Now

MI430X

Available 1H 2027

MI350P

Available Now

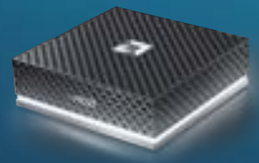
ROCm.ai

AMD
ROCm.AI

Available August

“Gorgon Halo”

DEVELOPER PLATFORM



Available 2H

Kria AI Robotics

SOM & DEVELOPER PLATFORM

Available Q4


AMD

Endnotes

9xx5-266: SPECrate®2017_int_base comparison based on published scores from www.spec.org as of 10/24/2025. 2P AMD EPYC 9965, 192C, 3230 SPECrate®2017_int_base, https://www.spec.org/cpu2017/results/res2025q2/cpu2017-20250324-47086.html SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

9xx6-001A: SPECrate®2017_int_base comparison based on AMD internal estimates for top of stack 2P 6th Generation EPYC CPU and 5th Gen EPYC measurements and pricing as of 07/22/2026. Preliminary performance estimates based on AMD engineering projections or measurements for respective product launch dates and subject to change. 2P AMD EPYC 9996, 512C total, 600W Default CPU Power, $14904, 16x64GB 12800 MT/s MRDIMM, AOCC 6.0, performance determinism, SPECrate®2017_int_base 4900, SPECrate®2017_int_base/CPU W 4.083, SPECrate®2017_int_base/CPU $ 0.329 2P AMD EPYC 9965, 384C total, 500W TDP, $ 11988, 12x64GB DDR5-6400, AOCC 5.0, performance determinism, SPECrate®2017_int_base 2740, SPECrate®2017_int_base/CPU W 2.740, SPECrate®2017_int_base/CPU $ 0.229 SPEC® and SPECint® are registered trademarks of Standard Performance Evaluation Corporation. Learn more at spec.org. Starting with the 6th Gen AMD EPYC™ server processor family, AMD uses Default CPU Power to describe processor power consumption, succeeding AMD's historical TDP reference. Default CPU Power reflects total power consumed across the processor's compute and I/O dies for the stated performance target. Default CPU Power and TDP may both serve as processor power references for product comparison, platform planning, and performance-per-watt analysis.

9xx6-004A: SPECrate®2017_int_base comparison based on AMD estimates for top of stack 2P 6th Gen AMD EPYC 9996 estimated SPECrate2017®_int-base score is 4900 (256C, 600W Default CPU Power, $14904 1Ku. Highest overall SPECrate2017®_int_base score as of July 22, 2026 is 3240 2P AMD EPYC 9965 (192C, 500W TDP, $11988 1ku) powered server, https://spec.org/cpu2017/results/res2026q2/cpu2017-20260518-51509.html). Highest published 2P Intel Xeon 6980P SPECrate2017_Int_base score as of July 22, 2026 is 2510 (128C, 500W TDP, $13955 1ku), https://spec.org/cpu2017/results/res2025q4/cpu2017-20251126-50522.html 2P Nvidia Vera performance based on AMD internal estimates as of July 22, 2026, estimated SPECrate2017_int_base score is 1459, (88C, 450W TDP, pricing not published), https://www.amd.com/content/dam/amd/en/documents/solutions/ai/methodology-description.pdf). ARM AGI performance based on AMD Internal estimates as of July 22, 2026, estimated SPACrate2017®_int_base score is 1944, (136C, 300W TDP, pricing not published) Preliminary performance estimates based on AMD engineering projections or measurements as of July 22, 2026 and subject to change. Pricing not published for Nvidia Vera, Intel pricing from ark.intel.com. Phoronix testing for Nvidia Vera at https://www.phoronix.com/review/nvidia-vera-benchmarks/. SPEC® and SPECint® are registered trademarks of Standard Performance Evaluation Corporation. Learn more at spec.org. Starting with the 6th Gen AMD EPYC™ server processor family, AMD uses Default CPU Power to describe processor power consumption, succeeding AMD's historical TDP reference. Default CPU Power reflects total power consumed across the processor's compute and I/O dies for the stated performance target. Default CPU Power and TDP may both serve as processor power references for product comparison, platform planning, and performance-per-watt analysis.,

9xx6-005A: vCPU support based on thread count, comparing Top of Stack 6th Gen EPYC 9996 CPU with 256 cores (512 threads) and 5th Gen EPYC 9965 with 192 cores (384 threads). In comparison, the highest thread count Intel® Xeon® processor is the Intel Xeon 6990E at 288 threads, and the highest thread count Nvidia CPU is the Nvidia Vera with 88 cores (176 threads). For reference, 6th Gen Intel Xeon CPUs support up to ~0.8x, ARM AGI is expected to support up to ~0.7x, and Nvidia Vera is expected to support ~0.5x the thread count of 5th Gen AMD EPYC CPUs. Intel specifications from ark.intel.com. ARM specifications from https://www.arm.com/products/cloud-datacenter/arm-agi-cpu#Specifications. Nvidia specifications from https://developer.nvidia.com/blog/nvidia-vera-cpu-delivers-high-performance-bandwidth-and-efficiency-for-ai-factories/.

9xx6-006A: 6th Gen EPYC CPUs can support up to 2.6X memory bandwidth (up to 16 channels of 12,800MT/s DDR5 for 1.6TB/s of memory bandwidth/socket compared to 614GB/s memory bandwidth/socket of 5th Gen EPYC CPUs (9005) with 12 channels of 6,400MT/s DDR5). 6th Gen EPYC CPUs can support up to 2X the IO bandwidth (PCIe Gen 6 supported at 64GT/s per lane, EPYC 9005 platforms support PCIe Gen 5 at 32GT/s per lane) Competitor memory bandwidth/socket: Vera: ~1.2 TB/s ARM AGI = 12 * 8 * 8800 = 844800 = ~0.8 TB/s 6980P = 12 * 8 * 8800 = 844800 = ~0.8 TB/s Intel specifications from ark.intel.com. ARM specifications from https://www.arm.com/products/cloud-datacenter/arm-agi-cpu#Specifications. Nvidia specifications from https://developer.nvidia.com/blog/nvidia-vera-cpu-delivers-high-performance-bandwidth-and-efficiency-for-ai-factories/. ARM AGI specifications from: https://www.arm.com/static/az/pdf/product-brief/arm-agi-cpu-product-brief.pdf 9xx6-007: AI inference throughput comparison based on AMD internal testing as of 6/26/2026, running GPT-OSS-20B at BF16 precision on vLLM with the ZenDNN/zentorch backend, offline serving, 1K input / 1K output, batch size 32, cumulative total token throughput (tokens/sec, higher is better) System configurations: 2P AMD EPYC 9996, 256C, 6th Gen, SMT disabled, 32 channels 128 GiB RDIMM @ 8000 MT/s, NPS1, BIOS WNIV6520N, microcode 0xb501005, Ubuntu 26.04 LTS, Linux 7.0.0-22-generic, vLLM + ZenDNN/zentorch. 2P Intel Xeon 6980P, 128C, Supermicro SYS-222HA-TN, SMT disabled, 24 channels 64 GiB RDIMM @ 8800 MT/s, BIOS 1.4, microcode 0x1000405, Ubuntu 26.04 LTS, Linux 7.0.0-22-generic, vLLM + ZenDNN/zentorch. 2P AMD EPYC 9965, 192C, 5th Gen, SMT disabled, 24 channels 64 GiB RDIMM @ 6400 MT/s, determinism Power, High Performance Mode, BIOS RVOT1003C, microcode 0xb101028, Ubuntu 26.04 LTS, Linux 7.0.0-22-generic, vLLM + ZenDNN/zentorch. 2P AMD EPYC 9755, 128C, 5th Gen, SMT disabled, 24 channels 64 GiB RDIMM @ 6400 MT/s, determinism Power, High Performance Mode, BIOS RVOT1003C, microcode 0xb00211e, Ubuntu 26.04 LTS, Linux 7.0.0-22-generic, vLLM + ZenDNN/zentorch. Results below are in the format of: [processor], [throughput tok/s], [ratio vs Intel Xeon 6980P] AMD EPYC 9996 (256C) 1527.81 1.37x Intel Xeon 6980P (128C) 1113.91 1.000 AMD EPYC 9965 (192C) 823.515 - AMD EPYC 9755 (128C) 765.47 - Results may vary due to factors including system configurations, software versions and BIOS settings.

Endnotes

9xx6-014: Comparison based on published Top-of-stack core counts and CPU W, Default CPU Power for 6th Gen EPYC, and TDPs for 5th Gen EPYC, Intel® Xeon®, and Nvidia Vera for estimated Highest Agents / CPU W across AMD EPYC™ 9956 (400W Default CPU Power), AMD EPYC™ 9965 (500W TDP), Nvidia Vera (450W TDP), ARM AGI (300W TDP), Intel® Xeon® 6980P (500W TDP), and Intel Xeon 6990E+ (450W TDP) powered servers as of 7/22/2026. 2 threads per core (SMT). 1 thread per core for ARM AGI. Agent counts are estimates derived from available CPU thread resources used as a proxy under a consistent theoretical workload. Actual agent capacity and throughput will vary based on workload, model, memory, software, orchestration, and system configuration. Starting with the 6th Gen AMD EPYC™ server processor family, AMD uses Default CPU Power to describe processor power consumption, succeeding AMD's historical TDP reference. Default CPU Power reflects total power consumed across the processor's compute and I/O dies for the stated performance target. Default CPU Power and TDP may both serve as processor power references for product comparison, platform planning, and performance-per-watt analysis. Intel Xeon TDP from ark.intel.com. Nvidia Vera TDP from https://developer.nvidia.com/blog/nvidia-vera-cpu-sets-a-new-standard-for-agentic-workloads-in-ai-factories/. ARM AGI Specifications from https://www.arm.com/products/cloud-datacenter/arm-agi-cpu#Specifications

9xx6-015: Comparison based on AMD Internal Estimates and Nvidia published specifications as of 07/22/2026. AMD EPYC 9956 (256C, 400W Default CPU Power) powered server racks are estimated to deliver up to 1.91x the cores per rack at 26% less power and 2.57x the cores / rack W compared to a Nvidia Vera (88C, 450W TDP) based server rack. AMD Rack specifications based on Supermicro Rack configurations and Gigabyte Rack configurations (Model H295-D80-CS1). Calculations reflect Total Sled power including CPU, Memory, Other System W with 15% loss factor. Nvidia specifications from: https://www.nvidia.com/en-us/data-center/products/vera-rack/ and https://developer.nvidia.com/blog/nvidia-vera-rubin-pod-seven-chips-five-rack-scale-systems-one-ai-supercomputer/. Supermicro Rack specifications provided by Supermicro. MGX SMC Rack 1 SMC Rack 2 Gigabyte Vera 88C 450W 9965 192c 500W 9956 256C 400W # 1U sleds 32 24 42 24 # CPUs/sled 8 8 4 8 # CPUs/rack 256 192 144 192 # core/CPU 88 192 256 256 Compute Rack Power Total 185104 144038.88 137461.8 269400 Relative Rack Power 1 0.778 0.743 1.455 Cores / Rack W 0.122 0.256 0.313 0.182 Relative Cores / Rack W 1 2.103 2.571 1.499 Cores/Rack 22528 36864 43008 49152 Cores/Rack 1 1.636 1.909 2.182 Starting with the 6th Gen AMD EPYC™ server processor family, AMD uses Default CPU Power to describe processor power consumption, succeeding AMD's historical TDP reference. Default CPU Power reflects total power consumed across the processor's compute and I/O dies for the stated performance target. Default CPU Power and TDP may both serve as processor power references for product comparison, platform planning, and performance-per-watt analysis. Results may vary based specific rack configuration used for comparison, including but not limited to wattage limits

9xx6-021: Rack-level general purpose performance comparison within a 100 kW rack power constraint based on AMD analysis as of July 2026 comparing 2P AMD EPYC™ “Venice” 9996 (512 total cores) vs 2P Nvidia Vera (176 total cores, Olympus).. Performance reflects geometric mean across 6 workloads: estimated SPECrate® 2017_int_base, Server-Side Java® Multi-JVM Max jOPS, Web serving (NGINX/WRK), Key-Value store (Redis), In-memory caching (Memcached), and Relational DB (TPROC-C). All platforms use 2-socket nodes. System manufacturers may vary configurations, yielding different results. See https://www.amd.com/content/dam/amd/en/documents/solutions/ai/methodology-description.pdf for full derivations and assumptions. TPROC-C workload is an open-source workload derived from TPC-Benchmark™ Standard, and as such is not comparable to published TPC-C™ results, as the results do not comply with the TPC-C Benchmark Standard. TPC, TPC Benchmark and TPC-C are trademarks of the Transaction Processing Performance Council. Preliminary performance estimates based on AMD engineering projections or measurements as of July 22, 2026 and subject to change. SPEC® and SPECint® are registered trademarks of Standard Performance Evaluation Corporation. Learn more at spec.org.

9xx6-022: Model-weight CPU-offload LLM inference output throughput comparison based on AMD internal testing measured on a PCIe Gen 5 x16 host link and estimated projected to PCIe Gen 6 x16 by host-to-GPU link bandwidth, as of 7/13/2026. Comparison is a 6th Gen AMD EPYC 96c HF processor host node (PCIe Gen 6, 5GHz) versus a 5th Gen AMD EPYC 9575F (PCIe Gen 5, 5GHz), with a 6th Gen Intel Xeon 6960P host node as a competitive reference. All measured systems were run 2-socket (full node) with SMT disabled. Workload: Qwen3-30B-A3B, BF16 weights (approximately 61 GB resident), tensor-parallel=1, 1024 input tokens, 256 output tokens, vLLM CPU-offload, median of n=3 iterations, measured on 1x NVIDIA B200. Measured Gen 5 output throughput for the 5th Gen AMD EPYC 9575F was 148.4, 114.8, 93.8, and 61.7 tokens per second at 6, 9, 12, and 20 GB of offloaded weights, with PCIe receive bandwidth pinned at 89 to 95 percent of the Gen 5 x16 ceiling and host DRAM read bandwidth below 10 percent of its 614 GB/s ceiling, establishing the PCIe link (not host memory) as the throughput limiter. Because decode is PCIe-transfer-bound, doubling host-to-GPU bandwidth (Gen 5 x16 at 64 GB/s per direction to Gen 6 x16 at 128 GB/s per direction) projects a 1.8x to 1.9x throughput ceiling for the 6th Gen AMD EPYC 96c HF processor. Measured Gen 5 output throughput for the 6th Gen Intel Xeon 6960P was 104.9 and 82.3 tokens per second at 9 and 12 GB of offloaded weights (median of n=3 iterations, same NVIDIA B200 host and vLLM configuration), versus 114.8 and 93.8 tokens per second for the 5th Gen AMD EPYC 9575F at the same offload points, so the 5th Gen AMD EPYC 9575F delivers 1.09x and 1.14x the 6th Gen Intel Xeon 6960P output throughput (up to 1.1x). Memory configuration was 24x 128 GB DDR5-6400 (AMD) and 24x 64 GB DDR5-6400 (Intel); host DRAM utilization remained below 10 percent throughout, so memory capacity was not the throughput limiter. Preliminary performance estimates based on AMD engineering projections or measurements as of 7/13/2026 and subject to change. PCIe is a registered trademark of PCI-SIG Corporation.

Endnotes

9xx6-036: FFmpeg video encoding and transcoding throughput comparison based on AMD internal testing as of 7/20/2026. 1P AMD EPYC 9996 (256C) delivers up to 2.7x the FFmpeg video encoding and transcoding throughput of 1P Intel Xeon 6980P (128C) Software: FFmpeg encode and transcode workload (raw to VP9, raw to H.264, VP9 to H.264, H.264 to VP9), 1080p source, one encode job per vCPU, KVM, Ubuntu 24.04. Configurations: 1P Intel Xeon 6980P: 1P Intel Xeon 6980P, 128 cores (256 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, Supermicro SYS-222HA-TN, BIOS v1.5, Max performance BIOS profile, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P Graviton5: 1P Graviton5 (AWS m9gd.metal-48xl), 192 cores (192 vCPU), SMT OFF, 768GB memory, Ubuntu 24.04, kernel Linux 6.17.0-1009-aws. 1P AMD EPYC 9965: 1P AMD EPYC 9965, 192 cores (384 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, AMD reference platform, BIOS RVOT1006C, TDP=500W, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P AMD EPYC 9996: 1P AMD EPYC 9996, 256 cores (512 vCPU), SMT ON, 64GB DDR5-8000 RDIMM fully populated, AMD reference platform, BIOS BKC10.3 90RC4, 600W Default CPU Power, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. Data presented: 1P Intel Xeon 6980P 3974348.83; 1P Graviton5 7381741; 1P AMD EPYC 9965 6942536.96; 1P AMD EPYC 9996 10607502.57. Results may vary. Results for AWS Graviton5 were obtained from a publicly available cloud instance. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. Results may vary due to factors including system configurations, software versions and BIOS settings. Starting with the 6th Gen AMD EPYC server processor family, AMD uses Default CPU Power to describe processor power consumption, succeeding AMD's historical TDP reference. Default CPU Power reflects total power consumed across the processor's compute and I/O dies for the stated performance target. Default CPU Power and TDP may both serve as processor power references for product comparison, platform planning, and performance-per-watt analysis.Results may vary. Results for AWS Graviton5 were obtained from a publicly available cloud instance. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. Results may not be directly comparable to physical server configurations due to differences in platform implementation, firmware, operating environment, memory configuration, and system tuning.

9xx6-037: Redis 8.8 Open Source Stable request-throughput comparison based on AMD internal testing as of 7/20/2026. 1P AMD EPYC 9996 (256C) delivers up to 2.9x the Redis 8.8 request throughput of 1P Intel Xeon 6980P (128C) Software: Redis 8.8 Open Source Stable, redis-benchmark SET and GET, one Redis server per 8 vCPUs aggregated, KVM, Ubuntu 24.04. Configurations: 1P Intel Xeon 6980P: 1P Intel Xeon 6980P, 128 cores (256 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, Supermicro SYS-222HA-TN, BIOS v1.5, Max performance BIOS profile, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P Graviton5: 1P Graviton5 (AWS m9gd.metal-48xl), 192 cores (192 vCPU), SMT OFF, 768GB memory, Ubuntu 24.04, kernel Linux 6.17.0-1009-aws. 1P AMD EPYC 9965: 1P AMD EPYC 9965, 192 cores (384 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, AMD reference platform, BIOS RVOT1006C, TDP=500W, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P AMD EPYC 9996: 1P AMD EPYC 9996, 256 cores (512 vCPU), SMT ON, 64GB DDR5-8000 RDIMM fully populated, AMD reference platform, BIOS BKC10.3 90RC4, 600W Default CPU Power, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. Data presented: 1P Intel Xeon 6980P 42636320, 1P Graviton5 50057444, 1P AMD EPYC 9965 76703115, 1P AMD EPYC 9996 124130835.50. Ratio 9996 vs 6980P: 124130835.50 / 42636320 = 2.91 (rounds to 2.9x). Results may vary. Results for AWS Graviton5 were obtained from a publicly available cloud instance. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. Starting with the 6th Gen AMD EPYC server processor family, AMD uses Default CPU Power to describe processor power consumption, succeeding AMD's historical TDP reference. Default CPU Power reflects total power consumed across the processor's compute and I/O dies for the stated performance target. Default CPU Power and TDP may both serve as processor power references for product comparison, platform planning, and performance-per-watt analysis. Results may not be directly comparable to physical server configurations due to differences in platform implementation, firmware, operating environment, memory configuration, and system tuning.

9xx6-038: NGINX request-throughput comparison based on AMD internal testing as of 7/20/2026. 1P AMD EPYC 9996 (256C) delivers up to 3.7x the NGINX request throughput of 1P Intel Xeon 6980P (128C) Software: NGINX served with wrk load generator (1KB file), one NGINX server per 8 vCPUs aggregated, KVM, Ubuntu 24.04. Configurations: 1P Intel Xeon 6980P: 1P Intel Xeon 6980P, 128 cores (256 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, Supermicro SYS-222HA-TN, BIOS v1.5, Max performance BIOS profile, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P Graviton5: 1P Graviton5 (AWS m9gd.metal-48xl), 192 cores (192 vCPU), SMT OFF, 768GB memory, Ubuntu 24.04, kernel Linux 6.17.0-1009-aws. 1P AMD EPYC 9965: 1P AMD EPYC 9965, 192 cores (384 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, AMD reference platform, BIOS RVOT1006C, TDP=500W, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P AMD EPYC 9996: 1P AMD EPYC 9996, 256 cores (512 vCPU), SMT ON, 64GB DDR5-8000 RDIMM fully populated, AMD reference platform, BIOS BKC10.3 90RC4, 600W Default CPU Power, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. Data presented: 1P Intel Xeon 6980P 10526326, 1P Graviton5 15929450, 1P AMD EPYC 9965 24907610, 1P AMD EPYC 9996 38451232. Results for AWS Graviton5 were obtained from a publicly available cloud instance. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. Starting with the 6th Gen AMD EPYC™ server processor family, AMD uses Default CPU Power to describe processor power consumption, succeeding AMD's historical TDP reference. Default CPU Power reflects total power consumed across the processor's compute and I/O dies for the stated performance target. Default CPU Power and TDP may both serve as processor power references for product comparison, platform planning, and performance-per-watt analysis. Results may not be directly comparable to physical server configurations due to differences in platform implementation, firmware, operating environment, memory configuration, and system tuning.

Endnotes

9xx6-039: MySQL 8.4.10 online transaction processing throughput comparison (HammerDB TPROC-C, a workload derived from the TPC-C Benchmark Standard. Results are based on TPC-C derived workloads and are not comparable to published TPC benchmark results.) based on AMD internal testing as of 7/20/2026. 1P AMD EPYC 9996 (256C) delivers up to 2.6x the MySQL 8.4.10 online transaction processing throughput (HammerDB TPROC-C) of 1P Intel Xeon 6980P (128C) Software: HammerDB TPROC-C on MySQL 8.4.10, 500 warehouse dataset, users scaled to vCPU count, KVM, Ubuntu 24.04. Configurations: 1P Intel Xeon 6980P: 1P Intel Xeon 6980P, 128 cores (256 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, Supermicro SYS-222HA-TN, BIOS v1.5, Max performance BIOS profile, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P Graviton5: 1P Graviton5 (AWS m9gd.metal-48xl), 192 cores (192 vCPU), SMT OFF, 768GB memory, Ubuntu 24.04, kernel Linux 6.17.0-1009-aws. 1P AMD EPYC 9965: 1P AMD EPYC 9965, 192 cores (384 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, AMD reference platform, BIOS RVOT1006C, TDP=500W, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P AMD EPYC 9996: 1P AMD EPYC 9996, 256 cores (512 vCPU), SMT ON, 64GB DDR5-8000 RDIMM fully populated, AMD reference platform, BIOS BKC10.3 90RC4, 600W Default CPU Power, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. Data presented: 1P Intel Xeon 6980P 15499304, 1P Graviton5 23075914, 1P AMD EPYC 9965 25168617, 1P AMD EPYC 9996 40804653. + Results for AWS Graviton5 were obtained from a publicly available cloud instance. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. Starting with the 6th Gen AMD EPYC™ server processor family, AMD uses Default CPU Power to describe processor power consumption, succeeding AMD's historical TDP reference. Default CPU Power reflects total power consumed across the processor's compute and I/O dies for the stated performance target. Default CPU Power and TDP may both serve as processor power references for product comparison, platform planning, and performance-per-watt analysis. Results may not be directly comparable to physical server configurations due to differences in platform implementation, firmware, operating environment, memory configuration, and system tuning.

9xx6-040: OpenSSL cryptographic throughput comparison based on AMD internal testing as of 7/20/2026. 1P AMD EPYC 9996 (256C) delivers up to 2.5x the OpenSSL cryptographic throughput of 1P Intel Xeon 6980P (128C) Software: OpenSSL cryptographic throughput workload, KVM, Ubuntu 24.04. Configurations: 1P Intel Xeon 6980P: 1P Intel Xeon 6980P, 128 cores (256 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, Supermicro SYS-222HA-TN, BIOS v1.5, Max performance BIOS profile, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P Graviton5: 1P Graviton5 (AWS m9gd.metal-48xl), 192 cores (192 vCPU), SMT OFF, 768GB memory, Ubuntu 24.04, kernel Linux 6.17.0-1009-aws. 1P AMD EPYC 9965: 1P AMD EPYC 9965, 192 cores (384 vCPU), SMT ON, 64GB DDR5-6400 RDIMM fully populated, AMD reference platform, BIOS RVOT1006C, TDP=500W, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. 1P AMD EPYC 9996: 1P AMD EPYC 9996, 256 cores (512 vCPU), SMT ON, 64GB DDR5-8000 RDIMM fully populated, AMD reference platform, BIOS BKC10.3 90RC4, 600W Default CPU Power, CPPC on, power determinism, Ubuntu 24.04, kernel Linux 6.17.0-35-generic. Data presented: 1P Intel Xeon 6980P 552433496.01, 1P Graviton5 545579628, 1P AMD EPYC 9965 1073669574.93, 1P AMD EPYC 9996 1389064585.24. Results for AWS Graviton5 were obtained from a publicly available cloud instance. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. Starting with the 6th Gen AMD EPYC™ server processor family, AMD uses Default CPU Power to describe processor power consumption, succeeding AMD's historical TDP reference. Default CPU Power reflects total power consumed across the processor's compute and I/O dies for the stated performance target. Default CPU Power and TDP may both serve as processor power references for product comparison, platform planning, and performance-per-watt analysis. Results may not be directly comparable to physical server configurations due to differences in platform implementation, firmware, operating environment, memory configuration, and system tuning.

9xx6-047: SPECrate®2017_int_base per-core comparison based on AMD internal projections and measurements as of 07/20/2026. 6th Gen AMD EPYC delivers leadership 1P SPECrate®2017_int_base per-core performance versus three competing server processors, each at the AMD core configuration noted. Versus NVIDIA Vera: 1P 96C 6th Gen AMD EPYC 9996 score 1350, per-core 1350 / 96 = 14.06, versus 1P 88C NVIDIA Vera score 729.5, per-core 729.5 / 88 = 8.29; 14.06 / 8.29 = 1.70x. Versus Intel Xeon 6980P: 1P 128C 6th Gen AMD EPYC score 1565, per-core 1565 / 128 = 12.23, versus 1P 128C Intel Xeon 6980P score 1255, per-core 1255 / 128 = 9.80; 12.23 / 9.80 = 1.25x. Versus Arm AGI: 1P 96C 6th Gen AMD EPYC 9996 score 1350, per-core 14.06, versus 1P 136C Arm AGI score 972, per-core 972 / 136 = 7.15; 14.06 / 7.15 = 1.97x. Configurations: 6th Gen AMD EPYC (96C and 128C) using AOCC 5.1 with 16x64GB 4R MRDIMM 12800 MT/s; Intel Xeon 6980P using OneAPI with 12x64GB MRDIMM 8800 MT/s; Arm AGI using GCC 13 with 12x64GB MRDIMM 8800 MT/s; NVIDIA Vera using GCC 13 with 8x96GB SOCAMM2. Preliminary performance estimates based on AMD engineering projections or measurements as of 07/20/2026 and subject to change. Results compare AMD internal testing of the listed AMD and Intel systems against testing performed on the listed AWS EC2 Graviton5 bare-metal instance. Cloud instance results may be affected by cloud service configuration, instance availability, regional deployment, hypervisor/Nitro behavior, storage/network configuration, and other cloud provider variables. SPEC® and SPECint® are registered trademarks of Standard Performance Evaluation Corporation. Learn more at spec.org.

9xx6-049: SPECrate®2026_int_base per-core comparison based on AMD internal estimates for the 6th Gen AMD EPYC 96c HF processor and NVIDIA Vera as of 7/22/2026. Preliminary performance estimates based on AMD engineering projections and subject to change. 2P 6th Gen AMD EPYC 96c HF processor, GCC 15.2, estimated SPECrate®2026_int_base score 1210 (6.3 per core). 2P NVIDIA Vera: 88 cores, GCC 15.2, estimated SPECrate®2026_int_base score 925 (5.3 per core). For: ~1.2x the performance / core Nvidia Vera performance from: https://nvdam.widen.net/s/nmw5vblpqd/nvidia_vera_cpu_architecture_whitepaper?nvid=nv-tblg-543584. SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. Results may vary.

Endnotes

9xx6-050: SPECrate®2026_int_base throughput comparison based on AMD internal estimates for the 2P 6th Gen AMD EPYC 9996 and 2P NVIDIA Vera as of 7/22/2026. Preliminary performance estimates based on AMD engineering projections and subject to change. 2P 6th Gen AMD EPYC 9996: GCC 15.2, estimated SPECrate®2026_int_base score 2070. NVIDIA Vera: 88 cores, GCC 15.2, estimated SPECrate®2026_int_base score 925. For ~2.2x the performance https://nvdam.widen.net/s/nmw5vblpqd/nvidia_vera_cpu_architecture_whitepaper?nvid=nv-tblg-543584. SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. Results may vary.

EPYC-025D: As of July 2026, 6th Gen EPYC 9996 has 256 cores and 512 threads with SMT enabled which is higher than any other publicly disclosed 1P CPU

EPYC-055D - Mercury Research Sell-In Revenue Shipment Estimates, Q1 2026. Revenue share of 46.2%, unit share of 33.2%.

EPYC-067 - Based on AMD internal analysis and modeling of representative infrastructure workloads underlying agentic AI systems as of June, 2026. Performance comparisons reflect estimated rack-scale throughput within a modeled 100kW deployment and include proxy workloads spanning web services, data infrastructure, and orchestration layers. Results are derived from a combination of AMD testing, third-party data, and projections, and may vary based on system configuration, software stack, and workload mix. See AMD Agentic AI rack-scale performance blog and AMD methodology description for details on assumptions, workloads, and measurement approach.

EPYC-068 - The AMD EPYC server CPU portfolio spans the industry's broadest ranges of data center deployments, from general-purpose enterprise, cloud, telecom, SMB, and HPC systems to emerging AI environments including sandboxed agentic AI deployments and GPU head node servers. AMD EPYC 6th Generation platforms extend this breadth by uniquely combining high core and thread density of up to 512 threads, advanced memory bandwidth of up to 16 channels of 12.8 GT/s MRDIMM support, next-generation PCIe® Gen 6 connectivity, and select SKUs with boost frequencies up to 5 GHz.

GD-181a: All performance and/or cost savings claims are provided by the 3rd party organization featured herein and have not been independently verified by AMD. Performance and cost benefits are impacted by a variety of variables. Results herein are specific to such 3rd party organization and may not be typical.

MI350-81: Testing by AMD Performance Labs as of July 7, 2026, measuring the inference performance in tokens per second (TPS) of a system configured with an AMD Instinct MI355x 8x GPU platform and AMD ROCm 7.0 software vs a similarly configured system using a preview version of AMD ROCm.ai (ROCm 7.2.2 with optimizations such as Optimized Kernels, Parallelism and Scheduling) running GLM-5, Kimi-K2.5, and DeepSeekk-R1-0528 models. Stated performance uplift is expressed as a combined average TPS over across the (3) models tested. Hardware Configuration Supermicro AS -4126GS-NMR-LCC (board H14DSG-OD) 8x AMD Instinct MI355X. BIOS AMI v1.4a (2025-04-16), GPU firmware SMC 04.86.11.02, TA RAS 27.69.00.10, TA XGMI 32.00.00.20, RLC43, MEC36, SDMA12, Ubuntu 22.04.2 LTS, kernel 5.15.0-70-generic, amdgpu driver 6.16.6, HOST ROCm 7.1.0 Software Configuration(s) GLM-5 ROCm Docker Image: rocm/sgl-dev:v0.5.8.post1-rocm700-mi35x-20260219 PYTorch Version 2.8.0, SGLang v0.5.8 Kimi ROCm Docker Image: vllm/vllm-openai-rocm:v0.16.0, vLLM version 0.16.0 DeepSeek-R1 Docker Image: rocm/7.0:...sgl-dev-v0.5.2-rocm7.0-mi35x-20250915, SGLang version V0.5.13 vs GLM-5 ROCm Docker Image: rocm/atom:rocm7.2.2_ubuntu24.04_py3.12_pytorch_release_2.10.0_atom0.1.2.post, ATOM v0.1.2.post Kimi ROCm Docker Image: vllm/vllm-openai-rocm:v0.22.0, vLLM version V0.22.0 DeepSeek-R1 Docker Image: lmsysorg/sglang-rocm:v0.5.13-rocm720-mi35x-20260612, SGLang version V0.5.13 Server manufacturers may vary configurations, yielding different results. Performance may vary based on configuration, software, vLLM version, and the use of the latest drivers and optimizations.

MI350-82: Testing by AMD Performance Labs as of July 7, 2026, measuring the training performance in tokens per second (TPS) of AMD ROCm 7.0 software vs a preview version of AMD ROCm.ai (ROCm 7.2.2 with optimizations such as Optimized Kernels, Parallelism and Scheduling),) using Megatron -LM on a system with 8x AMD Instinct MI355x 8x GPUs GPU platform running DeepSeek-V2-Lite, DeepSeek-V3-16B, and Qwen3-30B-A3B models. Stated performance uplift is expressed as the combined average TPS over across the (3) models tested. Hardware Configuration Supermicro AS -4126GS-NMR-LCC (board H14DSG-OD) 8x AMD Instinct MI355. BIOS AMI v1.4a (2025-04-16), GPU firmware SMC 04.86.11.02, TA RAS 27.69.00.10, TA XGMI 32.00.00.20, RLC 43, MEC 36, SDMA 12, Ubuntu 22.04.2 LTS, kernel 5.15.0-70-generic, amdgpu driver 6.16.6, HOST ROCm 7.1.0 Software Configuration(s) DeepSeek-V2-Lite, ROCm 7.2.1 + Primus v26.3 DeepSeek-V3-16B, ROCm 7.2.1 + Primus v26.3 Qwen3-30B-A3B, ROCm 7.2.1 + Primus v26.3 Server manufacturers may vary configurations, yielding different results. Performance may vary based on configuration, software, and the use of the latest drivers and optimizations.

Endnotes

MI350P-001: Based on AMD internal testing (July 2026) on a (1x) AMD Instinct MI350P GPU vs (1x) Nvidia RTX Pro 6000 Blackwell Server Edition, on the OpenAI GPT-OSS-120B (MXFP4) benchmark measuring online serving output-throughput (tok/s) at ISL/OSL 1024/1024 across concurrency levels 1, 4, 8, 16, 32, 64, 128, 256, 512; median of 3 runs per point. Stated result is the per-concurrency ratio reported as the peak of the per-concurrency ratio MI350P GPU served via AIMS ccontainer silogenai/aim-instinct-openai-gpt-oss-120b:0.12.0-rc14; NVIDIA RTX PRO 6000 served via NVIDIA NIM nvcr.io/nim/openai/gpt-oss-120b:2.0.3. Configurations: 8x AMD Instinct MI350P PCIe Card (CDNA4, gfx950, 128 CUs, 144 GB HBM3E, SPX compute / NPS1), vBIOS 113-350P-01-1K1-000A, GPU driver 6.19.13-2353916.24.04, ROCm 7.14.0 (AMD-SMI 26.5.0); host 2P AMD EPYC 9455 (48-core), Dell PowerEdge XE7745, BIOS 1.7.6, microcode 0xb002162, SMT Enabled, Ubuntu 24.04.4 LTS, Linux 6.8.0-124-generic, versus 8x NVIDIA RTX PRO 6000 Blackwell Server Edition, vBIOS 98.02.8D.00.01, GPU driver 595.45.04, CUDA 13.2; host 2P AMD EPYC 9455 (48-core), Dell PowerEdge XE7745, BIOS 1.6.4, microcode 0xb00215a, SMT Enabled, Ubuntu 24.04.4 LTS, Linux 6.8.0-124-generic. Server manufacturers may vary configurations, yielding different results. Results may vary due to factors including system configuration, software version and BIOS settings.

MI350P-003: Based on AMD internal testing (June 2028) on (1x) AMD Instinct MI350P GPU vs (1x) NVIDIA RTX PRO 6000 Blackwell Server Edition GPU on the Meta Llama-3.3-70B-Instruct (FP8) benchmark, measuring online serving output-throughput (tok/s) ., at ISL/OSL 1024/1024 across concurrency levels 1, 4, 8, 16, 32, 64, 128, 256, 512; median of 3 runs per point. Stated results are the peak per-concurrency ratios. MI350P GPU served via AIMS container silogenai/aim-instinct-meta-llama-llama-3-3-70b-instruct:0.12.0-rc6; NVIDIA RTX PRO 6000 GPU served via NVIDIA NIM nvcr.io/nim/meta/llama-3.3-70b-instruct:2.0.6. Configurations: 8x AMD Instinct MI350P PCIe Card (CDNA4, gfx950, 128 CUs, 144 GB HBM3E, SPX compute / NPS1), vBIOS 113-350P-01-1K1-000A, GPU driver 6.19.13-2353916.24.04, ROCm 7.14.0 (AMD-SMI 26.5.0); host 2P AMD EPYC 9455 (48-core), Dell PowerEdge XE7745, BIOS 1.7.6, microcode 0xb002162, SMT Enabled, Ubuntu 24.04.4 LTS, Linux 6.8.0-124-generic || NVIDIA RTX PRO 6000: 8x NVIDIA RTX PRO 6000 Blackwell Server Edition, vBIOS 98.02.8D.00.01, GPU driver 595.45.04, CUDA 13.2; host 2P AMD EPYC 9455 (48-core), Dell PowerEdge XE7745, BIOS 1.6.4, microcode 0xb00215a, SMT Enabled, Ubuntu 24.04.4 LTS, Linux 6.8.0-124-generic Server manufacturers may vary configurations, yielding different results. Results may vary due to factors including system configurations, software versions and BIOS settings.

MI350P-005: Based on AMD internal testing as of 6/23/2026 on an AMD Instinct MI350P GPU running Autonomous Threat Hunting workload. 1,488 threat-detection queries run per model using Gemma-4-31B and Qwen-3.6-35B-A2B (local, on MI350P) and Claude Opus 4.7 (cloud API). Success rate based on the model with best response quality. Average Query response completion time comparison of 2.4 seconds on local LLM vs 7.1 seconds on frontier cloud-only LLM. Response time used as tiebreaker. Cost per user across all queries: frontier cloud-only (all Opus) = $14.39; hybrid routing (each query to its winning model) = $8.23. On-prem cost basis $10,401.72/GPU/yr. Assumes Oracle routing to the winning model. Hardware and Cloud costs based on AMD Internal Estimates and published prices at https://www.anthropic.com/new/claude-opus-4-7. 1x AMD Instinct MI350P (CDNA4, gfx950, 128 CUs, 144 GB HBM3E, SPX compute / NPS1), vBIOS 113-350P-01-1K1-000A, GPU driver 6.19.13-2353916.24.04, ROCm 7.14.0 (AMD-SMI 26.5.0); host 2P AMD EPYC 9455 (48-core), Dell PowerEdge XE7745, BIOS 1.7.6, microcode 0xb002162, SMT Enabled, Ubuntu 24.04.4 LTS, Linux 6.8.0-124-generic || NVIDIA RTX PRO 6000: 1x NVIDIA RTX PRO 6000 Blackwell Server Edition, vBIOS 98.02.8D.00.01, GPU driver 595.45.04, CUDA 13.2; host 2P AMD EPYC 9455 (48-core), Dell PowerEdge XE7745, BIOS 1.6.4, microcode 0xb00215a, SMT Enabled, Ubuntu 24.04.4 LTS, Linux 6.8.0-124-generic. Server manufacturers may vary configurations, yielding different results. Results may vary based on configuration, workload, routing accuracy, and token pricing.

MI350P-006: Based on AMD internal testing (July 2026)), on a (1x) AMD Instinct MI350P GPU vs (1x) NVIDIA H200 NVL GPU on the Mistral Small 3.2 24B Instruct (FP16) benchmark, measuring online serving output-throughput (tok/s) comparison at ISL/OSL 1024/1024 across concurrency levels 1, 4, 8, 16, 32, 64, 128, 256, 512; median of 3 runs per point. Stated results are the peak per-concurrency ratios. MI350P served via AIMS container silogenai/aim-instinct-mistralai-mistral-small-3-2-24b-instruct-2506:0.12.0-rc11; RTX PRO 6000 via NVIDIA NIM nvcr.io/nvidia/vllm:26.05-py3. Configuration: 8x AMD Instinct MI350P PCIe Card (CDNA4, gfx950, 128 CUs, 144 GB HBM3E, SPX compute / NPS1), vBIOS 113-350P-01-1K1-000A, GPU driver 6.19.13-2353916.24.04, ROCm 7.14.0 (AMD-SMI 26.5.0); host 2P AMD EPYC 9455 (48-core), Dell PowerEdge XE7745, BIOS 1.7.6, microcode 0xb002162, SMT Enabled, Ubuntu 24.04.4 LTS, Linux 6.8.0-124-generic || NVIDIA RTX PRO 6000: 8x NVIDIA RTX PRO 6000 Blackwell Server Edition, vBIOS 98.02.8D.00.01, GPU driver 595.45.04, CUDA 13.2; host 2P AMD EPYC 9455 (48-core), Dell PowerEdge XE7745, BIOS 1.6.4, microcode 0xb00215a, SMT Enabled, Ubuntu 24.04.4 LTS, Linux 6.8.0-124-generic. Sever manufacturers may vary configurations, yielding different results. Results may vary due to factors including system configurations, software versions and BIOS settings.

MI350P-007: Based on AMD internal testing (July 2026), on a (1x) AMD Instinct MI350P GPU vs (1x) NVIDIA H200 NVL GPU on the Llama 3.3 70B Instruct (FP8) online serving output-throughput per dollar (tok/s/USD) comparison at ISL/OSL 1024/1024 across concurrency levels 1, 4, 8, 16, 32, 64, 128, 256, 512; median of 3 runs per point. MI350P based server internal AMD estimated pricing as $327,238.40 USD. RTX_PRO_6000 based server public list price reported on OEM website as $265,928.24 USD as of 7/16/2026. Stated results are the peak per-concurrency ratios: MI350P served via AIMS silogenai/aim-instinct-meta-llama-llama-3-3-70b-instruct:0.12.0-rc6; H200 NVL via NVIDIA NIM nvcr.io/nim/meta/llama-3.3-70b-instruct:2.0.6; RTX PRO 6000 via NVIDIA NIM nvcr.io/nim/meta/llama-3.3-70b-instruct:2.0.6. Configuration: 8x AMD Instinct MI350P PCIe Card (CDNA4, gfx950, 128 CUs, 144 GB HBM3E, SPX compute / NPS1), vBIOS 113-350P-01-1K1-000A, GPU driver 6.19.13-2353916.24.04, ROCm 7.14.0 (AMD-SMI 26.5.0); host 2P AMD EPYC 9455 (48-core), Dell PowerEdge XE7745, BIOS 1.7.6, microcode 0xb002162, SMT Enabled, Ubuntu 24.04.4 LTS, Linux 6.8.0-124-generic || NVIDIA RTX PRO 6000: 8x NVIDIA RTX PRO 6000 Blackwell Server Edition, vBIOS 98.02.8D.00.01, GPU driver 595.45.04, CUDA 13.2; host 2P AMD EPYC 9455 (48-core), Dell PowerEdge XE7745, BIOS 1.6.4, microcode 0xb00215a, SMT Enabled, Ubuntu 24.04.4 LTS, Linux 6.8.0-124-generic. Sever manufacturers may vary configurations, yielding different results. Results may vary due to factors including system configurations, software versions and BIOS settings.

Endnotes

MI400-001B: Based on engineering projections by AMD performance labs in July 2026, to estimate peak theoretical precision performance of an AMD Instinct™ MI430X GPU using FP64 (Vector) datatype vs. NVIDIA Rubin GPU with FP64 (Vector) datatype. Results subject to change when products are released in market.

MI400-004: Calculations by AMD Performance Labs in June 2026, based on published memory capacity and memory bandwidth of AMD Instinct™ MI455X and MI430X GPUs vs. the published memory capacity and memory bandwidth specifications of a NVIDIA Vera Rubin GPU. System manufacturers may vary configurations, yielding different results.

MI400-005: Based on calculations by AMD Performance Labs in June 2026, to determine the peak theoretical precision performance of an AMD Helios Rackscale Solution using peak Matrix FP16, BF16, INT8 datatypes and peak Open Compute Project MXFP6, MXFP8, FP8 and MXFP4 datatypes vs. NVIDIA Vera Rubin NVL72 with NVFP4 Dense datatype with FP8/FP6 datatypes. System manufacturers may vary configurations, yielding different results.

MI400-007: Calculations by AMD Performance Labs in June 2026 based on published memory capacity and memory bandwidth specifications of an AMD Helios rackscale solution vs. the published memory capacity and memory bandwidth specifications of a NVIDIA Vera Rubin NVL72 rack. System manufacturers may vary configurations, yielding different results.

MI400-019: Calculations by AMD Performance Labs in July 2026, based on published scale-out bandwidth of the AMD rackscale solution when comparing specifications vs. published NVIDIA Vera Rubin NVL72 Rack specs. System manufacturers may vary configurations, yielding different results.

MI400-020: Based on measurements and calculations by AMD Performance Labs in July 2026, for the AMD Instinct™ MI455X GPU to determine measured token throughput at high, medium and low interactivity points run on Deepseek V4 Flash with FP4 serving compared to AMD Instinct™ MI355X GPU. System manufacturers may vary configurations, yielding different results.

MI400-021: Based on modelling by AMD Performance Labs and Cerebras in July 2026 to determine tokens per second per kilowatt (TPS/kW) at a comparable interactivity point with Kimi 2.6 1T Model comparing an AMD Helios rackscale solution with Cerebras WSE to a Cerebras WSE-only configuration. System manufacturers may vary configurations, yielding different results.

MI400-022: Based on calculations by AMD Performance Labs in July 2026, on a system configured with AMD Instinct™ MI455X GPU to determine cost per million tokens at high, medium and low interactivity using $/hr cloud pricing and tokens per second compared to AMD Instinct™ MI355X GPU run on Deepseek V4 Flash with FP4 serving. System manufacturers and $/hr cloud pricing may vary, yielding different results.

MI400-023: Based on modelling by AMD Performance Labs in July 2026, on an AMD Helios rackscale solution to determine token throughput/GPU at high, medium and low interactivity points using Kimi K2 Thinking with an evaluated ISL/OSL combination of 32K/8K compared to the published modeled specifications for NVIDIA Vera Rubin NVL72 rack. System manufacturers may vary configurations, yielding different results.

MI400-025: Based on AMD Performance Labs estimates as of July 2026, tokens-per-dollar performance was calculated using the Kimi K2 Thinking workload (32K input / 8K output) on an AMD Helios rackscale solution compared to an NVIDIA Vera Rubin NVL72 rack. Results reflect estimated aggregate throughput across low, medium, and high-interactivity operating points and hourly pricing projection of system GPUs based on market conditions. System configurations may vary by manufacturer and may produce different results.

MI400-027 Based on testing and calculations by AMD Performance Labs in July 2026, on systems configured with a single AMD Instinct™ MI455X GPU powered by AMD ROCm.ai vs an AMD Instinct™ MI355X GPU to determine median bandwidth in TB/s using an industry standard Multi-Latent Attention (MLA) benchmark. System manufacturers may vary configurations, yielding different results. MI400-027

MI400-028: Based on calculations by AMD Performance Labs in July 2026, on a system configured with a single AMD Instinct™ MI455X GPU powered by AMD ROCm.ai to determine MXFP4 performance using an AITER GEMM kernel benchmark compared to the published specifications for a single AMD Instinct™ MI355X GPU. System manufacturers may vary configurations, yielding different results.

MI400-029: Based on calculations by AMD Performance Labs in July 2026, on systems configured with 4x AMD Instinct™ MI455X GPUs and 8x AMD Instinct™ MI355X GPUs, both powered by AMD ROCm.ai, to determine single GPU scale-up networking bandwidth using a common memory bandwidth benchmark with a 16GB transfer size. Median results from five runs are reported. System manufacturers may vary configurations, yielding different results.

Endnotes

MI400-030: Based on testing by AMD Performance Labs in July 2026, on systems powered by AMD ROCm.ai and configured with an AMD Instinct™ MI455X GPU and AMD Pensando™ Vulcano 800 AI NIC and AMD Instinct™ MI355X GPU with AMD Pensando™ Pollara 400 AI NIC to determine bidirectional scale-out networking bandwidth running a bandwidth benchmark. System manufacturers may vary configurations, yielding different results.

MI500-001: Based on engineering projections by AMD Performance Labs in December 2025, to estimate the peak theoretical precision performance of AMD Instinct™ MI500 Series GPU powered AI Rack vs. an AMD Instinct MI300X platform. Results subject to change when products are released in market.

REX-016: Based on the OpenNav Robotics Workload Benchmark commissioned by AMD, published by Open Navigation LLC on July 23, 2026, as measured on the GMKtec EVO-X2 AI Mini PC AMD Ryzen AI Max+ 395 configured to reflect Ryzen AI Embedded X199 specifications (5.1 GHz CPU, 2.9 GHz GPU, 120W TDP, LPDDR5X-7500), vs. the NVIDIA Jetson AGX Thor Developer Kit: https://opennav.org/news/opennav-robotics-workload-benchmark

REX-018: Based on the OpenNav Robotics Workload Benchmark commissioned by AMD, published by Open Navigation LLC on July 23, 2026, as measured on the GMKtec EVO-X2 AI Mini PC AMD Ryzen AI Max+ 395 configured to reflect Ryzen AI Embedded X199 specifications (5.1 GHz CPU, 2.9 GHz GPU, 120W TDP, LPDDR5X-7500), vs. the NVIDIA Jetson AGX Thor Developer Kit: https://opennav.org/news/opennav-robotics-workload-benchmark

REX-019: Based on the mimik whitepaper “Architectural Fit for Production-Scale Agentic AI on Heterogeneous SoCs” commissioned by AMD, published by mimik on July 23, 2026, based on a modeled sweep of 455 feasible agentic AI workflows across two device classes (X100, NVIDIA Jetson T5000), checking spare CPU, GPU, and memory. For more information see: https://www.mimik.com/agentix-compute-benchmarking

‘Fortune 100’ refers to the top 20% ranked companies in the 2025 Fortune 500 list, published in June 2025. Fortune and Fortune Media IP Limited are not affiliated with, and do not endorse products or services of Advanced Micro Devices, Inc.

General Disclaimer and Attribution

Timelines, roadmaps, and/or product release dates shown in these slides are plans only and subject to change.

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 set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18u.