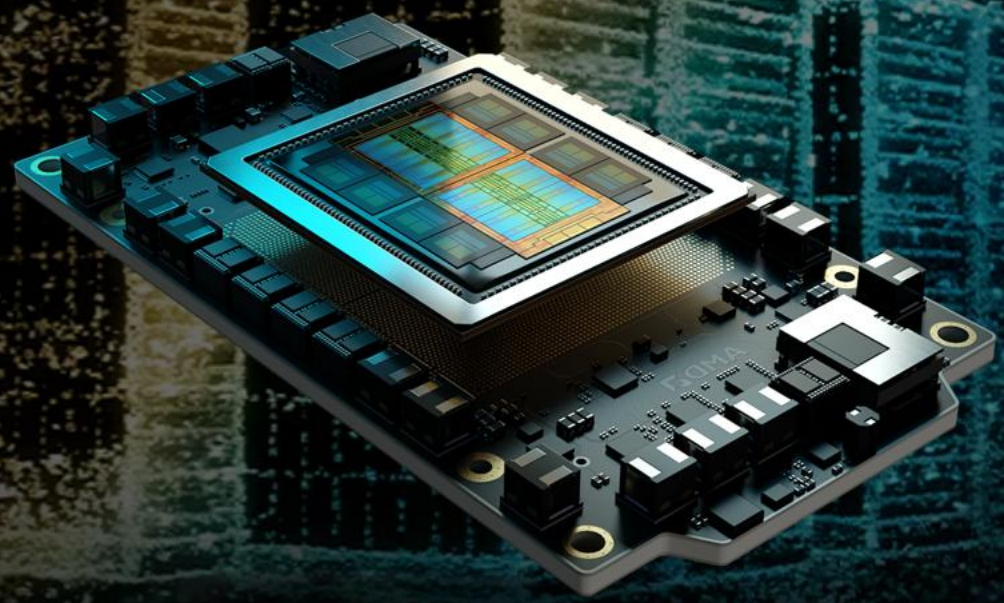
4th Gen AMD CDNA™ Generative AI Architecture Powering AMD Instinct™ MI350 Series GPUs and Platforms



Michael Steffen – AMD Fellow, Instinct SoC Architect

Michael Floyd – AMD Fellow, Instinct Validation Architect

Hot Chips 2025

AMD
together we advance_

Large Language Models: Explosive Growth

Trends effecting GPUs

Larger parameter count models

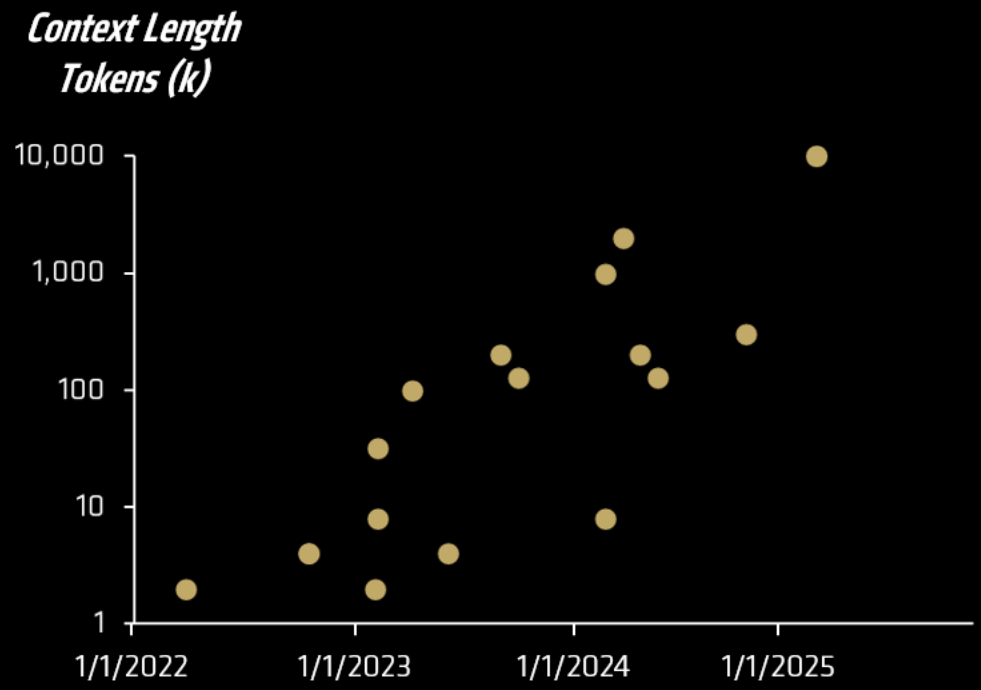
- Innovation in datatype formats to reduce memory footprint
- Richer input modalities (particularly video)
- Larger memory requirements for storing the parameters

Longer context lengths

- Chain of thought reasoning
- Images and video input modalities
- Longer text documents for Retrieval-Augmented Generation (RAG)

Increased Agentic AI processing

- Trend toward Reinforcement Learning (RL) methods
- Synthetic Data generation creating more adaptable models
- More MoE models using distributed computing parallelism



Reflects major models/vendors including OpenAI, Anthropic, Google®, Meta
*Assumes Llama 3.1 405B Arch. For estimate.
Source: Epoch.AI, Web Search, Team Analysis

AMD

Generative AI Needs

More Performant Solutions for AI Models

GPUs now require more:

- Memory bandwidth to keep them fed
- Memory capacity with low latency to house the models
- Performant ALUs with faster, lower precision data types
- Power-efficient architectures in the data fabric and system to boost their frequency
- Scale-up bandwidth to support larger multi-engine model training and inferencing

AMD Instinct™ MI350 Series

Keeping Leadership Roadmap Commitment In AI & HPC

2021

AMD CDNA 2

MI200 SERIES

IMPROVED HPC AND AI

Denser compute with larger memory capacity & bandwidth

MI250X "Frontier" Supercomputer

2023

AMD CDNA 3

MI300 SERIES

DATA CENTER APU & DISCRETE GPU

Improved AI performance, unified memory, and in-node networking

MI300A "El Capitan" Supercomputer

2024

AMD CDNA 3

MI325X

LARGER AI MODELS

Increased HBM memory capacity and bandwidth in the same form factor

2025

AMD CDNA 4

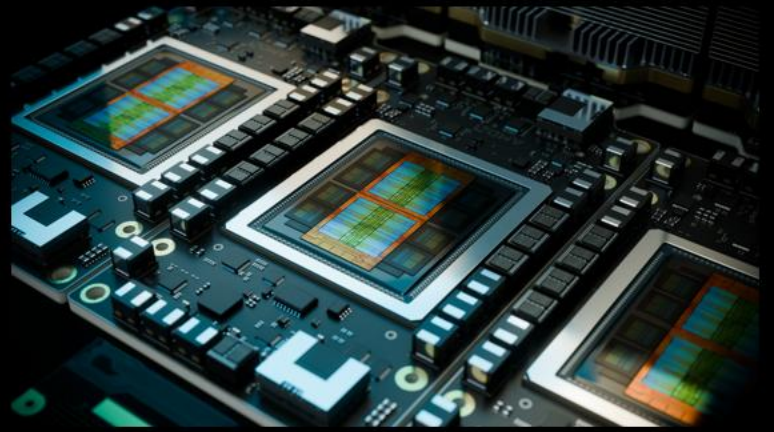
MI350 SERIES

GENERATIVE AI INFERENCING AND TRAINING

Continued advances in AI data formats, memory capacity & bandwidth, and networking

AMD Instinct™ MI350 Series Architecture Enhancements

Improving Performance and Efficiency of AI Workloads



> Extend HBM Bandwidth and Capacity leadership to accelerate memory bound applications

> Support faster training and inference on larger models with increased link speeds

> Enhance power efficiency and performance

Increases performance by reducing un-core power

- Wider Infinity Fabric enables higher bandwidth at more power efficient frequencies

Supports lower precision data formats

- Provide full access to FP8 (scaled & un-scaled) data types
- Add industry standard micro-scaled MXFP6 and MXFP4 data types

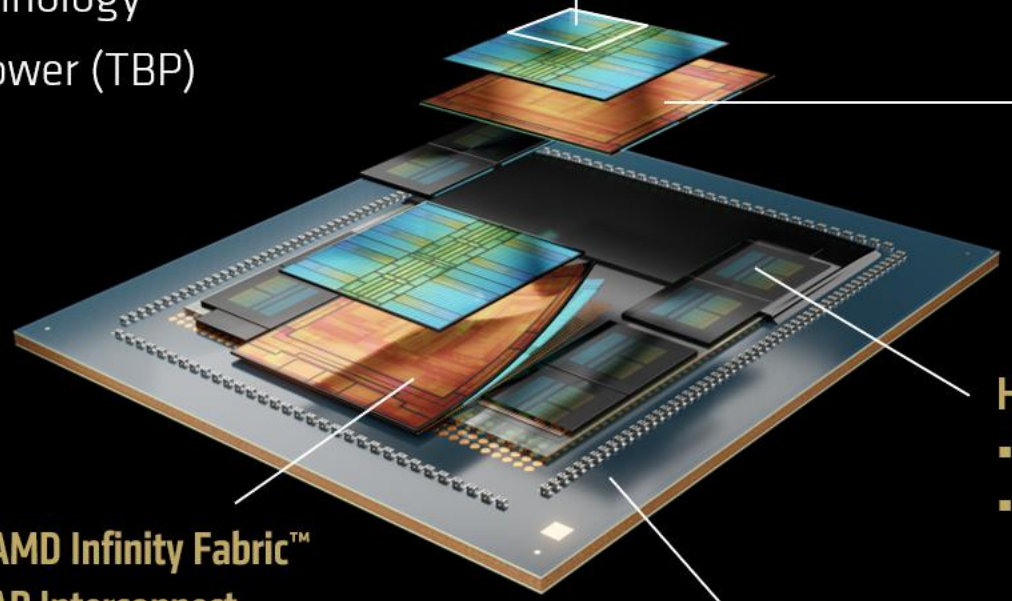
Per OAM Socket	MI300X	MI325X	MI350X	MI355X
Power (Watts, TBP)	750	1000	1000	1400
Compute Max Clock (MHz)	2100		2200	2400
Local Data Share (KB, per CU)	64		160 (2.5x*)	
HBM Capacity (GB)	192	256 (1.3x*)	288 (1.5x*)	
HBM peak BW (TB/s)	5.3	6.0 (1.15x*)	8.0 (1.5x*)	
Infinity Fabric peak BW (GB/s, bi-dir)	896		1075.2 (1.2x*)	

* Compared to MI300X

AMD Instinct™ MI350 Series GPU

State-of-the-Art Construction

- 185 Billion transistors
- 3D Multi-Chiplet (2 chiplet types + HBM3E)
- Heterogenous process technologies
- Proven COWOS-S packaging technology
- Up to 1,400 watts Total Board Power (TBP)



Accelerator Complex Die (XCD)

- Latest TSMC 3nm (N3P) Technology
- Latest process for compute perf/watt optimization
- Max 8 per package (Max 4 per IOD)

AMD I/O Base Die (IOD)

- TSMC 6nm FinFET (N6) Technology
- Mature process for optimal yield & cost
- Max 2 per package

HBM3E

- Max 4 stacks per IOD (Max 8 per package)
- Up to 288 GB of HBM memory

AMD Infinity Fabric™ AP Interconnect

Package

AMD Instinct™ MI350 Series GPU Chiplets

MI350 I/O Die (IOD)

HBM3E Memory

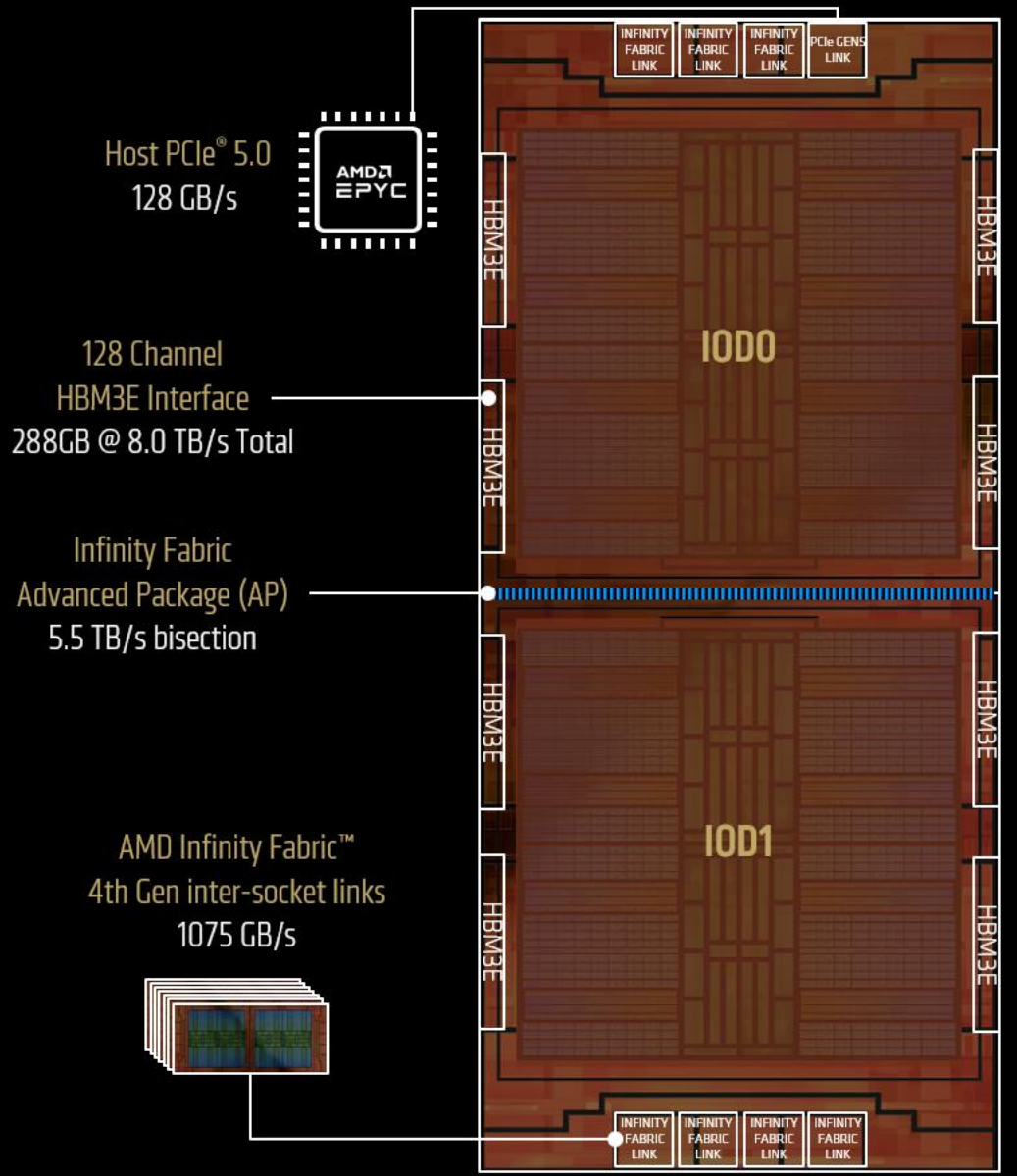
- 8 physical stacks with 128 Channel Interface
- 288GB total – 36GB per 12Hi stack at 8Gbps
- 8.0 TB/s total bandwidth

AMD Infinity Fabric™

- Supports full bandwidth, flat address space for all chiplets
  - using Infinity Fabric Advanced Package (AP) with 5.5 TB/s bisection
- 2 IOD provide enhanced power efficiency compared to MI300
  - Wider data pipelines enable higher bandwidth at lower frequencies
- 256MB AMD Infinity Cache™

I/O Connectivity

7 - 4th Gen AMD Infinity Fabric™Links between sockets
  153.6 GB/s per link bi-directional
  1075.2 GB/s bi-directional aggregate
1 – Gen5 x16 PCIe® Link to Host
  128 GB/s per link

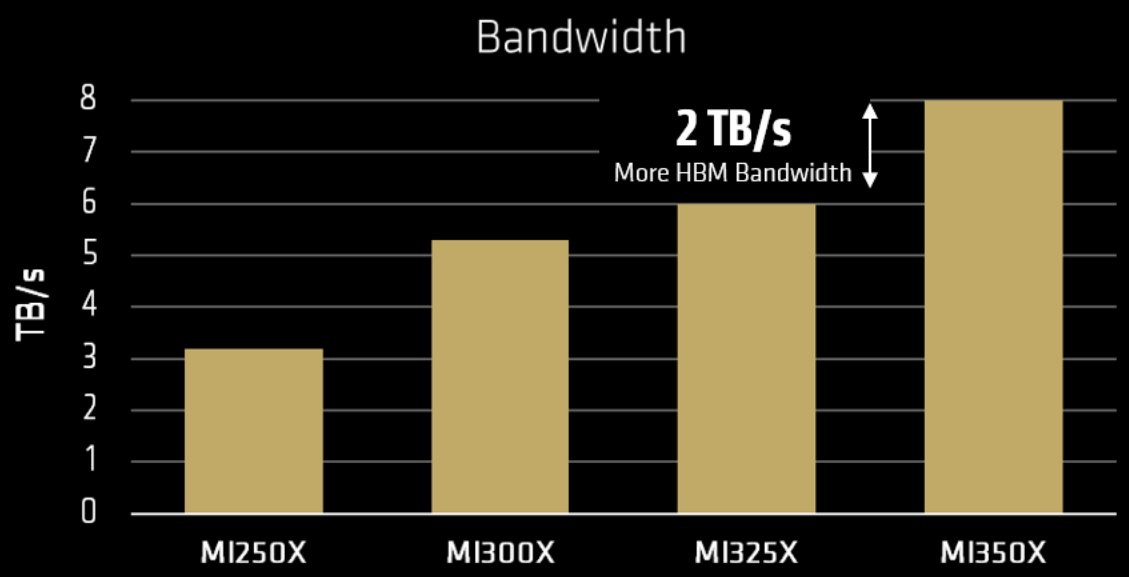


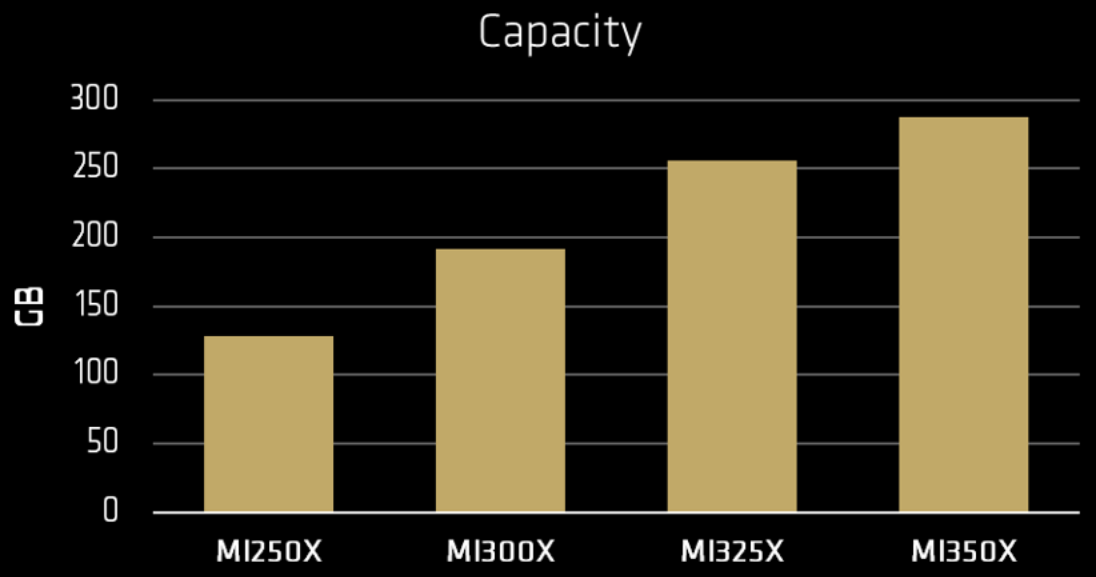
Peak theoretical wire rates; real workloads may have lower effective utilization

AMD

AMD Instinct™ MI350 Series GPU

Pushing AMD Instinct advantage on memory capacity and bandwidth





Max LLM size per system (MXFP4)

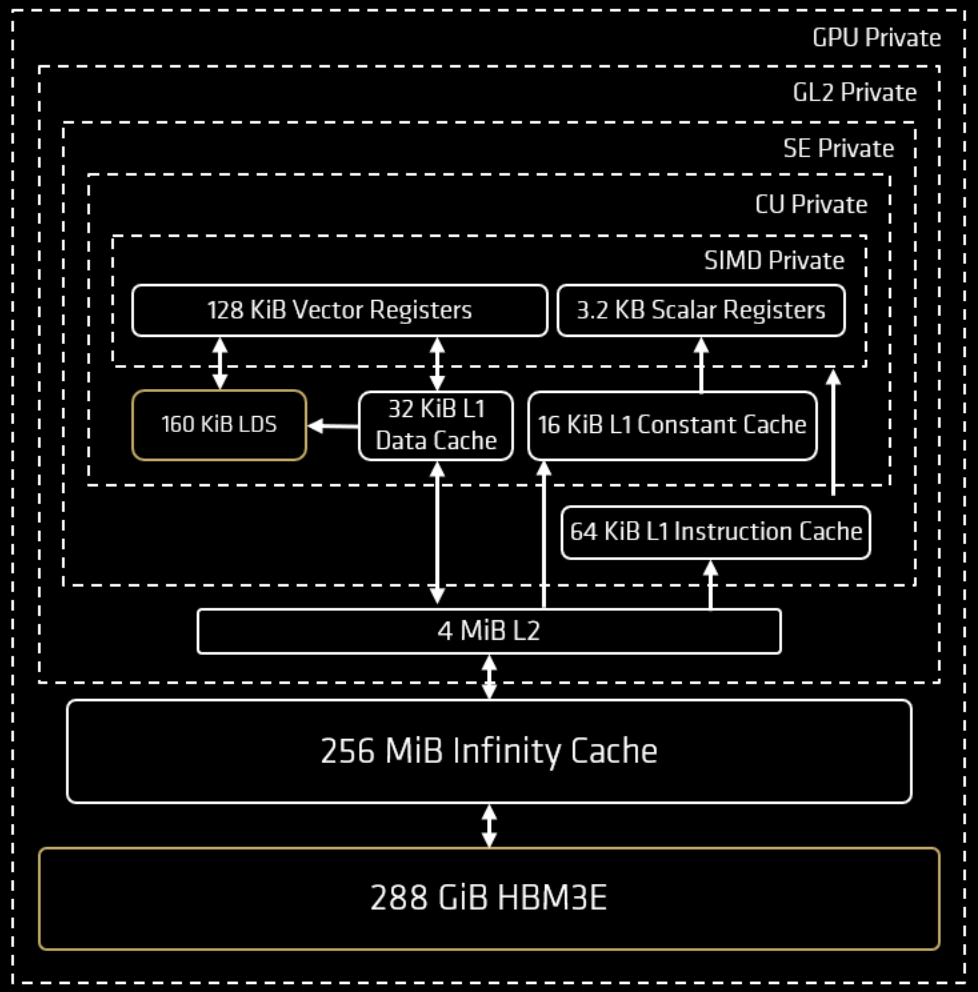
Single Nvidia B200 HGX		Single AMD Instinct MI355X Platform	
1.4 TB HBM3 @ 62 TB/s		2.3 TB HBM3E @ 64 TB/s	
Training	Inference	Training	Inference
~84B	~2600B	~135B	~4180B

See endnote MI350-059, MI350-060

AMD

AMD Instinct™ MI350 Series GPU

Cache and Memory Hierarchy



Level	Capacity
VGPR	128 KiB / SIMD
LDS	160 KiB / CU
L1 Data Cache	32 KiB / CU
L2 Cache	4 MiB / L2
Infinity Cache	256 MiB / GPU
HBM	288 GiB / GPU

Local Data Share (LDS) Enhancements

- Increased LDS banking for capacity and bandwidth
- 2.5x increase in capacity from 64 KB to 160 KB
- 2x increase in read bandwidth to 256 bytes per clock

Infinity Fabric Enhancements

- 5.5 TB/s peak C2C bisectional bandwidth (~1.3x MI300X)
- Two IOD chiplets (vs. four in MI300X)
- Widened NoC to deliver peak HBM bandwidth at Vmin
- Lower fabric power for improved compute performance

AMD Instinct™ MI350 Series GPU Chiplets

AMD Accelerator Complex Die (XCD)

Compute

32 AMD CDNA™ 4 Compute Units (CU) per XCD
Up to 8 XCD per package
256 CU Total
1,024 Matrix Cores
16,384 Stream Processors
2.4 GHz Peak Engine Clock

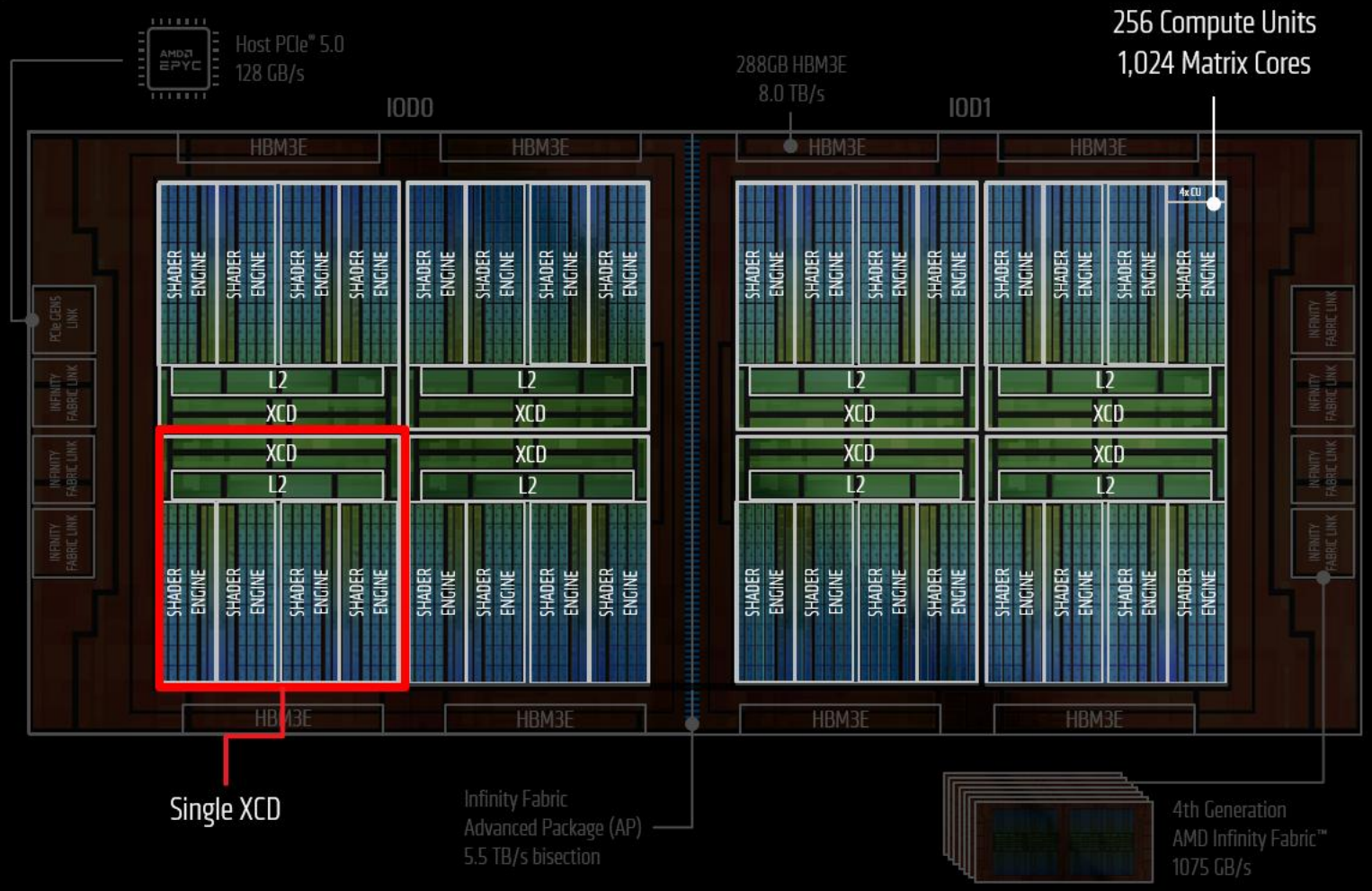
Memory

512 KiB Vector Registers/CU
160 KiB Local Data Storage (LDS) @ 537 GB/s/CU
32 KiB of L1 cache per CU
4 MiB Shared L2 cache per XCD

Compute Partitioning

- XCDs have their own GPU host interface using PCIe® Single Root I/O Virtualization (SR-IOV)
- Each XCD
  - can be its own GPU Compute Partition

  OR
  - can collaborate with other XCDs forming various sizes of GPU Compute Devices



Peak theoretical wire rates; real workloads may have lower effective utilization

AMD Instinct™ MI350 Series Compute Units (CU)

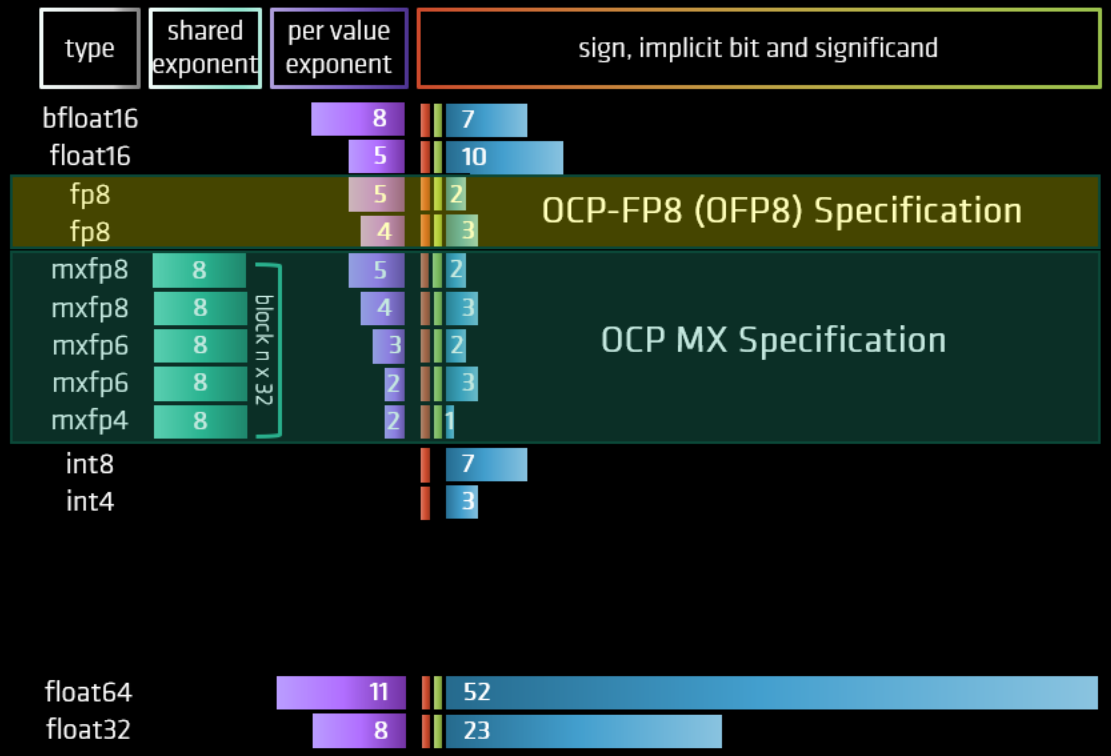
Supported Data Formats & Generational Performance Improvements

AI Lower Precision Data Types

Computation	MI350 Series (FLOPS/clock/CU)	MI355X (Peak Theoretical)	MI355X Peak Speedup‡
Vector FP16+	256	157.3 TFLOPs	~1.0x
Matrix FP16/BF16	4096	2.5 PFLOPs	1.9x
Matrix FP8	8192	5.0 PFLOPs✧	1.9x
Matrix INT8/INT4	8192	5.0 PFLOPs	1.9x
Matrix MXFP6/MXFP4	16834	10 PFLOPs	New to MI350*

High Performance Computing (HPC)

Vector FP64	128	78.6 TFLOPs	~1.0x
Matrix FP64	128	78.6 TFLOPs	~0.5x
Vector FP32+	256	157.3 TFLOPs	~1.0x
Matrix FP32	256	157.3 TFLOPs	~1.0x



Peak theoretical performance without Sparsity
‡ Compared to MI300X * AMD Instinct™ MI300X GPUs do not support MXFP6, MXFP4
+ Refers to non-packed vector instructions on AMD Instinct™ MI300X and MI350X Series GPUs
✧ AMD Instinct MI350 Series GPUs support both OFP8 format as well as OCP MX formats
◊ AMD Instinct™ MI350 Series GPUs support TF32 through software emulation

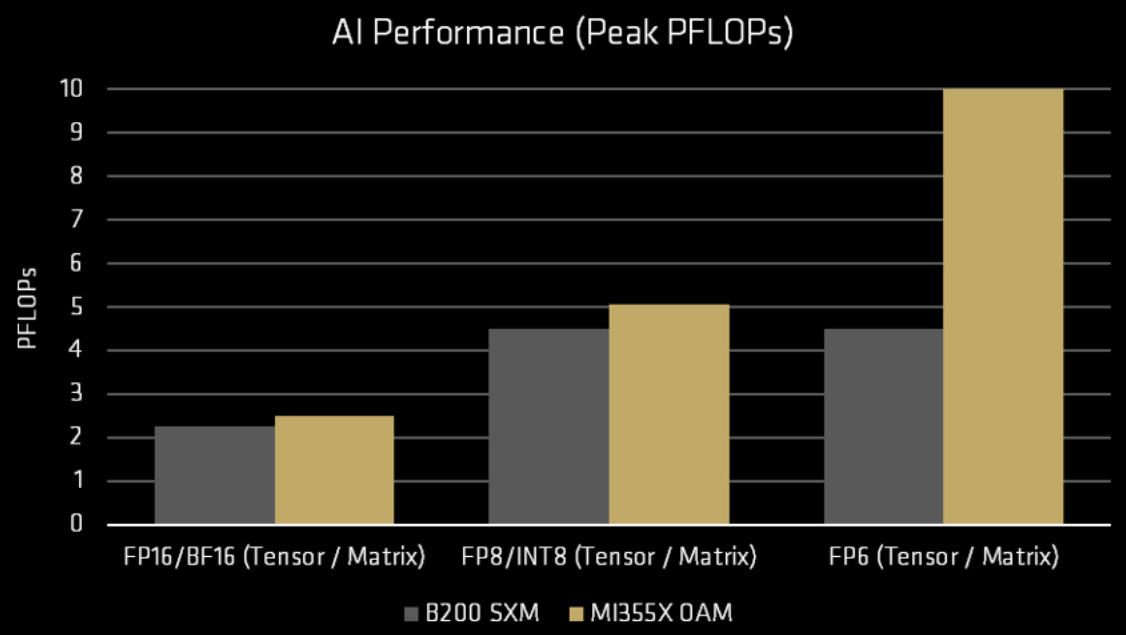
AMD

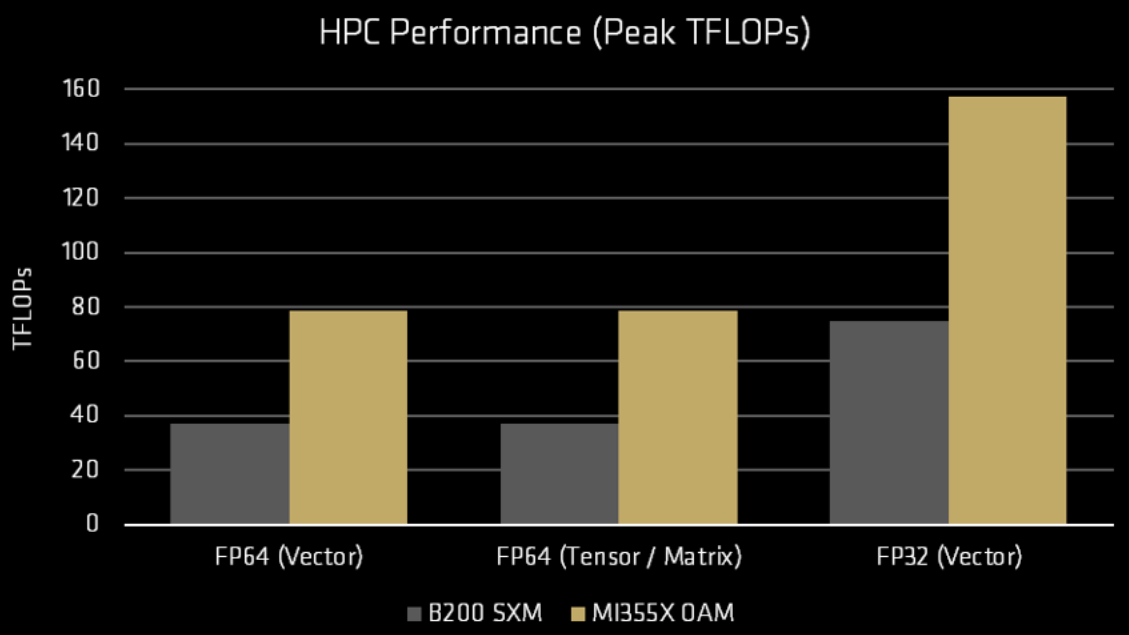
AMD Instinct™ MI350 Series Compute Units (CU)

Supported Data Formats & Performance Comparison

Peak math throughputs

- Up to 2.1X the AI performance vs. competitive accelerators
- Up to 2.1X the HPC performance vs. competitive accelerators

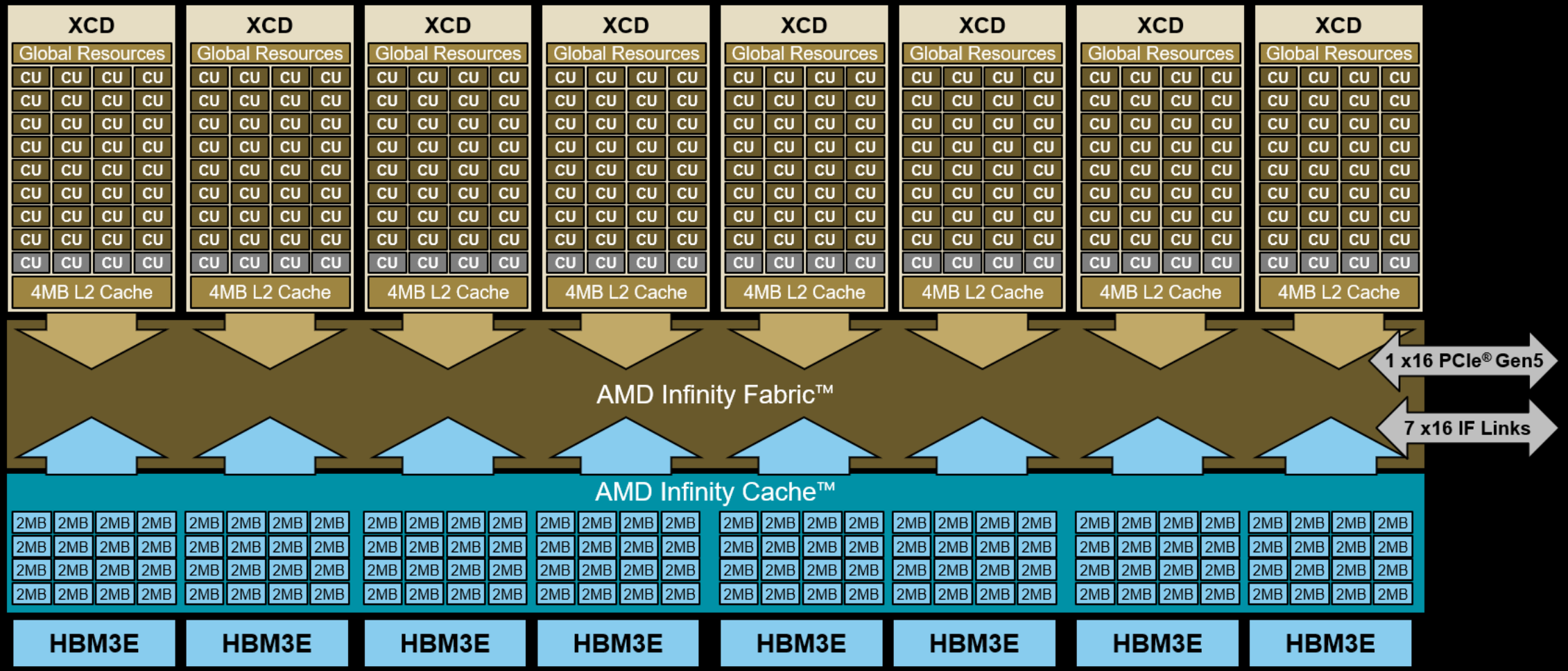




Peak theoretical performance without Sparsity
See endnote MI350-006, MI350-009

AMD Instinct™ MI350 Series GPU

SoC Block Diagram

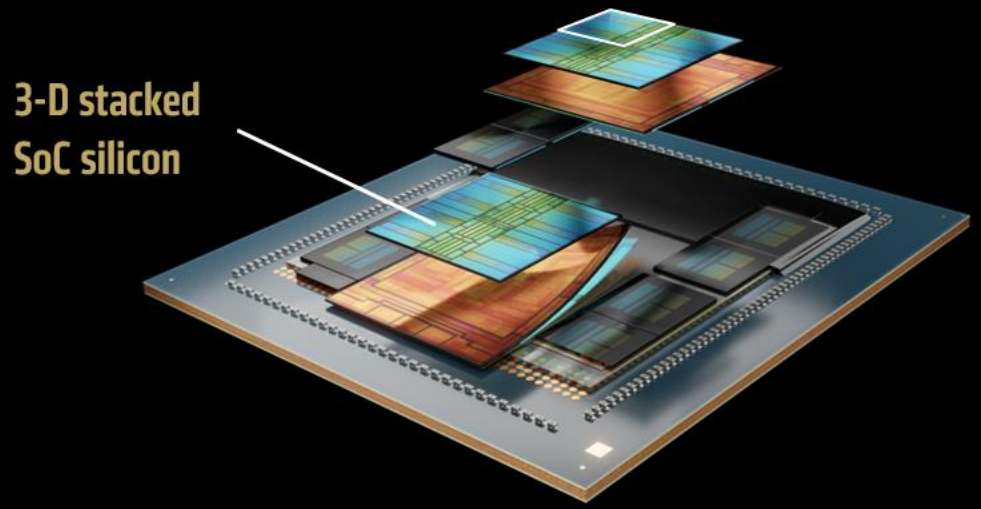


AMD Instinct™ MI350 Series GPU

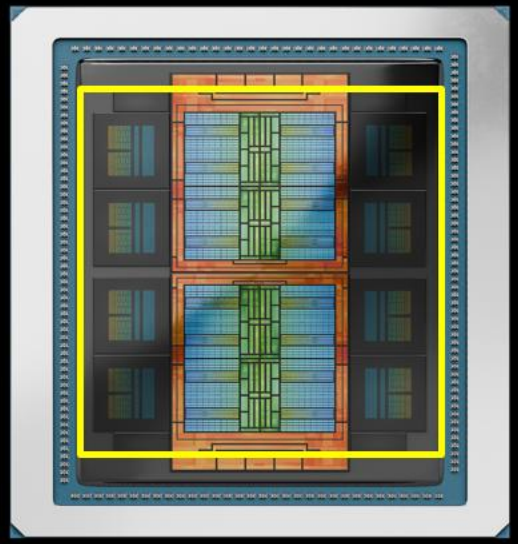
Flexible GPU Partitioning per Socket

Maximize GPU Utilization with:

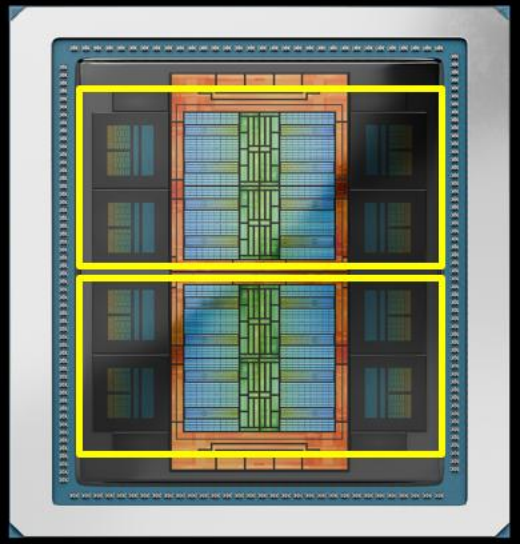
- One of two localized memory NUMA-Per-Socket modes NPS1 or NPS2



MI350 Series Memory Partitioning Options



Single partition
(NPS1)



Two partitions
(NPS2)

See endnote MI350-012, MI350-037

AMD

AMD Instinct™ MI350 Series GPU

Flexible GPU Partitioning per Socket

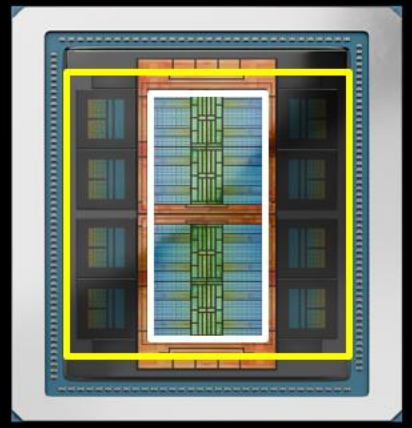
Maximize GPU Utilization with:

- One of two localized memory NUMA-Per-Socket modes NPS1 or NPS2
- Up to 8 spatial compute partitions per socket 1, 2, 4, or 8 logical GPUs

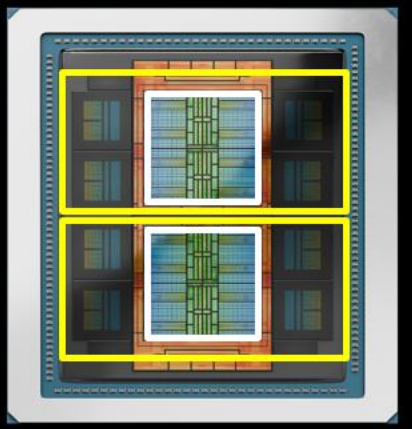
For example, Instinct MI350 Series GPUs can support:

- Up to 520B parameter AI Model in SPX+NPS1
- Up to 8x Instances of Llama 3.1 70B in CPX+NPS2

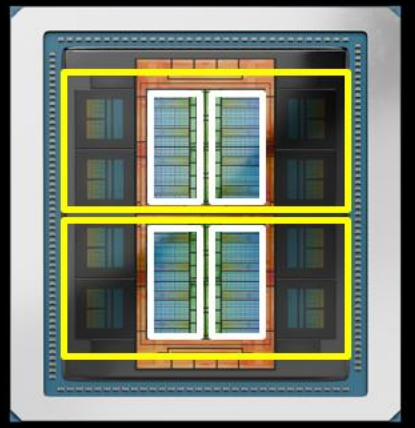
MI350 Series Partitioning Options



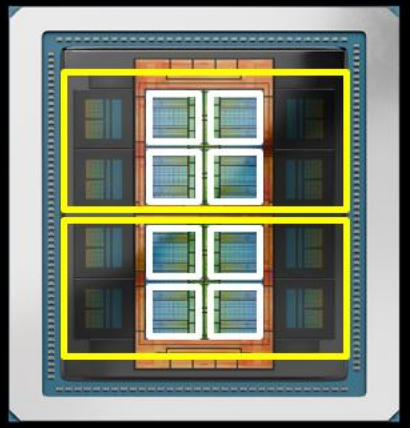
Single partition
(SPX+NPS1)



Two partitions
(DPX+NPS2)



Four partitions
(QPX+NPS2)



Eight partitions
(CPX+NPS2)

See endnote MI350-012, MI350-037

AMD

AMD Instinct™ MI350 Series Node GPU Interconnect

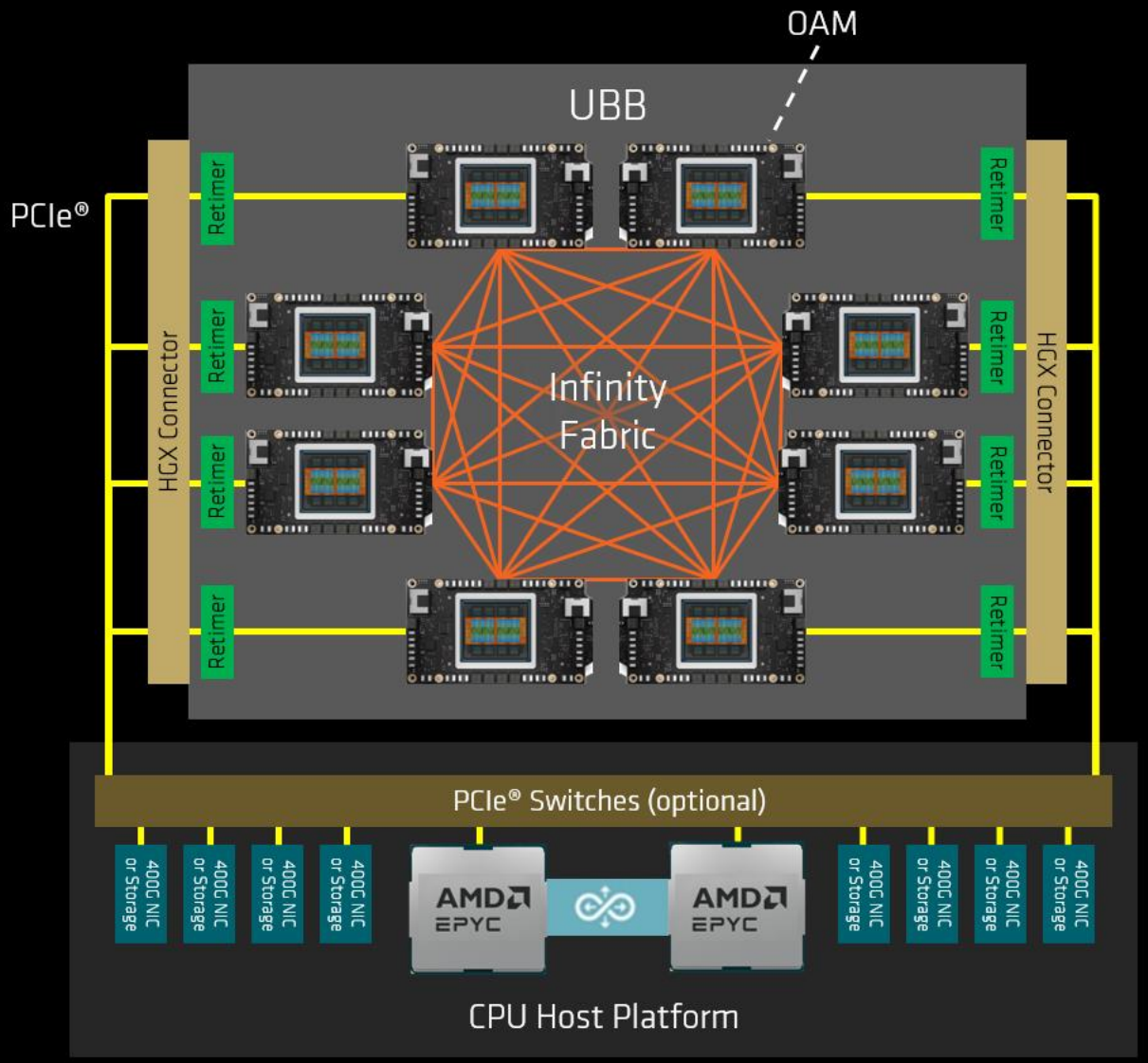
Infinity Platform

PCIe® Gen5 x16 per GPU

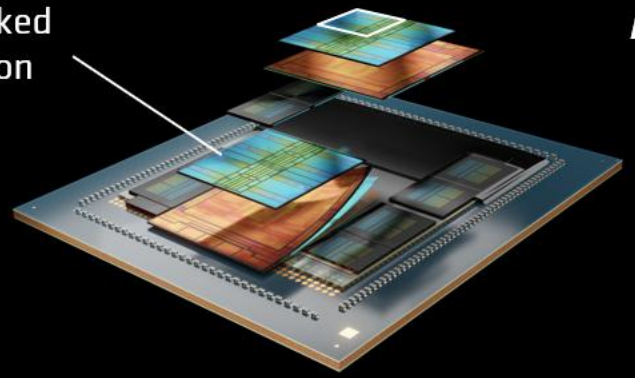
- For CPU host server connectivity and I/O
- HGX connectors with PCIe® retimers per link

AMD's Infinity Fabric™

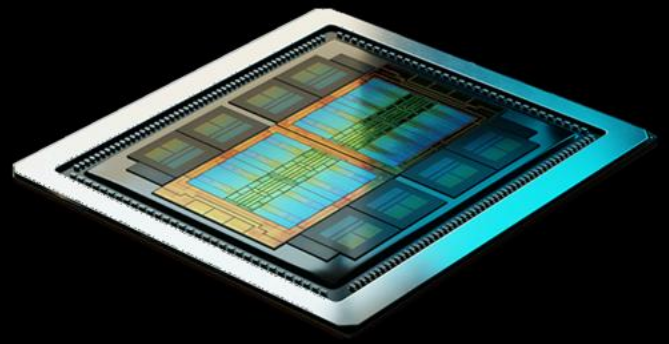
- Direct all-to-all connectivity between 8 GPUs
- Seven bi-directional links @ 154 GB/s
- 20% faster than previous Instinct offering (38.4 vs. 32 Gbps)



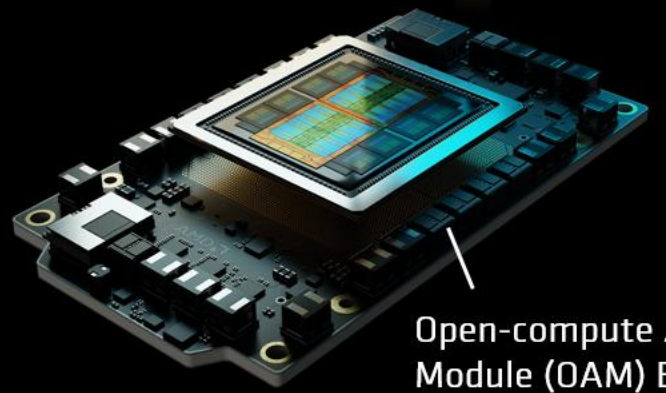
AMD Instinct™ MI350 Series Air Cooled OAM Assembly



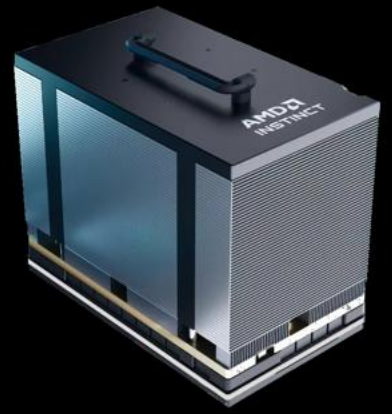
Package Assembly



OAM Assembly



OAM Heatsink Attach



AMD

AMD Instinct™ MI350 Series
Air Cooled UBB Platform Assembly

8 OAMs plug into UBB

Universal Base Board (UBB) 2.0

UBB plugs into Host Node Platform

Industry Standard Host Node

Host Node plugs into EIA Rack

AMD

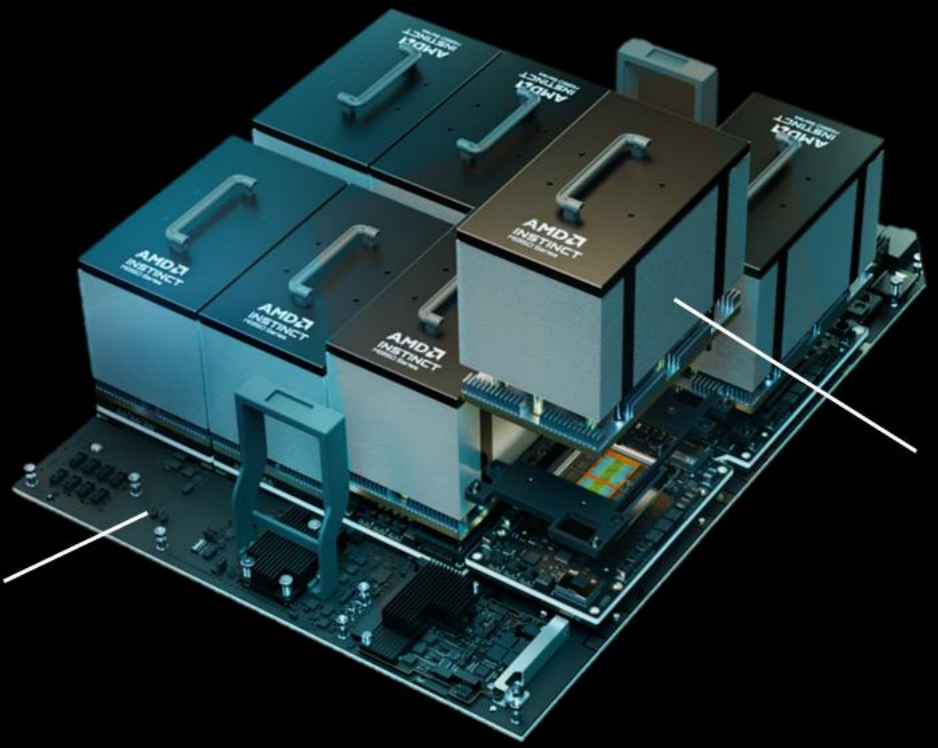
AMD Instinct™ MI350 Series Air-Cooled Platforms

Leverage Existing Datacenter Infrastructure

- PCIe® Gen 5 attach to CPU Host/NIC Interfaces
- Updated air-flow, inlet temp, and power subsystems

Existing:

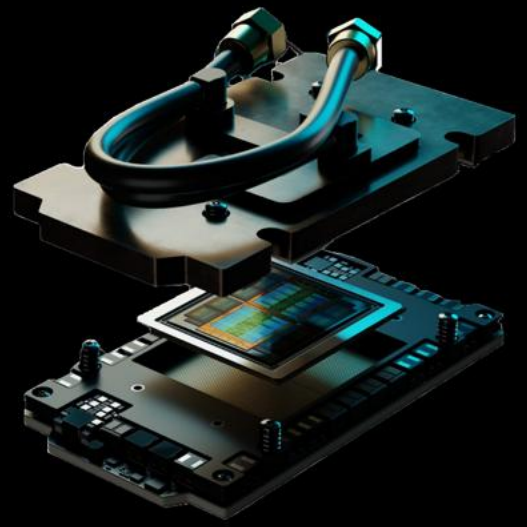
- MI300X AC 750W
- MI325X AC 1000W



1000W Products

4U Options Fit into existing UBB8

AMD Instinct™ MI350 Series
DLC OAM & UBB Platform Assembly



8 OAMs plug into UBB

Universal Base Board (UBB) 2.0

UBB plugs into Host Node Platform

Host Node plugs into EIA or "ORv3" Liquid Cooled Rack

AMD

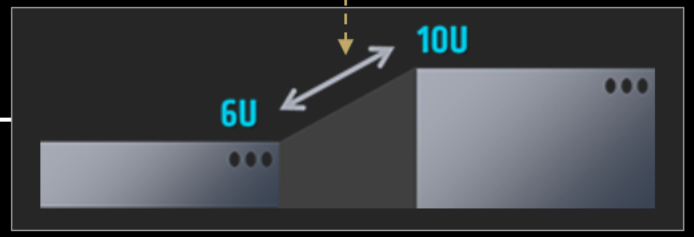
AMD Instinct™ MI350 Series Platforms

Flexible Platform Node Offerings

8x MI350X Series GPUs | 2.25TB HBM3E | 1075GB/s AMD Infinity Fabric™ Bandwidth

AMD Instinct™
MI350X Platform

36.9PF FP16/BF16 **73.8PF** FP8



Air Cooled (AC)
Compute Nodes

AMD Instinct™
MI355X Platform

40.2PF FP16/BF16 **80.5PF** FP8

Direct Liquid Cooled (DLC)
Compute Nodes

Peak theoretical performance with Sparsity

AMD

AMD Instinct™ MI350 Series Rack-Scale Solutions

Industry Standard Racks, Nodes & Platforms

MI355X DLC "ORv3" Rack

- 128 GPUs
- 36 TB HBM3E
- 1.3 EF FP8
- 2.6 EF FP6
- 2.6 EF FP4

2 OU

MI355X DLC EIA Rack

- 96 GPUs
- 27 TB HBM3E
- 1 EF FP8
- 2 EF FP6
- 2 EF FP4

3 OU

MI350X AC EIA Rack

- 64 GPUs
- 18 TB HBM3E
- 0.6 EF FP8
- 1.2 EF FP6
- 1.2 EF FP4

6 RU

Peak theoretical performance with Sparsity

AMD Instinct™ MI350 Series Rack Infrastructure

Fully Built on Open Standards with Industry-leading chips

x86 CPU
5th Gen EPYC™

Instinct GPU
MI350 series

UEC Scale-Out NIC
AMD Pollara NIC

OPEN Compute Project®

OCP Design
Industry Standard Design

Ultra Ethernet Consortium

UEC Supported
Open Network Solution

AMD ROCm™ 7 Software on AMD Instinct™ MI350 Series GPUs

Accelerating AI innovation and developer productivity

Open Source Advantage

- Faster advances than with closed source (e.g., TensorRT-LLM)
- Deeper engagement and collaboration with open source community from day one
- Out-of-the-box support for the latest OSS models from OpenAI
- Fosters innovation leading to breakthroughs, such as distributed inference techniques:
  - Expert parallelism with Load Balancing
  - Prefill/Decode Disaggregation

MI350 Series Optimizations

- Utilizing the AMD CDNA 4 architecture
- Provide flexibility of new FP4 & FP6 data types
- Incorporate HBM3E for improved data handling and performance

Latest Algorithms & Models

- Improve reasoning capabilities & attention mechanisms
- Implement sparse MoE models for enhanced computational efficiency

Scaling AI

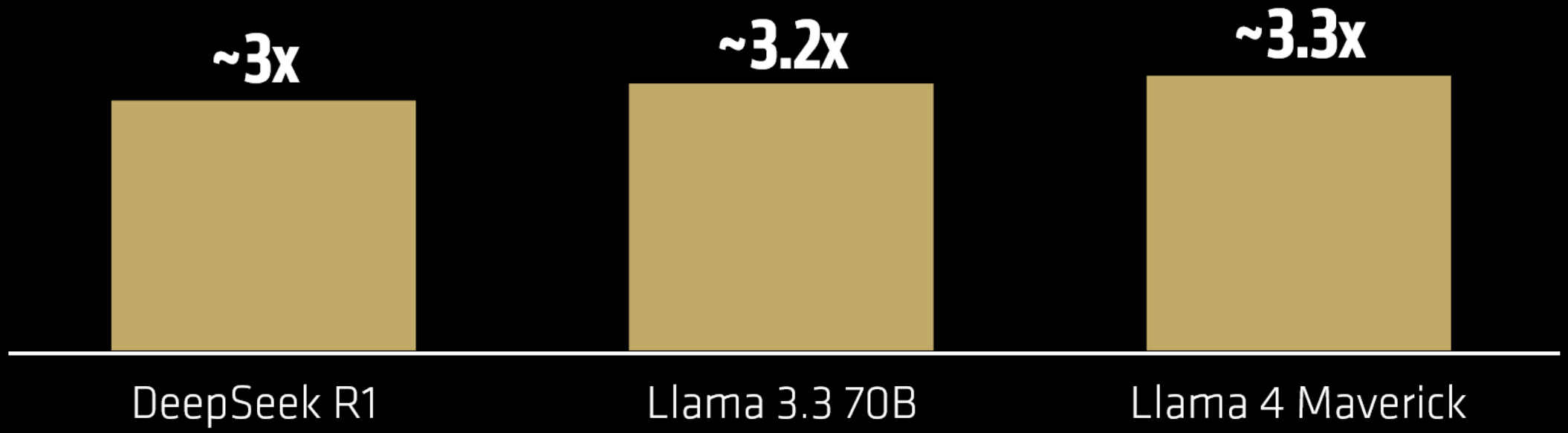
- Facilitate Rack-scale and distributed inference processes
- Enable large-scale reinforcement learning
- Promote effective resource allocation and performance scaling

Enterprise Solutions

- Offer streamlined tools for managing enterprise AI initiatives and cluster resources
- Designed for scalability across a variety of industry applications
- Plug into enterprise ops with full Kubernetes® and Slurm support

AMD Instinct™ MI350 Series GPU Inference

Increasing Performance For Most Popular Inference AI Models



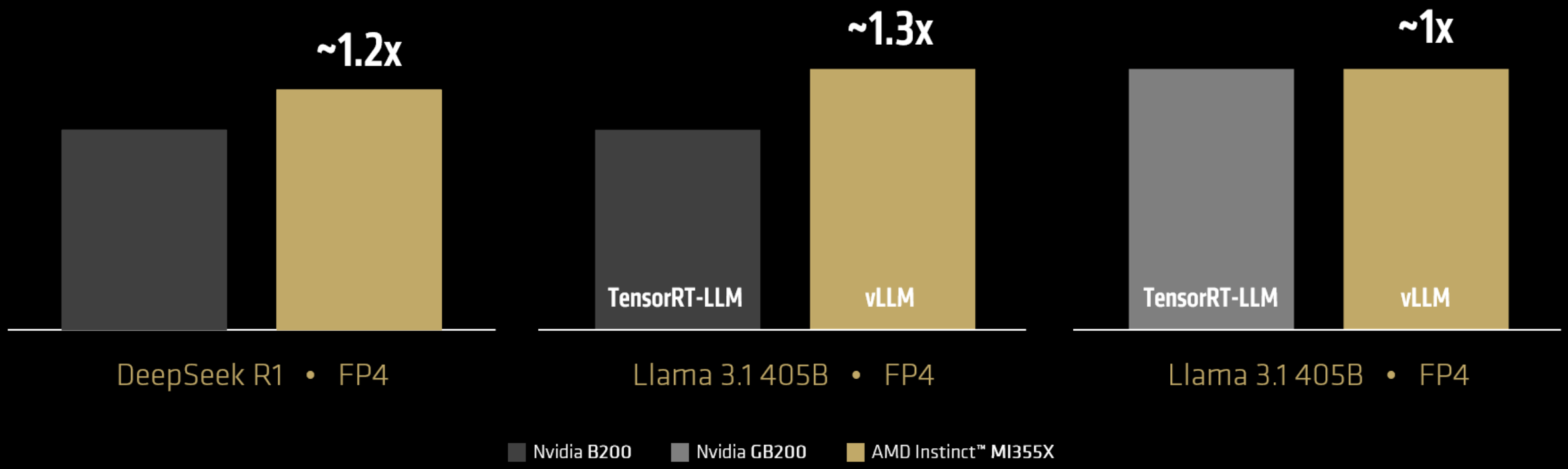
MI355X vs. MI300X

*Measured throughput, tokens/s MI300X (FP8) and MI355X (FP4)
See endnote MI350-041, MI350-043, MI350-048, MI350-049

AMD

AMD Instinct™ MI350 Series GPU Leadership

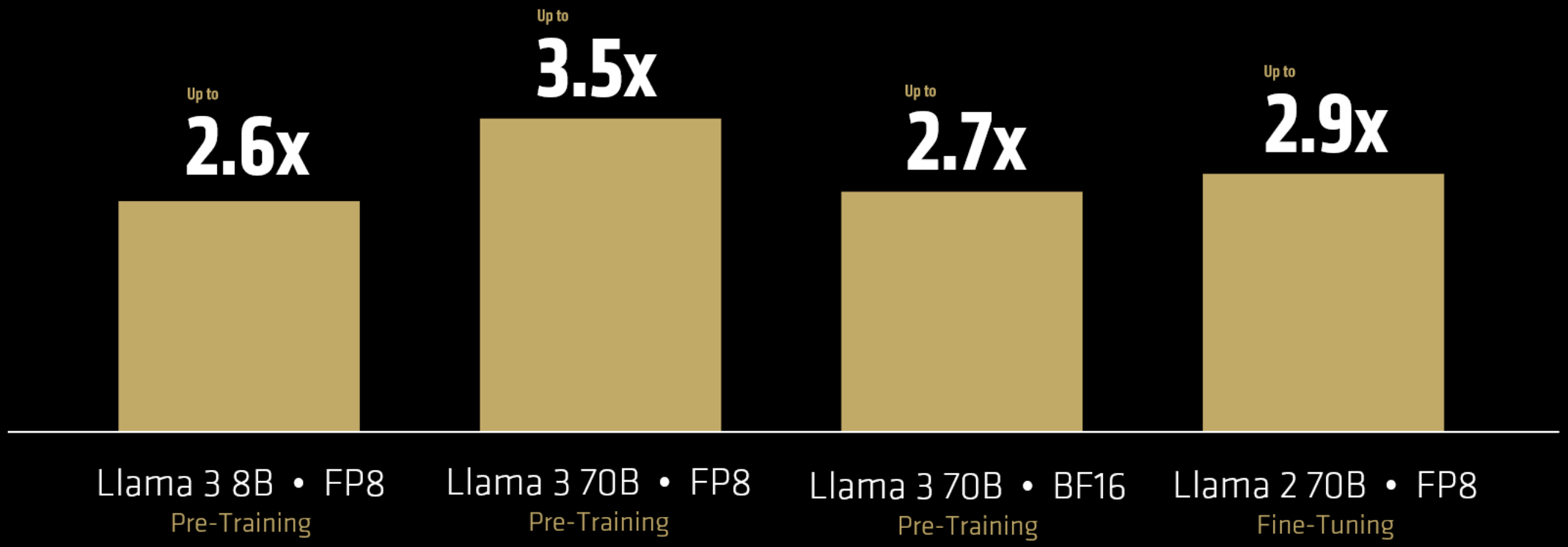
Leadership Performance For Large Inference Models



Equal GPU Count for Performance Compare with GB200
*Throughput See endnote MI350-038, MI350-039, MI350-040

AMD Instinct™ MI350 Series GPU Training

Increasing Pre-Training and Fine-Tuning Performance



MI355X vs. MI300X

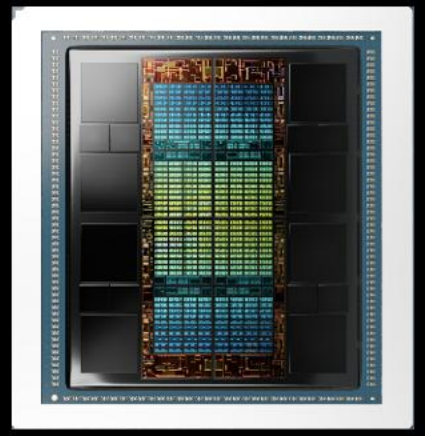
See endnote MI350-034, MI350-035

AMD

AMD INSTINCT

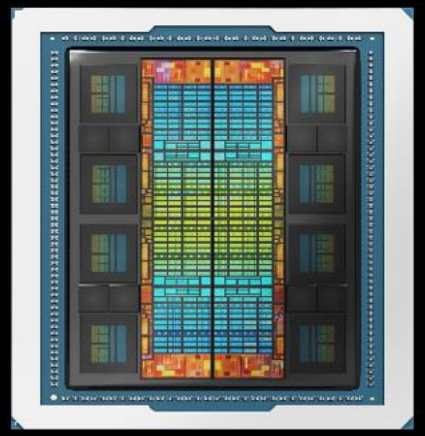
Delivering on Annual Roadmap Commitment

AMD Instinct™
MI300A/X



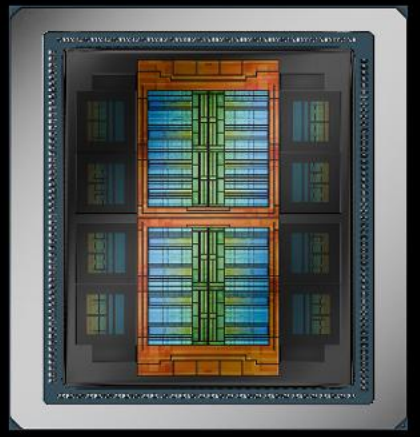
2023

AMD Instinct™
MI325X



2024

AMD Instinct™
MI350 SERIES



2025

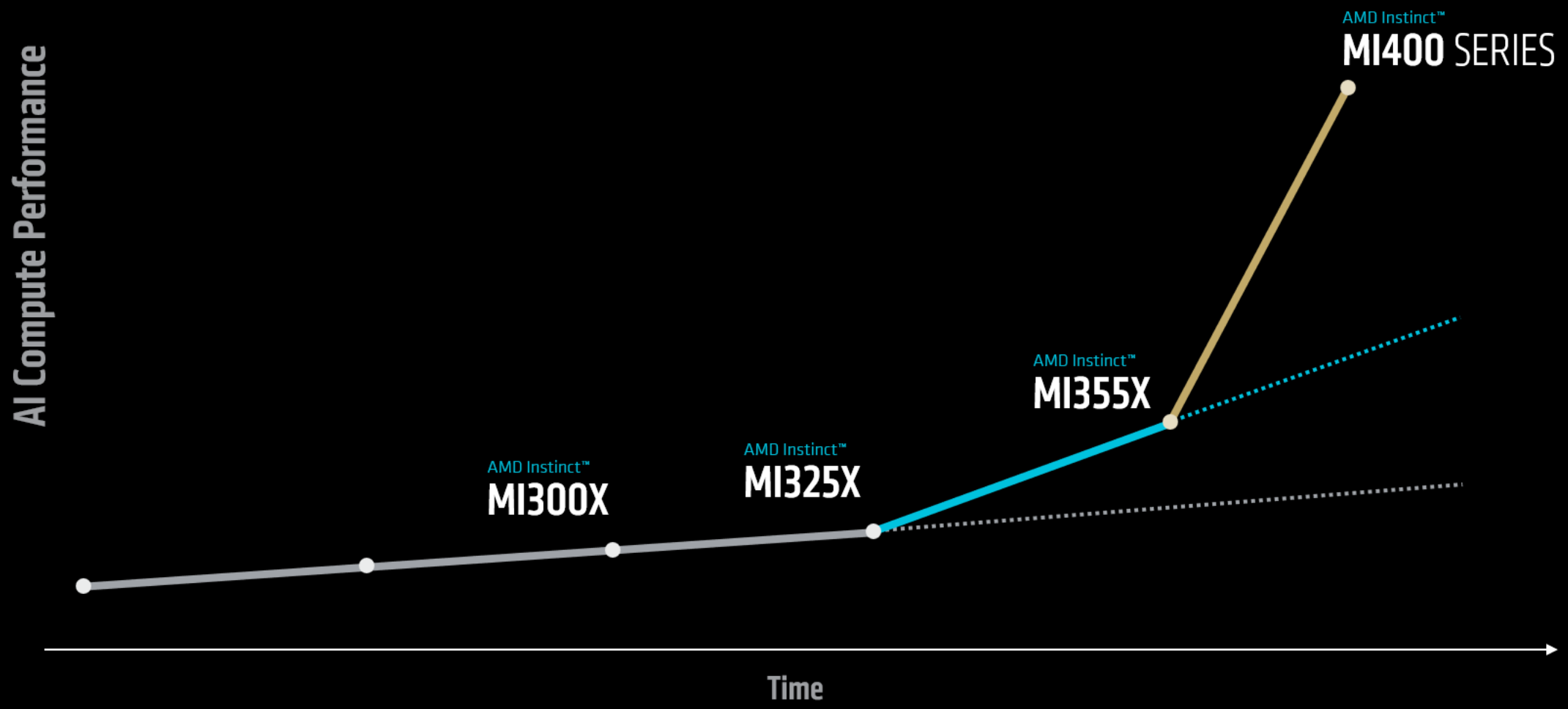
AMD Instinct™
MI400 SERIES

2026

Roadmap subject to change

AMD

Accelerating AI Compute Performance



See endnote: MI350-047

AMD

AMD Instinct™ MI400

Leadership Gen AI Accelerator

40 PF | 20 PF	432 GB	19.6 TB/s	300 GB/s
FP4 • FP8 Flops	HBM4 Memory Capacity	Memory Bandwidth	Scale Out Bandwidth / GPU

Coming in 2026

Specs based on engineering projections

Preview

AMD “Helios”

Optimized AI Rack Solution

AMD EPYC | AMD INSTINCT | AMD PENSANDO | AMD ROCm

Available in 2026

OPEN Compute Project | ULTRA ACCELERATOR LINK™ | Ultra Ethernet Consortium

AMD

AMD “Helios” AI Rack

Rack Scale AI Performance Leadership

	AMD Instinct™ MI400 Series “Helios”	vs. Vera Rubin Oberon
GPU DOMAIN	72	1.0x
SCALE UP BANDWIDTH	260 TB/s	1.0x
FP4 • FP8 FLOPS	2.9 EF • 1.4 EF	1.0x
HBM4 MEMORY CAPACITY	31 TB	1.5x
MEMORY BANDWIDTH	1.4 PB/s	1.5x
SCALE OUT BANDWIDTH	43 TB/s	1.5x

Source: AMD specs based on engineering projections, results subject to change

AMD | Advancing AI Infrastructure on an Annual Cadence

2025

AMD EPYC
"TURIN"

AMD Instinct
MI350 SERIES

AMD Pensando
POLLARA 400

2026

AMD EPYC
"VENICE"

AMD Instinct
MI400 SERIES

AMD Pensando
"VULCANO"

"Helios"

2027

AMD EPYC
"VERANO"

AMD Instinct
MI500 SERIES

AMD Pensando
"VULCANO"

Next Gen AI Rack

Roadmap subject to change

AMD

Endnotes

MI350-006: Calculations by AMD Performance Labs in May 2025, based on the published memory capacity specifications of AMD Instinct™ MI350X / MI355X OAM GPU vs. an NVIDIA Hopper H200 GPU. Server manufacturers may vary configurations, yielding different results. MI350-006

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

MI350-012: Based on calculations by AMD Performance Labs as of April 17, 2025, on the published memory specifications of the AMD Instinct MI350X / MI355X GPU (288GB) vs MI300X (192GB) vs MI325X (256GB). Calculations performed with FP16 precision datatype at (2) bytes per parameter, to determine the minimum number of GPUs (based on memory size) required to run the following LLMs: OPT (130B parameters), GPT-3 (175B parameters), BLOOM (176B parameters), Gopher (280B parameters), PaLM 1 (340B parameters), Generic LM (420B, 500B, 520B, 1.047T parameters), Megatron-LM (530B parameters), LLaMA (405B parameters) and Samba (1T parameters). Results based on GPU memory size versus memory required by the model at defined parameters plus 10% overhead. Server manufacturers may vary configurations, yielding different results. Results may vary based on GPU memory configuration, LLM size, and potential variance in GPU memory access or the server operating environment. *All data based on FP16 datatype. For FP8 = X2. For FP4 = X4. MI350-012.

MI350-037: Based on calculations by AMD Performance Labs as of June 6, 2025, on the published memory specifications of the AMD Instinct™ MI350X / MI355X GPUs (288GB) versus the memory required by an LLM, Meta Llama 3.1 (70B parameters), to determine the number of virtualized instances 1 GPU and 8 GPU serves can run on the model. Calculations performed with FP4 precision datatype at (0.5) bytes per parameter, plus overhead. [JB1] [GL2] Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of the latest drivers and optimizations. MI350-037.

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

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

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

Endnotes

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

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

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

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

MI350-049: Based on performance testing by AMD Labs as of 6/6/2025, measuring the text generated inference throughput on the LLaMA 3.1-405B model using the FP4 datatype with various combinations of input/output token length on the AMD Instinct MI355X 8x GPU platform and comparing to the published performance results for NVIDIA B200 HGX 8xGPU platform. Performance per dollar calculated with pricing for NVIDIA B200 available from the Coreweave website as of 6/10/25 and using expected Instinct MI355X GPU-based cloud instance pricing. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of the latest drivers and optimizations and is subject to change. MI350-049

MI350-059: Based on calculations by AMD Performance Labs as of July 2025, using the total memory capacity of a single AMD Instinct MI350X / MI355X 8x GPU Platform (2304 GB HBM3) vs a single Nvidia B200 8x GPU Platform (1440GB). Calculations performed with FP4 precision datatype at (.5) bytes per parameter, to determine the minimum number of 8x GPU platform(s) required to run a Generic Inference model at 4.19T parameters (AMD) and 2.62T parameters (NVIDIA). Stated results include a calculated 10% overhead. Server manufacturers may vary configurations, yielding different results. Results may vary based on GPU memory configuration, LLM size, and any potential variances in GPU memory access or the server operating environment. MI350-059

MI350-060: Based on calculations by AMD Performance Labs as of July 2025, using the total memory capacity of a single AMD Instinct MI350X / MI355X 8x GPU platform (2304GB HBM3) vs a single Nvidia B200 8x GPU Platform (1440GB). Calculations performed with FP4 precision datatype at (.5) bytes per parameter, optimizer states at 16 bytes, memory for weights and gradients at 135B (AMD) and 84B (NVIDIA), memory for activations at 2160GB (AMD) and 1334GB (NVIDIA), and total memory required to run the training model at 2295GB (AMD) and 1428GB (NVIDIA), to determine the minimum number of 8x GPU platform(s) required to run a Generic Training model at 135B parameters (AMD) and 84B parameters (NVIDIA). Calculations include a 10% overhead. Server manufacturers may vary configurations, yielding different results. Results may vary based on GPU memory configuration, LLM size, and any potential variances in GPU memory access or the server operating environment. MI350-060

Disclaimer and Attribution

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

together we advance_AI