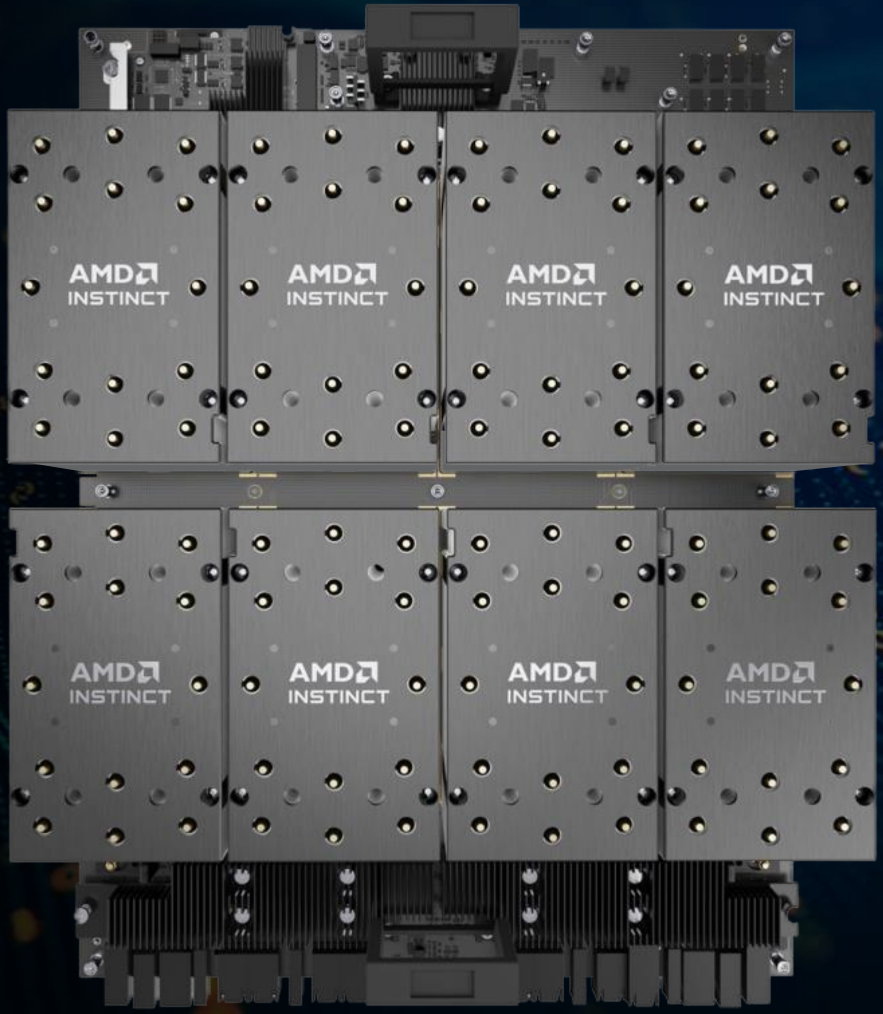
AMD Instinct MI300X Generative AI Accelerator and Platform Architecture

Alan Smith, Sr. Fellow, Instinct Lead SoC Architect

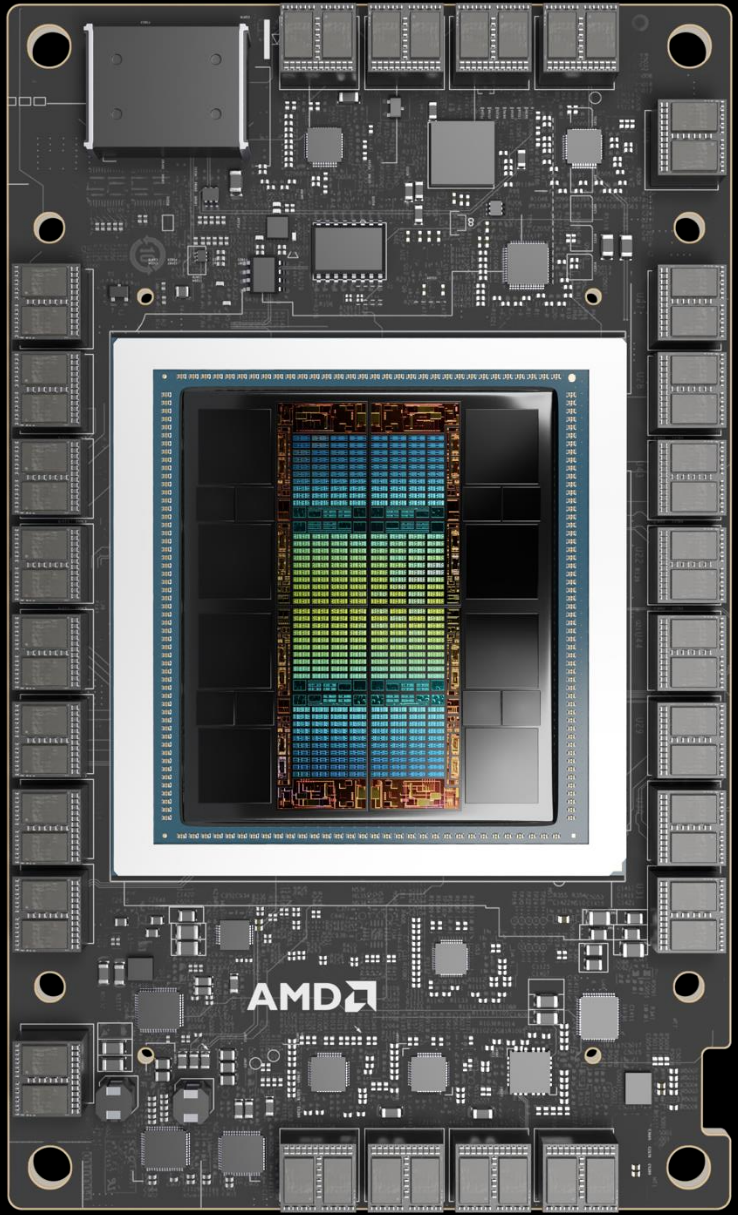
Vamsi Alla, Fellow, Instinct Chief Engineer

Hot Chips 2024



Agenda

- AMD Instinct™ MI300X Accelerator Overview
- AMD CDNA™ 3 Architecture
- Memory System Overview
- Spatial Partitioning
- 4th Gen Infinity Architecture
- System Architecture
- AMD Instinct™ MI300X Platform
- Application Performance



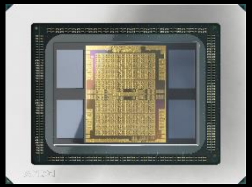
The AMD Instinct™ Accelerator Journey

Multiple generations of architecture focused advancing HPC & AI compute

MI100
AMD CDNA™

ECOSYSTEM GROWTH

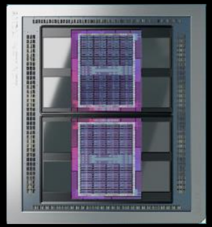
First purpose-built GPU architecture to accelerate FP64 and FP32 HPC workloads



MI200
AMD CDNA™ 2

DRIVING HPC AND AI TO A NEW FRONTIER

Denser compute architecture with leading memory capacity/bandwidth

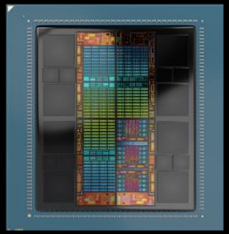


MI300
AMD CDNA™ 3

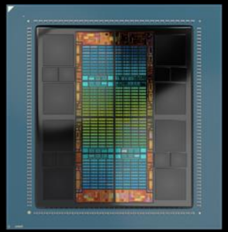
DATA CENTER APU & DISCRETE GPU

Focused improvements on unified memory, AI data format performance, and in-node networking

MI300A



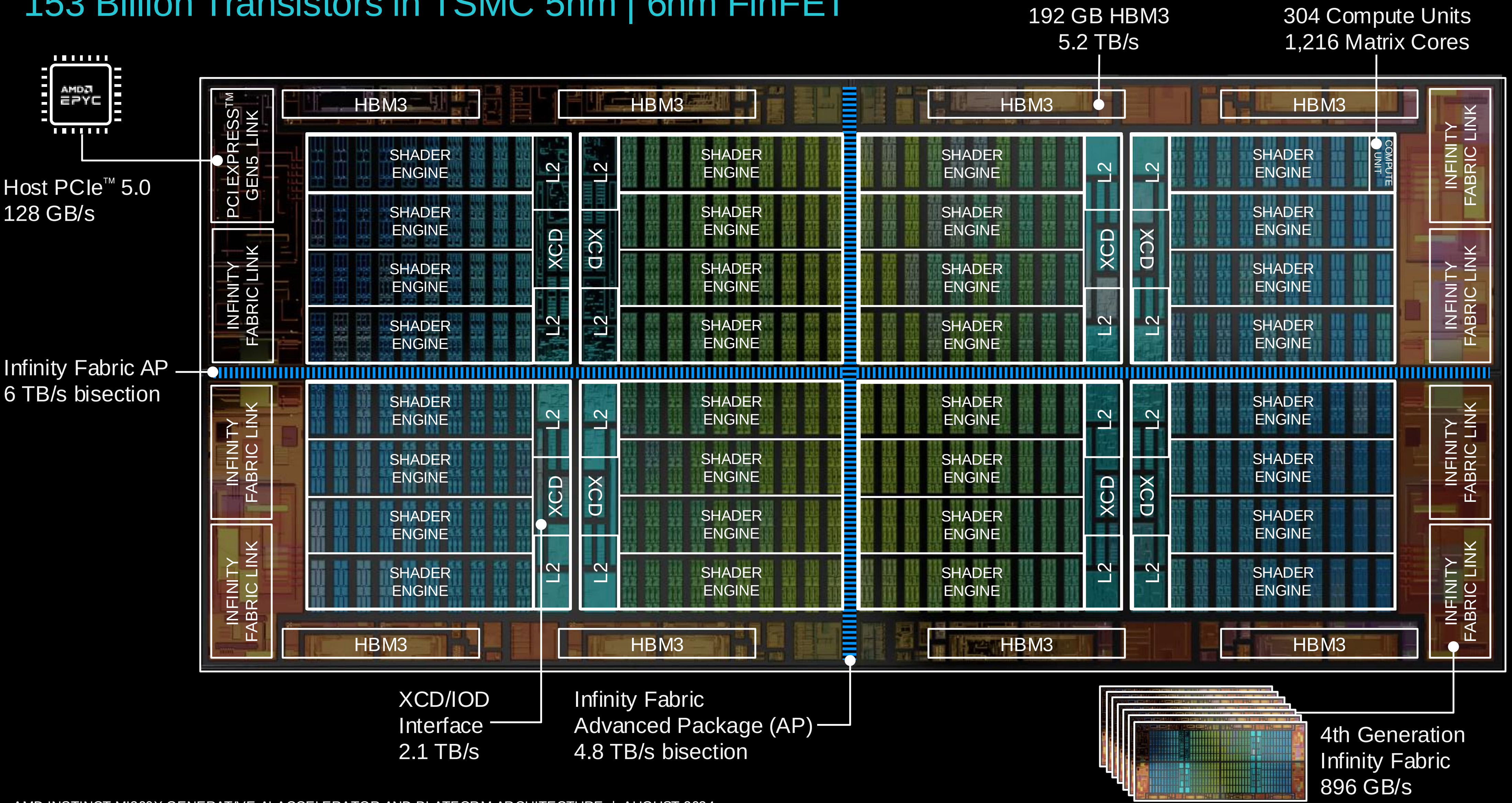
MI300X



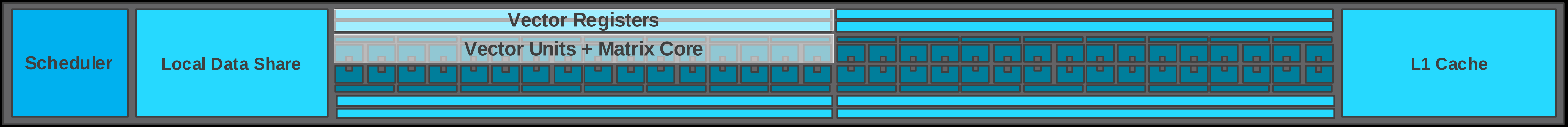
2020 → 2023

AMD Instinct™ MI300X Multi-chiplet Accelerator

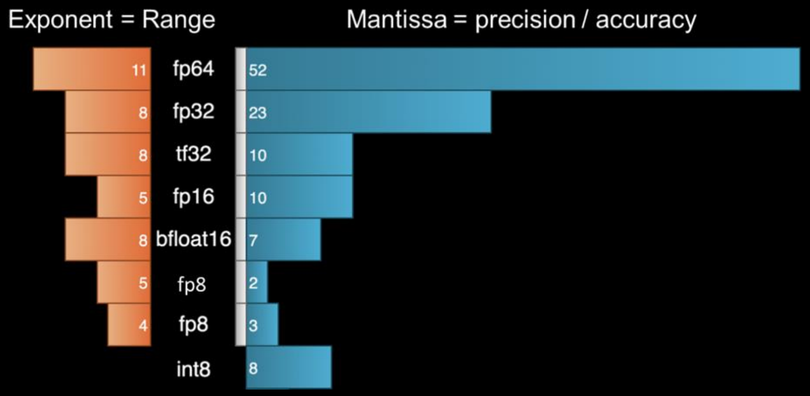
153 Billion Transistors in TSMC 5nm | 6nm FinFET



AMD CDNA™ 3 Architecture



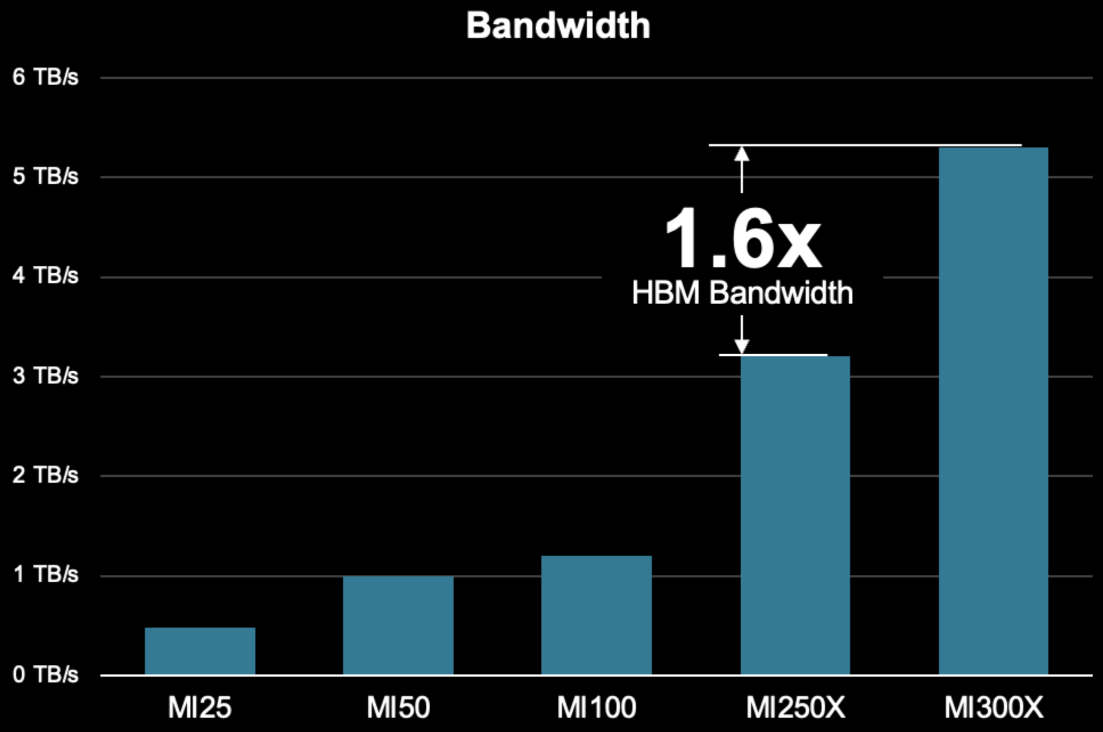
- Doubled low precision matrix ops/clk/cu
- 2:4 structured sparsity support for INT8, FP8, FP16, BF16
- Additional 2x performance with sparsity enabled
- TF32 and FP8 numerical format support
  - 2-bit mantissa and a 5-bit exponent for training (E5M2)
  - 3-bit mantissa with a 4-bit exponent for inference (E4M3)
  - OCP FP8 compliant
- Co-issue FP16 | FP32 | INT32 with FP16 | FP32 | FP64

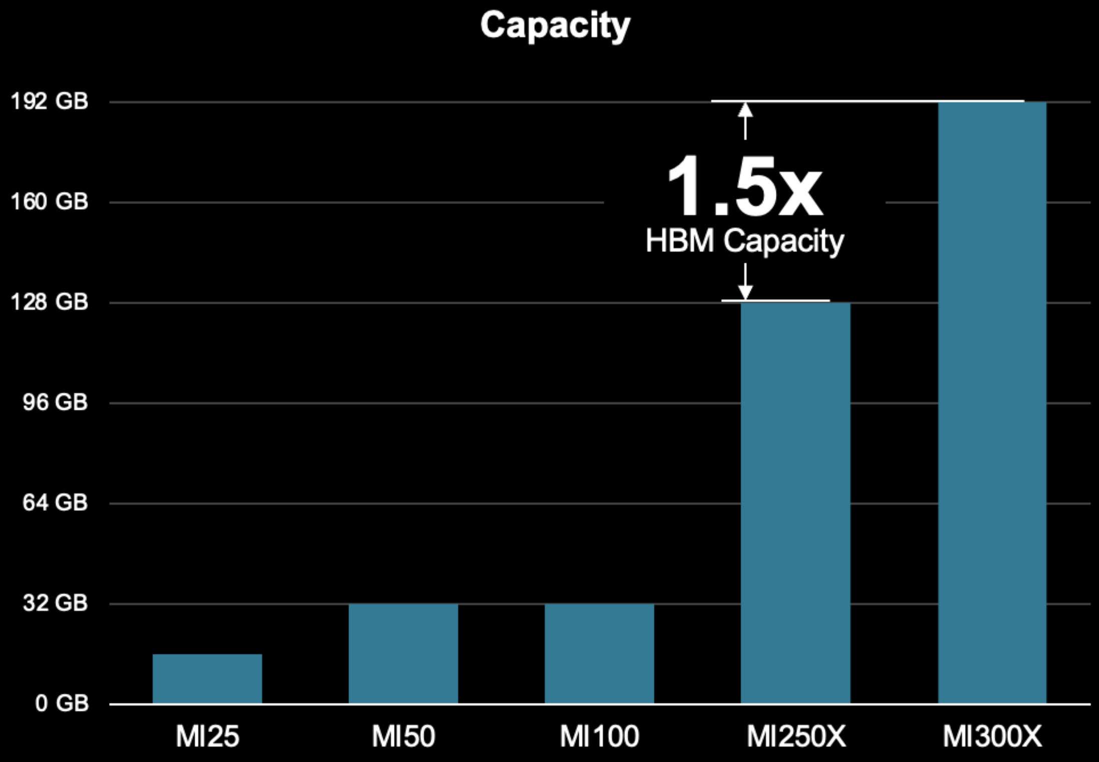


Computation	MI300X (FLOPS/clock/CU)	MI250X (FLOPS/clock/CU)	MI300X (Peak TFLOP/s)	MI250X (Peak TFLOP/s)	MI300X Peak Speedup
Vector FP64	128	128	81.7	47.9	1.7x
Vector FP32[2]	256	128	163.4	47.9	3.4x
Matrix FP64	256	256	163.4	95.7	1.7x
Matrix FP32	256	256	163.4	95.7	1.7x
Matrix TF32	1024	N/A	653.7	N/A[1]	N/A[1]
Matrix FP16[3]	2048	1024	1307.4	383	3.4x
Matrix BF16[3]	2048	1024	1307.4	383	3.4x
Matrix FP8[3]	4096	N/A*	2614.9	N/A[1]	N/A[1]
Matrix INT8[3]	4096	1024	2614.9	383	6.8x

1. AMD Instinct™ MI200 Series accelerators don't support FP8, TF32 or exploit structured sparsity
2. Refers to non packed vector instructions on AMD Instinct™ MI200 and MI300 Series accelerators
3. Refers to dense compute on AMD Instinct™ MI300 accelerators

Worlds First 8-Stack HBM3 Memory Architecture

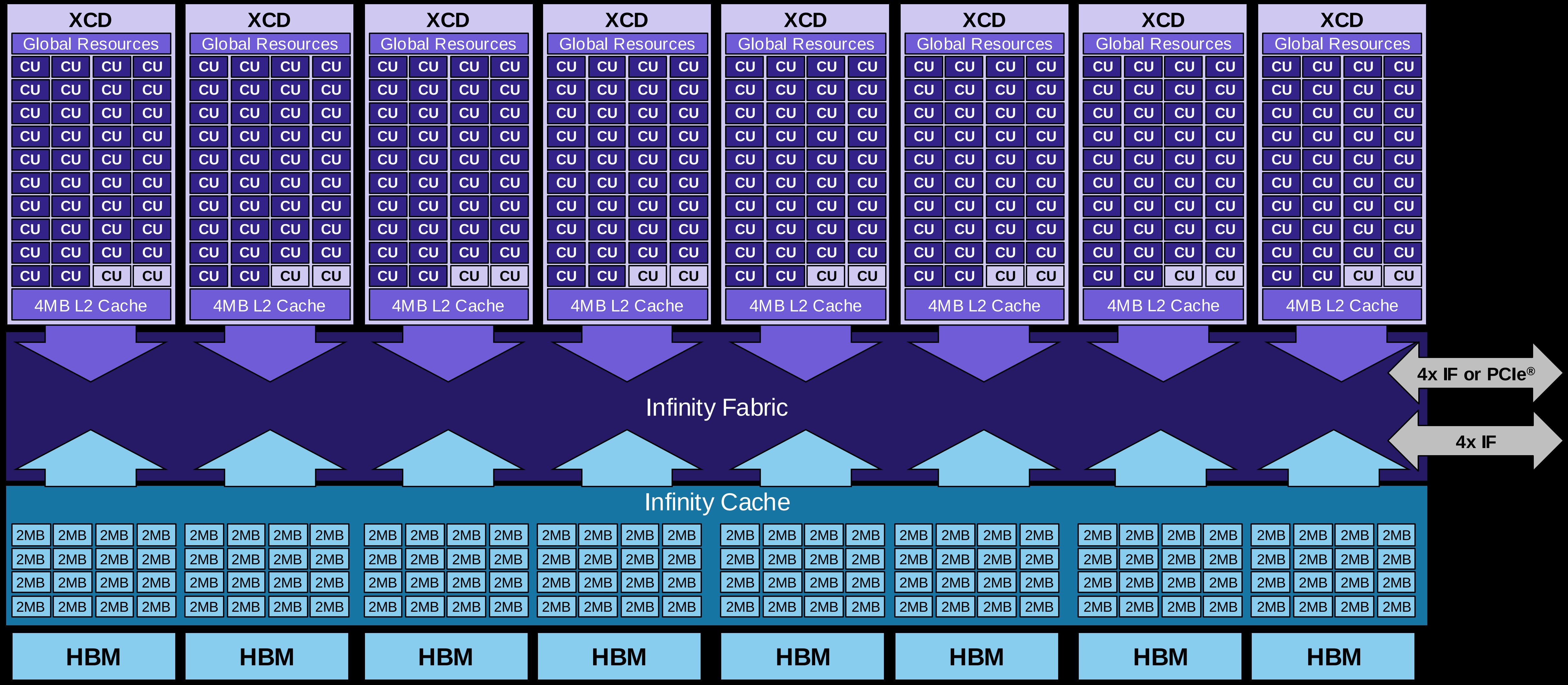




Max LLM size per system (FP16)			
Single Nvidia H100 HGX		Single AMD Instinct MI300X Platform	
640 GB HBM3 | 25.6 TB/s		1.5 TB HBM3 | 42.4 TB/s	
Training	Inference	Training	Inference
~30B	~290B	~70B	~680B

See endnote MI300-42

MI300X Block Diagram



XCD	XCD	XCD	XCD	XCD	XCD	XCD	XCD
Global Resources	Global Resources	Global Resources	Global Resources	Global Resources	Global Resources	Global Resources	Global Resources
CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU
CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU
CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU
CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU
CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU
CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU
CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU
CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU
CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU
CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU	CU CU CU CU
4MB L2 Cache	4MB L2 Cache	4MB L2 Cache	4MB L2 Cache	4MB L2 Cache	4MB L2 Cache	4MB L2 Cache	4MB L2 Cache

Infinity Fabric

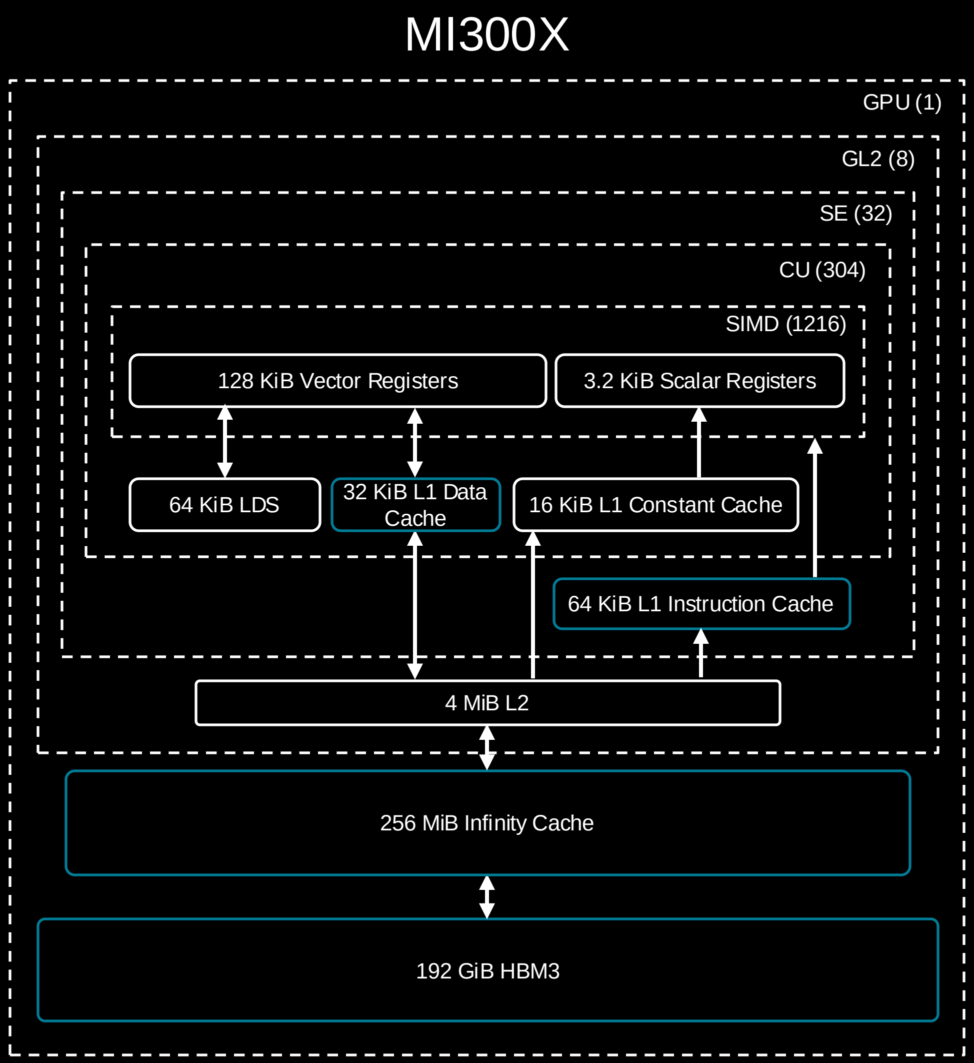
4x IF or PCIe®

4x IF

Infinity Cache

2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB
2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB
2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB
2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB	2MB 2MB 2MB 2MB
HBM	HBM	HBM	HBM	HBM	HBM	HBM	HBM

MI300X Cache and Memory Hierarchy



Level	Capacity	Number per GPU
VGPR	128 KiB	1216
LDS	64 KiB	304
L1 Data Cache	32 KiB	304
L2 Cache	4 MiB	8
Infinity Cache	256 MiB	1
HBM	192 GiB	1

CDNA3 Compute Unit Cache Optimizations

- L1 Data Cache 128B cache line, 32 KiB (vs. 16 KiB on MI200 Series)
- L1 Instruction cache 64 KiB, shared by two CU (vs. 32 KiB on MI200 Series)
- XCD private L2 cache
  - Write back, write allocate cache
  - Increase request coalescing and reduce spill
  - Each instance delivers 2048 Bytes/Clk
  - Agent scope coherent

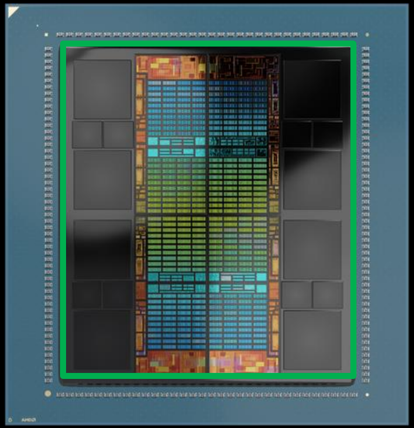
AMD Infinity Cache™ Benefits

- 256 MB at 14.7 TB/s peak BW
- Bandwidth amplification
- Power and latency reduction
- Device/System scope coherent

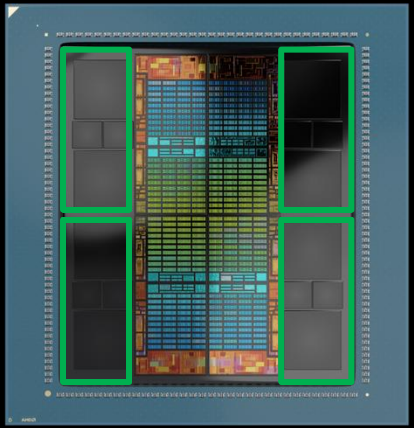
AMD Instinct™ MI300X GPU Spatial Partitioning

Flexible Partitions for Bare Metal and Virtualization

NPS1

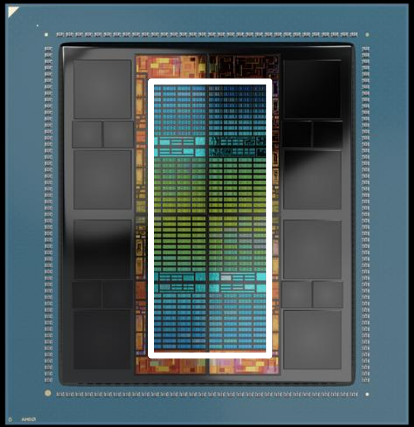


NPS4

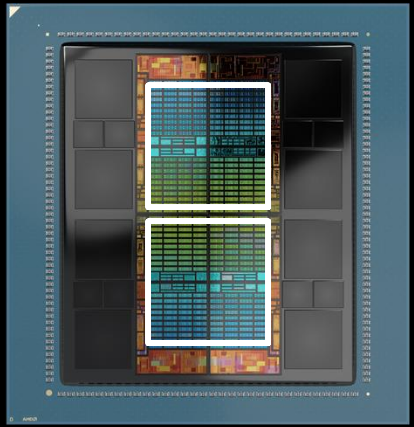


- All XCDs operate together to present the GPU as a single processor
- GPU can be spatially partitioned into as many partitions as XCDs
- XCDs can be grouped to appear as multiple GPUs
- Instinct MI300X supports Single Root IO Virtualization (SR-IOV)
  - Supports up to 64 VF's per platform and dynamic re-partitioning
  - Provides isolation of Virtual Functions (VFs) and protect a VF from accessing information or state of the Physical Function (PF) or another VF
- Each XCD could operate on a separate stream of input queries for inference
- NUMA partitions per socket or NPS, exploit data locality for partitions

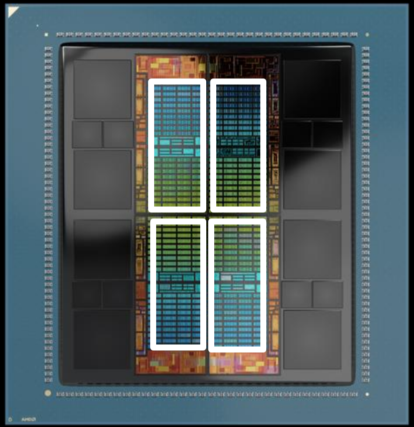
Single partition



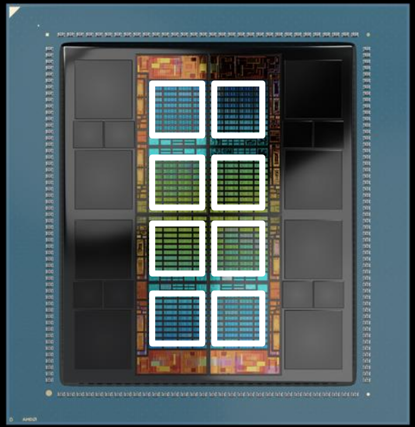
Two partitions



Four partitions



Eight partitions



AMD Instinct™ Platform

Industry leading generative AI platform

8x AMD Instinct MI300X Accelerator

Leadership memory capacity

4th Gen AMD Infinity Fabric™ Technology

Industry-standard design

The AMD Instinct™ System Journey

Enhancing system architecture to complement our silicon development

MI100

AMD CDNA™

Scale Up System to support large Model ML

MI200

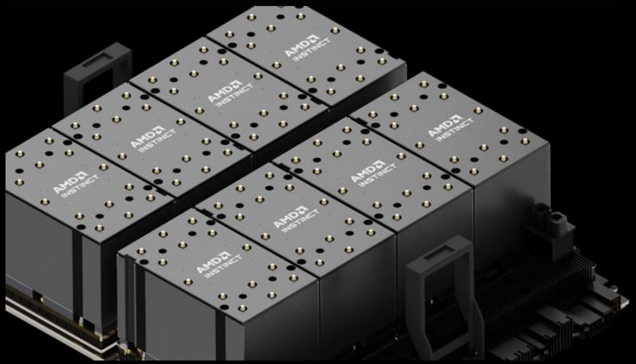
AMD CDNA™ 2

OAM to power a new frontier beyond PCIe™ Form Factor

MI300

AMD CDNA™ 3

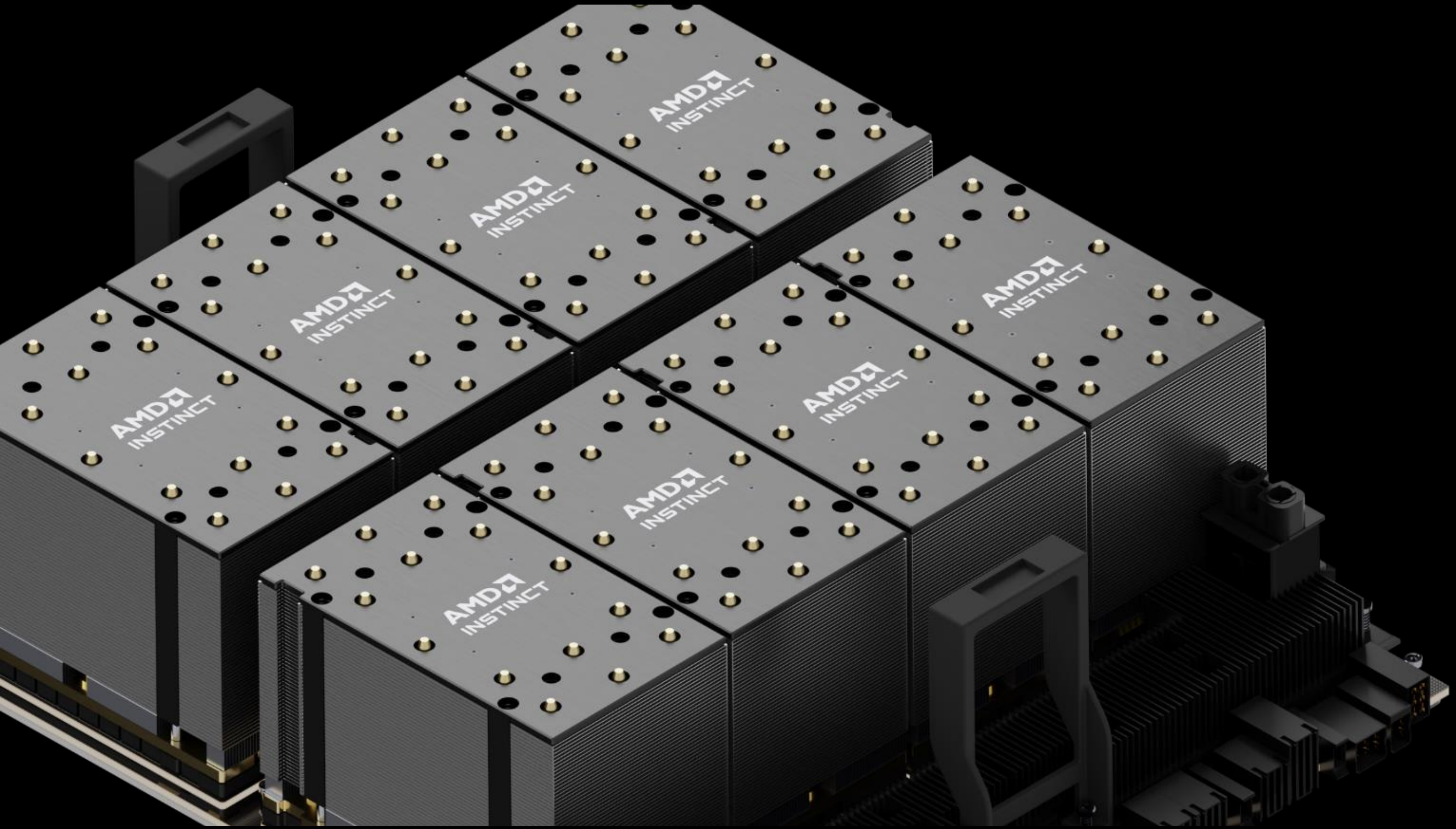
AI Subsystem to serve LLM's



2020 → 2023

AMD Instinct™ MI300X Platform

Industry-leading generative AI platform



8
AMD Instinct™ MI300X

~10.4 PF
BF16/FP16

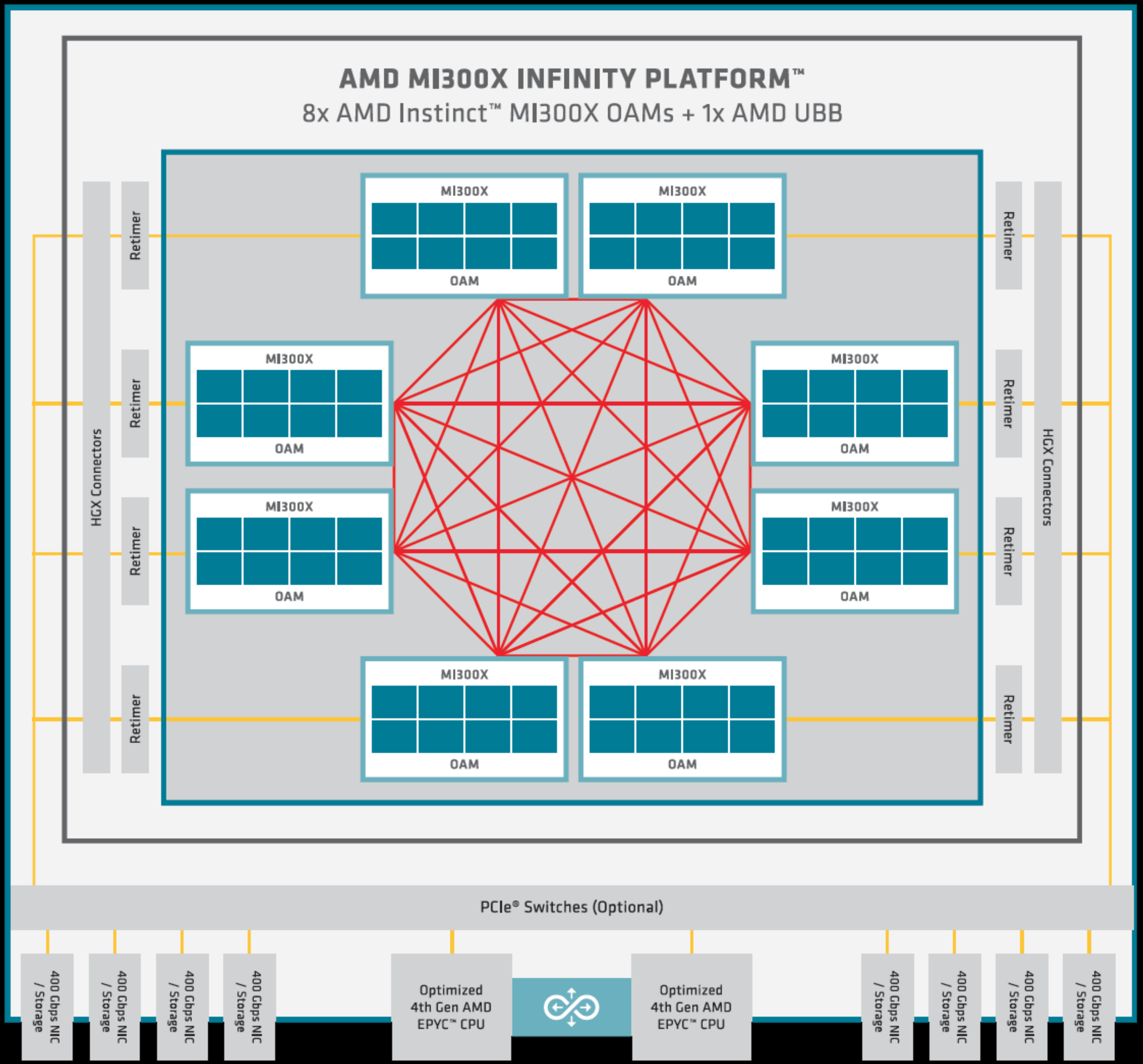
1.5 TB
HBM3

~896 GB/s
Infinity Fabric™ Bandwidth

Industry-Standard Design

MI300X Infinity Platform™

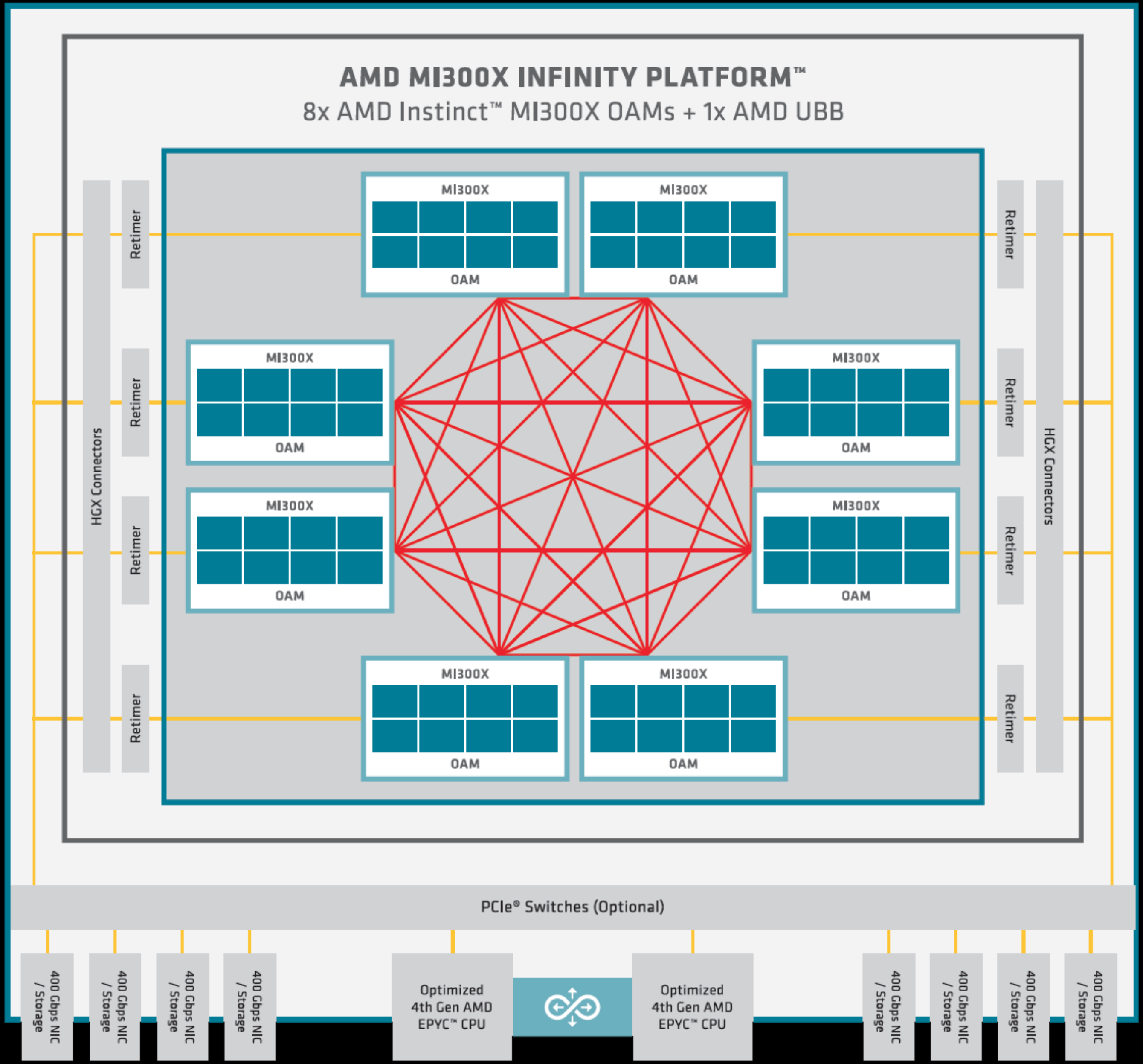
- Direct connectivity for 8 OAMs via AMD Infinity Fabric™
- Seven bi-directional links @ 128 GB/s
- PCIe® Gen 5 x16 per OAM for server connectivity and I/O
- 192 GB HBM for RDMA


AMD MI300X INFINITY PLATFORM™

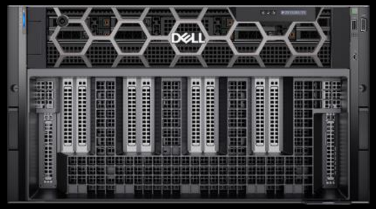
8x AMD Instinct™ MI300X OAMs + 1x AMD UBB

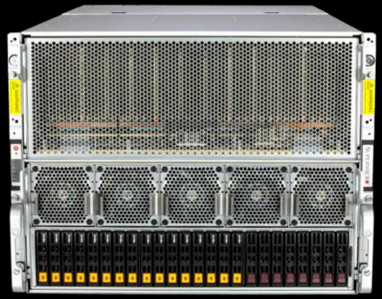
MI300X Infinity Platform™

- UBB 2.0™: Open standards, faster deployment, seamless datacenter integration
- Security: SPDM™ attestation, measurement
- RAS: Full-chip ECC memory, page retirement, page avoidance
- Telemetry: RedFish™, logs, notifications
- Firmware management: RedFish™ PLDM™ bundles, redundancy


AMD MI300X INFINITY PLATFORM™
8x AMD Instinct™ MI300X OAMs + 1x AMD UBB
MI300X
OAM
Retimer
HGX Connectors
PCIe® Switches (Optional)
400 Gbps NIC / Storage
Optimized 4th Gen AMD EPYC™ CPU
Optimized 4th Gen AMD EPYC™ CPU

MI300X Industry Standard OCP Server Designs

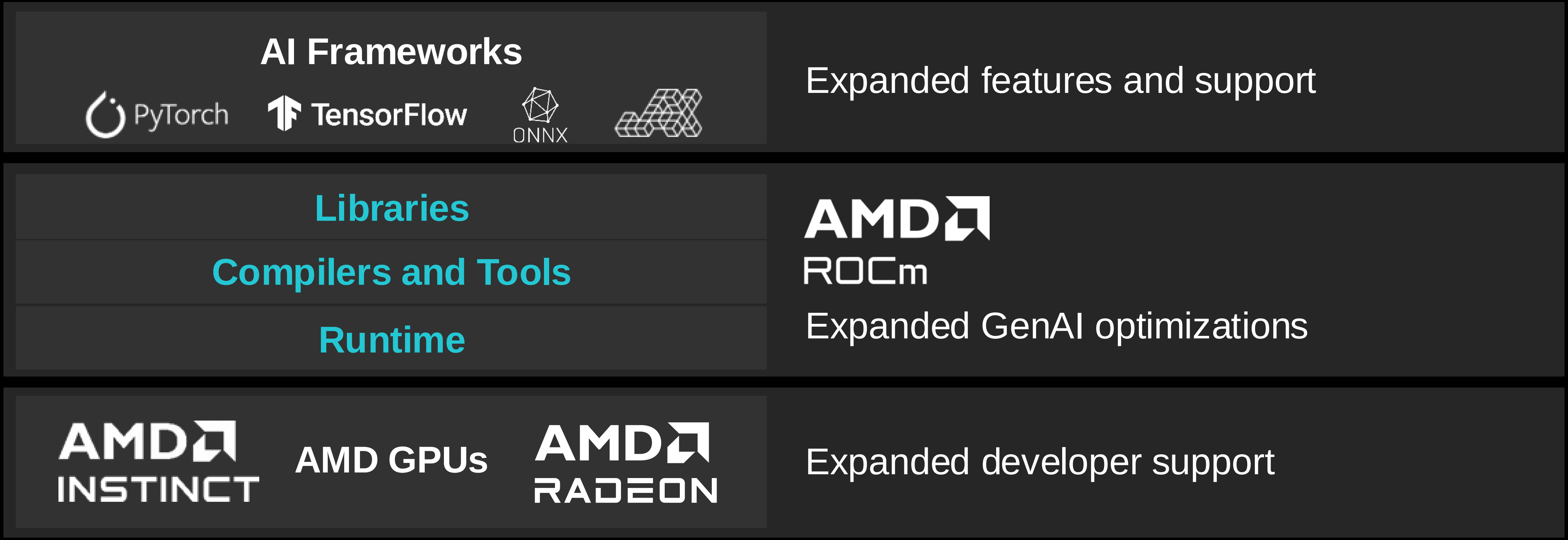




Dell **PowerEdge XE9680**	**GIGABYTE** **G593-ZX1-AAZ1**	**HPE** **CRAY SC XD675**	**Lenovo** **ThinkSystem SR685a V3 Rack Server**	**Super Micro** **AS-8125GS-TNMR2 Server**
UBB 2.0 6U	UBB 2.0 5U	UBB 2.0 8U	UBB 2.0 8U	UBB 2.0 8U
Dual CPUs with up to 56 cores per processor	Dual AMD EPYC™ 9004 Series Processors (with AMD 3D V-Cache™ Technology)	2x AMD EPYC up to 400W	2x 4th Gen AMD EPYC™ Processors	Dual AMD EPYC™ 9004 Series Processors
8x AMD Instinct MI300X Accelerators	8x AMD Instinct MI300X Accelerators	8x AMD Instinct MI300X Accelerators	8x AMD Instinct MI300X Accelerators	8x AMD Instinct MI300X Accelerators

AMD ROCm

Open software ecosystem



AI Frameworks

PyTorch TensorFlow ONNX

Expanded features and support

Libraries

Compilers and Tools

Runtime

AMD ROCm

Expanded GenAI optimizations

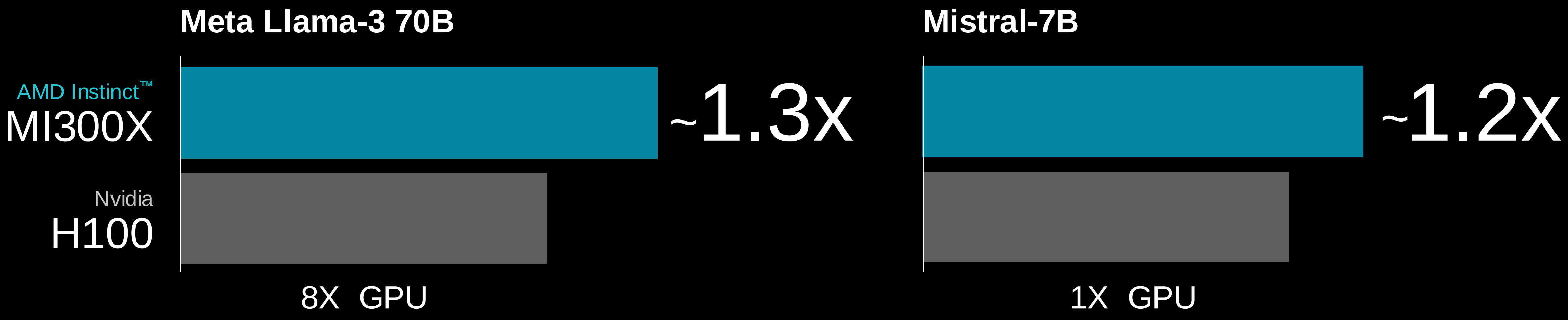
AMD INSTINCT AMD GPUs AMD RADEON

Expanded developer support

AMD
ROCm

ROCm™ 6 Software

Leadership performance for generative AI



Token Generation Throughput

Mi300-53, MI300-54

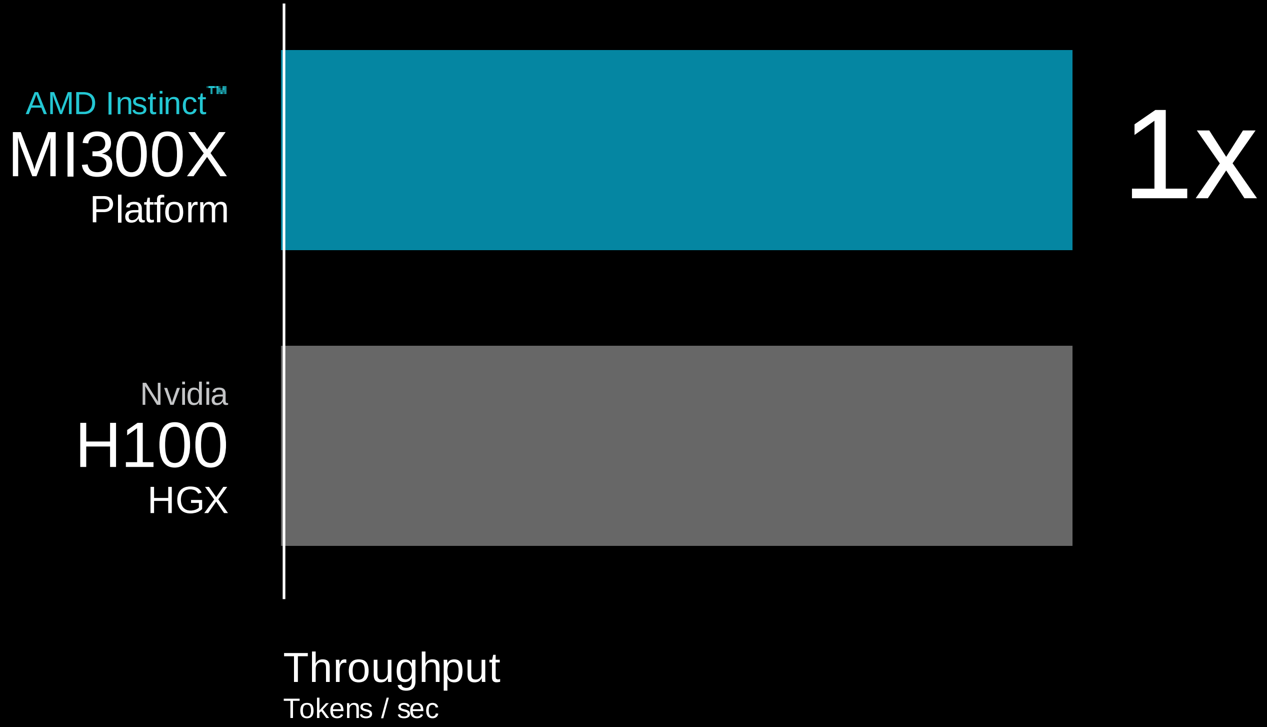
World Class Training Performance

Single Server 8x MI300X

MPT

Model size: 30B

Model Fine Tuning



Results may vary. See endnotes: MI300-40

Summary

Datacenter APU and Accelerator Architecture

- Optimal Efficiency through unified memory, AI focused data formats

8 stack HBM3 Memory System

- 192GB per MI300X and 1.5 TB per platform, enabling 680B parameter LLM inference

Instinct Platform™ Architecture

- Modular AI Subsystem enabling fast industry adoption

Industry Leading AI performance

AMD

Endnotes

MI300-53: Testing completed on 05/28/2024 by AMD performance lab attempting text generated throughput measured using Mistral-7B model comparison. Tests were performed using batch size 1 and 2048 input tokens and 2048 output tokens for Mistral-7B Configurations: 2P AMD EPYC 9534 64-Core Processor based production server with 8x AMD InstinctTM MI300X (192GB, 750W) GPU, Ubuntu® 22.04.1, and ROCm™ 6.1.1 Vs. 2P Intel Xeon Platinum 8468 48-Core Processor based production server with 8x NVIDIA Hopper H100 (80GB, 700W) GPU, Ubuntu 22.04.3, and CUDA® 12.2 Only 1 GPU on each system was used in this test. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest d rivers and optimizations. MI300-53

MI300-54: Testing completed on 05/28/2024 by AMD performance lab attempting text generated Llama3-70B using batch size 1 and 2048 input tokens and 128 output tokens for each system. Configurations: 2P AMD EPYC 9534 64-Core Processor based production server with 8x AMD InstinctTM MI300X (192GB, 750W) GPU, Ubuntu® 22.04.1, and ROCm™ 6.1.1 Vs. 2P Intel Xeon Platinum 8468 48-Core Processor based production server with 8x NVIDIA Hopper H100 (80GB, 700W) GPU, Ubuntu 22.04.3, and CUDA® 12.2 8 GPUs on each system was used in this test. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. MI300-54

MI300-40: Testing completed 11/28/2023 by AMD performance lab using MosaicML vllm-foundry to fine tune the MPT-30b model for 2 epochs using the MosaicML instruct-v3 dataset and a max sequence length of 8192 tokens using custom docker container for each system .Configurations: 2P Intel Xeon Platinum 8480C CPU server with 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, ROCm® 6.0 pre-release, PyTorch 2.0.1, MosaicML llm-foundry pre-release, Ubuntu 22.04.2.Vs.An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, CUDA 11.8, PyTorch 2.0.1., MosaicML llm-foundry, Ubuntu 22.04.3.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI300-34: Token generation throughput using DeepSpeed Inference with the Bloom-176b model with an input sequence length of 1948 tokens, and output sequence length of 100 tokens, and a batch size tuned to yield the highest throughput on each system comparison based on AMD internal testing using custom docker container for each system as of 11/17/2023.Configurations: 2P Intel Xeon Platinum 8480C CPU powered server with 8x AMD Instinct™ MI300X 192GB 750W GPUs, pre-release build of ROCm™ 6.0, Ubuntu 22.04.2.Vs.An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 80GB 700W GPUs, CUDA 12.0, Ubuntu 22.04.3.8 GPUs on each system were used in this test. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI300-39: Number of simultaneous text generating copies of the Llama2-70b chat model, using vLLM, comparison using custom docker container for each system based in AMD internal testing as of 11/26/2023.Configurations: 2P Intel Xeon Platinum 8480C CPU server with 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, ROCm® 6.0 pre-release, PyTorch 2.2.0, vLLM for ROCm, Ubuntu 22.04.2.Vs.An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, CUDA 12.1., PyTorch 2.1.0. vLLM v.02.2.2 (most recent), Ubuntu 22.04.3.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

M300-42: Measurements by internal AMD Performance Labs as of December 1, 2023 on current specifications and/or internal engineering calculations. Inference and training Large Language Model (LLM) run comparisons with FP16 precision to determine the largest Large Language model size that is expected to run on the 8x AMD Ins tinct™ MI300X (192GB) accelerator platform and on the Nvidia 8x H100 (80GB) GPUs DGX platform. Calculated estimates based on GPU-only memory size versus memory required by the model at defined parameters plus 10% overhead. Calculations rely on published and sometimes preliminary model memory sizes. Multiple LLMs and parameter sizes were analyzed. Max size determined by memory capacity of 8x platform. Configurations: 8x AMD Instinct™ MI300X (192GB HBM3, OAM Module) 750W accelerator at 2,100 MHz peak boost engine clock designed with 3rd Gen AMD CDNA™ 3 5nm FinFET process technology. Vs.8x Nvidia HGX H100 (80GB HBM3, SXM5) platform Nvidia memory specification at https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet. Results for Inferencing: Largest parameter size for 8X H100: MI300X GPUs H100 GPUs Gopher Deepmind (290B) 4 Calculated 8 Calculated Largest parameter size for 8x MI300X: MI300X GPUs H100 GPUsPaLM-1 (680B) 8 Calculated 19 Calculated Results for Training: Largest parameter size for 8X H100:
MI300X GPUs H100 GPUs Mosiac MPT-30B parameter 4 Calculated 8 CalculatedLargest parameter size for 8x MI300X: MI300X GPUs H100 GPUsMosiac MPT-70B parameter 7 Calculated 16 Calculated Assumptions: FP16 Datatype Batchsize 1Memory needs for model = 2Bytes per Parameter Memory size needs for activations and others = +10% Actual maximum LLM parameter size that can run on each platform may vary upon performance testing with physical servers.

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.