第23回PCクラスタシンポジウム

AMD Instinct™ と ROCm™ による AIイノベーションの加速化

2023年12月7日
日本AMD株式会社

大原久樹 (Hisaki.Ohara@amd.com)

AMD
together we advance_

Cautionary Statement

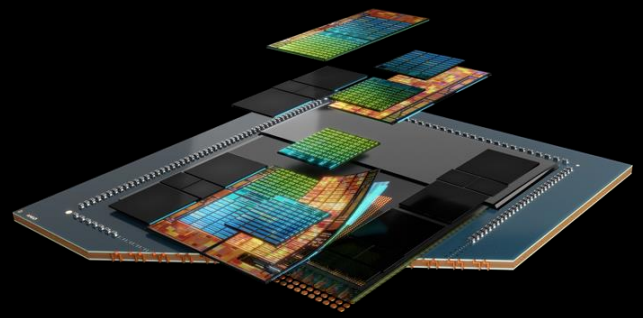
This presentation contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) such as the features, functionality, performance, availability, timing and expected benefits of AMD products as well as AMD product roadmaps, which are made pursuant to the Safe Harbor provisions of the 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.

本日の内容

1. AMD Instinct™ の振り返り
2. MI300X プラットフォーム
3. ROCmを用いたLLMの性能評価
4. MI300A プラットフォーム

AMD AI



Broad portfolio of training and inference compute engines

AMD ROCm

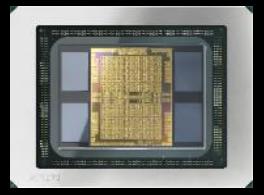
Open and proven software capabilities

PyTorch Foundation

Ultra Ethernet Consortium

Deep ecosystem of AI partners and co-innovation

OUR JOURNEY IN GPU ACCELERATION

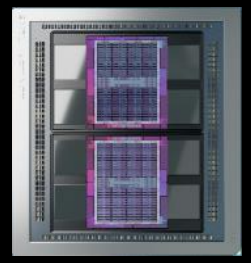


AMD Instinct™ MI100

AMD CDNA™

Ecosystem Growth

First purpose-built GPU architecture for the data center

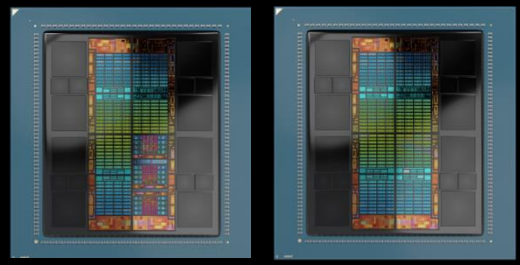


AMD Instinct™ MI200

AMD CDNA™ 2

Driving HPC and AI to a New Frontier

First purpose-built GPU powering discovery at Exascale



AMD Instinct™ MI300

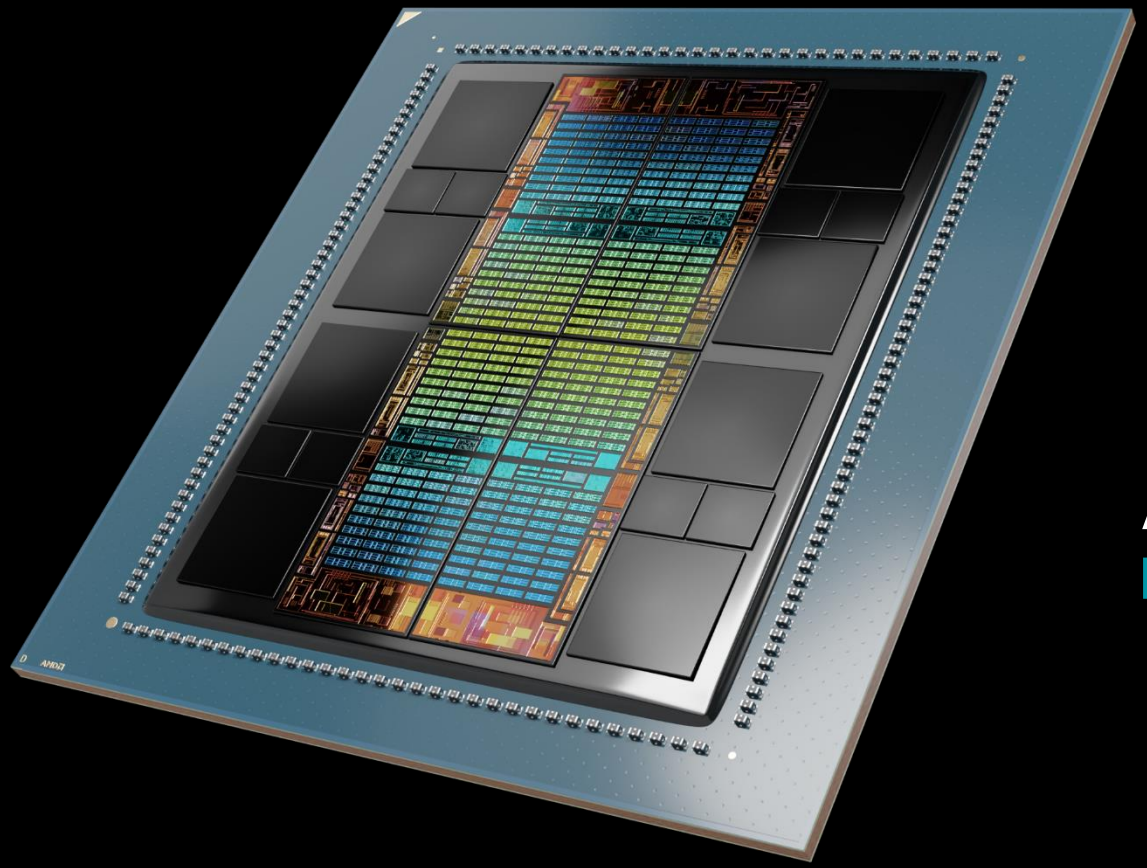
AMD CDNA™ 3

Data Center APU & Discrete GPU

Breakthrough architecture designed for leadership efficiency and performance for AI and HPC

2020 → 2024

Roadmaps Subject to Change

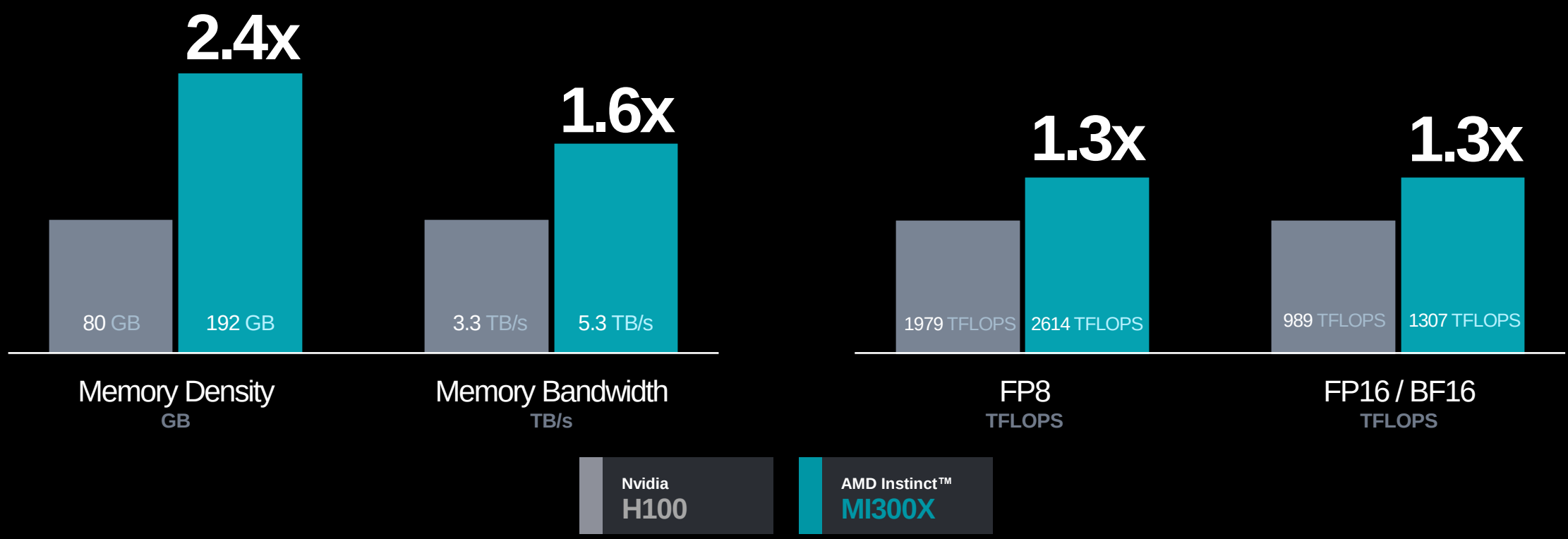


AMD Instinct™ MI300X

Leadership generative AI accelerator

AMD CDNA 3	192 GB HBM3	5.2 TB/s Memory Bandwidth	896 GB/s Infinity Fabric™ Bandwidth	153 B Transistors

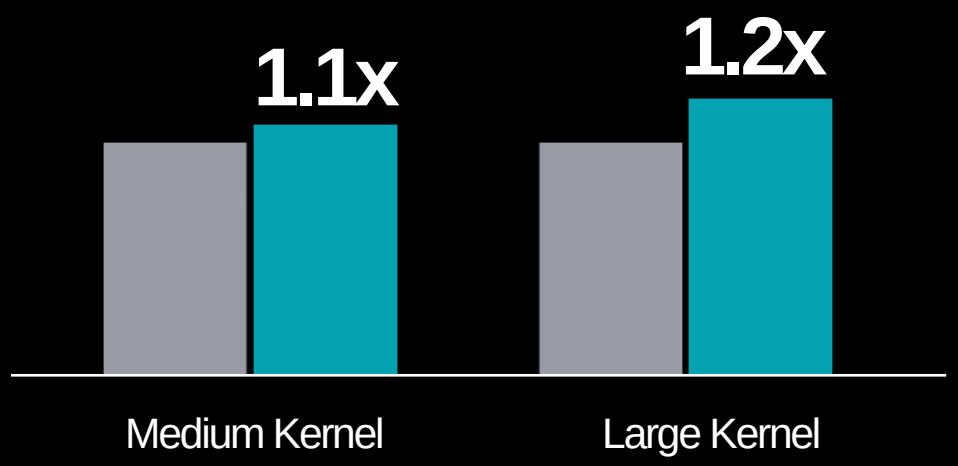
Generative AI Leadership



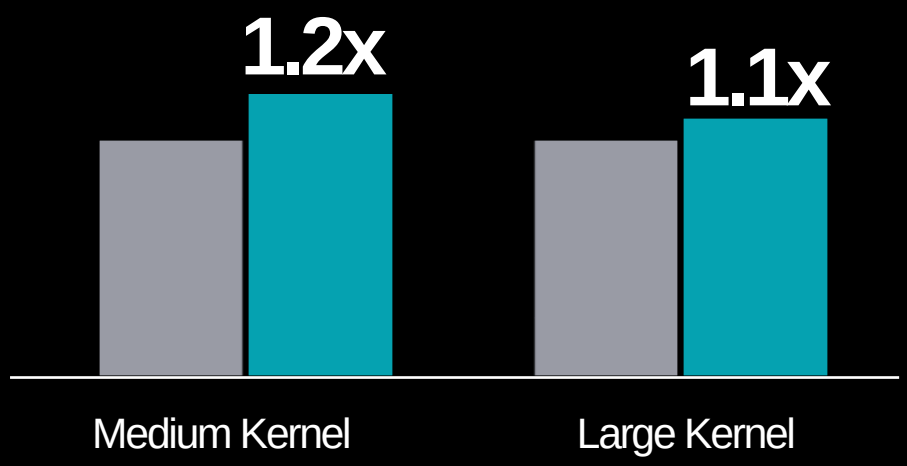
Theoretical peak

Key AI Kernel performance leadership

Common LLM kernels



Flash Attention 2



Llama2-13B kernel

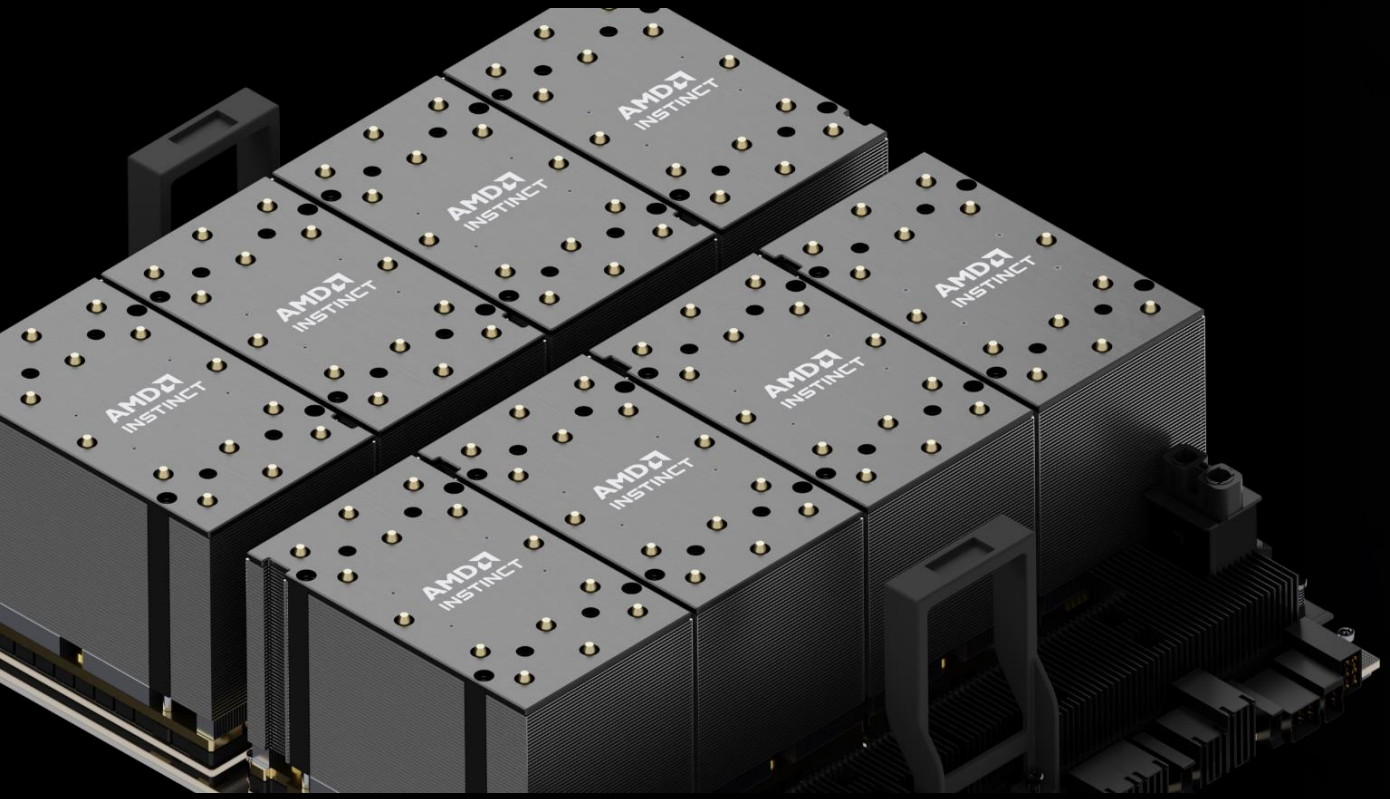
Nvidia H100

AMD Instinct™ MI300X

Add footnote

AMD Instinct™ MI300X Platform

Industry-leading generative AI platform



8
AMD Instinct™ MI300X

21 PF
BF16/FP16 w/ Sparsity

1.5 TB
HBM3

896 GB/s
Infinity Fabric™ Bandwidth

Industry-Standard OCP Design

AMD Instinct™ Platform

Infrastructure performance

AMD Instinct™ MI300X	Nvidia H100 HGX	AMD Instinct™ MI300X Advantage
1.5 TB HBM3 memory	**640** GB HBM3 memory	**2.3X** More memory
21 PF FP16 / BF16 FLOPS w/Sparsity	**15.8** PF FP16 / BF16 FLOPS	**1.3X** More Compute
896 GB/s Aggregate bi-directional bandwidth	**900** GB/s Aggregate bi-directional bandwidth	Comparable
448 GB/s Single node ring bandwidth	**450** GB/s Single node ring bandwidth	Comparable
Up to **400** GbE NIC / GPU	Up to **400** GbE NIC / GPU	Equivalent
PCIe® Gen 5 128 GB/s	**PCIe® Gen 5** 128 GB/s	Equivalent

AMD ROCm | Optimized AI software stack

AI Models and **Algorithms**

PyTorch TensorFlow ONNX

AI Ecosystem optimized for AMD

Libraries

Compilers and **Tools**

Runtime

AMD ROCm | A proven software stack

AMD Instinct™ GPU

AMD INSTINCT | Leadership performance

Strong developer ecosystem momentum

Hugging Face

62,000+ models running nightly
Fully integrated optimum library

PyTorch

From ‘port-to’ to ‘develop-on’
with latest platforms

Tensor Flow

Dynamo Inductor

JAX

OpenAI Triton

ONNX Runtime

OpenXLA

DeepSpeed

MLIR | IREE

Increasing open-source contributions
and expanding footprint

TRANSITIONING WORKLOADS: WHEN USING FRAMEWORKS

ZERO CODE CHANGES NEEDED

CUDA → ROCM

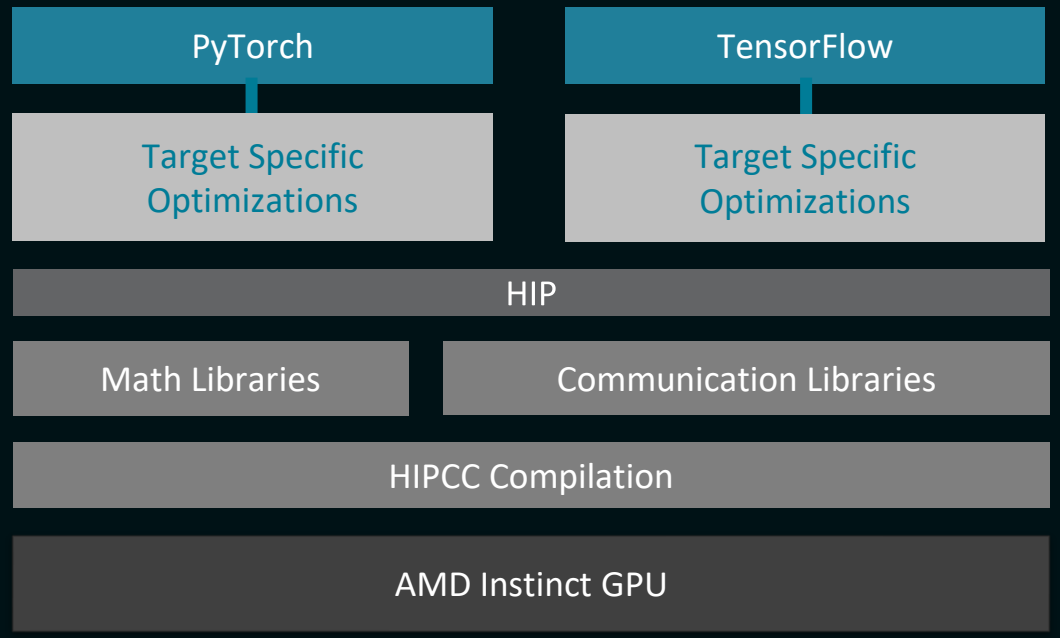
```
import torch
# Get cpu or gpu device for training.
device = "cuda" if torch.cuda.is_available() else "cpu"
print(f"Using {device} device")
# Define model
class Network(nn.Module):
    def __init__(self):
        super().__init__()
        self.flatten = nn.Flatten()
        self.linear_relu_stack = nn.Sequential(
            nn.Linear(28*28, 512),
            nn.ReLU(),
            nn.Linear(512, 512),
            nn.ReLU(),
            nn.Linear(512, 10)
        )
    def forward(self, x):
        x = self.flatten(x)
        logits = self.linear_relu_stack(x)
        return logits
model = Network().to(device)
print(model)
```

```
import torch
# Get cpu or gpu device for training.
device = "cuda" if torch.cuda.is_available() else "cpu"
print(f"Using {device} device")
# Define model
class Network(nn.Module):
    def __init__(self):
        super().__init__()
        self.flatten = nn.Flatten()
        self.linear_relu_stack = nn.Sequential(
            nn.Linear(28*28, 512),
            nn.ReLU(),
            nn.Linear(512, 512),
            nn.ReLU(),
            nn.Linear(512, 10)
        )
    def forward(self, x):
        x = self.flatten(x)
        logits = self.linear_relu_stack(x)
        return logits
model = Network().to(device)
print(model)
```

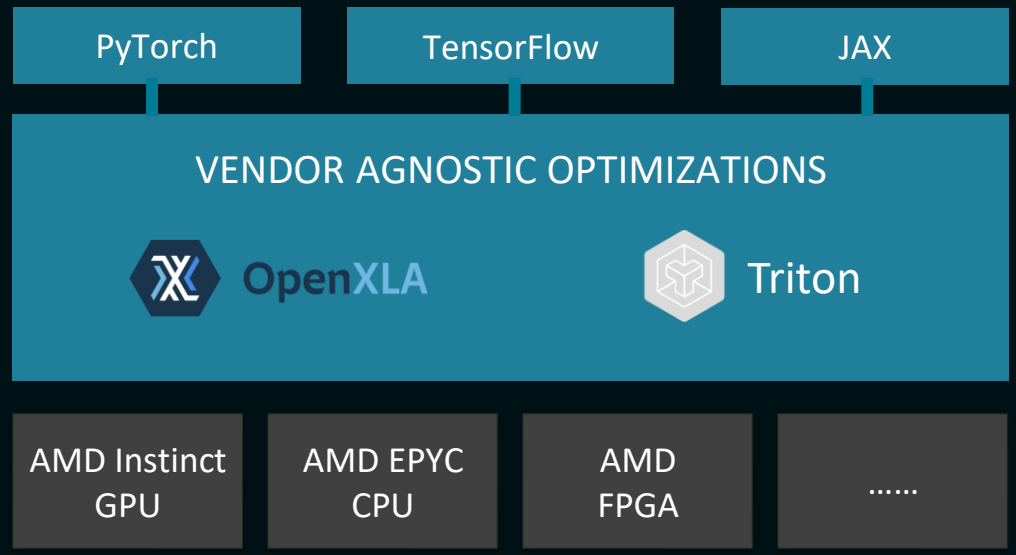
AI FRAMEWORKS: LIBRARY AND COMPILER-BASED OPTIMIZATIONS

AMD IS OPTIMIZING FOR BEST PERFORMANCE FOR BOTH APPROACHES

1 AMD OPTIMIZES SPECIFIC LIBRARIES FOR BEST PERFORMANCE ON FRAMEWORK OPERATORS



2 OPTIMIZED IR-BASED SOLUTIONS DELIVER PORTABLE PERFORMANCE ON AMD DEVICES



AMD

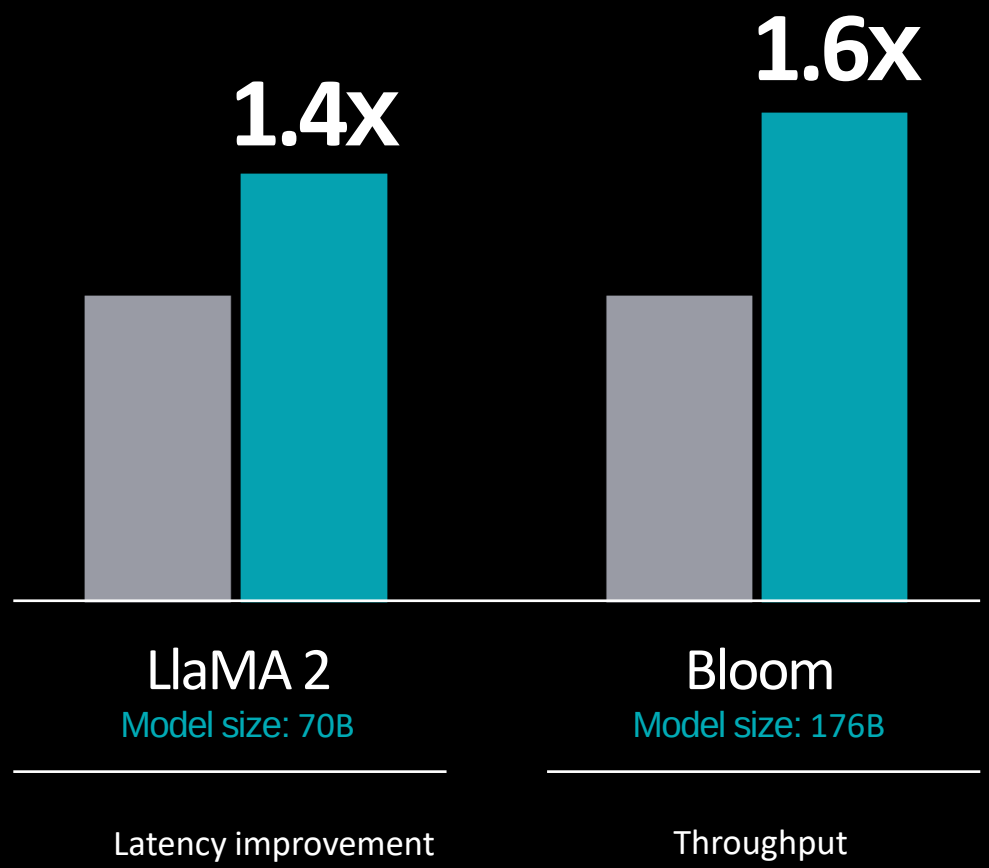
AI Inference

Performance Leadership

Inference Performance Leadership

Single server (8x GPU)

Nvidia	AMD Instinct™
H100	MI300X
HGX	Platform



AMD

AI Training
Performance Leadership

World class training performance

Single server 8x MI300

MPT

Model size: 30B

Add footnote

AMD Instinct™ MI300X
Ecosystem Partners

Cloud Service Providers

Microsoft Azure

ORACLE CLOUD

Platform Partners

DELL Technologies

Hewlett Packard Enterprise

Lenovo

SUPERMICRO

GIGABYTE™

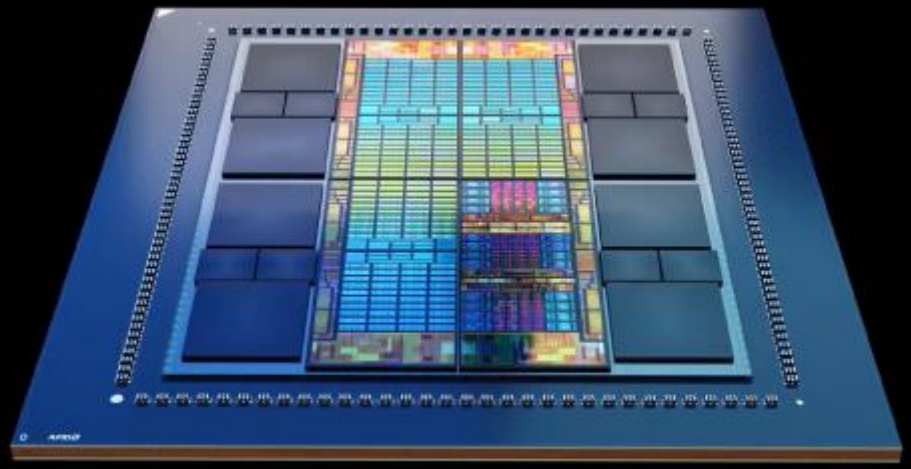
Inventec

QCT

ingrasys

wistron

AMD Instinct™ MI300A

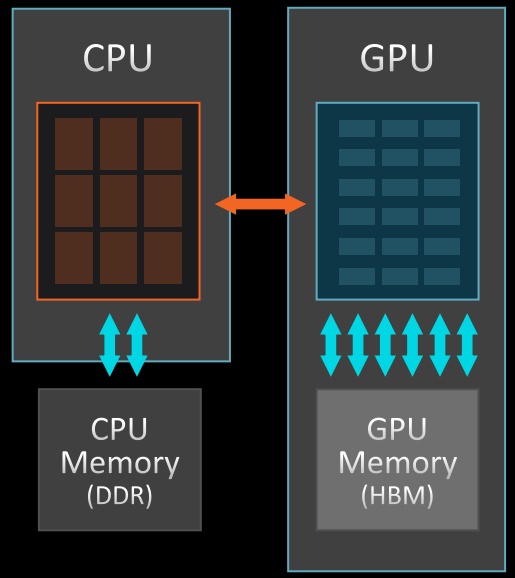


World's first APU for HPC and AI

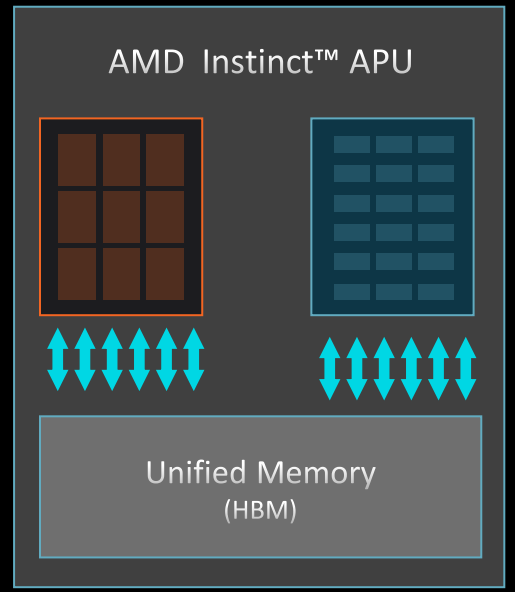
61 TF	122 TF	128 GB	5.3 TB/s	146B
FP64	FP32	HBM3	Memory Bandwidth	Transistors

MI300A UNIFIED MEMORY APU ARCHITECTURE BENEFITS

AMD CDNA™ 2 Coherent Memory Architecture → AMD CDNA™ 3 Unified Memory APU Architecture



- **Eliminate Redundant Memory Copies**
- **Does not need a programming distinction between host and device memory spaces**
- **High performance, fine-grained sharing between CPU and GPU processing elements**



AMD'S OPEN-SOURCE APPROACH TO SOFTWARE

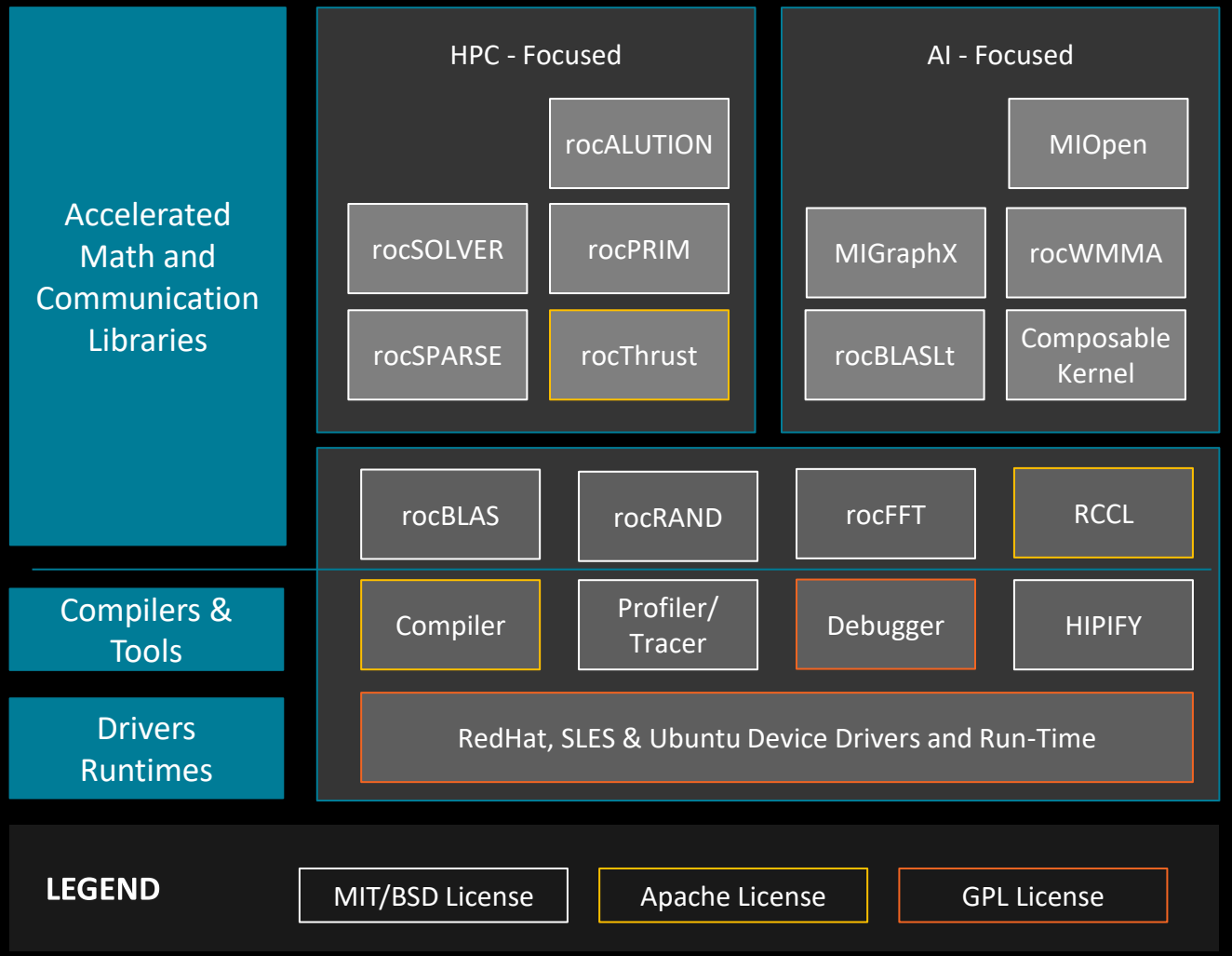
COMMITMENT TO AN OPEN-SOURCE ECOSYSTEM

LICENSE FREE AND ACCESSIBLE

- All source code is published on Github: includes drivers, tools and libraries. Build environment scripts (cmake) available to compile source for target devices.
- Users may modify and tune any component for their specific purposes. Allows end users to wrap all changes into a commercial service.

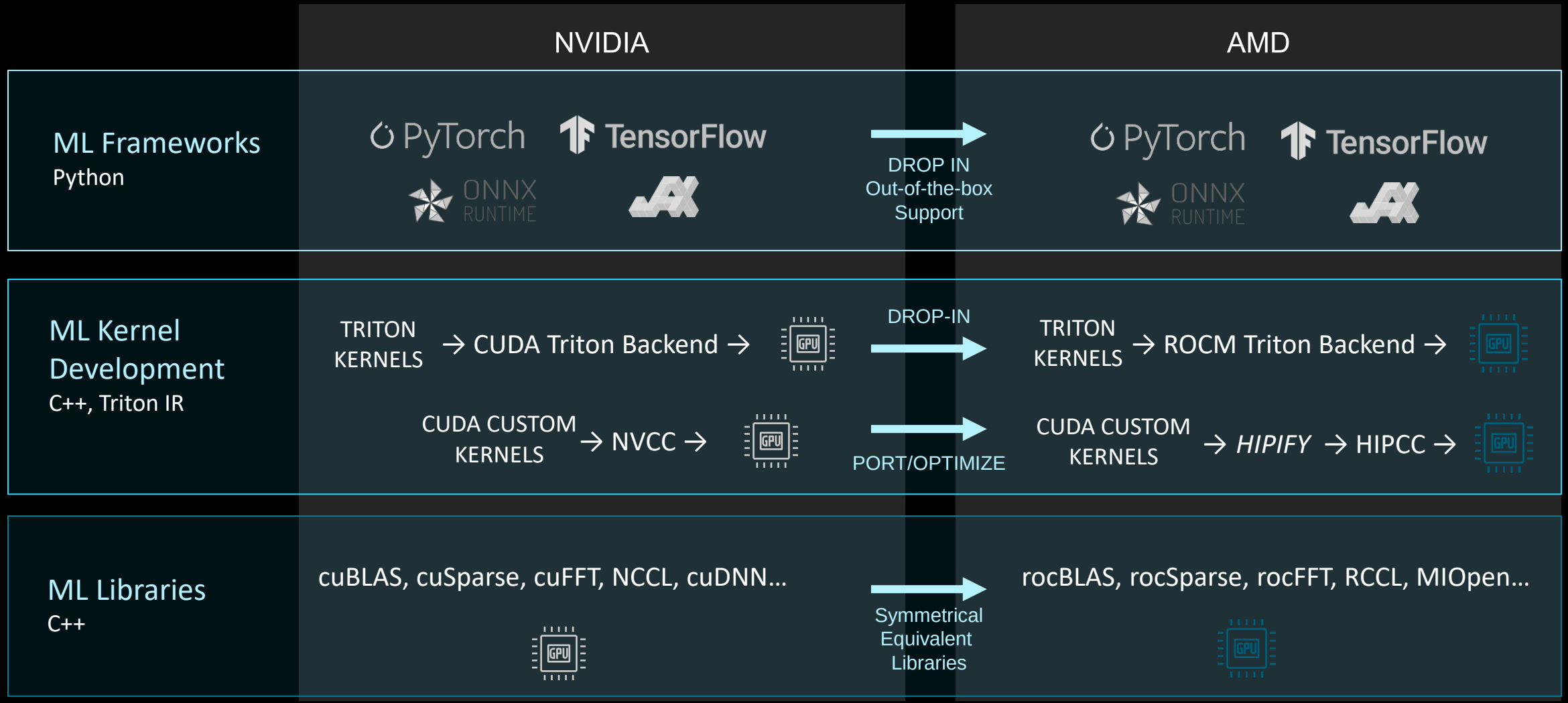
COMMUNITY ENGAGEMENT

- Increasing community involvement with multiple exascale systems accessed by 1000s of developers
- Opportunity for customers to contribute code into the ROCm stack via github - active involvement on github forums (for ROCm core libraries) to drive closure of user requests
- Engaging Linux distro owners for wider validation coverage across releases



TRANSITIONING WORKLOADS TO INSTINCT GPUS

LOW FRICTION SOFTWARE PORTING FOR EXISTING NVIDIA USERS TO AMD



AMD Instinct™ MI300A

OEM and Solution Partners

Hewlett Packard Enterprise

EVIDEN an atos business

GIGABYTE™

SUPERMICRO

References

- AMD Instinct™ Accelerators
  - https://www.amd.com/en/products/accelerators/instinct.html
- AMD Instinct™ MI300A APU Datasheet
  - https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/data-sheets/amd-instinct-mi300a-data-sheet.pdf
- AMD Instinct™ MI300X Datasheet
  - https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/data-sheets/amd-instinct-mi300x-data-sheet.pdf
- AMD Instinct™ MI300X Platform Datasheet
  - https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/data-sheets/amd-instinct-mi300x-platform-data-sheet.pdf
- AMD CDNA™ 3 Architecture
  - https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/white-papers/amd-cdna-3-white-paper.pdf
- AMD ROCm™ Documentation
  - https://rocm.docs.amd.com/en/latest/


AMD

ENDNOTES

- MI300-05A: Calculations conducted by AMD Performance Labs as of November 17, 2023, for the AMD Instinct™ MI300X OAM accelerator 750W (192 GB HBM3) designed with AMD CDNA™ 3 5nm FinFet process technology resulted in 192 GB HBM3 memory capacity and 5.325 TFLOPS peak theoretical memory bandwidth performance. MI300X memory bus interface is 8,192 and memory data rate is 5.2 Gbps for total peak memory bandwidth of 5.325 TB/s (8,192 bits memory bus interface * 5.2 Gbps memory data rate/8). The highest published results on the NVidia Hopper H200 (141GB) SXM GPU accelerator resulted in 141GB HBM3e memory capacity and 4.8 TB/s GPU memory bandwidth performance. https://nvdam.widen.net/s/nb5zzzsjdf/hpc-datasheet-sc23-h200-datasheet-3002446 The highest published results on the NVidia Hopper H100 (80GB) SXM5 GPU accelerator resulted in 80GB HBM3 memory capacity and 3.35 TB/s GPU memory bandwidth performance. https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet MI300-06: Calculations as of May 17th, 2023. AMD Instinct™ MI300X built on AMD CDNA™ 3 technology accelerators support AMD Infinity Fabric™ technology providing up to 128 GB/s peak total aggregate theoretical transport data GPU peer-to-peer (P2P) bandwidth per AMD Infinity Fabric link.AMD Instinct™ MI300X AMD CDNA 3 technology-based accelerators include up to eight AMD Infinity Fabric links providing up to 1,024 GB/s peak aggregate theoretical GPU peer-to-peer (P2P) transport rate bandwidth performance per GPU OAM card.AMD Instinct™ MI250/MI250X built on AMD CDNA™ 2 technology accelerators support AMD Infinity Fabric™ technology providing up to 100 GB/s peak total aggregate theoretical transport data GPU peer-to-peer (P2P) bandwidth per AMD Infinity Fabric link.AMD Instinct™ MI250/MI250X AMD CDNA 2 technology-based accelerators include up to eight AMD Infinity Fabric links providing up to 800 GB/s peak aggregate theoretical GPU peer-to-peer (P2P) transport rate bandwidth performance per GPU OAM card. Server manufacturers may vary configuration offerings yielding different results.
- MI300-06: Calculations as of May 17th, 2023. AMD Instinct™ MI300X built on AMD CDNA™ 3 technology accelerators support AMD Infinity Fabric™ technology providing up to 128 GB/s peak total aggregate theoretical transport data GPU peer-to-peer (P2P) bandwidth per AMD Infinity Fabric link.AMD Instinct™ MI300X AMD CDNA 3 technology-based accelerators include up to eight AMD Infinity Fabric links providing up to 1,024 GB/s peak aggregate theoretical GPU peer-to-peer (P2P) transport rate bandwidth performance per GPU OAM card.AMD Instinct™ MI250/MI250X built on AMD CDNA™ 2 technology accelerators support AMD Infinity Fabric™ technology providing up to 100 GB/s peak total aggregate theoretical transport data GPU peer-to-peer (P2P) bandwidth per AMD Infinity Fabric link.AMD Instinct™ MI250/MI250X AMD CDNA 2 technology-based accelerators include up to eight AMD Infinity Fabric links providing up to 800 GB/s peak aggregate theoretical GPU peer-to-peer (P2P) transport rate bandwidth performance per GPU OAM card. Server manufacturers may vary configuration offerings yielding different results.
- MI300-11: Measurements conducted by AMD Performance Labs as of November 11th, 2023 on the AMD Instinct™ MI300X (750W) GPU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 163.4 TFLOPS peak theoretical double precision (FP64 Matrix), 81.7 TFLOPS peak theoretical double precision (FP64), 163.4 TFLOPS peak theoretical single precision matrix (FP32 Matrix), 163.4 TFLOPS peak theoretical single precision (FP32), 653.7 TFLOPS peak theoretical TensorFloat-32 (TF32), 1307.4 TFLOPS peak theoretical half precision (FP16), 1307.4 TFLOPS peak theoretical Bfloat16 format precision (BF16), 2614.9 TFLOPS peak theoretical 8-bit precision (FP8), 2614.9 TOPs INT8 floating-point performance. The results calculated for the AMD Instinct™ MI250X (560W) 128GB HBM2e OAM accelerator designed with AMD CDNA™ 2 5nm FinFET process technology at 1,700 MHz peak boost engine clock resulted in 95.7 TFLOPS peak theoretical double precision (FP64 Matrix), 47.9 TFLOPS peak theoretical double precision (FP64), 95.7 TFLOPS peak theoretical single precision matrix (FP32 Matrix), 47.9 TFLOPS peak theoretical single precision (FP32), TF32 (N/A), 383.0 TFLOPS peak theoretical half precision (FP16), 383.0 TFLOPS peak theoretical Bfloat16 format precision (BF16), FP8 (N/A), 383.0 TOPs INT8 floating-point performance.AMD TFLOPS and TOPS calculations conducted with the following equation for AMD Instinct MI300X and MI250X GPUs: FLOPS calculations are performed by taking the engine clock from the highest DPM state and multiplying it by xx CUs per GPU. Then, multiplying that number by xx stream processors, which exist in each CU. Then, for MI300X that number is multiplied by 256 FLOPS per clock/CU for FP64 Matrix to determine TFLOPS, 128 FLOPS per clock/CU for FP64 to determine TFLOPS, 256 FLOPS per clock/CU for FP32 Matrix to determine TFLOPS, 256 FLOPS per clock/CU for FP32 to determine TFLOPS, 1024 FLOPS per clock/CU for TF32 to determine TFLOPS, 2048 FLOPS per clock/CU for FP16 to determine TFLOPS, 2048 FLOPS per clock/CU for BF16 to determine TFLOPS, 4096 FLOPS per clock/CU for FP8 to determine TFLOPS, 4046 FLOPS per clock/CU for INT8 to determine TOPS. Divide results by 100,000 to get TFLOPS.Then, for MI250X that number is multiplied by 256 FLOPS per clock/CU for FP64 Matrix to determine TFLOPS, 128 FLOPS per clock/CU for FP64 to determine TFLOPS, 256 FLOPS per clock/CU for FP32 Matrix to determine TFLOPS, 128 FLOPS per clock/CU for FP32 to determine TFLOPS, 1024 FLOPS per clock/CU for FP16 to determine TFLOPS, 1024 FLOPS per clock/CU for BF16 to determine TFLOPS. 1024 FLOPS per clock/CU for INT8 to determine TOPS. Divide results by 100,000 to get TFLOPS (TF32, FP8 Not Available). Divide results by 100,000 to get TFLOPS.Calculations:FP64 Matrix: 163.4 / 95.7 = 1.71x (71% faster)FP64: 81.7 / 47.9 = 1.71x (71% faster)FP32 Matrix: 163.4 / 95.7 = 1.71x (71% faster)FP32: 163.4 / 47.9 = 3.41x (241% faster)FP16: 1307.4 / 383.0 = 3.41x (241% faster)BF16: 1307.4 / 383.0 = 3.41x (241% faster)FP8: 2614.9 (FP8) / 383.0 (FP16) = 6.83x (583% faster)*INT8: 2614.9 / 383.0 = 6.83x (583% faster) * MI200 Series GPUs don't support TF32, FP8 or sparsity
- MI300-12: Calculations conducted by AMD Performance Labs as of November 7, 2023, for the AMD Instinct™ MI300A APU accelerator 760W (128 GB HBM3) designed with AMD CDNA™ 3 5nm FinFet process technology resulted in 128 GB HBM3 memory capacity and 5.325 TFLOPS peak theoretical memory bandwidth performance. MI300A memory bus interface is 8,192 (1024 bits x 8 die) and memory data rate is 5.2 Gbps for total peak memory bandwidth of 5.325 TB/s (8,192 bits memory bus interface * 5.2 Gbps memory data rate/8). The highest published results on the NVidia Hopper H200 (141GB) SXM GPU accelerator resulted in 141GB HBM3e memory capacity and 4.8 TB/s GPU memory bandwidth performance. https://nvdam.widen.net/s/nb5zzzsjdf/hpc-datasheet-sc23-h200-datasheet-3002446 The highest published results on the NVidia Hopper H100 (80GB) SXM GPU accelerator resulted in 80GB HBM3 memory capacity and 3.35 TB/s GPU memory bandwidth performance. https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet Memory Capacity:MI300A APU: 128GB HBM3 / H100 SXM5: 80GB HBM3 = 1.6X (60% more)Memory Bandwidth:MI300A APU: 5.325 TB/s / H200 SXM5: 4.8 TB/s = ~1.109X (up to 11% more)MI300A OAM: 5.325 TB/s / H100 SXM5: 3.352 TB/s = ~1.589X (up to 59% more)

ENDNOTES

- MI300-13: Calculations conducted by AMD Performance Labs as of November 7, 2023, for the AMD Instinct™ MI300X OAM accelerator 750W (192 GB HBM3) designed with AMD CDNA™ 3 5nm FinFet process technology resulted in 192 GB HBM3 memory capacity and 5.325 TFLOPS peak theoretical memory bandwidth performance. MI300X memory bus interface is 8,192 (1024 bits x 8 die) and memory data rate is 5.2 Gbps for total peak memory bandwidth of 5.325 TB/s (8,192 bits memory bus interface * 5.2 Gbps memory data rate/8). The AMD Instinct™ MI250 (500W) / MI250X (560W) OAM accelerators (128 GB HBM2e) designed with AMD CDNA™ 2 6nm FinFet process technology resulted in 128 GB HBM3 memory capacity and 3.277 TFLOPS peak theoretical memory bandwidth performance. MI250/MI250X memory bus interface is 8,192 (4,096 bits times 2 die) and memory data rate is 3.20 Gbps for total memory bandwidth of 3.277 TB/s ((3.20 Gbps*(4,096 bits*2))/8). Memory Capacity:MI300X OAM: 192GB HBM3 / MI250/MI250X OAMs: 128GB HBM2e = 1.5X (50% more)Memory Bandwidth:MI300X OAM: 5.325 TB/s / MI250/MI250X OAMs : 3.2 TB/s = ~1.66X (up to 66% more)
- MI300-16: Measurements conducted by AMD Performance Labs as of November 11th, 2023 on the AMD Instinct™ MI300X (750W) GPU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 653.7 TFLOPS peak theoretical TensorFloat-32 (TF32), 1307.4 TFLOPS peak theoretical half precision (FP16), 1307.4 TFLOPS peak theoretical Bfloat16 format precision (BF16), 2614.9 TFLOPS peak theoretical 8-bit precision (FP8), 2614.9 TOPs INT8 floating-point performance.The MI300X is expected to be able to take advantage of fine-grained structure sparsity providing an estimated 2x improvement in math efficiency resulting 1,307.4 TFLOPS peak theoretical TensorFloat-32 (TF32), 2,614.9 TFLOPS peak theoretical half precision (FP16), 2,614.9 TFLOPS peak theoretical Bfloat16 format precision (BF16), 5,229.8 TFLOPS peak theoretical 8-bit precision (FP8), 5,229.8 TOPs INT8 floating-point performance with sparsity. The results calculated for the AMD Instinct™ MI250X (560W) 128GB HBM2e OAM accelerator designed with AMD CDNA™ 2 5nm FinFET process technology at 1,700 MHz peak boost engine clock resulted in TF32* (N/A), 383.0 TFLOPS peak theoretical half precision (FP16), 383.0 TFLOPS peak theoretical Bfloat16 format precision (BF16), FP8* (N/A), 383.0 TOPs INT8 floating-point performance. *AMD Instinct MI200 Series GPUs don't support TF32, FP8 or sparsity.
- MI300-17: Measurements conducted by AMD Performance Labs as of November 11th, 2023 on the AMD Instinct™ MI300X (750W) GPU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 653.7 TFLOPS peak theoretical TensorFloat-32 (TF32), 1307.4 TFLOPS peak theoretical half precision (FP16), 1307.4 TFLOPS peak theoretical Bfloat16 format precision (BF16), 2614.9 TFLOPS peak theoretical 8-bit precision (FP8), 2614.9 TOPs INT8 floating-point performance.The MI300X is expected to be able to take advantage of fine-grained structure sparsity providing an estimated 2x improvement in math efficiency resulting 1,307.4 TFLOPS peak theoretical TensorFloat-32 (TF32), 2,614.9 TFLOPS peak theoretical half precision (FP16), 2,614.9 TFLOPS peak theoretical Bfloat16 format precision (BF16), 5,229.8 TFLOPS peak theoretical 8-bit precision (FP8), 5,229.8 TOPs INT8 floating-point performance with sparsity.Sparsity Calculations:TF32: 1,307.4 TFLOPS (653.7 TFLOPS x 2 = 1,307.4 TFLOPS)FP16: 2,614.9 TFLOPS (1,307.4 TFLOPS x 2 = 2,614.9 TFLOPS)BF16: 2,614.9 TFLOPS (1,307.4 TFLOPS x 2 = 2,614.9 TFLOPS)FP8: 5,229.8 TFLOPS (2614.9 TFLOPS x 2 = 5,229.8 TFLOPS)INT8: 5,229.8 TOPS (2614.9 TFLOPS x 2 = 5,229.8 TOPS) Published results on Nvidia H100 SXM (80GB) 700W GPU resulted in 989.4 TFLOPs peak TensorFloat-32 (TF32) with sparsity, 1,978.9 TFLOPS peak theoretical half precision (FP16) with sparsity, 1,978.9 TFLOPS peak theoretical Bfloat16 format precision (BF16) with sparsity, 3,957.8 TFLOPS peak theoretical 8-bit precision (FP8) with sparsity, 3,957.8 TOPs peak theoretical INT8 with sparsity floating-point performance.Nvidia H100 source: https://resources.nvidia.com/en-us-tensor-core/ AMD TFLOPS and TOPS calculations conducted with the following equation for AMD Instinct MI300X GPUs: FLOPS calculations are performed by taking the engine clock from the highest DPM state and multiplying it by xx CUs per GPU. Then, multiplying that number by xx stream processors, which exist in each CU. Then, for MI300X that number is multiplied by 1024 FLOPS per clock/CU for TF32 to determine TFLOPS, 2048 FLOPS per clock/CU for FP16 to determine TFLOPS, 2048 FLOPS per clock/CU for BF16 to determine TFLOPS, 4096 FLOPS per clock/CU for FP8 to determine TFLOPS, 4046 FLOPS per clock/CU for INT8 to determine TOPS. Divide results by 100,000 to get TFLOPS or TOPs.Calculations:TF32 Sparsity: MI300X 1,307.4 / H100 989.4 = 1.321x (32% faster) the floating-point performanceFP16 Sparsity: MI300X 2,614.9 / H100 1,978.9 = 1.321x (32% faster) the floating-point performanceBF16 Sparsity: MI300X 2,614.9 / H100 1,978.9 = 1.321x (32% faster) the floating-point performance FP8 Sparsity: MI300X 5,229.8 / H100 3,957.8 = 1.321x (32% faster) the floating-point performance INT8 Sparsity: MI300X 5,229.8 / H100 3,957.8 = 1.321x (32% faster) the floating-point performance
- MI300-18: Measurements conducted by AMD Performance Labs as of November 11th, 2023 on the AMD Instinct™ MI300X (750W) GPU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 163.4 TFLOPs peak theoretical double precision Matrix (FP64 Matrix), 81.7 TFLOPs peak theoretical double precision (FP64), 163.4 TFLOPs peak theoretical single precision Matrix (FP32 Matrix), 163.4 TFLOPs peak theoretical single precision (FP32), 653.7 TFLOPS peak theoretical TensorFloat-32 (TF32), 1307.4 TFLOPS peak theoretical half precision (FP16), 1307.4 TFLOPS peak theoretical Bfloat16 format precision (BF16), 2614.9 TFLOPS peak theoretical 8-bit precision (FP8), 2614.9 TOPs INT8 floating-point performance.Published results on Nvidia H100 SXM (80GB) GPU resulted in 66.9 TFLOPs peak theoretical double precision tensor (FP64 Tensor), 33.5 TFLOPs peak theoretical double precision (FP64), 66.9 TFLOPs peak theoretical single precision (FP32), 494.7 TFLOPs peak TensorFloat-32 (TF32)*, 989.4 TFLOPS peak theoretical half precision tensor (FP16 Tensor), 133.8 TFLOPS peak theoretical half precision (FP16), 989.4 TFLOPS peak theoretical Bfloat16 tensor format precision (BF16 Tensor), 133.8 TFLOPS peak theoretical Bfloat16 format precision (BF16), 1,978.9 TFLOPS peak theoretical 8-bit precision (FP8), 1,978.9 TOPs peak theoretical INT8 floating-point performance.Nvidia H100 source: https://resources.nvidia.com/en-us-tensor-core/ * Nvidia H100 GPUs don't support FP32 Tensor.

ENDNOTES

- MI300-20: Measurements conducted by AMD Performance Labs as of November 11th, 2023 on the AMD Instinct™ MI300A (760W) GPU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 122.6 TFLOPs peak theoretical double precision Matrix (FP64 Matrix), 61.3 TFLOPs peak theoretical double precision (FP64), 122.6 TFLOPs peak theoretical single precision Matrix (FP32 Matrix), 122.6 TFLOPs peak theoretical single precision (FP32), 490.29 TFLOPS peak theoretical TensorFloat-32 (TF32), 980.58 TFLOPS peak theoretical half precision (FP16), 980.58 TFLOPS peak theoretical Bfloat16 format precision (BF16), 1,961.16 TFLOPS peak theoretical 8-bit precision (FP8), 1,961.16 TOPs INT8 floating-point performance.Published results on Nvidia H100 SXM (80GB) 700W GPU resulted in 66.9 TFLOPs peak theoretical double precision tensor (FP64 Tensor), 33.5 TFLOPs peak theoretical double precision (FP64), 66.9 TFLOPs peak theoretical single precision (FP32), 494.7 TFLOPs peak TensorFloat-32 (TF32), 989.4 TFLOPS peak theoretical half precision tensor (FP16 Tensor), 133.8 TFLOPS peak theoretical half precision (FP16), 989.4 TFLOPS peak theoretical Bfloat16 tensor format precision (BF16 Tensor), 133.8 TFLOPS peak theoretical Bfloat16 format precision (BF16), 1,978.9 TFLOPS peak theoretical 8-bit precision (FP8), 1,978.9 TOPs peak theoretical INT8 floating-point performance.Nvidia H100 source: https://resources.nvidia.com/en-us-tensor-core/ AMD TFLOPS and TOPS calculations conducted with the following equation for AMD Instinct MI300A APUs: FLOPS calculations are performed by taking the engine clock from the highest DPM state and multiplying it by xx CUs per APU. Then, multiplying that number by xx stream processors, which exist in each CU. Then, for MI300A that number is multiplied by 1024 FLOPS per clock/CU for TF32 to determine TFLOPS, 2048 FLOPS per clock/CU for FP16 to determine TFLOPS, 2048 FLOPS per clock/CU for BF16 to determine TFLOPS, 4096 FLOPS per clock/CU for FP8 to determine TFLOPS, 4046 FLOPS per clock/CU for INT8 to determine TOPS. Divide results by 100,000 to get TFLOPS or TOPs.Calculations:FP64 Matrix | Tensor: MI300A 122.57 / H100 66.9 = 1.832x (83% faster) the floating-point performanceFP64: MI300A 61.3 / H100 33.5 = 1.830x (83% faster) the floating-point performanceFP32 Matrix: MI300A 122.57 / H100 (FP32) 66.9 = 1.832x (83% faster) the floating-point performance*FP32: MI300A 122.57 / H100 66.9 = 1.832x (83% faster) the floating-point performanceTF32: MI300A 490.29 / H100 494.7 = 0.991x (0.009% slower) the floating-point performanceFP16 (Tensor): MI300A 980.58 / H100 (FP64 Tensor) 989.4 = 0.991x (0.009% slower) the floating-point performanceFP16: MI300A 980.58 / H100 133.8 = 7.329x (633% faster) the floating-point performanceBF16 (Tensor): MI300A 980.58 / H100 (Tensor) 989.4 = 0.991x (0.009% slower) the floating-point performance BF16: MI300A 980.58 / H100 133.8 = 7.329x (633% faster) the floating-point performance FP8: MI300A 1,961.16 / H100 1,978.9 = 0.991x (0.009% slower) the floating-point performance INT8: MI300A 1,961.16 / H100 1,978.9 = 0.991x (0.009% slower) the floating-point performance* Nvidia H100 GPUs don't support FP32 Tensor.
- MI300-21: Measurements conducted by AMD Performance Labs as of November 11th, 2023 on the AMD Instinct™ MI300A (750W) APU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 490.29 TFLOPS peak theoretical TensorFloat-32 (TF32), 980.58 TFLOPS peak theoretical half precision (FP16), 980.58 TFLOPS peak theoretical Bfloat16 format precision (BF16), 1,961.16 TFLOPS peak theoretical 8-bit precision (FP8), 1,961.16 TOPs INT8 floating-point performance.The MI300A is expected to be able to take advantage of fine-grained structure sparsity providing an estimated 2x improvement in math efficiency resulting 980.58 TFLOPS peak theoretical TensorFloat-32 (TF32), 1,961.16 TFLOPS peak theoretical half precision (FP16), 1,961.16 TFLOPS peak theoretical Bfloat16 format precision (BF16), 3,922.33 TFLOPS peak theoretical 8-bit precision (FP8), 3,922.33 TOPs INT8 floating-point performance with sparsity.Published results on Nvidia H100 SXM5 (80GB) GPU resulted in 989.4 TFLOPs peak TensorFloat-32 (TF32) Tensor Core with sparsity, 1,978.9 TFLOPS peak theoretical half precision (FP16) Tensor Core with sparsity, 1,978.9 TFLOPS peak theoretical Bfloat16 format precision (BF16) Tensor Core with sparsity, 3,957.8 TFLOPS peak theoretical 8-bit precision (FP8) Tensor Core with sparsity, 3,957.8 TOPs peak theoretical INT8 Tensor Core with sparsity floating-point performance. Nvidia H100 source: https://resources.nvidia.com/en-us-tensor-core/ Server manufacturers may vary configuration offerings yielding different results.
- MI300-25: Measurements conducted by AMD Performance Labs as of November 18th, 2023 on the AMD Instinct™ MI300X (192 GB HBM3) 750W GPU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 1307.4 TFLOPS peak theoretical half precision (FP16), 1307.4 TFLOPS peak theoretical Bfloat16 format precision (BF16).The MI300X is expected to be able to take advantage of fine-grained structure sparsity providing an estimated 2x improvement in math efficiency resulting 2,614.9 TFLOPS peak theoretical half precision (FP16), 2,614.9 TFLOPS peak theoretical Bfloat16 format precision (BF16 floating-point performance with sparsity.Published results on Nvidia H100 SXM (80GB HBM3) 700W GPU resulted in 989.4 TFLOPS peak theoretical half precision (FP16 Tensor), 989.4 TFLOPS peak theoretical Bfloat16 format precision (BF16 Tensor), 1,978.9 TFLOPS peak theoretical half precision (FP16 Tensor) with sparsity, 1,978.9 TFLOPS peak theoretical Bfloat16 format precision (BF16 Tensor) with sparsity floating-point performance. Nvidia H100 source: https://resources.nvidia.com/en-us-tensor-core/ AMD Instinct™ MI300X AMD CDNA 3 technology-based accelerators include up to eight AMD Infinity Fabric links providing up to 1,024 GB/s peak aggregate theoretical GPU peer-to-peer (P2P) transport rate bandwidth performance per GPU OAM module.
- MI300-29: GROMACS STMV comparison based on AMD internal testing as of 11/18/2023 and published Nvidia data.Configurations: AMD Instinct™ MI300A bring-up platform with 1x AMD Instinct MI300A (128GB, 550W) APU, Pre-release build of ROCm® 6, Ubuntu® 22.04.2. GROMACS version 2022.0.Vs.Nvidia public claims https://developer.nvidia.com/hpc-application-performance, as of 11/17/2023. GROMACS version 2023.2. Only 1 GPU on each system was used in this test.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.
- MI300-30: HPCG benchmark comparison based on AMD internal testing as of 11/16/2023.Configurations: AMD Instinct™ MI300A bring-up platform with 4x AMD Instinct MI300A (128GB, 550W) APU, pre-release build of ROCm™ 6.0, Ubuntu® 22.04.2, HPCG rocm pre-release 6.0.Vs.Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, CUDA® 12.0, Ubuntu 22.04.3, CUDA container nvcr.io/nvidia/hpc-benchmarks:23.10Only 1 GPU on each system was used in this test.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

ENDNOTES

- MI300-31: Mini-Nbody benchmark comparison based on AMD internal testing as of 11/18/2021.Configurations: AMD Instinct™ MI300A bring-up reference platform with 1x AMD Instinct MI300A (128GB, 550W) APU , pre-release build of ROCm™ 6.0, Ubuntu® 22.04.2.Vs.Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, CUDA® 12.0, Ubuntu 22.04.3.Only 1 GPU on each system was used in testing mini-nbody (https://github.com/harrism/mini-nbody)BenchmarkServer manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.
- MI300-32: OpenFOAM® v2206 HPC Motorbike comparison based on AMD internal testing as of 11/15/2023.Configurations: AMD Instinct™ MI300A bring-up platform with 4x AMD Instinct MI300A (128GB, 550W) APU, pre-release build of ROCm™ 6.0, Ubuntu 22.04.2Vs.Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, CUDA 12.0, Ubuntu 22.04.3.Only 1 GPU on each system was used in this test.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.
- MI300-33: Text generated with Llama2-70b chat using input sequence length of 4096 and 32 output token comparison using custom docker container for each system based on AMD internal testing as of 11/17/2023.Configurations: 2P Intel Xeon Platinum CPU server using 4x AMD Instinct™ MI300X (192GB, 750W) GPUs, ROCm® 6.0 pre-release, PyTorch 2.2.0, vLLM for ROCm, Ubuntu® 22.04.2.Vs.2P AMD EPYC 7763 CPU server using 4x AMD Instinct™ MI250 (128 GB HBM2e, 560W) GPUs, ROCm® 5.4.3, PyTorch 2.0.0., HuggingFace Transformers 4.35.0, Ubuntu 22.04.6.4 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.
- MI300-34: 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.
- MI300-35: Flash Attention v2 forward kernel for inference, head_dim=128 and causal=false, comparison based on AMD internal testing as of 11/29/2023.Configurations: AMD MI300X bring-up platform with 1x AMD Ryzen™ 9 7950X, 1x AMD Instinct™ MI300X (192GB, 750W) GPU, Ubuntu® 22.04.3, Pre-release build of ROCm™ 6.0, Flash attention v2 forward kernel using an internal containerVsAn Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, Ubuntu® 22.04.3, CUDA® 12.2.2Flash attention v2 forward kernel using nvcr.io/nvidia/pytorch:23.10-py3 container.Only 1 GPU on each system was used in this test.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.
- MI300-36: Overall latency for text generation using the Llama2-13b chat model with vLLM comparison based on AMD internal testing as of 11/29/2023. Tests were performed using an input sequence length of 2048 input tokens and 128 output tokens.Configurations: AMD MI300X bring-up platform with 1x AMD Ryzen 9 7950X, 1x AMD Instinct™ MI300X (192GB, 750W) GPU, ROCm® 6.0 pre-release, Ubuntu® 22.04.2, AMD port of vLLM for ROCm.Vs.An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, CUDA 12.1, Ubuntu 22.04.3, vLLM v.0.2.2 (most recent).Only 1 GPU on each system was used in this test.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.
- MI300-37: Llama2-70b inference comparison, with Key GEMM kernels used, based on AMD internal testing as of 11/17/2023.Configurations: AMD MI300X bring-up platform with 1x AMD Ryzen 9 7950X, 1x AMD Instinct™ MI300X (192GB, 750W) GPU, ROCm® 6.0 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 12.2.2, Ubuntu 22.04.3Only 1 GPU on each system was used in this test.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.
- MI300-38: Overall latency for text generation using the Llama2-70b chat model with vLLM comparison using custom docker container for each system based on AMD internal testing as of 11/23/2023. Sequence length of 2048 input tokens and 128 output tokens.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.
- 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.

ENDNOTES

- 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.
- "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 Instinct™ 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 GPUsGopher Deepmind (290B) 4 Calculated 8 CalculatedLargest parameter size for 8x MI300X: MI300X GPUs H100 GPUsPaLM-1 (680B) 8 Calculated 19 CalculatedResults for Training:Largest parameter size for 8X H100: MI300X GPUs H100 GPUsMosiac MPT-30B parameter 4 Calculated 8 CalculatedLargest parameter size for 8x MI300X: MI300X GPUs H100 GPUsMosiac MPT-70B parameter 7 Calculated 16 CalculatedAssumptions:FP16 DatatypeBatchsize 1Memory needs for model = 2Bytes per ParameterMemory 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. M300-42“
- MI300-43: Measurements conducted by AMD Performance Labs as of December 4th, 2023 on the AMD Instinct™ MI300A (760W) APU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in:• 122.6 TFLOPs peak theoretical double precision Matrix (FP64 Matrix), • 61.3 TFLOPs peak theoretical double precision (FP64), • 122.6 TFLOPs peak theoretical single precision Matrix (FP32 Matrix), • 122.6 TFLOPs peak theoretical single precision (FP32), floating-point performance.Published results on Nvidia GH200 1000W GPU: • 67 TFLOPs peak theoretical double precision tensor (FP64 Tensor), • 34 TFLOPs peak theoretical double precision (FP64), • N/A FP 32 Tensor - Nvidia GH200 GPUs don’t support FP32 Tensor. Regular FP32 number used as proxy. • 67 TFLOPs peak theoretical single precision (FP32), floating-point performance.Nvidia GH200 source: https://resources.nvidia.com/en-us-grace-cpu/grace-hopper-superchip GH200 TFLOPs per Watt Calculations (peak wattage of 1000W used):• FP64 Matrix: 67 TFLOPs / 1000W = 0.067 TFLOPs per Watt• FP64: 34 TFLOPs / 1000W = 0.034 TFLOPs per Watt• FP32: 67 TFLOPs / 1000W = 0.067 TFLOPs per Watt* Nvidia GH200 GPUs don’t support FP32 Tensor. Actual performance and performance per watt may vary on production systems.
- MI300-44: Llama2-70b model vLLM, hip Graph and Flash Attention LLM Performance Optimization comparison using custom docker containers across sequence lengths from 512 to 7168 based on AMD internal testing as of 11/22/2023.Testing done by comparing baseline (LLama2-70b model vLLM, hip Graph and Flash Attention LLM) performance optimizations off. This performance was measured against the performance with each optimization turned on to determine the performance impact of the optimization. Configurations: 2P Intel Xeon Platinum 8480C CPU server using 4x AMD Instinct™ MI300X (192GB, 750W) GPUs, ROCm® 6.0 pre-release, PyTorch 2.2.0, Ubuntu® 22.04.2.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.