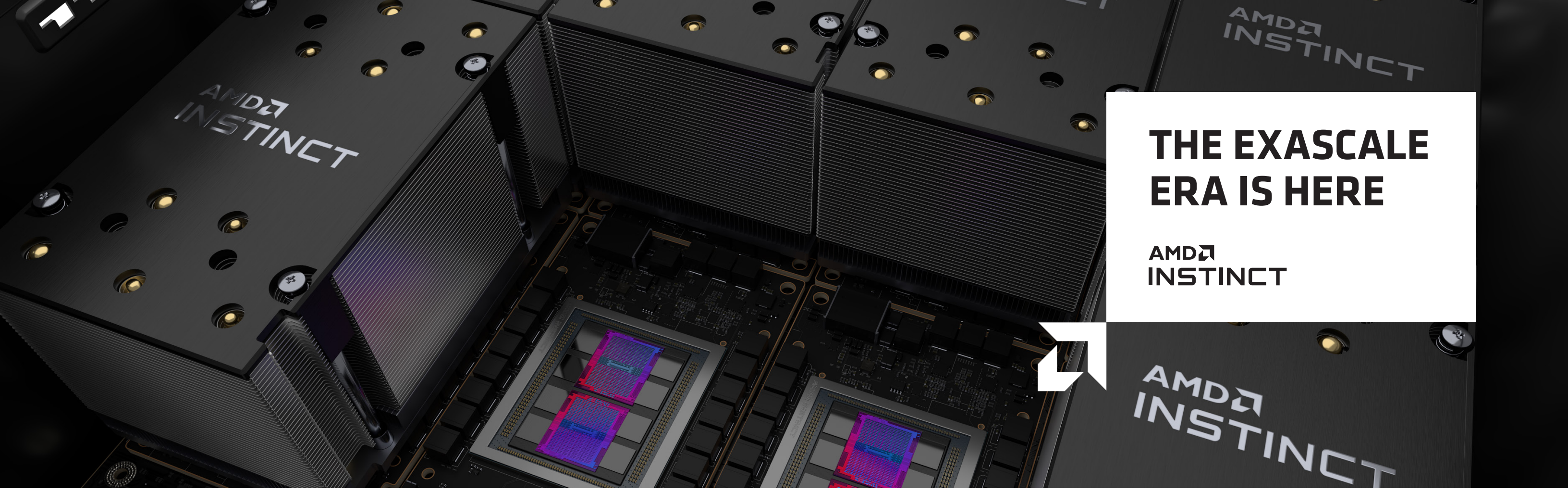


THE EXASCALE ERA IS HERE

AMD INSTINCT

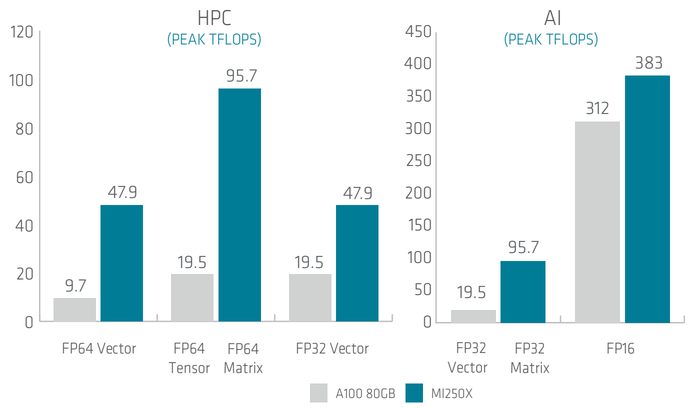
AMD INSTINCT™ MI200 SERIES ACCELERATOR

World's Fastest HPC and AI Accelerator[1]

The era of exascale is here. Immense computational power coupled with the fusion of HPC and AI is enabling researchers and scientists to tackle our most pressing challenges from climate change to vaccine research. With the AMD Instinct™ MI200 accelerators and ROCm™ 5.0 software ecosystem, innovators can tap the power of the world's most powerful HPC and AI data center GPUs to accelerate their time to science and discovery.[1]

Based on the 2nd Gen AMD CDNA™ architecture, AMD Instinct™ MI200 accelerators deliver a quantum leap in HPC and AI performance over competitive data center GPUs today. With an up to 4x advantage in HPC performance compared to competitive GPUs, the MI200 accelerator is the first data center GPU to deliver 383 teraflops of theoretical mixed precision FP16 performance for deep learning training, offering users a powerful platform to fuel the convergence of HPC and AI.[1]

Clean Sweep Performance Leadership



Graph 1: Peak TFLOPS across range of mixed-precision Compute[1]

Innovations Delivering Performance Leadership

AMD innovations in architecture, packaging and integration are pushing the boundaries of computing by unifying the most important processors in the data center, the CPU and the GPU accelerator. With industry-first multi-chip GPU modules along with 3rd Gen AMD Infinity Architecture, AMD is delivering performance, efficiency and overall system throughput for HPC and AI using AMD EPYC™ CPUs and AMD Instinct™ MI200 series accelerators

Key Features

PERFORMANCE	MI250	MI250X
Compute Units	208CU	220CU
Stream Processors	13,312	14,080
Peak FP64/FP32 Vector	45.3 TFLOPS	47.9 TFLOPS
Peak FP64/FP32 Matrix	90.5 TFLOPS	95.7 TFLOPS
Peak FP16/BF16	362.1 TFLOPS	383.0 TFLOPS
Peak INT4/INT8	362.1 TOPS	383.0 TOPS
MEMORY		
Memory Size	128GB HBM2e	128GB HBM2e
Memory Interface	8,192 bits	8,192 bits
Memory Clock	1.6GHz	1.6GHz
Memory Bandwidth	up to 3.2TB/sec[2]	up to 3.2TB/sec[2]

RELIABILITY	MI250	MI250X
ECC (Full-chip)	Yes	Yes
RAS Support	Yes	Yes
SCALABILITY		
Infinity Fabric™ Links	up to 6	up to 8
Coherency Enabled	No	Yes
OS Support	Linux™ 64 Bit	Linux 64 Bit
AMD ROCm™ Compatible	Yes	Yes
BOARD DESIGN		
Form Factor	OAM	OAM
Thermal	Passive & Liquid	Passive & Liquid
Max Power	500W & 560W TDP	500W & 560W TDP
Bus Interface	PCIe® Gen 4 Support	
Warranty	3 Year Limited[4]	

Ecosystem without Borders

AMD ROCm™ is an open software platform allowing researchers to tap the power of AMD Instinct™ accelerators to drive scientific discoveries. The ROCm platform is built on the foundation of open portability, supporting environments across multiple accelerator vendors and architectures. With ROCm 5.0, AMD extends its platform powering top HPC and AI applications with AMD Instinct MI200 series accelerators, increasing accessibility of ROCm for developers and delivering outstanding performance across key workloads.

HPC and MACHINE LEARNING APPLICATIONS

HPC | Life Sciences | Chemistry | Energy | Weather | Astrophysics | Automotive | Reinforcement Learning | Image | Object | Video Detection & Classification

OPEN PROGRAMING WITH CHOICE

OpenMP | HIP | OpenCL™ | Python

OPEN FRAMEWORKS

PyTorch | TensorFlow | ONNX | Kokkos | RAJA

OPTIMIZED LIBRARIES

BLAS | FFT | RNG | SPARSE | THRUST | MIOpen | RCCL

PROGRAMER AND SYSTEM TOOLS

Debuggers | Performance Analysis | System Management

2nd Generation AMD CNDA™ Architecture

The AMD Instinct™ MI200 series accelerator brings customers the compute engine selected for the first U.S. Exascale supercomputer. Powered by the 2nd Generation AMD CDNA™ architecture, the MI200 series accelerators deliver a quantum leap in HPC and AI performance over competitive data center GPUs today. The AMD Instinct MI200 series GPU delivers industry-leading double precision performance for HPC workloads with up to 47.9TFLOPS peak FP64 performance, enabling scientists and researchers across the globe to process HPC parallel codes more efficiently across several industries.[1]

AMD's Matrix Core technology delivers a full range of mixed precision operations bringing you the ability to work with large models and enhance memory-bound operation performance for whatever combination of AI and machine learning workloads you need to deploy. The MI200 offers optimized BF16, INT4, INT8, FP16, FP32, and FP32 Matrix capabilities bringing you supercharged compute performance to meet all your AI system requirements. The AMD Instinct MI200 accelerator handles large data efficiently for training and is the first data center GPU to deliver 383 teraflops of peak FP16 performance for deep learning training.[1]

AMD Infinity Fabric™ Link Technology

AMD Instinct MI200 series OAM accelerators with advanced peer-to-peer I/O connectivity through a maximum of eight AMD Infinity Fabric™ links deliver up to 800 GB/s I/O bandwidth performance.[3] With a cache coherency enabled solution using 3rd Gen AMD EPYC™ "Trento" CPU and MI250X accelerators, Infinity Fabric unlocks the promise of unified computing, enabling a quick and simple on-ramp for CPU codes to accelerated platforms.

Ultra-Fast HBM2e Memory

The AMD Instinct™ MI200 accelerators provide up to 128GB High-bandwidth HBM2e memory with ECC support at a clock rate of 1.6 GHz. and deliver an ultra-high 3.2 TB/s of memory bandwidth to help support your largest data sets and eliminate bottlenecks in moving data in and out of memory.[2] Combine this performance with the MI200's advanced I/O capabilities and you can push workloads closer to their full potential.

For More Information Visit:
AMD.com/INSTINCT | AMD.com/ROCm

[1] World's fastest data center GPU is the AMD Instinct™ MI250X. Calculations conducted by AMD Performance Labs as of Sep 15, 2021, for the AMD Instinct™ MI250X (128GB HBM2e OAM module) accelerator 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), 383.0 TFLOPS peak theoretical half precision (FP16), and 383.0 TFLOPS peak theoretical Bfloat16 format precision (BF16) floating-point performance. Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 (32GB HBM2 PCIe® card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak theoretical double precision (FP64), 46.1 TFLOPS peak theoretical single precision matrix (FP32), 23.1 TFLOPS peak theoretical single precision (FP32), 184.6 TFLOPS peak theoretical half precision (FP16) floating-point performance. Published results on the NVidia Ampere A100 (80GB) GPU accelerator, boost engine clock of 1410 MHz, resulted in 19.5 TFLOPS peak double precision tensor cores (FP64 Tensor Core), 9.7 TFLOPS peak double precision (FP64). 19.5 TFLOPS peak single precision (FP32), 78 TFLOPS peak half precision (FP16), 312 TFLOPS peak half precision (FP16 Tensor Flow), 39 TFLOPS peak Bfloat 16 (BF16), 312 TFLOPS peak Bfloat16 format precision (BF16 Tensor Flow), theoretical floating-point performance. The TF32 data format is not IEEE compliant and not included in this comparison. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf, page 15, Table 1. MI200-01

[2] Calculations conducted by AMD Performance Labs as of Sep 21, 2021, for the AMD Instinct™ MI250X and MI250 (128GB HBM2e) OAM accelerators designed with AMD CDNA™ 2 6nm FinFet process technology at 1,600 MHz peak memory clock resulted in 3.2768 TFLOPS peak theoretical memory bandwidth performance. MI250/MI250X memory bus interface is 4,096 bits times 2 die and memory data rate is 3.20 Gbps for total memory bandwidth of 3.2768 TB/s ((3.20 Gbps*(4,096 bits*2))/8). The highest published results on the NVidia Ampere A100 (80GB) SXM GPU accelerator resulted in 2.039 TB/s GPU memory bandwidth performance. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf MI200-07

[3] Calculations as of SEP 18th, 2021. AMD Instinct™ MI250 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, and include up to eight links providing up to 800GB/s peak aggregate theoretical GPU (P2P) transport rate bandwidth performance per GPU OAM card for 800 GB/s. AMD Instinct™ MI100 built on AMD CDNA technology accelerators support PCIe® Gen4 providing up to 64 GB/s peak theoretical transport data bandwidth from CPU to GPU per card, and include three links providing up to 276 GB/s peak theoretical GPU P2P transport rate bandwidth performance per GPU card. Combined with PCIe Gen4 support, this provides an aggregate GPU card I/O peak bandwidth of up to 340 GB/s. Server manufacturers may vary configuration offerings yielding different results. MI200-13

[4] The AMD Instinct™ accelerator products come with a three-year limited warranty. Please visit www.AMD.com/warranty page for warranty details on the specific graphics products purchased. Toll-free phone service available in the U.S. and Canada only, email access is global.

We thank the Computational Infrastructure for Geodynamics (http://geodynamics.org) which is funded by the National Science Foundation under awards EAR-0949446 and EAR-1550901.