
# BRINGING ADVANCED DATACENTER TECHNOLOGIES TO MAINSTREAM HPC & AI

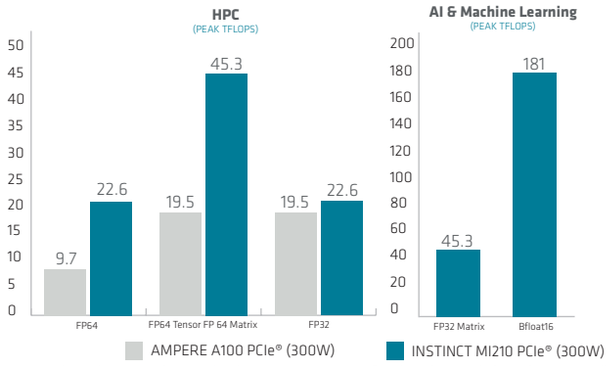
AMD INSTINCT

## AMD INSTINCT™ MI210 ACCELERATOR

### Advanced Technologies for the Data Center

The AMD Instinct™ MI210 accelerator extends AMD industry performance leadership in accelerated compute for double precision (FP64) on PCIe® form factors for mainstream HPC and AI workloads in the data center.[1,2] Built with the 2nd Gen AMD CDNA™ architecture, the MI210 enables scientists and researchers to tackle our most pressing challenges in HPC and AI. MI210 accelerators, combined with the AMD ROCm™ 6 software ecosystem, allow innovators to tap the power of HPC and AI data center PCIe® GPUs to accelerate their time to science and discovery.

**Superior Performance for HPC & AI**




Graph 1: Peak TFLOPS across range of mixed-precision Compute[2]

### Purpose-built Accelerators for HPC & AI Workloads

Powered by the 2nd Gen AMD CDNA™ architecture, AMD Instinct™ MI210 accelerator delivers HPC performance leadership over existing competitive PCIe® data center GPUs today with up to a 2.3x advantage over Nvidia Ampere A100 GPUs in FP64 performance delivering exceptional performance for a broad set of HPC & AI applications.[2] The MI210 accelerator is built to accelerate deep learning training, providing an expanded range of mixed-precision capabilities based on the AMD Matrix Core Technology, and delivers an outstanding 181 teraflops peak theoretical FP16 and BF16 performance to bring users a powerful platform to fuel the convergence of HPC and AI.

### 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 our innovative double-precision Matrix Core capabilities along with the 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™ MI210 accelerators.

### Key Features

| PERFORMANCE | MI210 |
|---|---|
| Compute Units | 104 CU |
| Stream Processors | 6,656 |
| Matrix Cores | 416 |
| Peak FP64/FP32 Vector | 22.6 TF |
| Peak FP64/FP32 Matrix | 45.3 TF |
| Peak FP16/BF16 | 181.0 TF |
| Peak INT4/INT8 | 181.0 TOPS |

| MEMORY | |
|---|---|
| Memory Size | 64GB HBM2e |
| Memory Interface | 4,096 bits |
| Memory Clock | 1.6GHz |
| Memory Bandwidth | up to 1.6 TB/sec[3] |

| RELIABILITY | MI210 |
|---|---|
| ECC (Full-chip) | Yes |
| RAS Support | Yes |

| SCALABILITY | |
|---|---|
| Infinity Fabric™ Links | up to 3 |
| Coherency Enabled | Yes (Dual \| Quad Hives) |
| OS Support | Linux™ 64 Bit |
| AMD ROCm™ Compatible | Yes |

| BOARD DESIGN | |
|---|---|
| Board Form Factor | Full-Height, Full-Length (Dual Slot) |
| Length | 4.5" x 10.5" (11.43 CM x 26.67 CM) |
| Bus Interface | PCIe® Gen4 \| Gen3 Support |
| SR-IOV Support | Yes (Passthrough Only) |
| Thermal | Passively Cooled |
| Max Power | 300W TDP (EPS12V, 8-pin) |
| Warranty | Three Year Limited[5] |

AMD

## 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 6, 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

### OPTIMIZED LIBRARIES

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

### OPEN FRAMEWORKS

PyTorch | TensorFlow | ONNX | Kokkos | RAJA

### PROGRAMER AND SYSTEM TOOLS

Debuggers | Performance Analysis | System Management

## 2nd Generation AMD CDNA™ Architecture

Powered by the 2nd Gen AMD CDNA™ architecture, the MI210 accelerator delivers outstanding performance for HPC and AI. The MI210 PCIe® GPU delivers superior double and single precision performance compared to the Nvidia Ampere A100 GPU for HPC workloads with up to 22.6 TFLOPS peak FP64|FP32 performance, enabling scientists and researchers around the globe to process HPC parallel codes more efficiently across several industries.[1]

AMD's Matrix Core technology delivers a broad 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 MI210 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 MI210 accelerator handles large data efficiently for training and delivers 181 teraflops of peak FP16 and bfloat16 floating-point performance for deep learning training.

## For More Information Visit:

## AMD.com/INSTINCT | AMD.com/ROCm

## AMD Infinity Fabric™ Link Technology

AMD Instinct MI210 GPUs provide advanced I/O capabilities in standard off-the-shelf servers with our AMD Infinity Fabric™ technology and PCIe® Gen4 support. The MI210 GPU delivers 64 GB/s CPU to GPU bandwidth without the need for PCIe® switches, and up to 300 GB/s of Peer-to-Peer (P2P) bandwidth performance through three Infinity Fabric links.[4] The AMD Infinity Architecture enables platform designs with dual and quad, direct-connect, GPU hives with high-speed P2P connectivity and delivers up to 1.2 TB/s of total theoretical GPU bandwidth within a server design.[4] Infinity Fabric helps unlock the promise of accelerated computing, enabling a quick and simple on-ramp for CPU codes to accelerated platforms.

## Ultra-Fast HBM2e Memory

AMD Instinct MI210 accelerators provide up to 64GB High-bandwidth HBM2e memory with ECC support at a clock rate of 1.6 GHz. and deliver an ultra-high 1.6 TB/s of memory bandwidth to help support your largest data sets and eliminate bottlenecks moving data in and out of memory.[3] Combine this performance with the MI210's advanced Infinity Fabric I/O capabilities and you can push workloads closer to their full potential.

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 Jan 14, 2022, for the AMD Instinct™ MI210 (64GB HBM2e PCIe® card) accelerator at 1,700 MHz peak boost engine clock resulted in 45.3 TFLOPS peak theoretical double precision (FP64 Matrix), 22.6 TFLOPS peak theoretical double precision (FP64), and 181.0 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), and 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) and 39 TFLOPS peak Bfloat16 format precision (BF16), theoretical floating-point performance. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf, page 15, Table 1. MI200-41

3. Calculations conducted by AMD Performance Labs as of Jan 27, 2022, for the AMD Instinct™ MI210 (64GB HBM2e) accelerator (PCIe®) designed with AMD CDNA™ 2 architecture 6nm FinFet process technology at 1,600 MHz peak memory clock resulted in 64 GB HBM2e memory capacity and 1.6384 TFLOPS peak theoretical memory bandwidth performance. MI210 memory bus interface is 4,096 bits and memory data rate is 3.20 Gbps for total memory bandwidth of 1.6384 TB/s ((3.20 Gbps*(4,096 bits))/8). Calculations conducted by AMD Performance Labs as of Sep 18, 2020, for the AMD Instinct™ MI100 (32GB HBM2) accelerator (PCIe®) designed with AMD CDNA™ architecture 7nm FinFet process technology at 1,502 MHz peak clock resulted in 32 GB HBM2 memory capacity and 1.2288 TFLOPS peak theoretical memory bandwidth performance. MI210 memory bus interface is 4,096 bits and memory data rate is 2.40 Gbps for total memory bandwidth of 1.2288 TB/s ((2.40 Gbps*(4,096 bits))/8). MI200-42

4. Calculations as of JAN 27th, 2022. AMD Instinct™ MI210 built on AMD CDNA™ 2 technology accelerators support PCIe® Gen4 providing up to 64 GB/s peak theoretical data bandwidth from CPU to GPU per card. AMD Instinct™ MI210 CDNA 2 technology-based accelerators include three Infinity Fabric™ links providing up to 300 GB/s peak theoretical GPU to GPU or Peer-to-Peer (P2P) bandwidth performance per GPU card. Combined with PCIe Gen4 support, this provides an aggregate GPU card I/O peak bandwidth of up to 364 GB/s. Dual-GPU hives: One dual-GPU hive provides up to 300 GB/s peak theoretical P2P performance. Four-GPU hives: One four-GPU hive provide up to 600 GB/s peak theoretical P2P performance. Dual four GPU hives in a server provide up to 1.2 TB/s total peak theoretical direct P2P performance per server. AMD Infinity Fabric link technology not enabled: One four-GPU hive provide up to 256 GB/s peak theoretical P2P performance with PCIe® 4.0. 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. AMD Instinct™ MI100 CDNA technology-based accelerators include three Infinity Fabric™ links providing up to 276 GB/s peak theoretical GPU to GPU or Peer-to-Peer (P2P) 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. One four-GPU hive provides up to 552 GB/s peak theoretical P2P performance. Dual four-GPU hives in a server provide up to 1.1 TB/s total peak theoretical direct P2P performance per server. AMD Infinity Fabric link technology not enabled: One four-GPU hive provides up to 256 GB/s peak theoretical P2P performance with PCIe® 4.0. Server manufacturers may vary configuration offerings yielding different results. MI210-43

5. 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-