5 REASONS WHY AMD INSTINCT™ MI210 ACCELERATORS SPEED RESEARCH AND INNOVATION

AT A GLANCE

How can more researchers and enterprises tap into breakthrough performance for fast results in HPC and AI applications? AMD Instinct MI210 accelerators are built on exascale-class technologies, supercharging the ability of every prospective HPC user to reduce time to insights and new discoveries. With a wide-ranging ecosystem of open software, support for a broad set of workloads and streamlined programmability between CPU and GPU, AMD Instinct MI210 accelerators help unlock a higher level of value for your organization.

1

MAXIMIZE YOUR WORKLOADS

Experience better performance and complete your jobs incredibly fast

AMD Instinct MI210 accelerators extend AMD performance leadership in accelerated compute for double precision (FP64) on PCIe® form factor cards for mainstream HPC and AI workloads in the data center.[1,2] Powered by 2nd Gen AMD CDNA™ (Compute DNA) architecture, AMD Instinct 200 series GPUs offer the industry's most powerful platform for HPC.[1]

2

LEVERAGE HIGHER LEVELS OF EFFICIENCY

Achieve more from HPC and AI investments

The AMD CDNA 2 architecture is designed to leverage the 6nm fabrication process, a next-gen architecture that delivers up to 2.3× the FP64 performance per watt of the Nvidia™ A100 (PCIe® 80 GB, 300W) GPU[3], for outstanding efficiency when running the most demanding applications.

3

RELY ON AMD'S CONSISTENT ROADMAP EXECUTION

Help assure your research and innovation future

With the Instinct MI210 accelerator, AMD brings exascale-class technologies to the mainstream and continues to break ground in the GPU arena. AMD has a solid track record of executing on its long-term roadmap, yielding outstanding performance and efficiency across multiple generations of AMD Instinct and AMD EPYC™ processors.

4

ADAPT YOUR SYSTEMS TO SCALE, PURPOSE AND WORKLOADS AS NEEDED

Use a CPU + GPU platform designed expressly for compute acceleration

3rd Gen AMD Infinity Architecture includes the AMD Infinity Fabric™ interconnect between AMD EPYC processors and AMD Instinct MI210 accelerators. This can simplify programmability to unlock unprecedented performance and enable more adaptive high-performance architectures.

5

FOCUS ON YOUR GOALS, NOT THE SILICON

Avoid always having to code down to the subcomponent

ROCm™ 5 software extends AMD's open platform for HPC and AI, optimizing at the device level with a freely shared open-source codebase that can evolve as systems and their parts evolve.

Continue reading for more technical detail

TECHNICAL DEEP DIVE

#1 MAXIMIZE YOUR WORKLOADS

- Powered by 2nd Gen AMD CDNA™ architecture, AMD Instinct™ MI210 accelerators deliver HPC performance leadership over existing competitive PCIe® data center GPUs with up to a 2.3× double-precision (FP64) advantage in HPC performance over Nvidia™ Ampere A100 GPUs for a broad set of HPC applications.[2]
- With 64GB high speed HBM2e memory and 1.6 TB/s peak theoretical memory bandwidth, the AMD Instinct MI210 accelerator provides 33% more bandwidth and 2× the memory capacity of the previous generation of AMD Instinct GPU compute products.[4]
- AMD CDNA 2 architecture and the Instinct MI210's new Matrix Cores deliver up to 3.9× the peak theoretical FP64 Matrix performance vs. peak theoretical FP64 of previous-generation AMD Instinct GPU compute products.[2]
- The Instinct MI210, based on AMD Matrix Core Technology, provides an expanded range of mixed-precision capabilities to accelerate deep learning training. Expect an outstanding 181 teraflops peak theoretical FP16 and BF16 performance for a platform that fuels the convergence of HPC and AI.

#2 LEVERAGE HIGHER LEVELS OF EFFICIENCY

- Get more performance per core. The Instinct MI210 achieves FP64 performance improvements with 1,024 fewer stream processors.[5]
- The Instinct MI210 scales memory and performance without compromise, doubling the memory capacity and FP64 performance in the same power envelope compared to previous generations, to achieve outstanding performance and efficiency for HPC and AI deployments.[2,4]

#3 RELY ON AMD'S CONSISTENT ROADMAP EXECUTION

- The Instinct MI210 is designed with the same architecture used to power the next generation of supercomputers. Visit AMD.com/Exascale to learn more.
- AMD's innovations in architecture, packaging and integration push the boundaries of advanced computing by unifying the CPU and the GPU accelerator.
- The Instinct MI210 represents continued value and development of AMD Instinct accelerators, now in their third generation.

#4 ADAPT YOUR SYSTEMS TO SCALE, PURPOSE AND WORKLOADS AS NEEDED

- AMD Instinct MI210 provides key support for a unified computing platform in the data center that can scale from single-server solutions up to the world's largest supercomputers for multiple purposes and workloads.
- 3rd Gen AMD Infinity Fabric technology also brings advanced platform connectivity and scalability, enabling fully connected dual and quad peer-to-peer (P2P) GPU hives through three AMD Infinity Fabric links per card. This delivers up to 300 GB/s (dual) and 600 GB/s (quad) of total aggregate peak P2P theoretical I/O bandwidth within hives.[6]
- Take a test drive today on the private AMD Accelerator Cloud (AAC) environment to help ensure AMD Instinct MI210 accelerators are a fit for your workloads.

#5 FOCUS ON YOUR GOALS, NOT THE SILICON

- AMD ROCm 5 is the first open-source software development platform for HPC/hyperscale-class GPU computing. It brings the UNIX philosophy of choice, minimalism and modular software development to GPU computing at the device level.
- ROCm 5's open ecosystem offers portability to support heterogenous environments with multiple GPU vendors and architectures.
- ROCm offers upstream support of key industry frameworks such as TensorFlow, PyTorch and ONNX-RT.
- ROCm 5 gives developers an acceleration continuum from the data center to the desktop with workstation support for AMD Radeon™ Pro W6800 GPUs.
- AMD Infinity Hub provides researchers, data scientists and end users a quick and straightforward way to find, download and install optimized containers for HPC applications and machine learning frameworks supported on AMD Instinct MI200 series accelerators and ROCm 5.

[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 83.0 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 Feb 15, 2022, for the AMD Instinct MI210 (64GB HBM2e PCIe card) 300-watt 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), 22.6 TFLOPS peak theoretical single precision (FP32), 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), floating-point performance. AMD TFLOPS calculations conducted with the following equation for AMD Instinct MI210 and MI100 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 MI210 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, 128 FLOPS per clock/CU for FP32 to determine TFLOPS, 1024 FLOPS per clock/CU for BF16 to determine TFLOPS. Divide results by 100,000 to get TFLOPS. Then, for MI100 that number is multiplied by 64 FLOPS per clock/CU for FP64 to determine TFLOPS. Divide results by 100,000 to get TFLOPS. Published results on the Nvidia Ampere A100 (80GB) 300 Watt PCIe 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), 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.

- MI210 45.3 TFLOPS FP64 Matrix divided by 300 Watts = 0.151 TFLOPS (151 GFLOPS) per watt
- MI210 22.6 TFLOPS FP64 divided by 300 Watts = 0.0753 TFLOPS (75 GFLOPS) per watt
- MI210 22.6 TFLOPS FP32 divided by 300 Watts = 0.0753 TFLOPS (75 GFLOPS) per watt
- MI210 181.0 TFLOPS BF16 divided by 300 Watts = 0.6033 TFLOPS (603 GFLOPS) per watt
- A100 19.5 TFLOPS FP64 Tensor Core divided by 300W = .065 TFLOPS (65 GFLOPS) per watt
- A100 9.7 TFLOPS FP64 divided by 300W = .0323 TFLOPS (32 GFLOPS) per watt
- A100 19.5 TFLOPS FP32 divided by 300W = .065 TFLOPS (65 GFLOPS) per watt
- A100 39 TFLOPS BF16 divided by 300W = .13 TFLOPS (130 GFLOPS) per watt

- .151/.065 = 2.3x the/1.3x better peak perf/watt (FP64 Matrix)
- .0753/.0323 = 2.3x the /1.3x better peak perf/watt (FP64)
- .0753/.065 = 1.15x the/0.15x better peak perf/watt (FP32)
- 0.6033/.13 = 4.6x the/3.6 better peak perf/watt (BF16)

MI200-44

4 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

5 The AMD Instinct™ MI210 PCIe accelerator has 104 compute units (CUs) and 6,656 stream cores. The AMD Instinct™ MI100 accelerator has 120 compute units (CUs) and 7,680 stream cores. Calculations conducted by AMD Performance Labs as of Feb 15, 2022, for the AMD Instinct™ MI210 (64GB HBM2e PCIe card) 300 Watt 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), 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), floating-point performance. AMD TFLOPS calculations conducted with the following equation for AMD Instinct MI210 and MI100 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 MI210 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. Divide results by 100,000 to get TFLOPS. Then, for MI100 that number is multiplied by 64 FLOPS per clock/CU for FP64 to determine TFLOPS. Divide results by 100,000 to get TFLOPS.

- MI210 45.3 TFLOPS FP64 Matrix divided by 104 CUs = 0.436 TFLOPS (436 GFLOPS) per CU
- MI210 22.6 TFLOPS FP64 divided by 104 CUs = 0.217 TFLOPS (217 GFLOPS) per CU
- MI100 11.54 TFLOPS FP64 divided by 120 CUs = 0.096 TFLOPS (96 GFLOPS) per CU

- 0.436/0.096 = 4.5x the/3.5x better peak perf/CU (FP64 Matrix)
- 0.217/0.096 = 2.3x the /1.3x better peak perf/CU (FP64)

MI200-045

6 Calculations as of Jan 27, 2022. AMD Instinct MI210 built on AMD CDNA2 technology accelerators support PCIe Gen4 providing up to 64 GB/s peak theoretical data bandwidth from CPU to GPU per card. AMD Instinct MI210 CDNA2 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. MI200-43