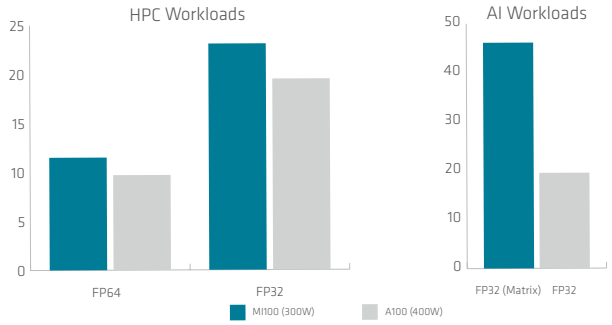
AMD INSTINCT™ MI100 ACCELERATOR

World's Fastest HPC GPU[1]

The era of exascale is here. Immense computational power coupled with the fusion of HPC and AI is enabling researchers to tackle grand challenges once thought beyond reach.

AMD Instinct™ MI100 accelerator is the world's fastest HPC GPU, engineered from the ground up for this new era of computing.[1] Powered by the first AMD CDNA architecture, the MI100 accelerators deliver a giant leap in compute and interconnect performance, offering a nearly 3.5x (FP32 Matrix) performance boost for HPC and a nearly 7x (FP16) performance boost for AI throughput compared to AMD's prior generation accelerators.[2]

Superior Performance for HPC & AI (Peak TFLOPS)



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

Heterogenous Computing Reimagined for the Exascale Era

HPC and AI are at the dawn of a new era. Disruptive technologies are needed to drive industries forward, and AMD is at the center of this computing revolution. The AMD Instinct™ MI100 accelerator has been designed in lock-step with AMD's award winning 2nd Gen AMD EPYC™ processors, built on our Infinity Architecture, to deliver true heterogeneous compute capabilities for HPC and AI. Combine these innovations with our partners' system offerings and the open and portable AMD ROCm™ programming ecosystem, and you gain access to a powerful computing solution that can meet your biggest challenges in HPC and AL.

Key Features

PERFORMANCE

Compute Units	120
Stream Processors	7,680
Peak BFLOAT16	Up to 92.3 TFLOPS
Peak INT4 | INT8	Up to 184.6 TOPS
Peak FP16	Up to 184.6 TFLOPS
Peak FP32 Matrix	Up to 46.1 TFLOPS
Peak FP32	Up to 23.1 TFLOPS
Peak FP64	Up to 11.5 TFLOPS
Bus Interface	PCIe® Gen 3 and Gen 4 Support[3]

MEMORY

Memory Size	32GB HBM2
Memory Interface	4,096Bits
Memory Clock	1.2 GHz
Memory Bandwidth	Up to 1.2 TB/s

RELIABILITY

ECC (Full-chip)	Yes[4]
RAS Support	Yes[5]

SCALABILITY

Infinity Fabric™ Links	3
OS Support	Linux® 64-bit
AMD ROCm™ Compatible	Yes

BOARD DESIGN

Board Form Factor	Full-Height, Dual Slot
Length	10.5" Long
Thermal	Passively Cooled
Max Power	300W TDP
Warranty	Three Year Limited[6]

CHOICE – Code Once, Use It Everywhere

The AMD ROCm™ ecosystem provides a software platform that is open-source, portable and accessible for HPC and machine learning accelerated compute. The AMD ROCm platform brings developers and customers programing choice, minimalism, and a modular software development environment designed to maximize developer's productivity when working on accelerated workloads.

HPC and MACHINE LEARNING APPLICATIONS


Cloud / Hyperscale


Financial Services


Energy


Reinforcement Learning


Life Sciences


Automotive


HPC


Image | Object | Video Detection & Classification

OPEN PROGRAMING WITH CHOICE

OpenMP | HIP | OpenCL™ | Python

OPTIMIZED LIBRARIES

MIOpen | FFT, RNG | BLAS, SPARSE | Eigen

OPEN FRAMEWORKS

PyTorch | TensorFlow | Kokkos | RAJA

PROGRAMER AND SYSTEM TOOLS

Debuggers | Performance Analysis | System Management

All-New Matrix Core Technology for Machine Learning

The AMD Instinct™ MI100 GPU brings customers all-new Matrix Core Technology with superior performance for 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 machine learning workloads you need to deploy. The MI100 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 MI100 handles large data efficiently for training complex neural networks used in deep learning and delivers a nearly 7x boost for AI (FP16) performance compared to AMD's prior generation accelerators.[2]

Ultra-Fast HBM2 Memory

The AMD Instinct™ MI100 GPU provides 32GB High-bandwidth HBM2 memory at a clock rate of 1.2 GHz and delivers an ultra-high ~1.2 TB/s of memory bandwidth to support your largest data sets and help eliminate bottlenecks in moving data in and out of memory. Combine this performance with the MI100's advanced I/O capabilities and you can push workloads closer to their full potential.[8]

For More Information Visit:

amd.com/INSTINCTMI100

AMD Infinity Fabric™ Link Technology

AMD Instinct™ MI100 GPUs provide advanced I/O capabilities in standard off-the-shelf servers with our Infinity Fabric™ technologies and PCIe® Gen4 support. The MI100 GPU delivers 64GB/s CPU to GPU bandwidth without the need for PCIe® switches, and up to 276 GB/s of peer-to-peer (P2P) bandwidth performance through three Infinity Fabric™ Links designed with AMD's 2nd Gen Infinity architecture.[7] AMD's Infinity technologies allow for platform designs with dual direct-connect quad GPU hives enabling superior P2P connectivity and delivering up to 1.1 TB/s of total theoretical GPU bandwidth within a server design.[7]

Industry's Latest PCIe® Gen 4.0

The AMD Instinct MI100 GPU is designed to support the latest PCIe Gen 4.0 technology which provides up to 64GB/s peak theoretical transport data bandwidth from CPU to GPU per card.

Leading FP64 Performance for HPC Workloads

The AMD Instinct™ MI100 GPU delivers industry-leading double precision performance with up to 11.5 TFLOPS peak FP64 performance, enabling scientists and researchers across the globe to more efficiently process HPC parallel codes across several industries including life sciences, energy, finance, academics, government, defense and more.[1]

---

1. 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 double precision (FP64), 46.1 TFLOPS peak single precision matrix (FP32), 23.1 TFLOPS peak single precision (FP32), 184.6 TFLOPS peak half precision (FP16) peak theoretical, floating-point performance. AMD TFLOPS calculations conducted with the following equation for AMD Instinct MI100 GPUs: FLOPS calculations are performed by taking the peak boost 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, that number is multiplied by 2 FLOPS per clock for FP32, 4 FLOPS per clock for FP16, to determine TFLOPS. The FP64 TFLOPS rate for MI100 is calculated using 1/2 rate per clock to calculate the FP64/FMA64 TFLOPS. External results on the NVidia Ampere A100 (40GB) GPU accelerator resulted in 9.7 TFLOPS peak double precision (FP64). 19.5 TFLOPS peak single precision (FP32), 78 TFLOPS peak half precision (FP16) theoretical, floating-point performance. Results found at: https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf MI100-03. **2**. Measurements conducted by AMD Performance Labs as of Sep 18, 2020 on the AMD Instinct™ MI100 accelerator designed with AMD CDNA 7nm FinFET process technology with 1,502 MHz boost engine clock resulted in 184.57 TFLOPS peak half precision (FP16), 46.14 TFLOPS peak single precision (FP32 Matrix) floating-point performance. The results calculated for Radeon Instinct™ MI50 GPU designed with Vega 7nm FinFET process technology with 1,725 MHz engine clock resulted in 26.5 TFLOPS peak half precision (FP16), 13.25 TFLOPS peak single precision (FP32 Matrix) floating-point performance. MI100-04. **3**.Works with PCIe® Gen 4.0 and Gen 3.0 compliant motherboards. Performance may vary from motherboard to motherboard. Refer to system or motherboard provider for individual product performance and features. **4**. ECC support on AMD Instinct™ GPU cards, based on the "AMD CDNA" technology includes full-chip ECC including HBM2 memory and internal GPU structures. **5**. Expanded RAS (Reliability, availability and serviceability) attributes have been added to the AMDs Instinct™ "AMD CDNA" technology-based GPU cards and their supporting ecosystem including software, firmware and system level features. AMD's remote manageability capabilities using advanced out-of-band circuitry allow for easier GPU monitoring via I2C, regardless of the GPU state. For full system RAS capabilities, refer to the system manufacturer's guidelines for specific system models. **6**. 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. **7**. As of SEP 18th, 2020. 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. Peak theoretical transport rate performance is calculated by Baud Rate * width in bytes *# directions = GB/s per card. PCIe Gen4: 16 * 2 * 2 = 64 GB/s. 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) transport rate bandwidth performance per GPU card. Combined with PCIe Gen4 compatibility providing an aggregate GPU card I/O peak bandwidth of up to 340 GB/s. Infinity Fabric link technology peak theoretical transport rate performance is calculated by Baud Rate * width in bytes *# directions *# links = GB/s per card. Infinity Fabric Link: 23 * 2 * 2 = 92 GB/s. MI100s have three links: 92 GB/s * 3 links per GPU = 276 GB/s. Four GPU hives provide up to 552 GB/s peak theoretical P2P performance: 276GB/s * 4 GPUs = 1,104 GB/s, divided by 2 = 552 GB/s. Dual 4 GPU hives in a server provide up to 1.1 TB/s total peak theoretical direct P2P performance per server: 552 GB/s per 4 GPU hive * 2 = 1,104 GB/s. AMD Infinity Fabric link technology not enabled: Four GPU hives provide up to 256 GB/s peak theoretical P2P performance with PCIe® 4.0: 64GB/s per GPU * 4 GPUs = 256 GB/s. Radeon Instinct™ MI50 "Vega 7nm" technology-based accelerators support PCIe® Gen 4.0* providing up to 64 GB/s peak theoretical transport data bandwidth from CPU to GPU per card. Radeon Instinct™ MI50 "Vega 7nm" technology-based accelerators include dual Infinity Fabric™ Links providing up to 184 GB/s peak theoretical GPU to GPU or Peer-to-Peer (P2P) transport rate bandwidth performance per GPU card. Combined with PCIe Gen 4 compatibility providing an aggregate GPU card I/O peak bandwidth of up to 248 GB/s. MI50 based four GPU hives provide up to 368 GB/s peak theoretical P2P performance: 184 GB/s * 4 GPUs = 736 GB/s, divided by 2 = 368 GB/s. Dual 4 GPU hives in a server provide up to 736 GB/s total peak theoretical direct P2P performance per server: 368 GB/s per 4 GPU hive * 2 = 736 GB/s.Performance guidelines are estimated only and may vary. Refer to server manufacture PCIe Gen4 compatibility and performance guidelines for potential peak performance of the specified server model numbers. Server manufacturers may vary configuration offerings yielding different results. https://pcisig.com/ , https://www.chipestimate.com/PCI-Ex- press-Gen-4-a-Big-Pipe-for-Big-Data/Cadence/Technical-Article/2014/04/15, https://www.tomshardware.com/news/pcie-4.0-power-speed-express,32525.html AMD has not independently tested or verified external/third party results/data and bears no responsibility for any errors or omissions therein. MI100-06. **8**. Calculations by AMD Performance Labs as of OCT 5th, 2020 for the AMD Instinct™ MI100 accelerator designed with AMD CDNA 7nm FinFET process technology at 1,200 MHz peak memory clock resulted in 1.2288 TFLOPS peak theoretical memory bandwidth performance. MI100 memory bus interface is 4,096 bits and memory data rate is 2.40 Gbps. The results calculated for Radeon Instinct™ MI50 GPU designed with "Vega" 7nm FinFET process technology with 1,000 MHz peak memory clock resulted in 1.024 TFLOPS peak theoretical memory bandwidth performance. MI50 memory bus interface is 4,096 bits and memory data rate is 2.00 Gbps. AMD memory bandwidth TFLOPs calculations, (x.xx Gbps memory data rate * xxxx bits memory bus interface memory bus interface)/8. MI100 memory bandwidth = (2.40 Gbps*4,096 bits)/8 = 1.2288 TB/s. MI50 memory bandwidth = (2.00 Gbps*4,096 bits)/8 = 1.024 TB/s. MI100 1.2288/ MI50 1.024 = 1.2x (or 20% higher). CDNA-04. **9**. The AMD Instinct™ MI100 accelerator has 120 compute units (CUs) and 7,680 stream cores. The Radeon Instinct™ MI50 GPU has 60 CUs and 3,840 stream cores. MI100 120 CUs divided by MI50 60 CUs = 2x CUs CDNA-02. **10**. Calculations by AMD Performance Labs as of Oct 5, 2020 using the AMD Instinct™ MI100 accelerator with CDNA 7nm FinFet process technology at 1,502 MHz peak boost engine clock resulted in 11.535 TFLOPS peak theoretical double precision (FP64) floating-point performance. The results calculated for Radeon Instinct™ MI50 GPU designed with "Vega" 7nm FinFET process technology with 1,725 MHz peak engine clock resulted in 6.62 TFLOPS peak theoretical double precision (FP64) floating-point performance. CDNA-03

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.