AMD

AMD Announces World's Fastest HPC Accelerator for Scientific Research[1]

– AMD Instinct™ MI100 accelerators revolutionize high-performance computing (HPC) and AI with industry-leading compute performance –

– First GPU accelerator with new AMD CDNA architecture engineered for the exascale era –

SANTA CLARA, Calif., Nov. 16, 2020 (GLOBE NEWSWIRE) -- [AMD](#) (NASDAQ: AMD) today announced the new AMD Instinct™ MI100 accelerator – the world's fastest HPC GPU and the first x86 server GPU to surpass the 10 teraflops (FP64) performance barrier.[1] Supported by new accelerated compute platforms from Dell, Gigabyte, HPE, and Supermicro, the MI100, combined with AMD EPYC™ CPUs and the ROCm™ 4.0 open software platform, is designed to propel new discoveries ahead of the exascale era.

Built on the new AMD CDNA architecture, the AMD Instinct MI100 GPU enables a new class of accelerated systems for HPC and AI when paired with 2nd Gen AMD EPYC processors. The MI100 offers up to 11.5 TFLOPS of peak FP64 performance for HPC and up to 46.1 TFLOPS peak FP32 Matrix performance for AI and machine learning workloads[2]. With new AMD Matrix Core technology, the MI100 also delivers a nearly 7x boost in FP16 theoretical peak floating point performance for AI training workloads compared to AMD's prior generation accelerators.[3]

"Today AMD takes a major step forward in the journey toward exascale computing as we unveil the AMD Instinct MI100 – the world's fastest HPC GPU," said Brad McCredie, corporate vice president, Data Center GPU and Accelerated Processing, AMD. "Squarely targeted toward the workloads that matter in scientific computing, our latest accelerator, when combined with the AMD ROCm open software platform, is designed to provide scientists and researchers a superior foundation for their work in HPC."

Open Software Platform for the Exascale Era

The AMD ROCm developer software provides the foundation for exascale computing. As an open source toolset consisting of compilers, programming APIs and libraries, ROCm is used by exascale software developers to create high performance applications. ROCm 4.0 has been optimized to deliver performance at scale for MI100-based systems. ROCm 4.0 has upgraded the compiler to be open source and unified to support both OpenMP® 5.0 and HIP. PyTorch and Tensorflow frameworks, which have been optimized with ROCm 4.0, can now achieve higher performance with MI100[7,8]. ROCm 4.0 is the latest offering for HPC, ML and AI application developers which allows them to create performance portable software.

"We've received early access to the MI100 accelerator, and the preliminary results are very encouraging. We've typically seen significant performance boosts, up to 2-3x compared to

other GPUs," said Bronson Messer, director of science, Oak Ridge Leadership Computing Facility. "What's also important to recognize is the impact software has on performance. The fact that the ROCm open software platform and HIP developer tool are open source and work on a variety of platforms, it is something that we have been absolutely almost obsessed with since we fielded the very first hybrid CPU/GPU system."

Key capabilities and features of the AMD Instinct MI100 accelerator include:

- **All-New AMD CDNA Architecture-** Engineered to power AMD GPUs for the exascale era and at the heart of the MI100 accelerator, the AMD CDNA architecture offers exceptional performance and power efficiency
- **Leading FP64 and FP32 Performance for HPC Workloads** - Delivers industry leading 11.5 TFLOPS peak FP64 performance and 23.1 TFLOPS peak FP32 performance, enabling scientists and researchers across the globe to accelerate discoveries in industries including life sciences, energy, finance, academics, government, defense and more.[1]
- **All-New Matrix Core Technology for HPC and AI** – Supercharged performance for a full range of single and mixed precision matrix operations, such as FP32, FP16, bFloat16, Int8 and Int4, engineered to boost the convergence of HPC and AI.
- **2nd Gen AMD Infinity Fabric™ Technology** – Instinct MI100 provides ~2x the peer-to-peer (P2P) peak I/O bandwidth over PCIe® 4.0 with up to 340 GB/s of aggregate bandwidth per card with three AMD Infinity Fabric™ Links.[4] In a server, MI100 GPUs can be configured with up to two fully-connected quad GPU hives, each providing up to 552 GB/s of P2P I/O bandwidth for fast data sharing.[4]
- **Ultra-Fast HBM2 Memory**– Features 32GB High-bandwidth HBM2 memory at a clock rate of 1.2 GHz and delivers an ultra-high 1.23 TB/s of memory bandwidth to support large data sets and help eliminate bottlenecks in moving data in and out of memory.[5]
- **Support for Industry's Latest PCIe® Gen 4.0** – Designed with the latest PCIe Gen 4.0 technology support providing up to 64GB/s peak theoretical transport data bandwidth from CPU to GPU.[6]

Available Server Solutions
The AMD Instinct MI100 accelerators are expected by end of the year in systems from major OEM and ODM partners in the enterprise markets, including:

Dell
"Dell EMC PowerEdge servers will support the new AMD Instinct MI100, which will enable faster insights from data. This would help our customers achieve more robust and efficient HPC and AI results rapidly," said Ravi Pendekanti, senior vice president, PowerEdge Servers, Dell Technologies. "AMD has been a valued partner in our support for advancing innovation in the data center. The high-performance capabilities of AMD Instinct accelerators are a natural fit for our PowerEdge server AI & HPC portfolio."

Gigabyte
"We're pleased to again work with AMD as a strategic partner offering customers server hardware for high performance computing," said Alan Chen, assistant vice president in NCBU, GIGABYTE. "AMD Instinct MI100 accelerators represent the next level of high-performance computing in the data center, bringing greater connectivity and data bandwidth for energy research, molecular dynamics, and deep learning training. As a new accelerator

in the GIGABYTE portfolio, our customers can look to benefit from improved performance across a range of scientific and industrial HPC workloads."

Hewlett Packard Enterprise (HPE)
"Customers use HPE Apollo systems for purpose-built capabilities and performance to tackle a range of complex, data-intensive workloads across high-performance computing (HPC), deep learning and analytics," said Bill Mannel, vice president and general manager, HPC at HPE. "With the introduction of the new HPE Apollo 6500 Gen10 Plus system, we are further advancing our portfolio to improve workload performance by supporting the new AMD Instinct MI100 accelerator, which enables greater connectivity and data processing, alongside the 2nd Gen AMD EPYC™ processor. We look forward to continuing our collaboration with AMD to expand our offerings with its latest CPUs and accelerators."

Supermicro
"We're excited that AMD is making a big impact in high-performance computing with AMD Instinct MI100 GPU accelerators," said Vik Malyala, senior vice president, field application engineering and business development, Supermicro. "With the combination of the compute power gained with the new CDNA architecture, along with the high memory and GPU peer-to-peer bandwidth the MI100 brings, our customers will get access to great solutions that will meet their accelerated compute requirements and critical enterprise workloads. The AMD Instinct MI100 will be a great addition for our multi-GPU servers and our extensive portfolio of high-performance systems and server building block solutions."

MI100 Specifications

Compute Units	Stream Processors	FP64 TFLOPS (Peak)	FP32 TFLOPS (Peak)	FP32 Matrix TFLOPS (Peak)	FP16/FP16 Matrix TFLOPS (Peak)	INT4 | INT8 TOPS (Peak)	bFloat16 TFLOPS (Peak)	HBM2 ECC Memory	Memory Bandwidth
120	7680	Up to 11.5	Up to 23.1	Up to 46.1	Up to 184.6	Up to 184.6	Up to 92.3 TFLOPS	32GB	Up to 1.23 TB/s

Supporting Resources

- Learn more about AMD Instinct™ Accelerators

- Learn more about AMD HPC Solutions

- AMD HPC Solutions Hub

- Learn more about AMD CDNA

- Learn more about the AMD 2nd Gen EPYC™ Processor

- Become a fan of AMD on Facebook

- Follow AMD on Twitter

About AMD
For more than 50 years AMD has driven innovation in high-performance computing, graphics and visualization technologies — the building blocks for gaming, immersive platforms and the data center. Hundreds of millions of consumers, leading Fortune 500 businesses and cutting-edge scientific research facilities around the world rely on AMD technology daily to improve how they live, work and play. AMD employees around the world are focused on building great products that push the boundaries of what is possible. For more information

about how AMD is enabling today and inspiring tomorrow, visit the AMD (NASDAQ: AMD) website, blog, Facebook and Twitter pages.

CAUTIONARY STATEMENT
This press release contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) such as the features, functionality, performance, availability, timing and expected benefits of AMD products including the AMD Instinct™ MI100 accelerator, 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 press release are based on current beliefs, assumptions and expectations, speak only as of the date of this press release 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. Material factors that could cause actual results to differ materially from current expectations include, without limitation, the following: Intel Corporation's dominance of the microprocessor market and its aggressive business practices; the ability of third party manufacturers to manufacture AMD's products on a timely basis in sufficient quantities and using competitive technologies; expected manufacturing yields for AMD's products; the availability of essential equipment, materials or manufacturing processes; AMD's ability to introduce products on a timely basis with features and performance levels that provide value to its customers; global economic uncertainty; the loss of a significant customer; AMD's ability to generate revenue from its semi-custom SoC products; the impact of the COVID-19 pandemic on AMD's business, financial condition and results of operations; political, legal, economic risks and natural disasters; the impact of government actions and regulations such as export administration regulations, tariffs and trade protection measures; the impact of acquisitions, joint ventures and/or investments on AMD's business, including the announced acquisition of Xilinx, and the failure to integrate acquired businesses; AMD's ability to complete the Xilinx merger; the impact of the announcement and pendency of the Xilinx merger on AMD's business; potential security vulnerabilities; potential IT outages, data loss, data breaches and cyber-attacks; uncertainties involving the ordering and shipment of AMD's products; quarterly and seasonal sales patterns; the restrictions imposed by agreements governing AMD's notes and the revolving credit facility; the competitive markets in which AMD's products are sold; market conditions of the industries in which AMD products are sold; AMD's reliance on third-party intellectual property to design and introduce new products in a timely manner; AMD's reliance on third-party companies for the design, manufacture and supply of motherboards, software and other computer platform components; AMD's reliance on Microsoft Corporation and other software vendors' support to design and develop software to run on AMD's products; AMD's reliance on third-party distributors and add-in-board partners; the potential dilutive effect if the 2.125% Convertible Senior Notes due 2026 are converted; future impairments of goodwill and technology license purchases; AMD's ability to attract and retain qualified personnel; AMD's ability to generate sufficient revenue and operating cash flow or obtain external financing for research and development or other strategic investments; AMD's indebtedness; AMD's ability to generate sufficient cash to service its debt obligations or meet its working capital requirements; AMD's ability to repurchase its outstanding debt in the event of a change of control; the cyclical nature of the semiconductor industry; the impact of modification or interruption of AMD's internal business processes and information systems; compatibility of AMD's products with some or all industry-standard software and hardware; costs related to defective products;

the efficiency of AMD's supply chain; AMD's ability to rely on third party supply-chain logistics functions; AMD's stock price volatility; worldwide political conditions; unfavorable currency exchange rate fluctuations; AMD's ability to effectively control the sales of its products on the gray market; AMD's ability to adequately protect its technology or other intellectual property; current and future claims and litigation; potential tax liabilities; and the impact of environmental laws, conflict minerals-related provisions and other laws or regulations. 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 Quarterly Report on Form 10-Q for the quarter ended September 26, 2020.

───────────────────────────────

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. Published 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. Server manufacturers may vary configuration offerings yielding different results. MI100-03

2. Calculations performed by AMD Performance Labs as of Sep 3, 2020 on the AMD Instinct™ MI100 (32GB HBM2 PCIe® card) accelerator at 1,502 MHz peak engine clock resulted in 46.1 TFLOPS peak theoretical single precision (FP32 Matrix) Math floating-point performance. The Nvidia Ampere A100 (40GB) GPU accelerator published results are 19.5 TFLOPS peak single precision (FP32) floating-point performance. Nvidia results found at: https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf. Server manufacturers may vary configuration offerings yielding different results. MI100-01

3. Calculations performed by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 accelerator at 1,502 MHz peak boost engine clock resulted in 184.57 TFLOPS peak theoretical half precision (FP16) and 46.14 TFLOPS peak theoretical single precision (FP32 Matrix) floating-point performance. The results calculated for Radeon Instinct™ MI50 GPU at 1,725 MHz peak engine clock resulted in 26.5 TFLOPS peak theoretical half precision (FP16) and 13.25 TFLOPS peak theoretical single precision (FP32 Matrix) floating-point performance. Server manufacturers may vary configuration offerings yielding different results. MI100-04

4. Calculations as of SEP 18th, 2020. AMD Instinct™ MI100 built on AMD CDNA technology accelerators supporting PCIe® Gen4 providing up to 64 GB/s peak theoretical transport data bandwidth from CPU to GPU per card. AMD Instinct™ MI100 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 support providing an aggregate GPU card

I/O peak bandwidth of up to 340 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. Dual 4 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: Four GPU hives provide up to 256 GB/s peak theoretical P2P performance with PCIe® 4.0. Server manufacturers may vary configuration offerings yielding different results. MI100-07

5. 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. 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. CDNA-04

6. 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.

7. Testing Conducted by AMD performance labs as of October 30th, 2020, on three platforms and software versions typical for the launch dates of the Radeon Instinct MI25 (2018), MI50 (2019) and AMD Instinct MI100 GPU (2020) running the benchmark application Quicksilver. MI100 platform (2020): Gigabyte G482-Z51-00 system comprised of Dual Socket AMD EPYC™ 7702 64-Core Processor, AMD Instinct™ MI100 GPU, ROCm™ 3.10 driver, 512GB DDR4, RHEL 8.2. MI50 platform (2019): Supermicro® SYS-4029GP-TRT2 system comprised of Dual Socket Intel Xeon® Gold® 6132, Radeon Instinct™ MI50 GPU, ROCm 2.10 driver, 256 GB DDR4, SLES15SP1. MI25 platform (2018): Supermicro SYS-4028GR-TR2 system comprised of Dual Socket Intel Xeon CPU E5-2690, Radeon Instinct™ MI25 GPU, ROCm 2.0.89 driver, 246GB DDR4 system memory, Ubuntu 16.04.5 LTS. MI100-14

8. Testing Conducted by AMD performance labs as of October 30th, 2020, on three platforms and software versions typical for the launch dates of the Radeon Instinct MI25 (2018), MI50 (2019) and AMD Instinct MI100 GPU (2020) running the benchmark application TensorFlow ResNet 50 FP 16 batch size 128. MI100 platform (2020): Gigabyte G482-Z51-00 system comprised of Dual Socket AMD EPYC™ 7702 64-Core Processor, AMD Instinct™ MI100 GPU, ROCm™ 3.10 driver, 512GB DDR4, RHEL 8.2. MI50 platform (2019): Supermicro® SYS-4029GP-TRT2 system comprised of Dual Socket Intel Xeon® Gold® 6254, Radeon Instinct™ MI50 GPU, ROCm 3.0.6 driver, 338 GB DDR4, Ubuntu® 16.04.6 LTS. MI25 platform (2018): a Supermicro SYS-4028GR-TR2 system comprised of Dual Socket Intel Xeon CPU E5-2690, Radeon Instinct™ MI25 GPU, ROCm 2.0.89 driver, 246GB DDR4 system memory, Ubuntu 16.04.5 LTS. MI100-15

Contacts:
Gary Silcott
AMD Communications
+1 512-602-0889
Gary.Silcott@amd.com

Laura Graves
AMD Investor Relations

+1 408-749-5467
Laura.Graves@amd.com

AMD

Source: Advanced Micro Devices, Inc.