RADEON INSTINCT MI8

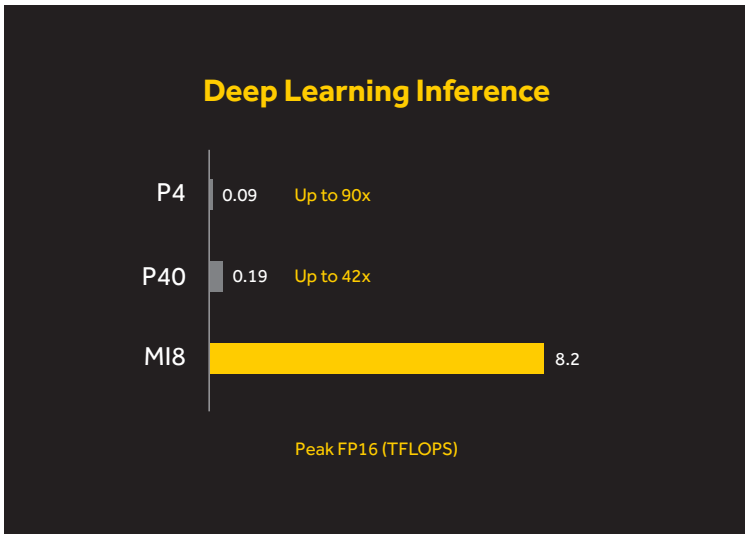
Open Ecosystem For Machine Intelligence

DEEP LEARNING INFERENCE ACCELERATOR WITH 8.2 TFLOPS SINGLE PRECISION COMPUTE PERFORMANCE IN EFFICIENT SCALE-OUT DESIGN[1]

Datacenter designers deploying Machine Learning and AI systems today require systems capable of running more complex workloads that are massively parallel in nature, while continuing to improve system efficiencies. Improvements in the capabilities of accelerators over the last decade, along with the advancement in software, are providing designers with the option to build more efficient heterogeneous computing systems to help them meet today's challenges.

The Radeon Instinct™ MI8 server accelerator is a highly efficient, cost-effective inference and HPC solution delivering 8.2 TFLOPS of peak performance with 4GB of ultrafast HBM1 memory, making it the perfect solution for running deep learning inference applications, where lots of new smaller data set inputs are being run at half or single precision against trained neural networks to discover new knowledge.[1] The MI8 accelerator is also the perfect open solution for general purpose HPC systems deployed in Financial, Energy, Life Science, Automotive, Academic (Research & Teaching), Government Labs and other

AMD's Radeon Instinct™ MI8 accelerator delivers superior half precision performance with 4GB of high-bandwidth HBM1 memory in an efficient, passively cooled, scale out accelerator form factor.[1]



Highlights

- 8.2 TFLOPS FP16 or FP32 Performance[1]
- Up To 47 GFLOPS Per Watt FP16 or FP32 Performance[2]
- 4GB HBM1 on 512-bit Memory Interface
- Passively Cooled Server Accelerator
- Large BAR Support for Multi GPU Peer to Peer
- ROCm Open Platform for HPC-Class Rack Scale
- Optimized MIOpen Libraries for Deep Learning
- MxGPU SR-IOV Hardware Virtualization

Key Features

GPU Architecture:	**AMD "Fiji"**
Stream Processors:	**4,096**
Performance:	
Half-Precision (FP16)	**8.2 TFLOPS**
Single-Precision (FP32)	**8.2 TFLOPS**
Double-Precision (FP64)	**512 GFLOPS**
GPU Memory:	**4GB HBM1**
Memory Bandwidth:	**Up to 512 GB/s**
Bus Interface:	**PCIe® Gen 3 x16**
MxGPU Capability:	**Yes**
Board Form Factor:	**Full-Height, Dual-Slot**
Length:	**6"**
Thermal Solution:	**Passively Cooled**
Standard Max Power:	**175W TDP**
Warranty:	**Three Year Limited[3]**
OS Support:	**Linux® 64-bit**
ROCm Software Platform:	**Yes**
Programing Environment:	
ISO C++, OpenCL™, CUDA (via AMD's HIP conversion tool) and Python[4] (via Anaconda's NUMBA)	

For more information, visit: Radeon.com/Instinct

RADEON INSTINCT MI8

AMD

OUTSTANDING SINGLE AND HALF PRECISION PERFORMANCE

The Radeon Instinct MI8 accelerator based on AMD's 3rd generation "Fiji" architecture with improved data-parallel processing and ultra-fast HBM1 memory delivers 8.2 TFLOPS of peak performance with up to 512 GB/s of memory bandwidth in a single, passively cooled GPU card.[1] The MI8 accelerator, combined with AMD's ROCm open software platform, is the perfect solution for cost sensitive system deployments for Machine Intelligence, Deep learning and HPC workloads, where performance and efficiency are key system requirements.

3RD GENERATION "FIJI" ARCHITECTURE

The Radeon Instinct™ MI8 server accelerator is based on the "Fiji" architecture which is built with AMD's 3rd Generation Graphics Core Next (GCN) packing 64 compute units (CU) with 64 stream processor per CU delivering 8.2 TFLOPS FP16 or FP32 compute performance in a single GPU card.[1]

HBM1: ULTRAFAST MEMORY BANDWIDTH

4GB of ultrafast HBM1 GPU memory delivering up to 512 GB/s of memory bandwidth. HBM1 is a modern type of memory design with low power consumption and ultra-wide communication lanes.

MxGPU SR-IOV HARDWARE VIRTUALIZATION

Design with support of AMD's MxGPU SR-IOV hardware virtualization technology for optimized datacenter cycle utilization, the Radeon Instinct MI8 provides a virtualization solution with dedicated user GPU resources, data security and version control, a cost effective licensing model with no additional hardware licensing fees, and a simplified native driver model ensuring operating system and application compatibility.

PASSIVELY COOLED

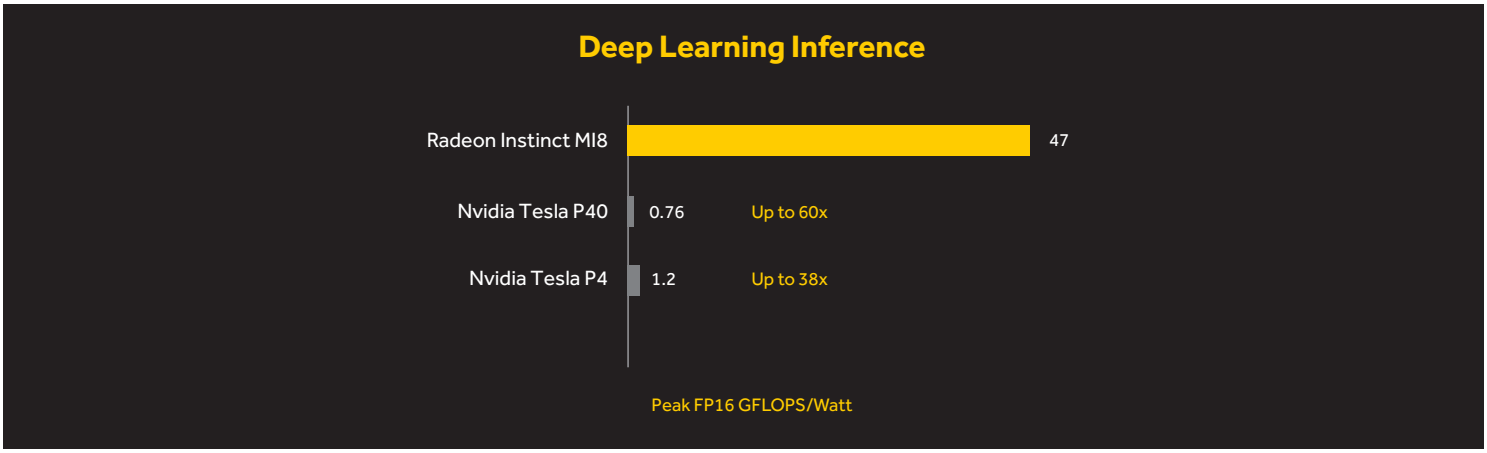
The Radeon Instinct MI8 design is a passively-cooled accelerator design for large-scale server deployments.

ROCm OPEN SOFTWARE PLATFORM

AMD's ROCm platform provides a scalable, fully open source software platform optimized for large-scale heterogeneous system deployments with an open source headless Linux driver, HCC compiler, rich runtime based on HSA, tools and libraries.

For more information, visit:
Radeon.com/Instinct
ROCm.github.io

SUPERIOR FP16 PERFORMANCE FOR INFERENCE[2]



FOOTNOTES

1. Measurements conducted by AMD Performance Labs as of June 2, 2017 on the Radeon Instinct™ MI8 "Fiji" architecture based accelerator. Results are estimates only and may vary. Performance may vary based on use of latest drivers. PC/system manufacturers may vary configurations yielding different results. The results calculated for MI8 resulted in 8.2 TFLOPS peak half precision (FP16) performance and 8.2 TFLOPS peak single precision (FP32) floating-point performance.

AMD TFLOPS calculations conducted with the following equation: 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, that number is multiplied by 2 FLOPS per clock for FP32. To calculate TFLOPS for FP16, 4 FLOPS per clock were used.
Measurements on the Nvidia Tesla P40 resulted in 0.19 TFLOPS peak half precision (FP16) peak floating-point performance with 250w TDP GPU card from external source.

Sources:
https://devblogs.nvidia.com/parallelforall/mixed-precision-programming-cuda-8/
http://images.nvidia.com/content/pdf/tesla/184427-Tesla-P40-Datasheet-NV-Final-Letter-Web.pdf
Measurements on the Nvidia Tesla P4 resulted in 0.09 TFLOPS peak half precision (FP16) floating-point performance with 75w TDP GPU card from external source.

Sources:
https://devblogs.nvidia.com/parallelforall/mixed-precision-programming-cuda-8/
http://images.nvidia.com/content/pdf/tesla/184457-Tesla-P4-Datasheet-NV-Final-Letter-Web.pdf
AMD has not independently tested or verified external and/or third party results/data and bears no responsibility for any errors or omissions therein.
RIF-1

2. Measurements conducted by AMD Performance Labs as of June 2, 2017 on the Radeon Instinct™ MI8 "Fiji" architecture based accelerator. Results are estimates only and may vary. Performance may vary based on use of latest drivers. PC/system manufacturers may vary configurations yielding different results. The results calculated for Radeon Instinct MI8 resulted in 47 GFLOPS/watt peak half precision (FP16) performance and 47 GFLOPS/watt peak single precision (FP32) floating-point performance.

AMD GFLOPS per watt calculations conducted with the following equation: 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, that number is multiplied by 2 FLOPS per clock for FP32. To calculate TFLOPS for FP16, 4 FLOPS per clock were used.
Once the TFLOPs are calculated, the number is divided by the 175w TDP power and multiplied by 1,000.
Measurements on the Nvidia Tesla P40 based on 0.19 TFLOPS peak FP16 with 250w TDP GPU card result in 0.76 GFLOPS/watt peak half precision (FP16) performance.

Sources for Nvidia Tesla P40 FP16 TFLOPs number:
https://devblogs.nvidia.com/parallelforall/mixed-precision-programming-cuda-8/
http://images.nvidia.com/content/pdf/tesla/184427-Tesla-P40-Datasheet-NV-Final-Letter-Web.pdf

Measurements on the Nvidia Tesla P4 based on 0.09 TFLOPS peak FP16 with 75w TDP GPU card result in 1.2 GFLOPS/watt peak half precision (FP16) performance.

Sources for Nvidia Tesla P40 FP16 TFLOPs number:
https://devblogs.nvidia.com/parallelforall/mixed-precision-programming-cuda-8/
http://images.nvidia.com/content/pdf/tesla/184457-Tesla-P4-Datasheet-NV-Final-Letter-Web.pdf

AMD has not independently tested or verified external and/or third party results/data and bears no responsibility for any errors or omissions therein.
RIF-2

3. The Radeon Instinct GPU accelerator products come with a three year limited warranty. Please visit www.AMD.com/warranty for details on the specific graphics products purchased. Toll-free phone service available in the U.S. and Canada only, email access is global.
4. Support for Python is planned, but still under development.