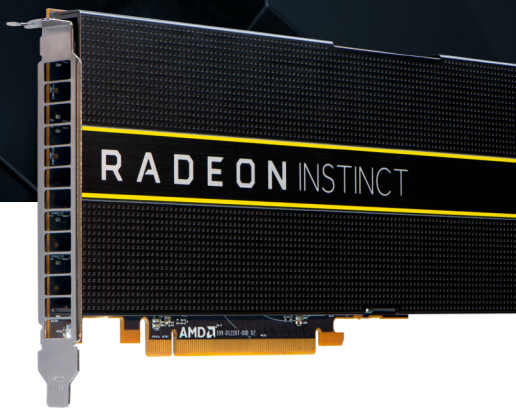
RADEON INSTINCT MI6

AMD

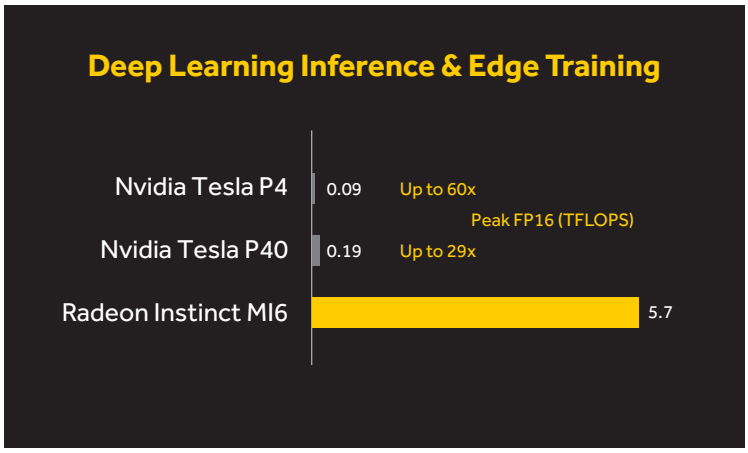


VERSATILE ACCELERATOR FOR MACHINE AND DEEP LEARNING EDGE-TRAINING AND INFERENCE APPLICATIONS

Today's compute workloads are more complex than ever before and require working with vast amounts of increasingly complex data. System designs need to be more open and to be more flexible to address these applications that have become more parallel in nature, while continuing to improve overall system efficiencies when deploying large systems. AMD's Radeon Instinct™ family of products have been distinctly designed and optimized to help address these growing needs with open heterogeneous computing solutions bringing about the next era of compute and machine intelligence.

The Radeon Instinct MI6 is a unique, versatile accelerator for deep learning edge-training and inference applications delivering 5.7 TFLOPS of peak half or single precision performance in a single-slot GPU card equipped with 16GB of ultra-fast memory at under 150 watts of TDP.[4] The Radeon Instinct MI6 accelerator's distinctive capabilities, combined with AMD's open ROCm software platform and MIOpen machine learning framework libraries, provide customers with a highly versatile, cost-sensitive compute solution for today's most demanding deep learning edge-training and inference applications.

AMD's Radeon Instinct™ MI6 delivers exceptional half precision performance with 16GB of memory in an efficient single-slot, passively cooled, scale out accelerator form factor.[4]



Highlights

- 5.7 TFLOPS FP16 or FP32 Performance[4]
- Up To 38 GFLOPS Per Watt Peak FP16 or FP32 Performance[5]
- 16GB Ultra-Fast GDDR5 Memory on 256-bit Memory Interface
- Passively Cooled Server Accelerator
- Large BAR Support for Multi-GPU Peer to Peer
- ROCm Open Platform for HPC-Class Scale Out
- Optimized MIOpen Libraries for Deep Learning
- MxGPU SR-IOV Hardware Virtualization

Key Features

GPU Architecture:	**AMD "Polaris"**
Stream Processors:	**2,304**
Performance:	
Half-Precision (FP16)	**5.7 TFLOPS**
Single-Precision (FP32)	**5.7 TFLOPS**
Double-Precision (FP64)	**358 GFLOPS**
GPU Memory:	**16GB GDDR5**
Memory Bandwidth:	**Up to 224 GB/s**
Bus Interface:	**PCIe® Gen 3 Compliant Motherboard[1]**
MxGPU Capability:	**Yes**
Board Form Factor:	**Full-Height, Single-Slot**
Length:	**9.5"**
Thermal Solution:	**Passively Cooled**
Standard Max Power:	**150W TDP**
Warranty:	**Three Year Limited[2]**
OS Support:	**Linux® 64-bit**
ROCm Software Platform:	**Yes**
Programing Environment:	
ISO C++, OpenCL™, CUDA (via AMD's HIP conversion tool) and Python[3] (via Anaconda's NUMBA)	

For more information, visit: Radeon.com/Instinct

RADEON
INSTINCT MI6

AMD

EXCEPTIONAL HALF OR SINGLE PRECISION PERFORMANCE WITH LARGE MEMORY

The Radeon Instinct MI6 accelerator based on AMD's new 4th generation "Polaris" architecture built on a 14nm FinFET process has exceptional data parallel processing capabilities and ultra-fast GDDR5 memory delivering 5.7 TFLOPS of peak performance with 16GB GDDR5 memory and up to 224 GB/s of memory bandwidth in a single, passively cooled GPU card.[4] The MI6 accelerator, combined with AMD's ROCm open software platform, is the perfect solution for efficiency and cost-sensitive inference and edge-training system deployments for Machine Intelligence and Deep learning, along with HPC workloads, where performance with large memory and efficiency are main system solution drivers.

GCN 4th GENERATION "POLARIS" ARCHITECTURE

The Radeon Instinct™ MI6 server accelerator is based on the new "Polaris" architecture which is built with AMD's 4th Generation Graphics Core Next (GCN) on 14nm FinFET process packing 36 compute units with 64 stream processor per CU delivering 5.7 TFLOPS FP16 or FP32 compute performance in a single GPU card.[4]

GDDR5: ULTRA-FAST MEMORY BANDWIDTH

16GB of ultra-fast GDDR5 GPU memory delivering up to 224 GB/s of memory bandwidth. GDDR5 helps accelerate applications and process computationally complex workflows, especially when working with large amounts of data.

MxGPU SR-IOV HARDWARE VIRTUALIZATION

Design with support of AMD's MxGPU SRI-OV hardware virtualization technology for optimized datacenter cycle utilization, the Radeon Instinct MI6 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

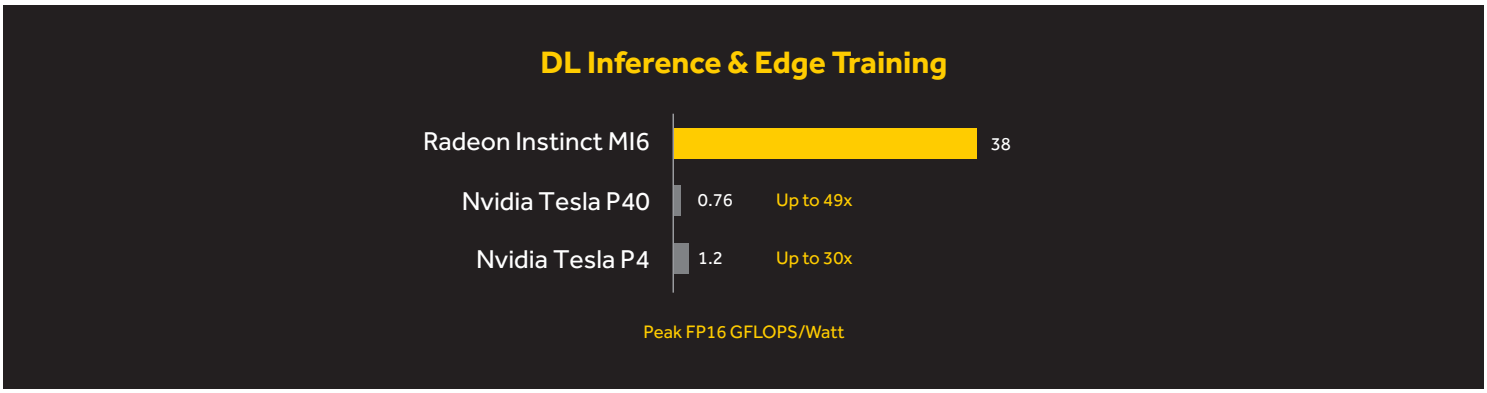
The Radeon Instinct MI6 design is a passively-cooled accelerator design for large-scale server deployments.

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 PER WATT IN SINGLE-SLOT GPU CARD[5]



AMD

FOOTNOTES

1. Works with PCIe 3.0 compliant motherboard.
2. 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.
3. Support for Python is planned, but still under development.
4. Measurements conducted by AMD Performance Labs as of June 2, 2017 on the Radeon Instinct™ MI6 "Polaris" 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 MI6 resulted in 5.7 TFLOPS peak half precision (FP16) performance and 5.7 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) floating-point performance with 250w TDP GPU card from external source.

Source:
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.

Source:
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. RIP-1

5. Measurements conducted by AMD Performance Labs as of June 2, 2017 on the Radeon Instinct™ MI6 "Polaris" 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 MI6 resulted in 38 GFLOPS/watt peak half precision (FP16) performance and 38 GFLOPS 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 150w 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:
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:
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/third party results/data and bears no responsibility for any errors or omissions therein. RIP-2