AMD's ROCm Software Helps Accelerate HPC & Deep Learning

A HETEROGENEOUS COMPUTE ACCELERATION PLATFORM DESIGNED FOR THE FUTURE

Executive Summary

Advanced Micro Devices (AMD) recently launched a software initiative called Radeon Open Compute Platform (ROCm) to help support GPUs in high performance computing (HPC) and deep learning applications. ROCm is an open source suite of drivers, tools, and libraries designed for a variety of programming models, including programs written to the NVIDIA CUDA proprietary programming interface. This paper examines AMD's HPC software strategy and the capabilities of AMD's product portfolio, and it makes recommendations for users considering gear for HPC and deep learning applications.

AMD has a history in HPC, traditionally in the CPU space where earlier generations of AMD Opteron CPUs were alternatives to Intel Xeon CPUs. AMD's GPUs have provided price / performance alternatives to NVIDIA for workstations, gaming, and virtual reality (VR). However, these markets are primarily Microsoft Windows-based, so naturally AMD's software stack had been optimized for the Microsoft Windows environment.

AMD has revamped its software stack for Linux HPC server applications to help them run on AMD GPU hardware. ROCm provides compilers, libraries, and development tools and an extended version of the Heterogeneous System Architecture (HSA) to provide a run-time execution environment for discrete GPUs. The result is designed for easier migration, favorable price / performance, and programmatic control of and orchestration of GPU resources and user processing queues.

AMD designed ROCm to enable a single source programming environment familiar to today's programmers for developing parallel applications that then can be compiled for CPUs, APUs, or discrete GPUs. With refreshed AMD GPU hardware and AMD's next generation x86 "Zen" processor core coming in 2017, ROCm merits evaluation for users seeking to improve performance while minimizing migration expenses to realize more cost-effective acceleration. AMD's approach also demonstrates a commitment to openness that could become a competitive differentiator in an environment that historically has been relatively closed and proprietary. While the ROCm software appears to establish a solid foundation, AMD will now need to nurture the ecosystem, launch more compute-centric GPUs, and work closely with leading customers to reenter the demanding HPC and deep learning markets.

ROCm: Open HPC for Heterogeneous Computation

Overview

AMD has offered GPU hardware since acquiring ATI in 2006, but in servers it has lagged NVIDIA on software tailored to Linux and HPC applications. So the AMD research team engaged with the HPC and CUDA developer community to understand how to ease the adoption of AMD hardware. As a result, AMD decided to transition from its historically closed source model to a fully open source model to help enable the integration of new products into customer workflows.

AMD's software stack, under the banner of the Radeon Open Compute Platform (ROCm), is the fruit of this effort and includes the following open source components:

- Server-focused, headless Linux driver for execution of HPC class applications
  - **ROCk**: New open source headless Linux driver available for most common Linux distributions, including Canonical / Ubuntu and Red Hat's Fedora with Red Hat Enterprise Linux (REHL) / Centos expected to follow
  - **ROCr**: System runtime libraries support HSA extensions to enable a unified memory address space between CPU and integrated or discrete GPU, supporting low-latency dispatch and data movement
- Development tools designed to help developers meet their application needs
  - **HCC**: High performance compute compiler, provides single source C, C++, and OpenMP environment for both CPU and GPU coding
  - **HIP**: Heterogeneous compute interface for portability, provides a bridge for CUDA codes, can be used to prepare code for subsequent compilation using AMD HCC or NVIDIA NVCC
  - OpenCL 1.2+ and Continuum Analytics' Anaconda
- Deep learning frameworks and optimized for fast-growing market for deep neural networks (DNN), mlOpen as well as common DNN frameworks
- Common HPC libraries and frameworks tuned to AMD GPUs

ROCk: Open Source Linux Kernel for HPC

AMD chose to build its first ROCk drivers for its newest products, the high bandwidth memory (HBM) equipped Fiji-based GPUs, which are available as a two-GPU PCIe card in the FirePro S9300 x2. Recently AMD released ROCm 1.2, adding GFX7 Hawaiian Islands family of double-precisions GPUs, the W9100, S9150, and S9170. These drivers will all share the same low-latency dispatch design characteristics needed to maximize performance in HPC and other parallel applications. Figure 1 shows

reduced latency with ROCk versus the Catalyst driver on Radeon R9 Nano, which was optimized for Windows not computational Linux workloads.

FIGURE 1: AMD ROCM IMPROVES LATENCY TO COMPUTE

(Source: AMD)

ROCR: HETEROGENEOUS SYSTEM ARCHITECTURE RUNTIME SUPPORT ON DISCRETE GPUS

ROCr is a system run-time execution environment. According to AMD, it provides base capabilities to the programmer though his favorite programing language as well as benefits to HPC users.

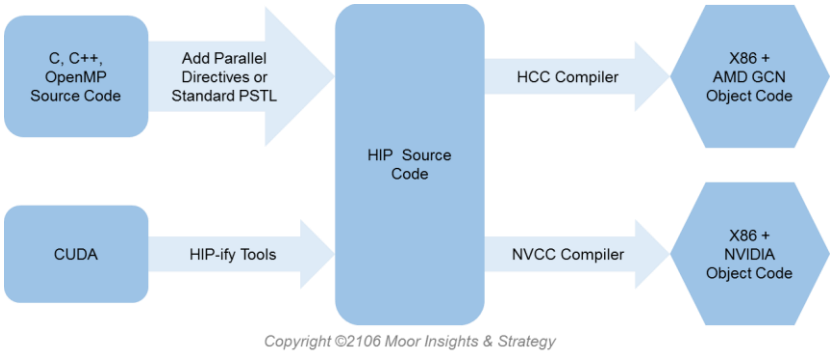
- Reduces latency to compute
  - User mode queues to enable low latency kernel dispatching on GPU
  - User mode DMA for low latency data transfers to the discrete GPU
- Increases GPU utilization and effectiveness
  - Provides process concurrency and preemption
  - Includes the HSA signals and atomic operations
- Provides native multi-GPU capabilities for scale up and scale out support
  - Multi-GPU memory management API
  - Multi-GPU coarse-grain shared virtual memory
  - Peer-to-peer multi-GPU dispatch and synchronization with RDMA
  - OpenUCX and HWLOC for OpenMPI, OpenSHEMM, and GASnet support

AMD HCC: Single Source Heterogeneous Compute Compiler

AMD's leadership in the HSA Foundation and its contributions to the US Department of Energy (DOE) Fast Forward program led the company to a simple thesis: Let programmers concentrate on their algorithms, and leave the code generation and optimization to the compiler, even for highly parallel code segments. Programmers can write code once, identify areas where parallel execution can be exploited, then let the compiler produce code optimized to the CPU, APU, or GPU for execution.

According to AMD, the Heterogeneous Compute Compiler allows the programmer to remain in his comfort zone of C, C++, and OpenMP using Clang and LLVM compiler technologies to handle the compilation. Programmers can identify and exploit parallelism via two mechanisms. The first is through an explicit directive syntax such as parallel_for_each, with control built around the C++ lambda construct. The second, which may become the preferred model, uses the Parallel Standard Templates Library (PSTL), which is expected to become an industry standard for C++ 17. AMD HCC gives users access to features designed to improve performance of virtually any accelerator, such as the ability to pre-fetch data into the accelerator's fast memory store, dispatching asynchronous compute kernels, and managing scratch-pad memories.

Figure 2: AMD Single Source Development Environment



(Source: Moor Insights & Strategy)

The AMD HCC compiler and HIP tools are designed to enable single source development regardless of the eventual execution target environment. Note that the x86 object code can run on Intel Xeons or AMD's future Zen-based processors.

HIP: Heterogeneous-Compute Interface for Portability

If AMD had stopped with a solid LINUX driver and a fast parallelizing compiler, it could have won fans looking for a competitive GPU platform. But AMD went further to provide a tool designed to automatically convert most NVIDIA CUDA codes to a higher-level

construct, called HIP, which can then be compiled with the HCC compiler. HIP is intended to allow programmers who prefer CUDA to stay in their comfort zone and continue to develop their code in CUDA. Developers can then compile their ported ("HIP-ified") code for execution on AMD GPUs, or according to AMD, compile it using the NVCC compiler to run on NVIDIA hardware.

OpenCL

AMD has long promoted OpenCL, and AMD plans to continue support for OpenCL for existing customers and those looking for broad industry portable programing language. AMD intends to extend OpenCL to support its ROCr runtime release with HSA extensions and support the common GCN ISA LLVM code generator. Thus OpenCL will support the broad optimizations that have been added to the LLVM native back end and will support assembler and disassembler as the other compilers do. This approach also supports off-line compilation and code caching for run-time dynamic compilation.

OpenCL 1.2+ will be a fully open source toolset moving forward and will support the following new features as part of the HSA+ environment.

- Coarse grain shared virtual memory between the CPU & GPU
- Generic address space
- C11 atomics
- OpenCL 2.0 images support
- User mode DMA: Dual engines with ASYNC transfer & user-mode queue support

Anaconda Python Acceleration with Numba from Continuum

Python has become a popular programming language in HPC, as it provides programmer productivity and a self-documenting coding style. AMD has worked with Continuum Analytics to bring Numba acceleration and integrate it into the native GCN compiler foundation. AMD also works with Anaconda to bring support for optimized math libraries to speed application such as Pandas. The updated Continuum Anaconda product provides a rich set of options to optimize for APUs, as well as AMD discrete GPUs, and supports asynchronous execution, shared memory, and implicit data transfer from DRAM to the faster discrete GPU memory.

ROCm Libraries

To improve performance—and in the open source spirit that is the basis for the ROCm platform—AMD is providing an initial set of math libraries, third-party frameworks, benchmarks, and applications on the ROCm portal. These include GPU Optimized Core

Math Libraries / Solvers & Software Primitives Libraries as well as third-party parallel programing frameworks such as Kokkos C++ and CHARM++. ROCm also includes a suite of popular applications for molecular dynamics, chroma dynamics, cosmology, bioinformatics, weather, and deep learning. Finally, AMD included a suite of optimized benchmarks such as SHOC, HPL, and Lulesh to simplify performance comparisons.

MLOPEN FOR MACHINE LEARNING ON AMD HARDWARE

AMD wants to offer GPU choices for accelerating the training of deep neural networks, a fast growing and computationally intensive workload. To help neural network researchers use AMD hardware acceleration for training DNNs, AMD provides the mlOpen solver for convolution networks. According to AMD, the impact of this solver for OpenCL CAFFE for CIFAR 10 is 9 times better performance than the current open source CAFFE distribution on the same hardware.

AMD's mlOpen currently provides:

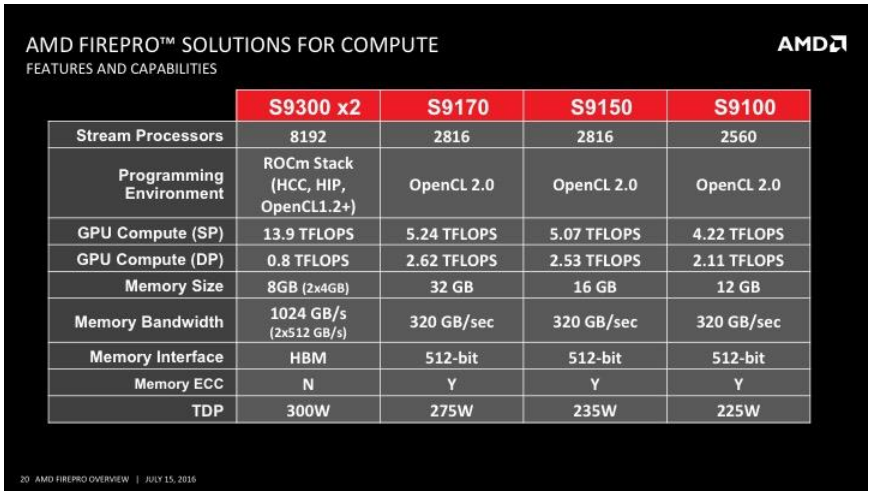
- Optimized solver for convolution based neural network
  - Direct convolution forward and backward propagation
  - Pooling (max, random, ave, total), normalization, dropout, softmax, entropy, log regression
  - Learning types: fixed, linear, exp_step, exp_inv, total
  - Neuron types: Passthru, logistic, TANH, RELU, BRELU, SOFTRELU, ABS, square, square root, linear, power
- Control over nodes in fully connected multilayer nets

AMD HARDWARE FOR COMPUTE-INTENSIVE WORKLOADS

AMD COMPUTE GPU OVERVIEW

AMD GPUs for compute workloads have traditionally focused on delivering more double-precision floating point and more memory than competitors at a lower price. AMD's new FirePro S9300 x2 card focuses on single-precision applications with high memory bandwidth, leveraging silicon from AMD Radeon gaming chips. Figure 3 summarizes the features and performance of current compute GPUs from AMD. Note the large number of stream processors, high SP (single-precision) performance, and memory bandwidth of the Fiji-based S9300 x2 product versus the high DP (double-precision) performance and memory capacity of the Hawaii-based S91xx products. AMD also supports ROCm on the Radeon R9 Nano cards. AMD has optimized each GPU for specific workloads.

FIGURE 3: AMD RADEON S SERIES COMPUTE GPUS



	S9300 x2	S9170	S9150	S9100
Stream Processors	8192	2816	2816	2560
Programming Environment	ROCm Stack (HCC, HIP, OpenCL1.2+)	OpenCL 2.0	OpenCL 2.0	OpenCL 2.0
GPU Compute (SP)	13.9 TFLOPS	5.24 TFLOPS	5.07 TFLOPS	4.22 TFLOPS
GPU Compute (DP)	0.8 TFLOPS	2.62 TFLOPS	2.53 TFLOPS	2.11 TFLOPS
Memory Size	8GB (2x4GB)	32 GB	16 GB	12 GB
Memory Bandwidth	1024 GB/s (2x512 GB/s)	320 GB/sec	320 GB/sec	320 GB/sec
Memory Interface	HBM	512-bit	512-bit	512-bit
Memory ECC	N	Y	Y	Y
TDP	300W	275W	235W	225W

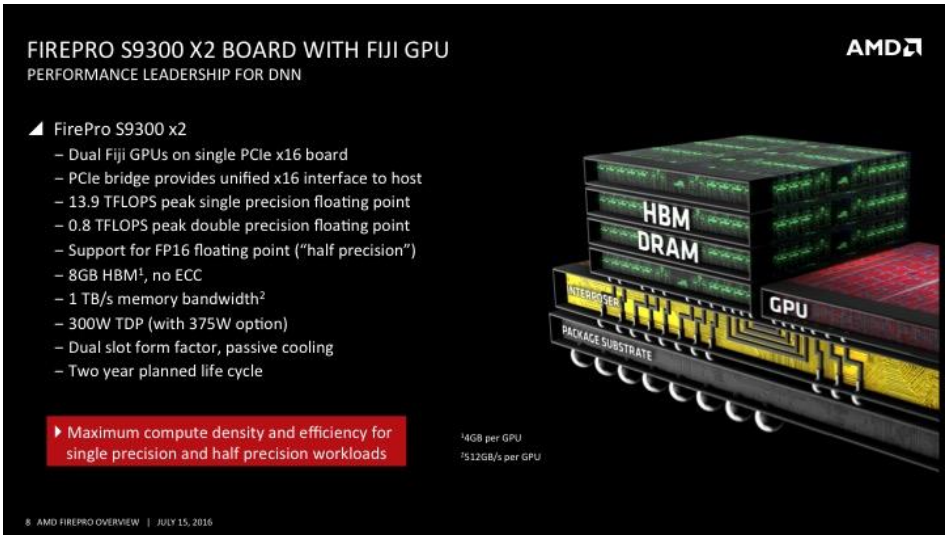
(Source: AMD)

FIREPRO S9300 X2

In 2015, AMD introduced the Radeon R9 Nano based on the Fiji chip. According to AMD, it was the first in the industry to use the HBM technology, which delivers 512 GB/s access to 4GB of on-package memory. Then, earlier in 2016, AMD launched the S9300 x2, which consists of two of these chips on a FirePro S Series card, offering industry-leading single-precision floating point and memory bandwidth with 1 TB/s of memory bandwidth to the HBM stacked die memory and 13.9 TFlops single-precision and half-precision performance. This move was aggressive, because the chip was initially a gaming chip without features expected from the AMD FirePro brand, such as ECC memory and double-precision FP. Note that since the higher volume gaming market turns products over much more rapidly that the server market, the S9300 x2 has a shorter lifespan than the professional GPU market often expects. But AMD wanted to leverage its head start in HBM memory, especially in the deep learning and oil and gas markets, where NVIDIA's high bandwidth PASCAL P100 product is not expected to ship until the end of 2016 and where gaming-class devices are often deployed.

For markets that use single-precession math, like oil and gas, deep learning, rendering, or molecular dynamics, this card is designed to offer value for single-precision workloads. Since it is the first product to support the AMD ROCm software stack, it can be used as a test and development platform for the HBM architecture and the ROCm software environment. AMD is likely to focus on this HBM approach with future 14 nm products to provide additional mainstream compute platforms.

FIGURE 4: FIREPRO S9300 X2 BOARD WITH FIJI GPU

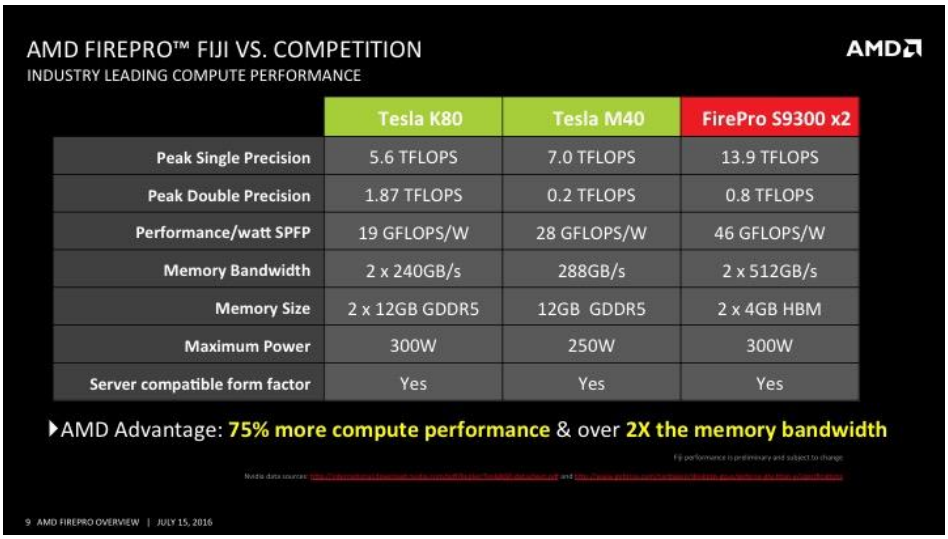


(Source: AMD)

COMPETITIVE COMPARISONS

AMD's HBM-equipped FirePro S9300 x2 is priced similarly to NVIDIA's Tesla K80 dual chip GPU. However, AMD's data shows over twice the single-precision performance and twice the memory bandwidth, albeit with a much smaller memory footprint.

FIGURE 5: AMD FIREPRO FIJI VS. COMPETITION



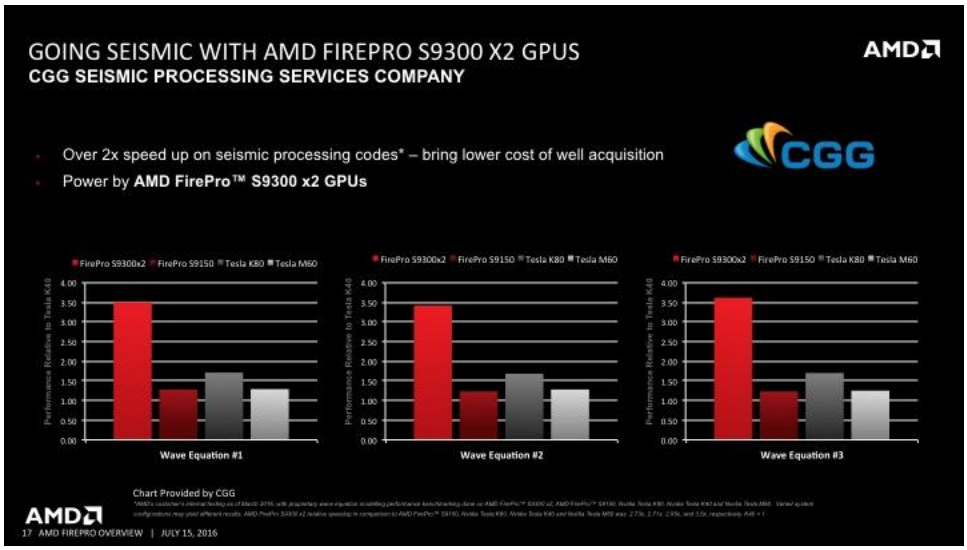
	Tesla K80	Tesla M40	FirePro S9300 x2
Peak Single Precision	5.6 TFLOPS	7.0 TFLOPS	13.9 TFLOPS
Peak Double Precision	1.87 TFLOPS	0.2 TFLOPS	0.8 TFLOPS
Performance/watt SPFP	19 GFLOPS/W	28 GFLOPS/W	46 GFLOPS/W
Memory Bandwidth	2 x 240GB/s	288GB/s	2 x 512GB/s
Memory Size	2 x 12GB GDDR5	12GB GDDR5	2 x 4GB HBM
Maximum Power	300W	250W	300W
Server compatible form factor	Yes	Yes	Yes

(Source: AMD)

Peak Flops only reveals so much about how a processor really performs. Figure 6 shows application performance provided by AMD from French oil services company

CGG. Benchmark results show the S9300 x2 outperforming the NVIDIA Tesla K80 GPU by over 2x for CGG's RTM application.

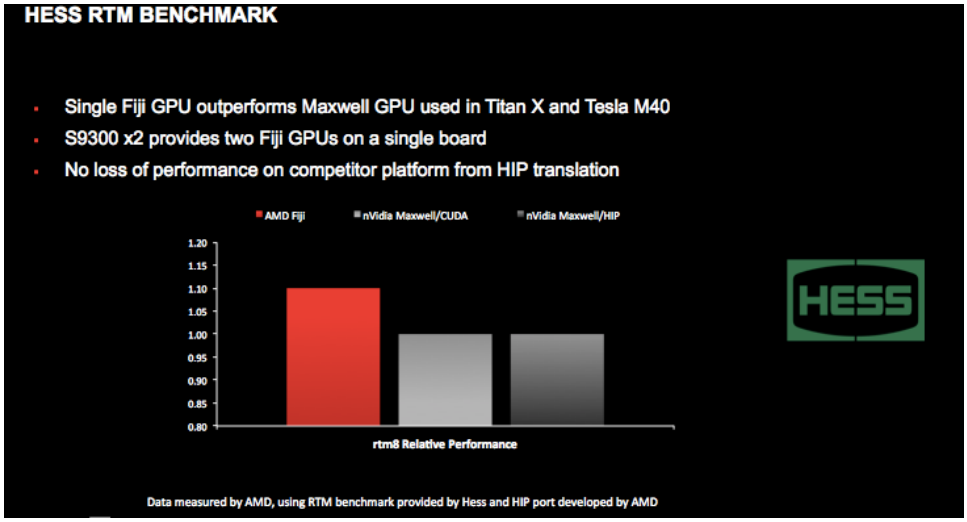
FIGURE 6: SEISMIC BENCHMARKS FOR CUDA CODES



(Source: AMD)

AMD also provided evidence of its HIP translator performance for CUDA codes. AMD ran the Reverse Time Migration codes from Hess, another energy services company, and found that the code did not suffer from the "HIP-ify" process. According to AMD, the code ran about the same on the NVIDIA M40 with and without the HIP translation layer, (Figure 7) and ran significantly faster on the AMD single chip Fiji Radeon card.

FIGURE 7: HESS RTM BENCHMARK



(Source: AMD)

OPPORTUNITIES & CHALLENGES

ROCm is designed to improve AMD's position for GPU compute with an open software environment and tools to ease the porting of NVIDIA CUDA codes. With this new software in place, AMD has opportunities but must address remaining challenges.

OPPORTUNITIES

- AMD could court early adopters in the rapidly growing market for training deep neural networks for machine learning.
- AMD could become a lower cost second-source supplier for cloud-based GPU instances in public and private clouds.
- AMD's relationships with the US DOE could help them position partnerships with other leading research institutions.
- AMD could gain traction in single-precision workloads such as reverse time migration in oil and gas exploration and in molecular dynamics in the pharmaceutical industry.
- The HPC accelerator industry is a small but growing market. The research, development, and end user communities should appreciate and embrace an open source HPC acceleration stack.

Although ROCm reduces some of the hurdles facing potential adopters of AMD GPU hardware, several challenges remain.

CHALLENGES

- While ROCm looks attractive, it is a completely new platform. Like all complex software, it will take time to mature and become accepted in the industry. Close relationships with leading universities and institutions will be essential to AMD's fledgling HPC ecosystem.
- NVIDIA's CUDA platform supports hundreds of applications, thanks to NVIDIA's technology and investments in customer-facing market development and support resources. AMD needs to invest in recruiting and tuning target market applications to address the large enterprise compute market.
- AMD will need to deploy business development and application engineering resources to help end-users adopt and optimize AMD hardware and the new ROCm stack. NVIDIA has a large global team in place today with deep customer relationships and workload specific knowledge.

CONCLUSIONS & RECOMMENDATIONS

AMD's hardware for HPC application acceleration has had limited market traction. AMD understands that good hardware only goes so far if it is not also easy to port and use in a production environment. AMD designed ROCm to be easy to port existing CUDA codes to take advantage large GDDR5 memory capacity and new HBM technologies. ROCm is open source and enables single source code development for virtually any underlying deployment hardware: CPU, AMD GPU, or NVIDIA GPU. ROCm and updated 14nm based GPUs help AMD vie for the fastest-growing segment in the industry: training of neural networks for machine learning. AMD will need to invest talent and resources to build a viable ecosystem and gain market share, though. When AMD expands its HPC portfolio with its future 14nm GPUs, the new software could enable them to compete in HPC segments currently dominated by NVIDIA Tesla products today. AMD publishes further information here.

For customers seeking competitive GPUs for compute-intensive applications, MI&S recommends obtaining an evaluation unit to test and benchmark using their own code. Customers with double-precision applications should contact AMD to determine when they can get beta and production drivers that support those cards. Customers with single-precision applications should consider testing the S9300 x2 card to determine the appropriateness of this technology for specific applications.

IMPORTANT INFORMATION ABOUT THIS PAPER

AUTHOR
Karl Freund, Senior Analyst at Moor Insights & Strategy

PUBLISHER
Patrick Moorhead, President & Principal Analyst at Moor Insights & Strategy

EDITOR
Scott McCutcheon, Director of Research at Moor Insights & Strategy

INQUIRIES
Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS
This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES
This paper was commissioned by Advanced Micro Devices, Inc. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.