It’s time to ROC

Gregory Stoner
Senior Director, Radeon Open Compute
April 4, 2016
Radeon Technology Group Presents

- New Path Forward for HPC and Ultrascale Computing Markets
- Focused Commitment to Meet Customer Computing Needs
- Open Foundation For Development, Discovery and Education

ROCM: Radeon Open Compute Platform
ROCm Platform: A New Stage to Play

Announcing revolution in GPU computing

ROCK - Headless Linux® 64-bit Kernel Driver and ROCr: HSA+ Runtime

- Open Source from the metal up
- Focus on overall latency to compute
- Optimized for Node and Rack Scale Multi-GPU Compute
- Foundation to explore GPU Compute
ROCm gives you a rich foundation for a “new sound”

Bringing new capabilities you requested

- Peer to Peer Multi-GPU
  - Process Concurrency & Preemption
  - HSA Signals and Atomics
  - Profiler Trace and Event Collection API
  - Peer to Peer with RDMA
    - User Mode DMA
    - Multi-GPU Coarse-grain Shared Virtual Memory
    - Multi-GPU Memory Management API
  - Standardized loader and Code Object Format

- Native GCN ISA Code Generation
- User Mode DMA
- HIP Runtime
- GCN ISA Assembler and Disassembler
- Low latency dispatch
- Large BAR
  - Low Overhead PCIe® data transfers
- Docker© Containerization Support
  - Large Memory Single Allocation
  - Systems Management API and Tools
  - Offline Compilation Support
  - HCC C++ and OpenMP C/C++ compiler
  - Continuum IO Anaconda with NUMBA
ROCM Keeping the Signal Clean

User Mode DMA: Drives Down Latency of the Transfer SHOC

S->D Measured Bandwidth (GB/s)

D->S Measured Bandwidth (GB/s)

SHOC Version 1.1.5 ROC RC2 vs Radeon Software Crimson Edition 15.12 on R9 Nano
It's about making “premium sound” on the ROCm stage

HCC (Heterogeneous Compute Compiler) Mainstream Standard Languages for GPU Acceleration

- HCC is a single source ISO C++ 11/14 compiler for both the CPU and GPU
  - OPENMP 3.1 C & C++ Support Today for CPU
- C++17 “Parallel Standard Template Library”
- Built on rich compiler infrastructure CLANG/LLVM and libC++
- Performance Optimization for Accelerators
  - Low level memory placement controls: pre-fetch, discard data movement
  - Asynchronous compute kernels
  - Scratchpad memories support
Bringing rhythm to today’s developers

HIP = “Heterogeneous-Compute Interface for Portability”

- Port from CUDA to a common C++ programming model
- HIP code runs through either CUDA NVCC or HCC
- HiPify tools simplify porting from CUDA to HIP
- Builds on HCC Compiler
  - Host and device code can use templates, lambdas, advanced C++ features
  - C-based runtime APIs (hipMalloc, hipMemcpy, hipKernelLaunch and more)
Playing “Deep Bass” with Anaconda

Embracing Python Developer Community with Heterogeneous Acceleration

Numba speed up your applications with high performance functions written directly in Python

Rich set of options to optimize for APU or discrete GPU
- Async execution, specify group size, use shared memory
- Data transfer can be performed implicitly based on kernel arguments

ANACONDA on ROCm: a New Class of Performance for Python
Expanding Set of Cross Platform Tools

Software Landscape Through 2015

Radeon Open Compute Platform (ROCm)

OpenCL

Catalyst™

CUDA

ISO C++
No GPU Acceleration

Improved Performance

One Code Base
Multiple Platforms

Simplest Path to
GPU Acceleration

Radeon Open Compute Platform (ROCm)

OpenCL

Catalyst™

CUDA

ISO C++
AMD HCC Compiler

Improved Performance

One Code Base
Multiple Platforms

Simplest Path to
GPU Acceleration

OpenCL TM

ROCL Runtime + OpenCL

C++
AMD HCC Compiler

ROC Runtime + HIP

ISO C++
AMD HCC Compiler

ROC Runtime + C++ 11/14 + PSTL

OpenCL and the OpenCL logo are trademarks of Apple Inc, used by permission by Khronos.
Amp up the signal

Focusing on Solution & Building Out Key Foundations to Support Libraries, Frameworks and Applications via GPUopen
ROCm - A Stage For Your "Songs"

Build a Rich Foundation of Libraries and Applications

Core math libraries
- BLAS (Skinny, Non-Square, Square) SGEMM & DGEMM
- FFT
- SPARSE
- RAND

Third-party Frameworks
- Kokkos C++ Framework
- CHARM++ Frameworks
- HPX Framework (in development)

Benchmarks and Applications
- SHOC
- Parboil
- Lullesh
- MiniFE
- Cloverleaf
- NAMD
- LeanMD
- CoMD
- MiniAMR
- HPGMG
- XSBench
Rocking the neural pathways

Instinctive Computing foundation for Machine Learning and Neural Networks

- Supporting Key Neural Network Frameworks
  - *Torch 7* and *Caffe*

- mlOpen
  - *Optimized Convolution Neural Network for NN Frameworks*

- OpenVX with Graph Optimizer
  - *Foundation for rich Machine Learning*
Chime telescope

Three-Dimensional Mapping Of The Universe

- Solving one of most puzzling new mystery in astronomy: Fast Radio Bursts (FRB)

“CHIME has a truly novel design. No moving parts! [...] Moreover, it will have 2048 antennas and a massive software correlator that allows it to ‘point’ in different directions all in software.”

- Astrophysicist Victoria Kaspi, Gerhard Herzberg 2016 Prize Laureate

Multi Pflops AMD FirePro S9300 x2 cluster

Image: Prof. Keith Vanderlinde, Dunlap Institute, University of Toronto.
Gear for the ROCm stage

512 GB/s memory bandwidth
8.19 TFlops Single Precision

1 TB/s memory bandwidth
13.9 TFlops Single Precision
Radeon™ R9 NANO BRING A SOLID SOUND
ROOFLINE FROM MIXBENCH* HIP PORT

Single Precision Gflops

Double Precision Gflops

https://github.com/ekondis/mixbench  Mixbench on ROCm RC2 Radeon R9 NANO ("FIJI")
Going seismic with AMD FirePro S9300 x2 GPUs

CGG Seismic Processing Services Company

- Over 2x speed up on seismic processing codes* – bring lower cost of well acquisition
- Power by AMD FirePro™ S9300 x2 GPUs

*AMD's customer's internal testing as of March 2016, with proprietary wave equation modelling performance benchmarking done on AMD FirePro™ S9300 x2, AMD FirePro™ S9150, Nvidia Tesla K80, Nvidia Tesla K40 and Nvidia Tesla M60. Varied system configurations may yield different results. AMD FirePro S9300 x2 relative speedup in comparison to AMD FirePro™ S9150, Nvidia Tesla K80, Nvidia Tesla K40 and Nvidia Tesla M60 was 2.73x, 2.71x, 2.05x, and 3.5x, respectively. K40 = 1
Going epic with Supermicro and Mellanox

- **Login/Storage Node**
  - 2U server
  - 2 Xeon E5-2640 v3
  - 64 GB DDR4
  - 64 TB

- **Storage/Admin Node**
  - 2U server
  - 2 Xeon E5-2640 v3
  - 64 GB DDR4
  - 64 TB

- **Compute Node**
  - 1U Compute Server
  - 2 Xeon E5-2640 v3
  - 64 GB DDR4
  - 2 x 960 GB Samsung SSD

- **Management Network**
  - 1 GigE Ethernet

- **Compute and Storage Network**
  - Infiniband FDR or EDR

- **Mellanox FDR 36 Port Switch**

- **48-Port Gigabit**

- **Radeon™ NANO Cluster**
  - 36 Compute nodes
  - 144 Radeon™ R9 Nano Compute GPU’s
  - 46 GFLOP/Watt
  - 1.14 Pflops in < 30 KW
We are looking to build out a worldwide band

**How to Join**

- Get started today developing with ROCm - GPUOpen ROCm Getting Started  [http://bit.ly/1ZTIk82](http://bit.ly/1ZTIk82)
- Engage In the development of ROCm @ GitHub RadeonOpenCompute
- Show case your applications, libraries and tools on to ROCm via GPUOpen

"The power of one, if fearless and focused, is formidable, but the power of many working together is better."

– Gloria Macapagal Arroyo
Disclaimer & Attribution

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

© 2016 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, FirePro and combinations thereof are trademarks of Advanced Micro Devices, Inc. in the United States and/or other jurisdictions. Other names are for informational purposes only and may be trademarks of their respective owners. PCIe and PCI Express are registered trademarks of PCI-SIG Corporation. OpenCL and the OpenCL logo are trademarks of Apple, Inc. and used by permission of Khronos.