RADEON OPEN COMPUTE PLATFORM

DMITRY KOZLOV

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

AMD

RADEON
TECHNOLOGIES GROUP

HARDWARE FOR THE ROCM STAGE

512 GB/s Memory Bandwidth
8.19 TFlops Single Precision

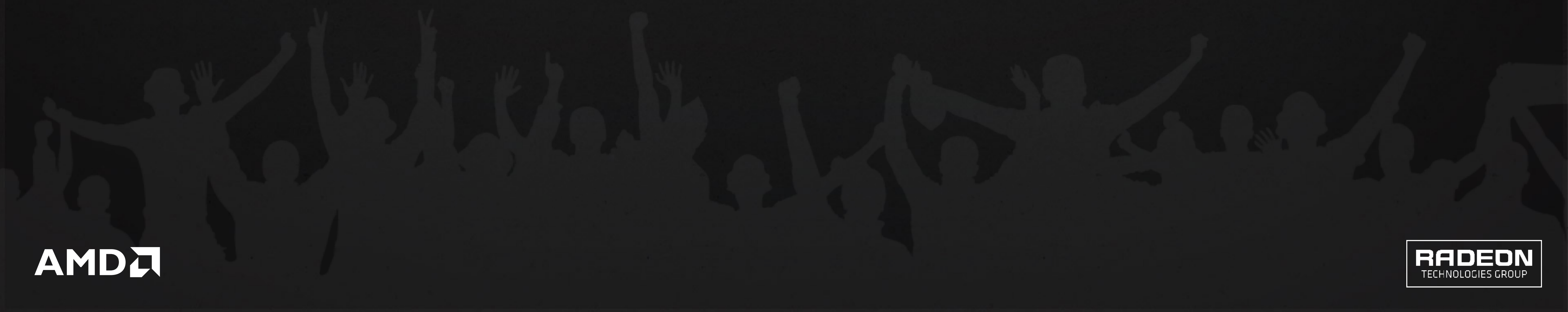
1 TB/s Memory Bandwidth
13.9 TFlops Single Precision

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

Native GCN ISA Code Generation

Peer to Peer Multi-GPU

User Mode DMA

Low latency dispatch

Process Concurrency & Preemption

HIP Runtime

HSA Signals and Atomics

GCN ISA Assembler and Disassembler

Profiler Trace and Event Collection API

Large BAR

Low Overhead PCIe® data transfers

Peer to Peer with RDMA

Docker© Containerization Support

User Mode DMA

Large Memory Single Allocation

Multi-GPU Coarse-grain Shared Virtual Memory

Systems Management API and Tools

Offline Compilation Support

Multi-GPU Memory Management API

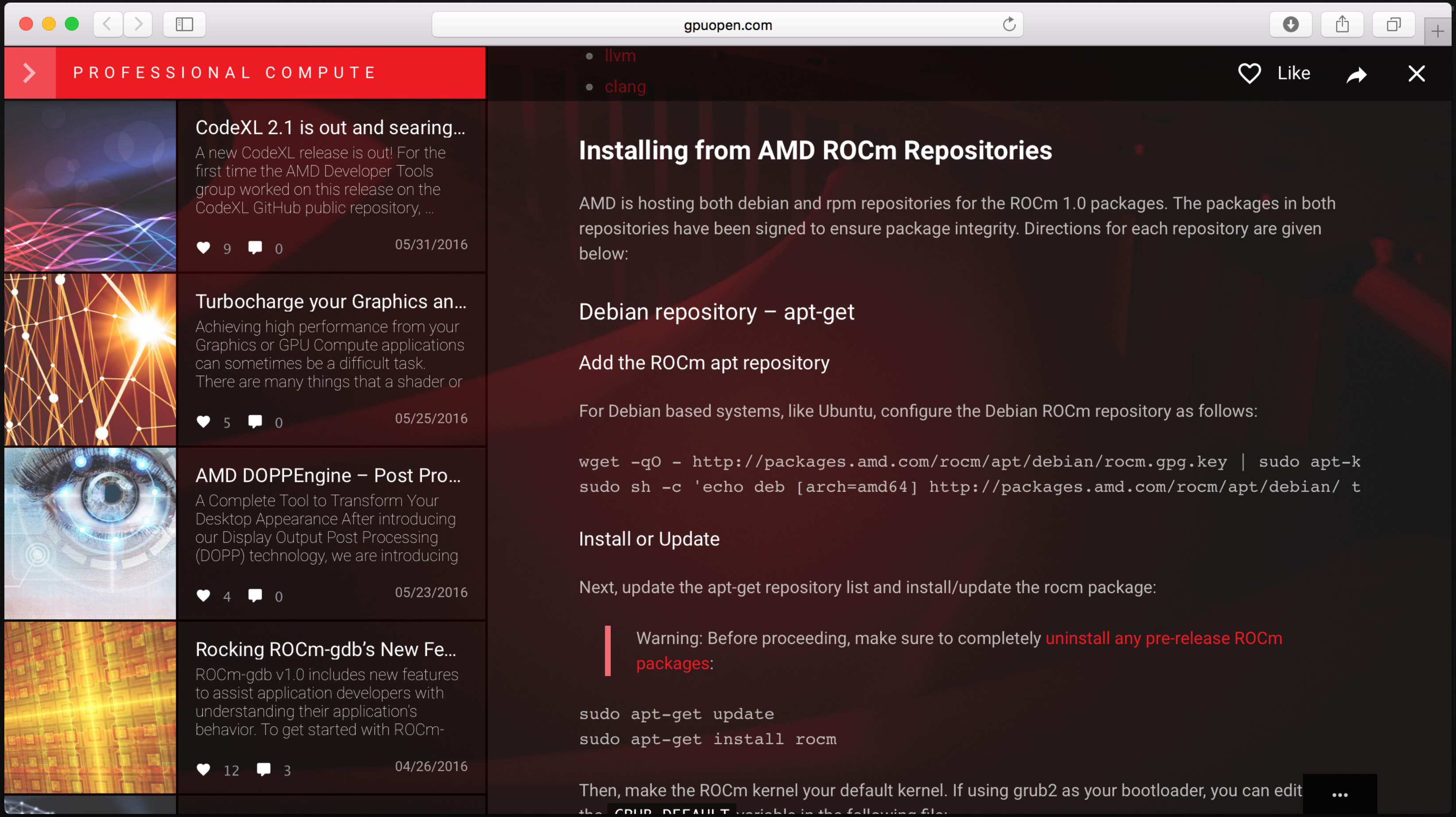
HCC C++ and OpenMP C/C++ compiler

Standardized loader and Code Object Format

Continuum IO Anaconda with NUMBA

ROCM PLATFORM: A NEW STAGE TO PLAY

Announcing revolution in GPU computing



CodeXL 2.1 is out and searing...

A new CodeXL release is out! For the first time the AMD Developer Tools group worked on this release on the CodeXL GitHub public repository, ...

♥ 9 💬 0 05/31/2016

Turbocharge your Graphics an...

Achieving high performance from your Graphics or GPU Compute applications can sometimes be a difficult task. There are many things that a shader or

♥ 5 💬 0 05/25/2016

AMD DOPPEngine – Post Pro...

A Complete Tool to Transform Your Desktop Appearance After introducing our Display Output Post Processing (DOPP) technology, we are introducing

♥ 4 💬 0 05/23/2016

Rocking ROCm-gdb's New Fe...

ROCm-gdb v1.0 includes new features to assist application developers with understanding their application's behavior. To get started with ROCm-

♥ 12 💬 3 04/26/2016

- llvm
- clang

Installing from AMD ROCm Repositories

AMD is hosting both debian and rpm repositories for the ROCm 1.0 packages. The packages in both repositories have been signed to ensure package integrity. Directions for each repository are given below:

Debian repository – apt-get

Add the ROCm apt repository

For Debian based systems, like Ubuntu, configure the Debian ROCm repository as follows:

```
wget -qO - http://packages.amd.com/rocm/apt/debian/rocm.gpg.key | sudo apt-k
sudo sh -c 'echo deb [arch=amd64] http://packages.amd.com/rocm/apt/debian/ t
```

Install or Update

Next, update the apt-get repository list and install/update the rocm package:

> Warning: Before proceeding, make sure to completely uninstall any pre-release ROCm packages:

```
sudo apt-get update
sudo apt-get install rocm
```

Then, make the ROCm kernel your default kernel. If using grub2 as your bootloader, you can edit the `GRUB_DEFAULT` variable in the following file:

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
- 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*

IT’S ABOUT MAKING PREMIUM SOUND ON THE ROCM STAGE

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

```
const float a = 100.0f;

float x[N];

float y[N];

…

for (int i = 0; i < N; i++) {

    y[i] = a * x[i] + y[i];

}
```

ROCM

ON

IT’S ABOUT MAKING PREMIUM SOUND ON THE ROCM STAGE

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

```
#include <hc.hpp>

hc::array_view<float, 1> av_x(N, x);

hc::array_view<float, 1> av_y(N, y_gpu);

// launch a GPU kernel to compute the saxpy
in parallel

hc::parallel_for_each(hc::extent<1>(N),
[=](index<1> i) [[hc]] {

    av_y[i] = a * av_x[i] + av_y[i];

});
```

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)*

Bringing rhythm to today’s developers

HIP = “Heterogeneous-Compute Interface for Portability”

```
git clone https://github.com/GPUOpen-ProfessionalCompute-Tools/HIP HIP

hipify square.cu > square.cpp
```

Bringing rhythm to today’s developers

HIP = “Heterogeneous-Compute Interface for Portability”

```
template <typename T>
__global__ void
vector_square(T *C_d, const T *A_d, size_t N)
{
    size_t offset = (blockIdx.x * blockDim.x + threadIdx.x);
    size_t stride = blockDim.x * gridDim.x ;


    for (size_t i=offset; i<N; i+=stride) {
             C_d[i] = A_d[i] * A_d[i];
       }
}
```

Bringing rhythm to today’s developers

HIP = “Heterogeneous-Compute Interface for Portability”

```
/*
* Square each element in the array A and write to array C.
*/
template <typename T>
__global__ void
vector_square(T *C_d, const T *A_d, size_t N)
{
    size_t offset = (hipBlockIdx_x * hipBlockDim_x + hipThreadIdx_x);
    size_t stride = hipBlockDim_x * hipGridDim_x ;

    for (size_t i=offset; i<N; i+=stride) {
        C_d[i] = A_d[i] * A_d[i];
    }
}
```

Bringing rhythm to today’s developers

HIP = “Heterogeneous-Compute Interface for Portability”

```
CHECK(hipMalloc(&A_d, Nbytes));
CHECK(hipMalloc(&C_d, Nbytes));

CHECK ( hipMemcpy(A_d, A_h, Nbytes, hipMemcpyHostToDevice));

const unsigned blocks = 512;
const unsigned threadsPerBlock = 256;

hipLaunchKernel(HIP_KERNEL_NAME(vector_square), dim3(blocks),
dim3(threadsPerBlock), 0, 0, C_d, A_d, N);

CHECK ( hipMemcpy(C_h, C_d, Nbytes, hipMemcpyDeviceToHost));
```

Bringing rhythm to today’s developers

HIP = “Heterogeneous-Compute Interface for Portability”

```
Fiji1:~/hip/samples/square$ hipcc
square.cpp -o square.hip.out
Fiji1:~/hip/samples/square$
./square.hip.out
info: running on device Fiji
info: allocate host mem (  7.63 MB)
info: allocate device mem (  7.63 MB))
info: copy Host2Device
info: launch 'vector_square' kernel
info: copy Device2Host
info: check result
PASSED!
```

```
TITAN1:~/ben/hip/samples/square$ hipcc
square.cpp -o square.hip.out
TITAN1:~/ben/hip/samples/square$
./square.hip.out
info: running on device GeForce GTX TITAN X
info: allocate host mem (  7.63 MB)
info: allocate device mem (  7.63 MB)
info: copy Host2Device
info: launch 'vector_square' kernel
info: copy Device2Host
info: check result
PASSED!
```

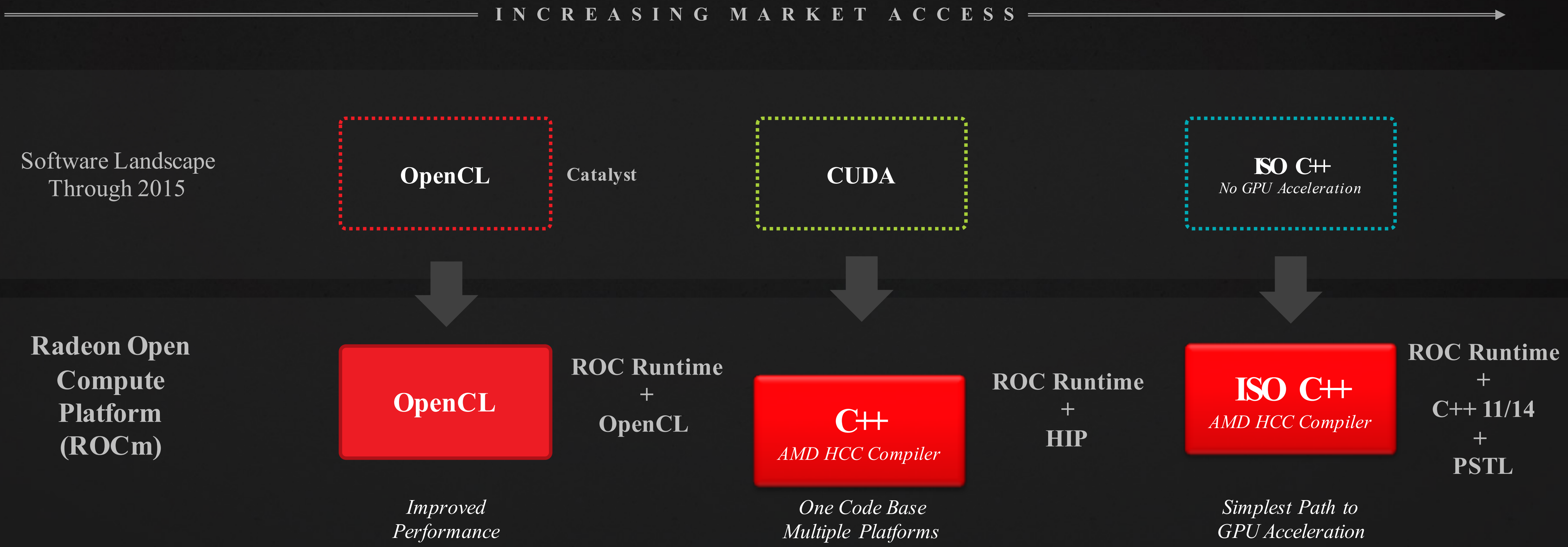
HIP AUTOMATED CONVERSION TOOLS

AMD INTERNAL TESTS , NON-FINAL HIP TOOL , JANUARY 2016

Application	LOC	CUDA to HIP	Unconverted APIs	Code Changed %	Conversion %
FinanceBench	34,820	457	0	1%	100%
Barracuda	17,269	222	6	1%	97%
Libgeodecomp	123,503	851	17	1%	98%
NVBio	276,523	4,255	125	2%	97%
Magma-1.7.0	677,620	21,318	259	3%	99%
Hoomd-v1.1.1	76,155	2,525	112	3%	96%
cuNN	6,820	540	0	8%	100%
cuTorch	14,320	752	30	5%	96%
Caffe	75,528	503	31	1%	94%
Gpubiotools	15,550	906	29	6%	97%
Arrayfire	144,097	2,201	77	2%	97%
quda	355,689	6,954	1,064	2%	87%
Stella	137,097	1,375	38	1%	97%
SHOC	19,038	1,860	38	10%	98%

Going Global

Expanding Set of Cross Platform Tools



INCREASING MARKET ACCESS

Software Landscape Through 2015

OpenCL

Catalyst

CUDA

ISO C++
No GPU Acceleration

Radeon Open Compute Platform (ROCm)

OpenCL

ROC Runtime
+
OpenCL

Improved Performance

C++
AMD HCC Compiler

ROC Runtime
+
HIP

One Code Base Multiple Platforms

ISO C++
AMD HCC Compiler

ROC Runtime
+
C++ 11/14
+
PSTL

Simplest Path to GPU Acceleration

AMP UP THE SIGNAL

Focusing on Solution & Building Out Key Foundations to Support Libraries, Frameworks and Applications via GPUopen

SECURITY SENSING

BIO-MEDICAL SYNTHESIS

WIRELESS COMMUNICATIONS

EXTRACTION MACHINE LEARNING

DISCOVERY MEDICAL IMAGING

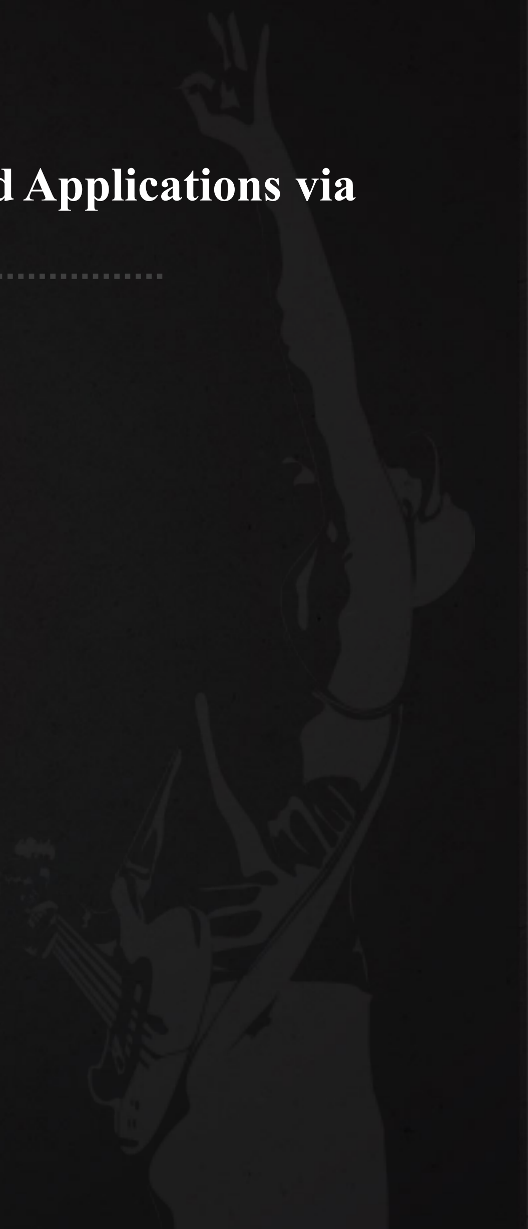
SPEECH BIO-IT FORENSICS OIL & GAS

MODELLING SECURITY

LEARNING ASTRONOMY

FINANCE ANALYSIS AUDIO

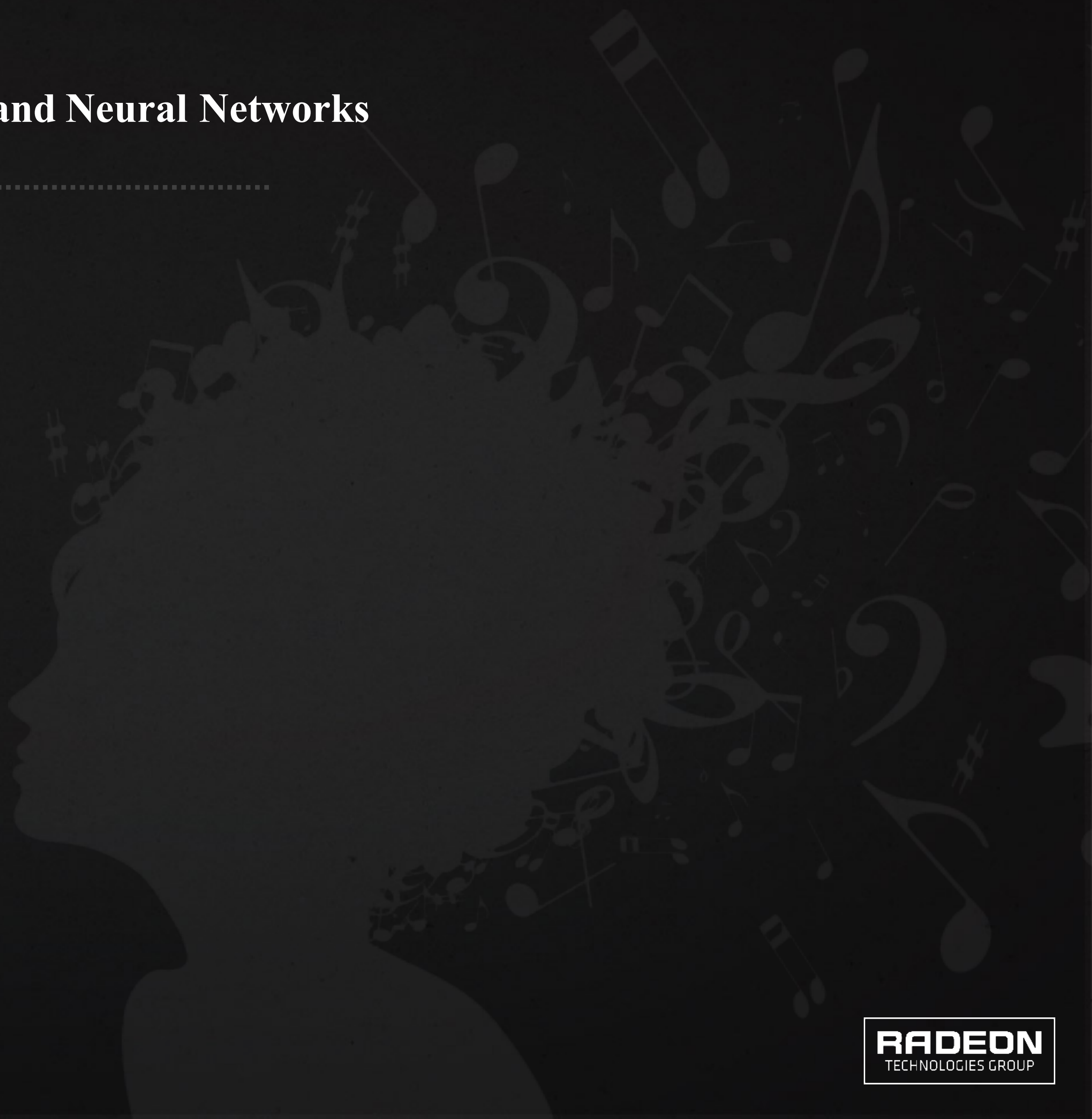
ACQUISITION



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 mysteries 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

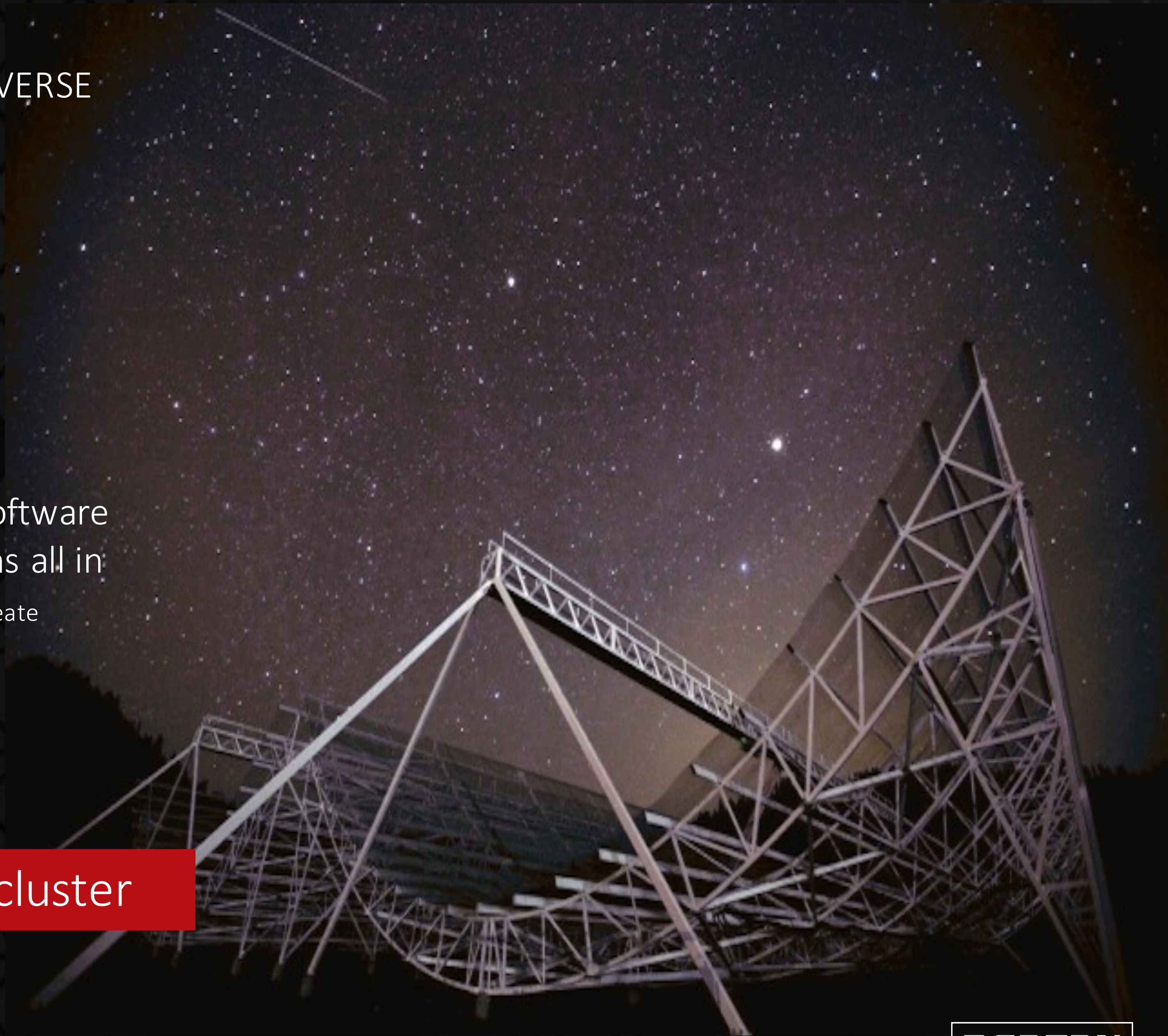
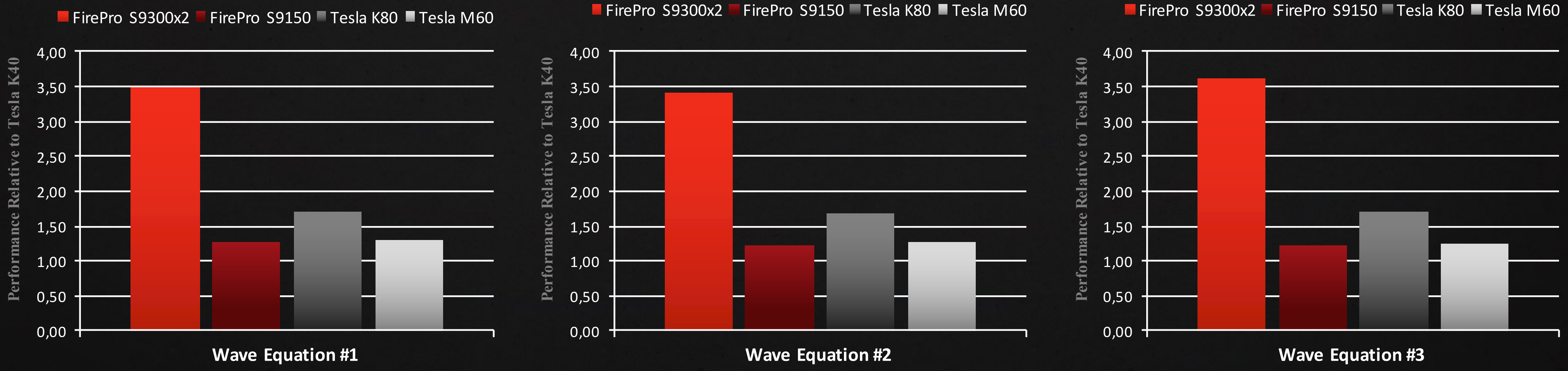


Image: Prof. Keith Vanderlinde, Dunlap Institute, University of Toronto.

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

OVER 2X SPEEDUP ON SEISMIC PROCESSING CODES

CGG SEISMIC PROCESSING SERVICES COMPANY

- Oil crisis? A reality!

  …but opportunities for the agile ones
- AMD helps CGG finding solutions
- AMD FirePro S9300 x2 GPU
  - – 2x NVIDIA Tesla K80
  - - 2.6x NVIDIA Tesla M60
  - - 3.5x NVIDIA Tesla K40

1 TB/s memory bandwidth, more than 2x competition[2]

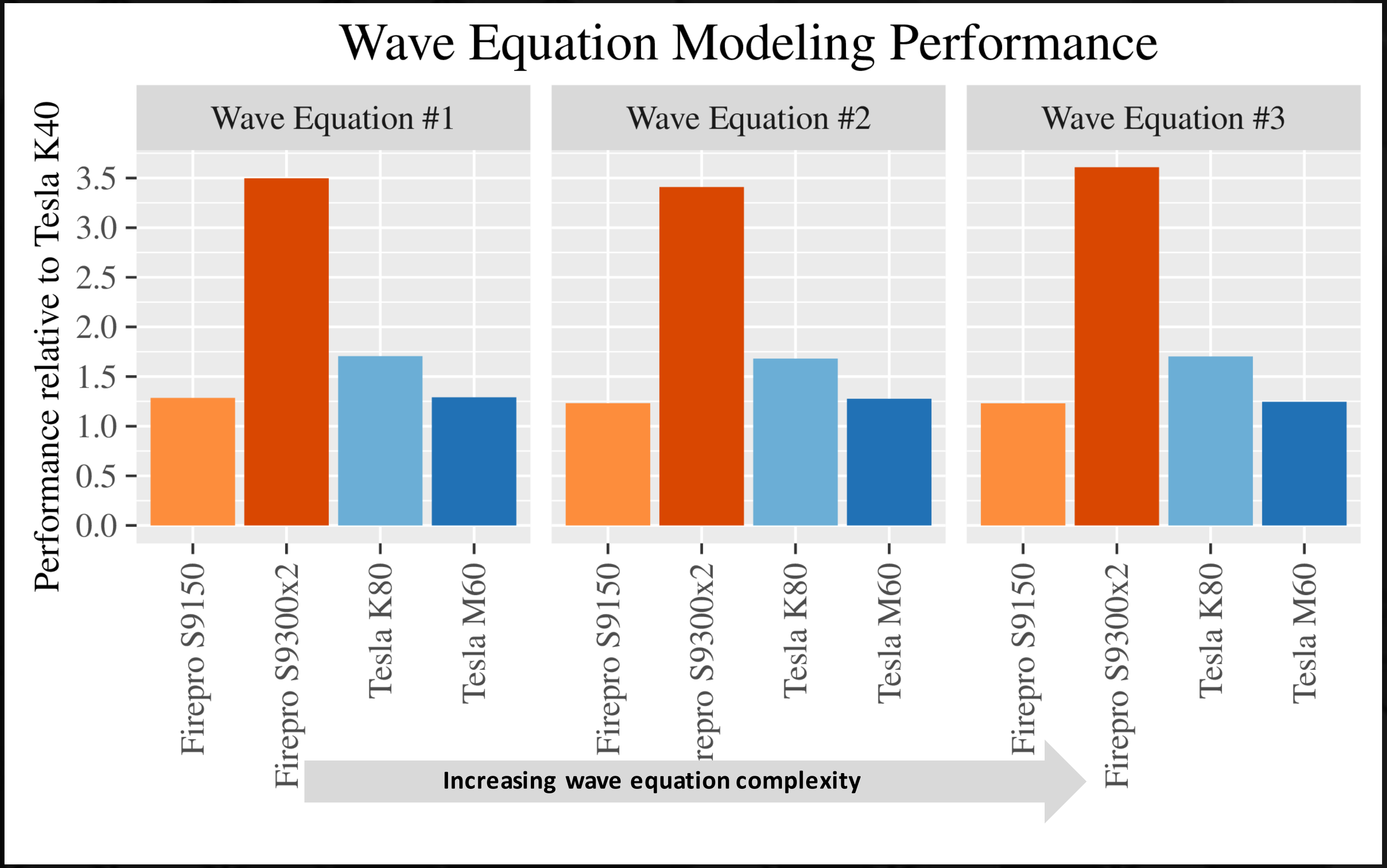
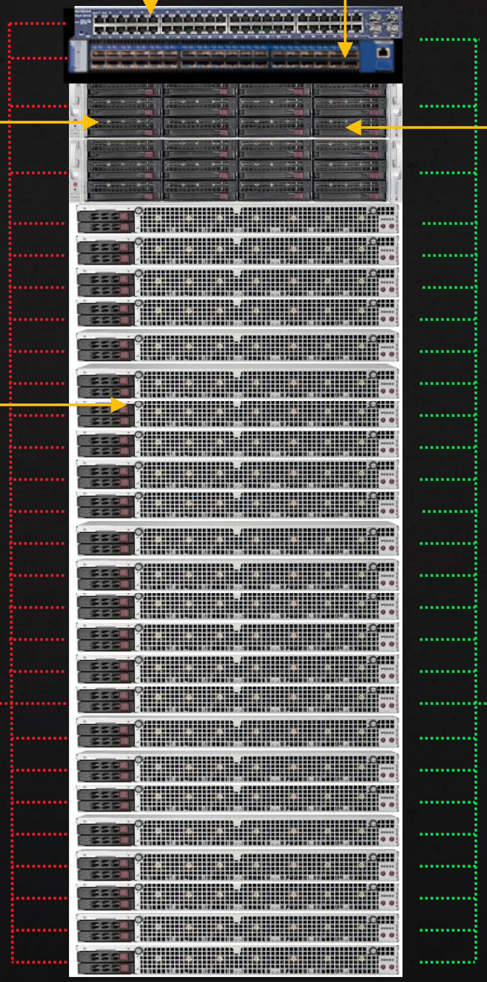


Chart Provided by CGG

GOING EPIC WITH SUPERMICRO

- **Radeon NANO Cluster**
- **36 Compute nodes**
- **144 R9 Nano Compute GPU's**
- **46 GFLOP/Watt**
- **1.14 Pflops in < 30 KW**



48-Port Gigabit

Mellanox FDR 36 Port Switch

Login/Storage Node

- 2U server
- 2 Xeon E5-2640 v3
- 128 GB DDR4
- 64 TB

Storage/Admin Node

- 2U server
- 2 Xeon E5-2640 v3
- 64 GB
- 64 TB storage

Compute Node

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

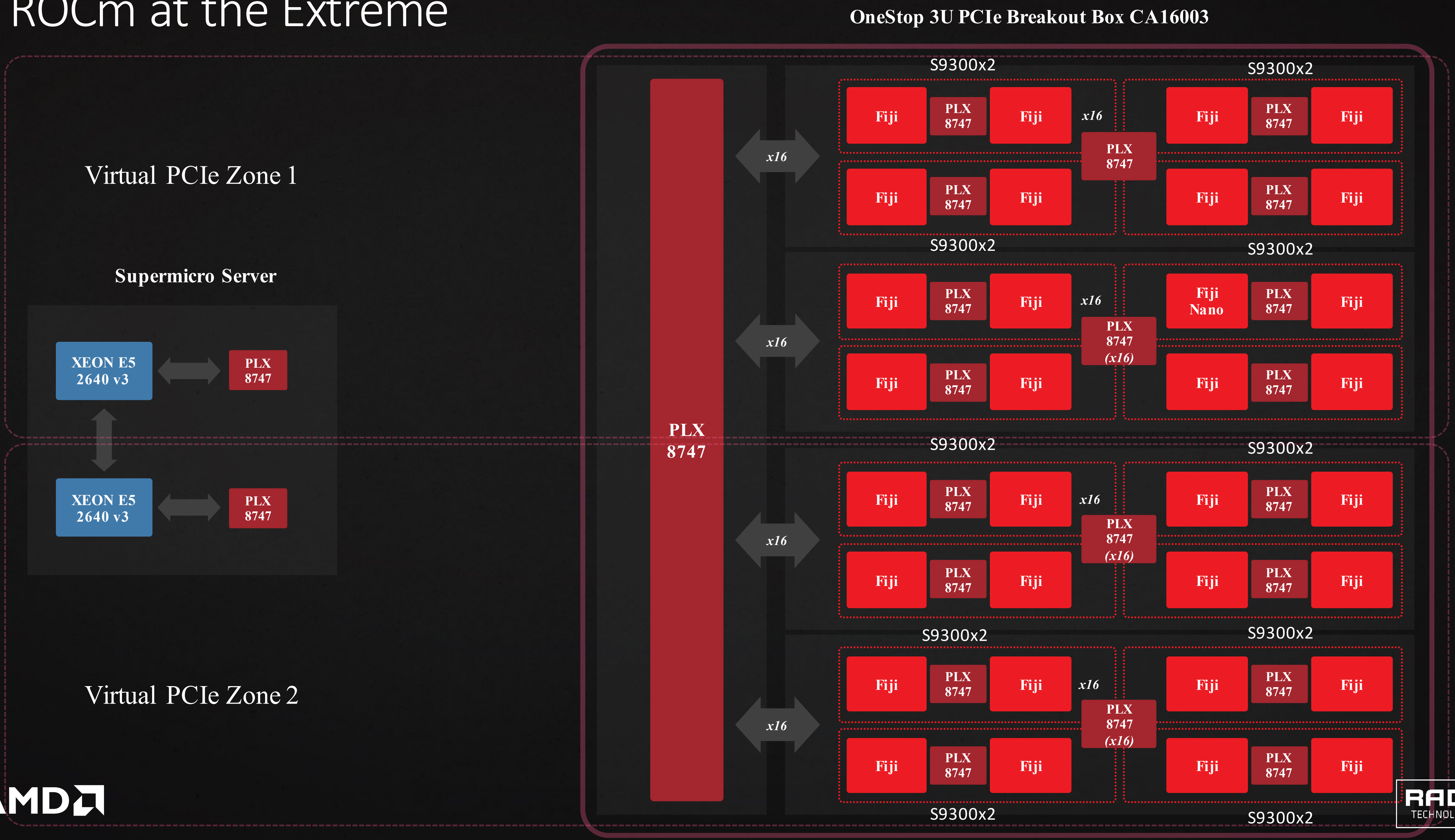
Management Network

- 1 GigE Ethernet

Compute and Storage Network

Infiniband FDR or EDR

ROCm at the Extreme



HPC SIGNAL PROCESSING

NEED FOR MAXIMUM MEMORY BANDWITH ALONG WITH MAXIMUM COMPUTE PERFORMANCE

WIRELESS COMMUNICATIONS

DISCOVERY

BIO-MEDICAL

SECURITY

MACHINE LEARNING

MEDICAL IMAGING

SENSING

SPEECH

AUDIO

BIO-IT

LEARNING

FORENSICS

SYNTHESYS

ACQUISITION

ANALYSIS

MODELLING

ASTRONOMY

OIL & GAS

FINANCE

EXTRACTION

The World’s Fastest Single-Precision GPU Accelerator[1]

AMD FIREPRO™ S9300 X2 GPU

1 TB/s memory bandwidth
13.9 TFLOPS 32-bit

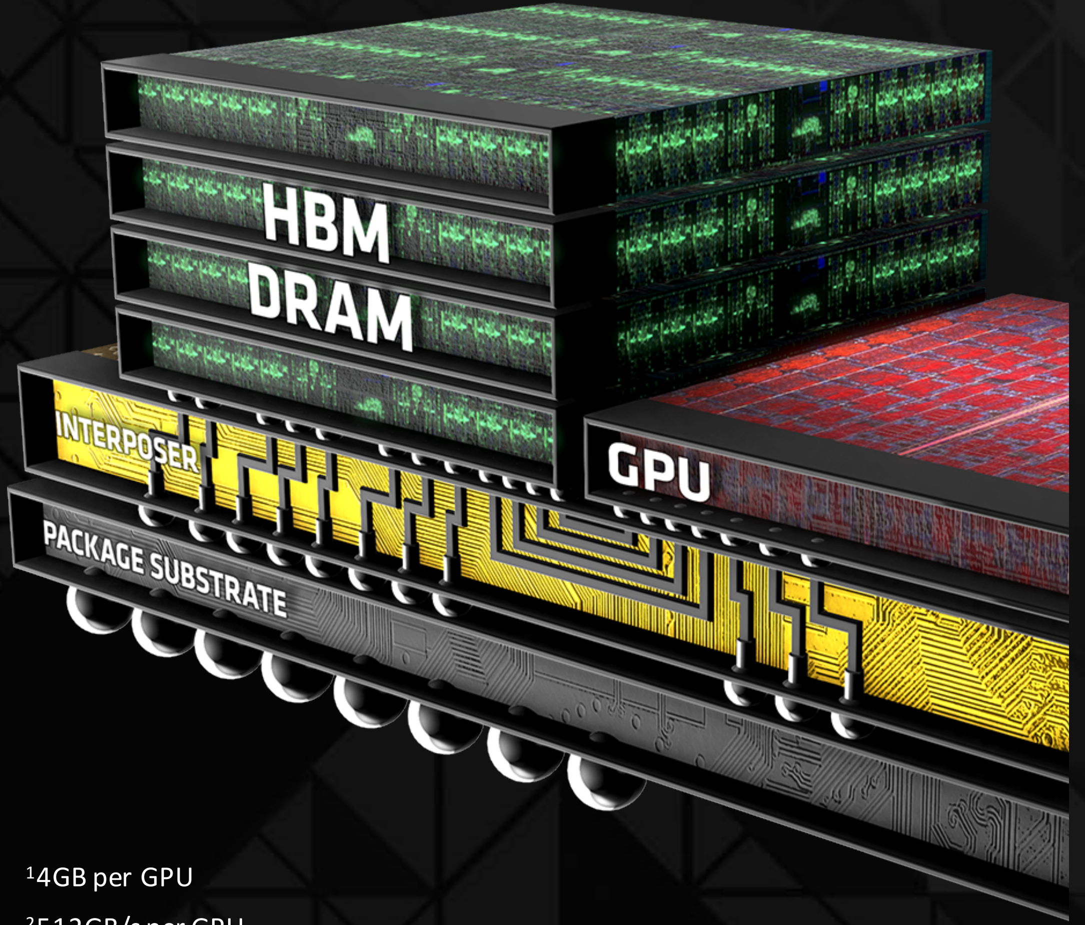
The World's Fastest Single-Precision GPU Accelerator[1]

AMD FIREPRO™ S9300 X2 BOARD WITH AMD "FIJI" GPU

THE FIRST DATA CENTER GPU WITH HBM

- Dual "Fiji" GPUs on single PCIe® x16 board
- PCIe bridge provides unified x16 interface to host
- 13.9 TFLOPS peak single precision floating point
- 0.8 TFLOPS peak double precision floating point
- Support for FP16 floating point ("half precision")
- 8GB HBM[1]
- 1 TB/s memory bandwidth[2]
- 300W TDP
- Dual slot form factor, passive cooling

Massive compute density and efficiency for single precision and half precision workloads



[1] 4GB per GPU

[2] 512GB/s per GPU

AMD FIREPRO™ S9300 X2 GPU VS. COMPETITION

INDUSTRY LEADING COMPUTE PERFORMANCE

	Tesla K80	Tesla M60	FirePro S9300 x2
Peak Single Precision	5.6 TFLOPS	7.4 TFLOPS	13.9 TFLOPS
Performance/watt SPFP	19 GFLOPS/W	25 GFLOPS/W	46 GFLOPS/W
Memory Bandwidth	480GB/s	320GB/s	1024GB/s
Memory Size	2 x 12GB GDDR5	2x 8GB GDDR5	2 x 4GB HBM
Maximum Power	300W	300W	300W
Server compatible form factor	Yes	Yes	Yes

- AMD Advantage: over **2X the compute performance** & over **2X the memory bandwidth**

Nvidia data sources: http://international.download.nvidia.com/pdf/kepler/TeslaK80-datasheet.pdf and http://www.geforce.com/hardware/desktop-gpus/geforce-gtx-titan-x/specifications

AMD FIREPRO™ S-SERIES VALUE PROP FOR HPC

Accessibility

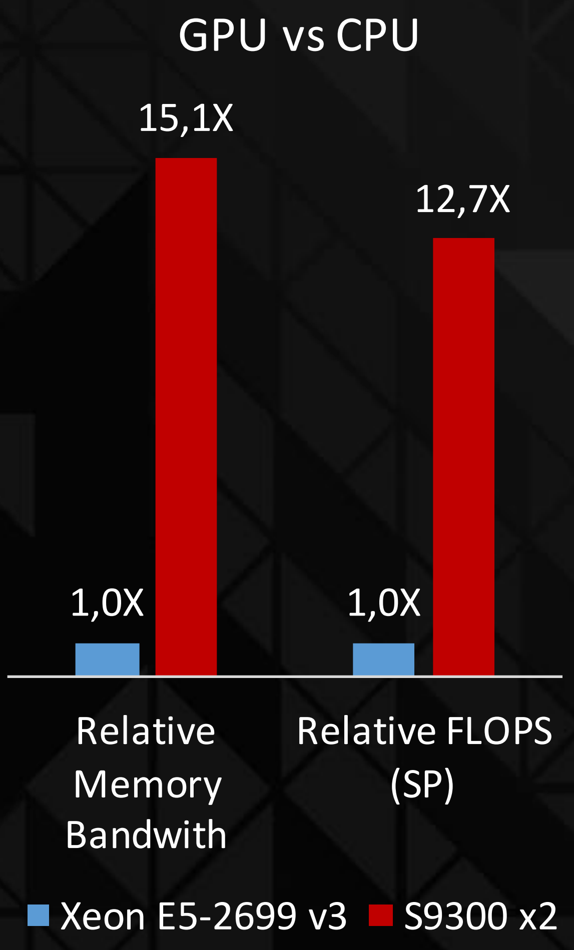
- Standardized programming: OpenCL, C++
- Open source driver and tools
- Code portability

Innovation

- Largest GPU memory
- First with HBM
- 1 TB/s memory bandwidth

Performance

- Leadership in compute performance
- Power efficiency



AMD FIREPRO™ S9300 X2 GPU FOCUS SEGMENTS

- Deep Neural Networks / Machine Learning
- Geoscience
- Molecular Dynamics
- Data processing and Analysis
- Development platforms for Exascale computing

WHAT THEY ARE SAYING…….

"We're very pleased with the AMD FirePro™ compute clusters," said Jean-Yves Blanc, chief IT architect, CGG. "We're also impressed by the 1TB/s memory bandwidth of the AMD FirePro S9300 x2, a board which delivers over 2x the performance of any other server GPU boards on CGG Wave Equation Modeling codes."

"As a believer in open source solutions for parallel computing and high performance clusters, I applaud AMD for its many contributions and ongoing efforts," said Simon McIntosh-Smith, head of the Microelectronics Group, Bristol University. "The combination of the innovative ROCm software as part of the GPUOpen efforts with the 1 TB/s memory bandwidth of the new AMD FirePro™ S9300 x2 Server GPU is creating excitement throughout the research and commercial communities."

HPC CASE STUDY – PARTICLE PHYSICS USING LATTICE QCD

- L-CSC supercomputer in Darmstadt Germany is used for leading edge particle physics research
  - Cluster consists of 160 nodes, each with four AMD FirePro™ S9150 GPUs
  - Total computational power of 3.2 PetaFLOPS
- Workload
  - Lattice Quantum Chromo Dynamics (LQCD) computations
  - Computation requires large sparse vector multiplication
  - Very high demand on memory bandwidth
  - Theoretical results are correlated with experimental data from FAIR (Facility for Anti-Proton and Ion Research)
- Considerations in platform definition
  - Performance – memory bandwidth and floating point computation
  - Operating cost – electrical power a significant portion of TCO
  - Vendor and platform independent software – drove use of OpenCL

Additional details at http://insidehpc.com/2015/04/interview-with-dr-david-rohr/

AMD FIREPRO™ SOLUTIONS FOR COMPUTE

FEATURES AND CAPABILITIES

	S9300 x2	S9170	S9150	S9100
Cooling	Passive	Passive	Passive	Passive
Stream Processors	8192	2816	2816	2560
OpenCL™ Support	Y	Y	Y	Y
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
Total Memory Size	8GB[1]	32 GB	16 GB	12 GB
Total Memory Bandwidth	1024 GB/s[2]	320 GB/sec	320 GB/sec	320 GB/sec
Memory Interface	HBM	512-bit	512-bit	512-bit
Memory ECC	N	Y	Y	Y
PCI Express Bandwidth	32 GB/sec	32 GB/sec	32 GB/sec	32 GB/sec
TDP	300W	275W	235W	225W

[1] 2x4GB

[2] 2x512 GB/s

AMD FIREPRO™ GRAPHICS

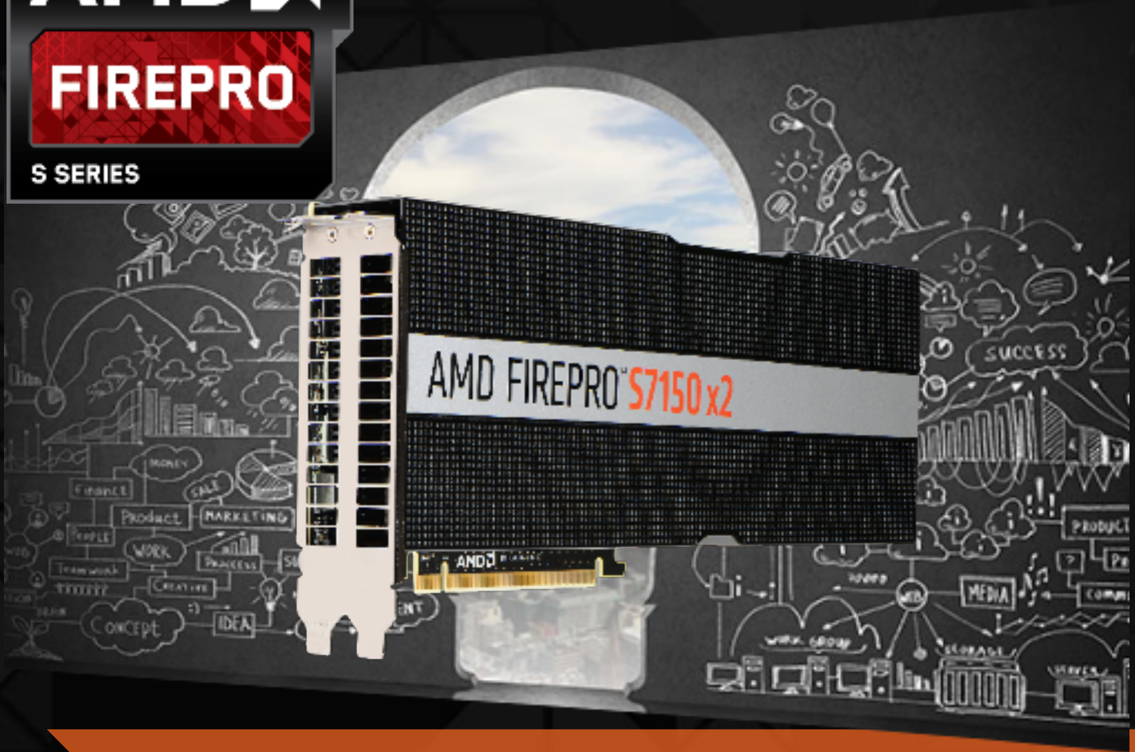
- Professional workstation graphics, VDI and HPC solutions serving multiple enterprise markets – mobile, desktop, server
- Designed for business and technical users who demand the highest quality, reliability and application performance
- Award-winning products and strong ISV partnerships
- Shipping in workstations and servers from Tier 1 OEMs including Apple, Dell, HP, and Lenovo

DELIVERING THE RIGHT AMD FIREPRO™ SOLUTIONS

AMD FIREPRO

W SERIES

Workstations



AMD FIREPRO

S SERIES

Data Centers

USE CASES FOR GPUS IN THE DATACENTER

HIGH PERFORMANCE COMPUTING (HPC) AND VIRTUALIZED DESKTOP INFRASTRUCTURE (VDI)

HPC

- GPU used for computation
- Almost completely Linux® OS
- Multiple GPUs per node (2-16)
- Multiple nodes per site (20-20,000)

VDI

- Deliver high performance graphics to remote users
- Software stack includes hypervisor, guest OS, and remoting protocol
- Multiple GPUs per node (2-8)
- Multiple nodes per site. Easily scalable

Hardware Requirements

- Passive cooling
- Out-of-band temperature monitoring
- Physical size (<10.5") and TDP (<300W) meet server requirements
- No physical display output
- Hardware virtualization

DOUBLE PRECISION COMPUTE PRODUCTS

PERFORMANCE LEADERSHIP, PROVEN TECHNOLOGY

- AMD FirePro™ S9170 GPU
  - 5.2 TFLOPS single precision floating point
  - 2.6 TFLOPS double precision floating point
  - 32GB GDDR5 graphics memory, with ECC
  - Improved power efficiency
  - 275W TDP (with 235W option)
  - Dual slot form factor, passive cooling
- AMD FirePro™ S9150 GPU
  - 16GB GDDR5, with ECC
  - 235W
- AMD FirePro™ S9100 GPU
  - 12GB GDDR5, with ECC
  - 225W

- Ideal for academic clusters, demanding double precision and large memory footprint workloads

WE ARE LOOKING TO BUILD OUT A WORLDWIDE BAND

How to Join the Band

- Get started today developing with ROCm - GPUopen ROCm Getting Started http://bit.ly/1ZTlk82
- Engage In the develop 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**

RADEON
TECHNOLOGIES GROUP

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