The Rise of Open Programming Frameworks

JC BARATAULT

IWOCL May 2015

1,000+ OpenCL projects

SourceForge

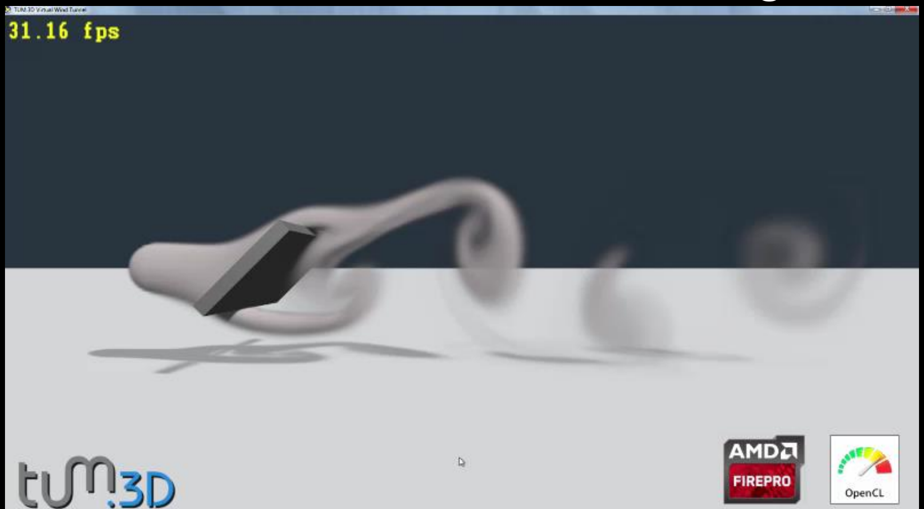
GitHub

Google Code

BitBucket

Interactive GPU Navier-Stokes fluid simulation implemented in OpenCL

1 million fluid cells in a 256x64x64 grid



TUM.3D Virtual Wind Tunnel

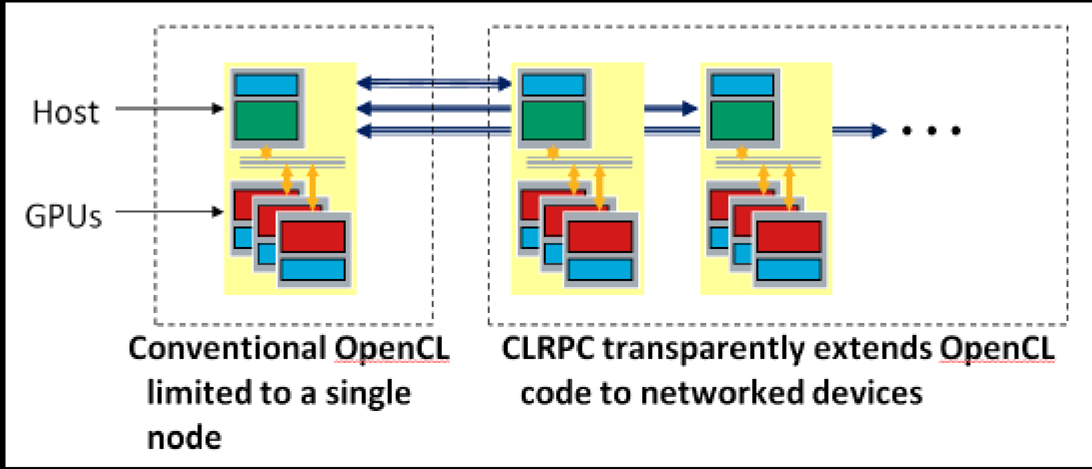
- 10K C++ lines of code, 30 GPU kernels
- CUDA 5.0 to OpenCL 1.2 port in less than a day
- 30 fps with one FirePro S9150
- Multi-GPU & Linux version in June

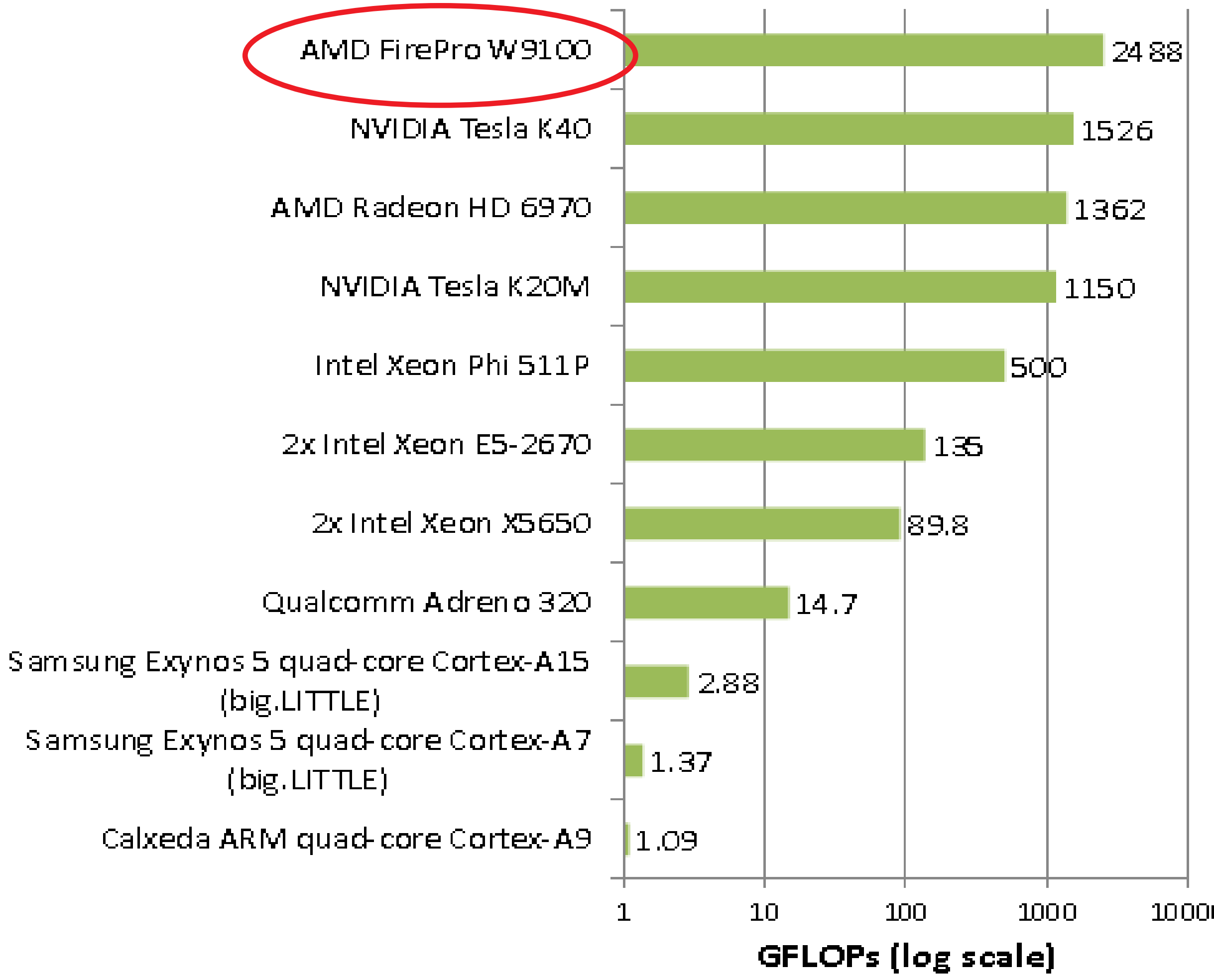
US Army Research Lab

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION IS UNLIMITED

Parallel Programmability on Heterogeneous Architectures

- Explore programming methodologies for the next generation hardware to achieve performance portability in current, emerging, and tomorrow's computational resources





N-body performance on wide range of architectures

DNNs Everywhere

Baidu 百度

Supercomputers	Datacenters	Tablets, smartphones	Wearable devices IoTs
			
1000s GPUs	100k-1m servers	700m (in China)	Billions?

Supercomputer used for training
Trained DNNs then deployed to data centers (cloud), smartphones, and even wearables and IoTs

Dr. Ren WU, BAIDU – "DEEP LEARNING MEETS HETEROGENEOUS COMPUTING"

OpenCL-based Open ECO-SYSTEM

- Diverse industry participation, <u>from cell phones to supercomputers</u>
  - Processor vendors, system OEMs, middleware vendors, application developers.
- OpenCL is the industry standard embraced by many companies.

* Courtesy of Simon McIntosh-Smith and Tom Deakin

Third party names are the property of their owners.

DNN – Anywhere, Anytime

Baidu 百度

- DNN-based image recognition on mobile device
- No connectivity needed
- Real time, directly works on video stream
- Everything is done within the device
- What you point is what you get

- OpenCL based, highly optimized
- Large deep neural network models
- Thousands of objects, flowers, dogs, and bags etc
- Unleashed the full potential of the device hardware

- World's first in-place mobile DNN app?
- And the best!

Dr. Ren WU, BAIDU – "DEEP LEARNING MEETS HETEROGENEOUS COMPUTING"

Open source clBLAS

github.com/clMathLibraries/clBLAS

AMD FirePro S9150

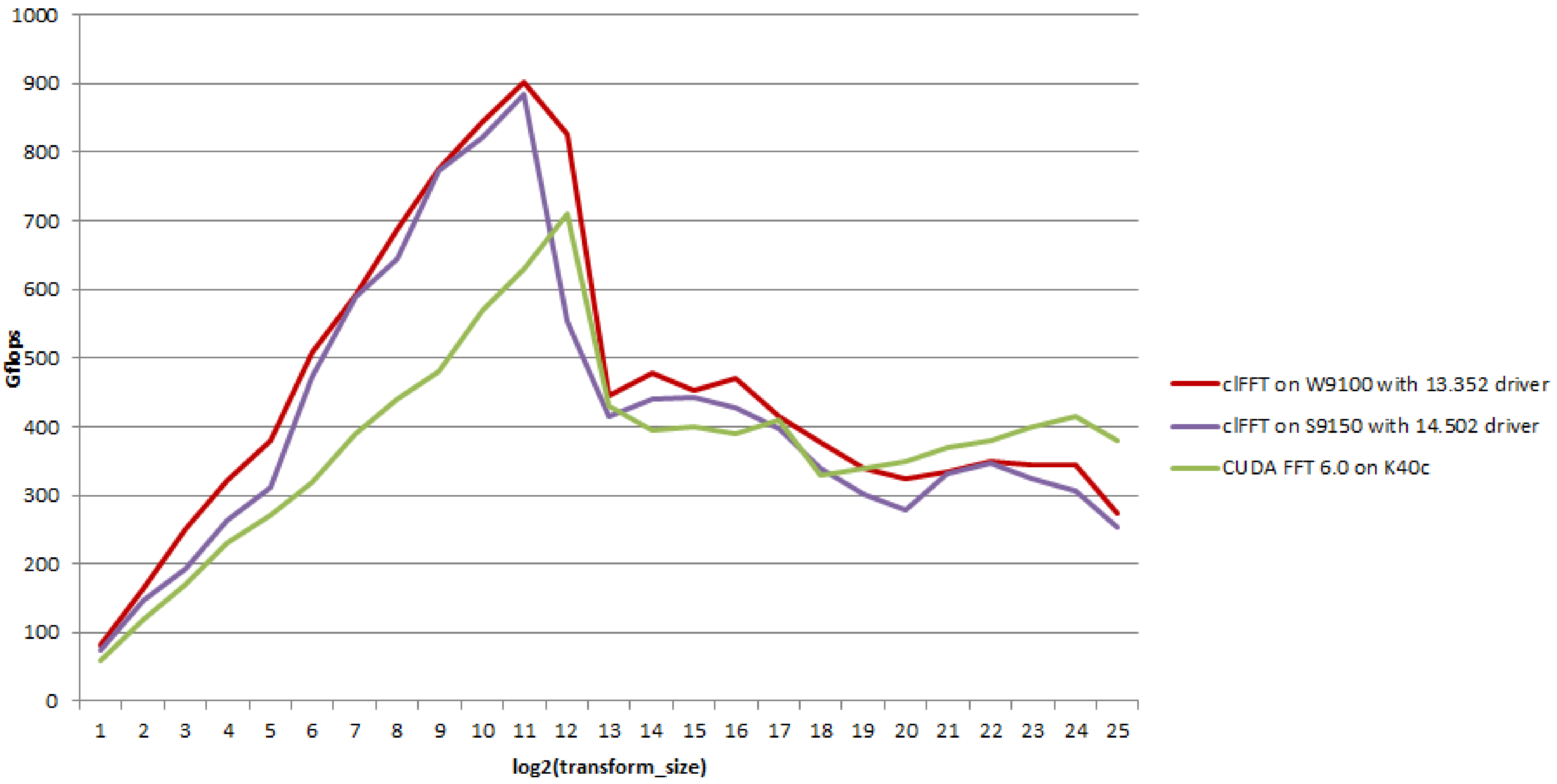
- 16GB GDDR5
- 320 GB/s memory bandwidth
- Full OpenCL + OpenGL

- 4 TFlops SGEMM
- 2 Tflops DGEMM

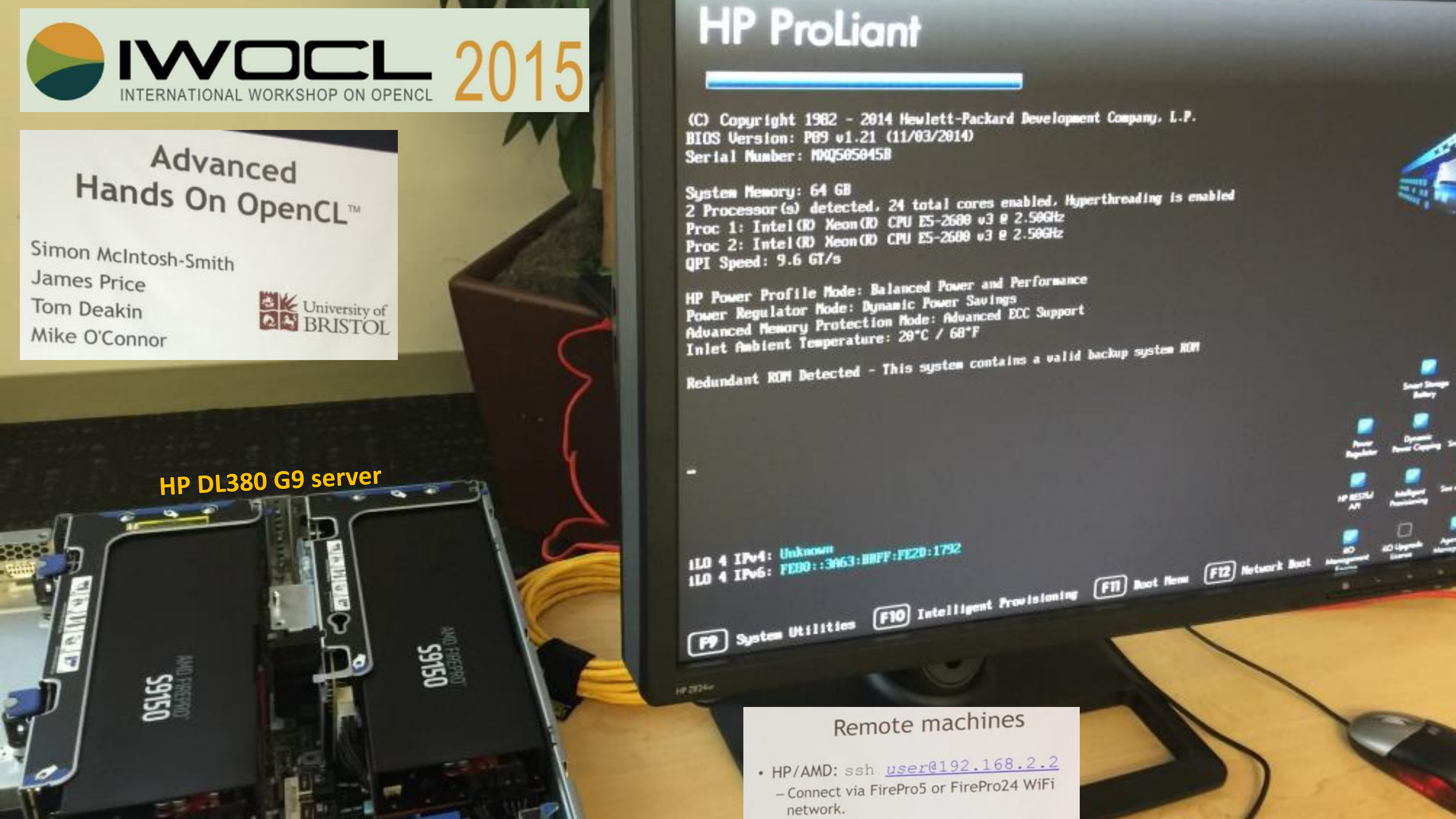
>80% efficiency

#1 GSI L-CSC cluster
600 FirePro S9150
5.27 GFlops/W

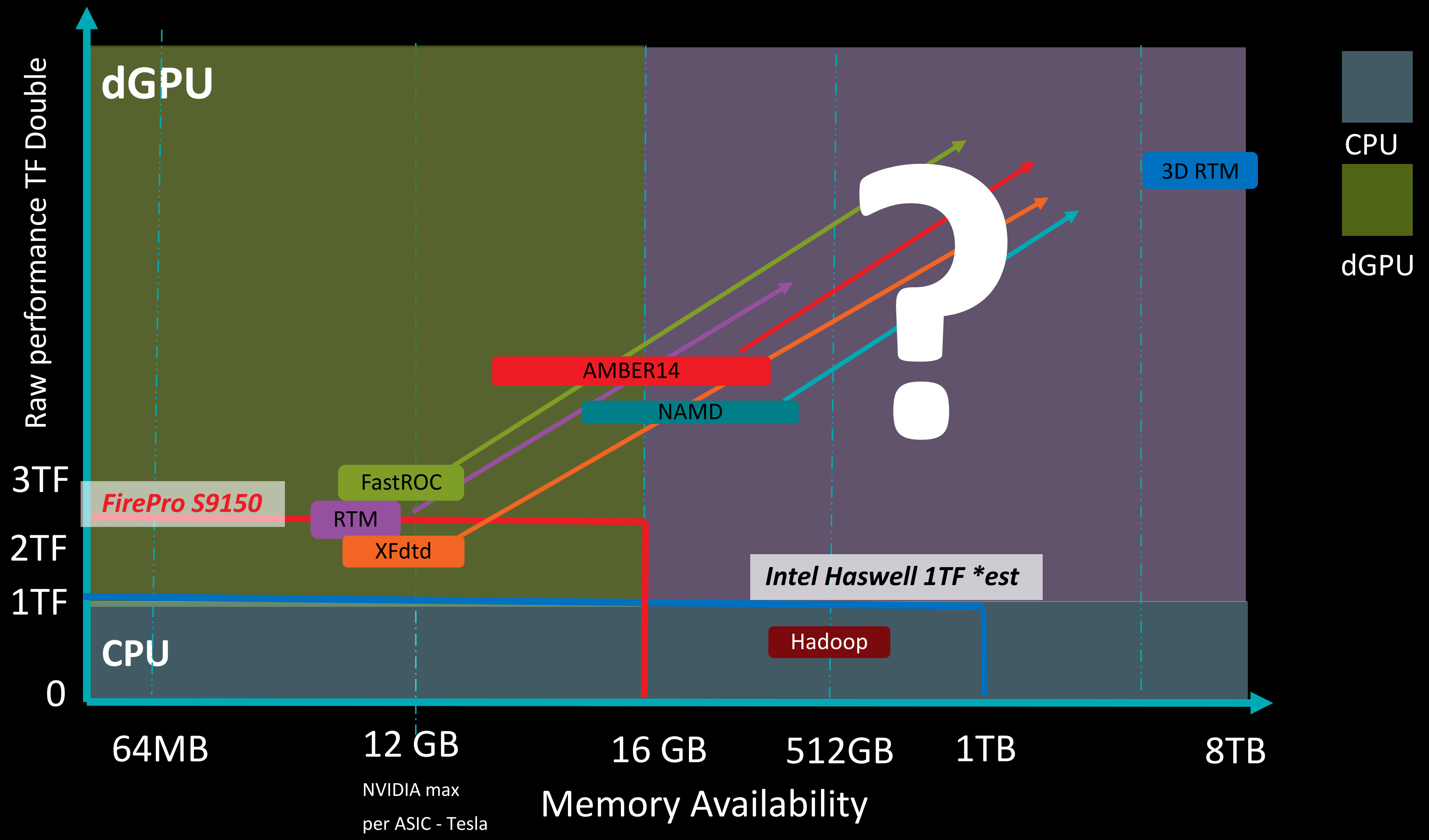
AMD clFFT vs Nvidia cuFFT 6.0
on AMD W9100 & S9150 vs Nvidia K40c
1D single precision complex batched FFTs



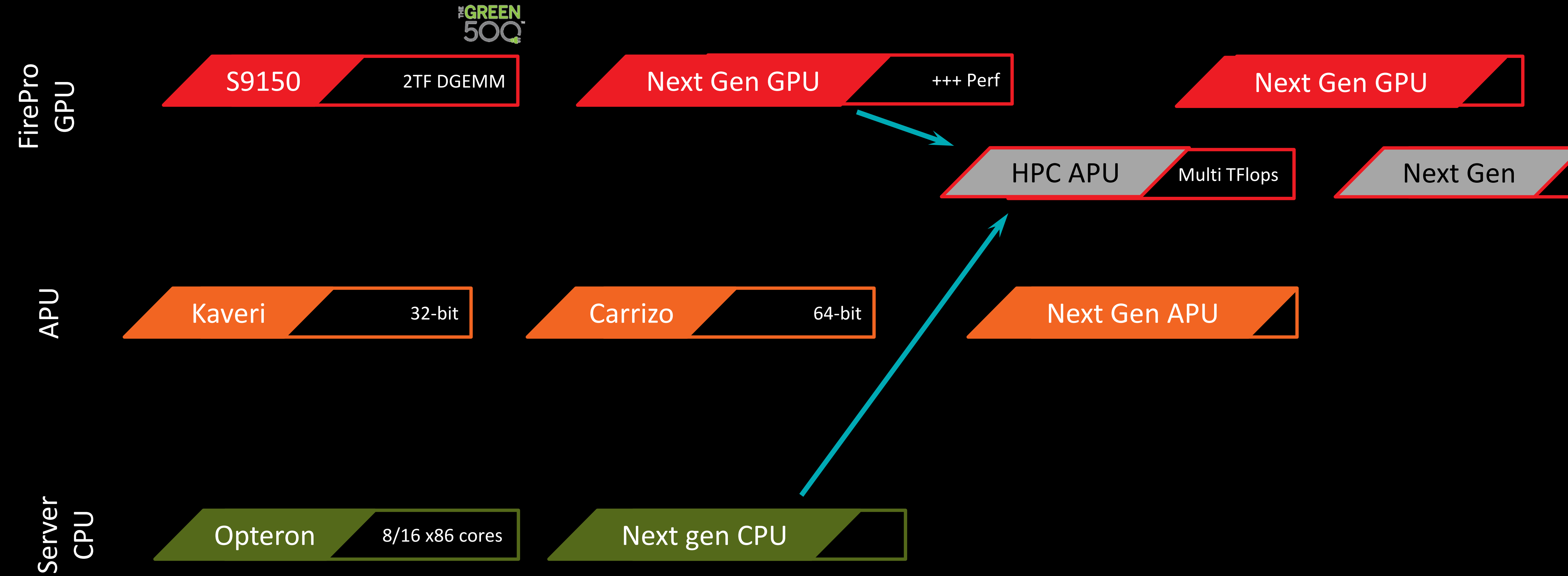
FirePro W9100 for workstation
FirePro S9150 for server


IWOCL 2015
INTERNATIONAL WORKSHOP ON OPENCL
Advanced Hands On OpenCL™
Simon McIntosh-Smith
James Price
Tom Deakin
Mike O'Connor
University of BRISTOL
HP DL380 G9 server
HP ProLiant
(C) Copyright 1982 - 2014 Hewlett-Packard Development Company, L.P.
BIOS Version: P89 v1.21 (11/03/2014)
Serial Number: MXQ5050458
System Memory: 64 GB
2 Processor(s) detected, 24 total cores enabled, Hyperthreading is enabled
Proc 1: Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz
Proc 2: Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz
QPI Speed: 9.6 GT/s
HP Power Profile Mode: Balanced Power and Performance
Power Regulator Mode: Dynamic Power Savings
Advanced Memory Protection Mode: Advanced ECC Support
Inlet Ambient Temperature: 20°C / 68°F
Redundant ROM Detected - This system contains a valid backup system ROM
iLO 4 IPv4: Unknown
iLO 4 IPv6: FE80::3A63:BBFF:FE2D:1792
F9 System Utilities
F10 Intelligent Provisioning
F11 Boot Menu
F12 Network Boot
S9150
AMD FIREPRO
Remote machines
• HP/AMD: ssh user@192.168.2.2
– Connect via FirePro5 or FirePro24 WiFi network.

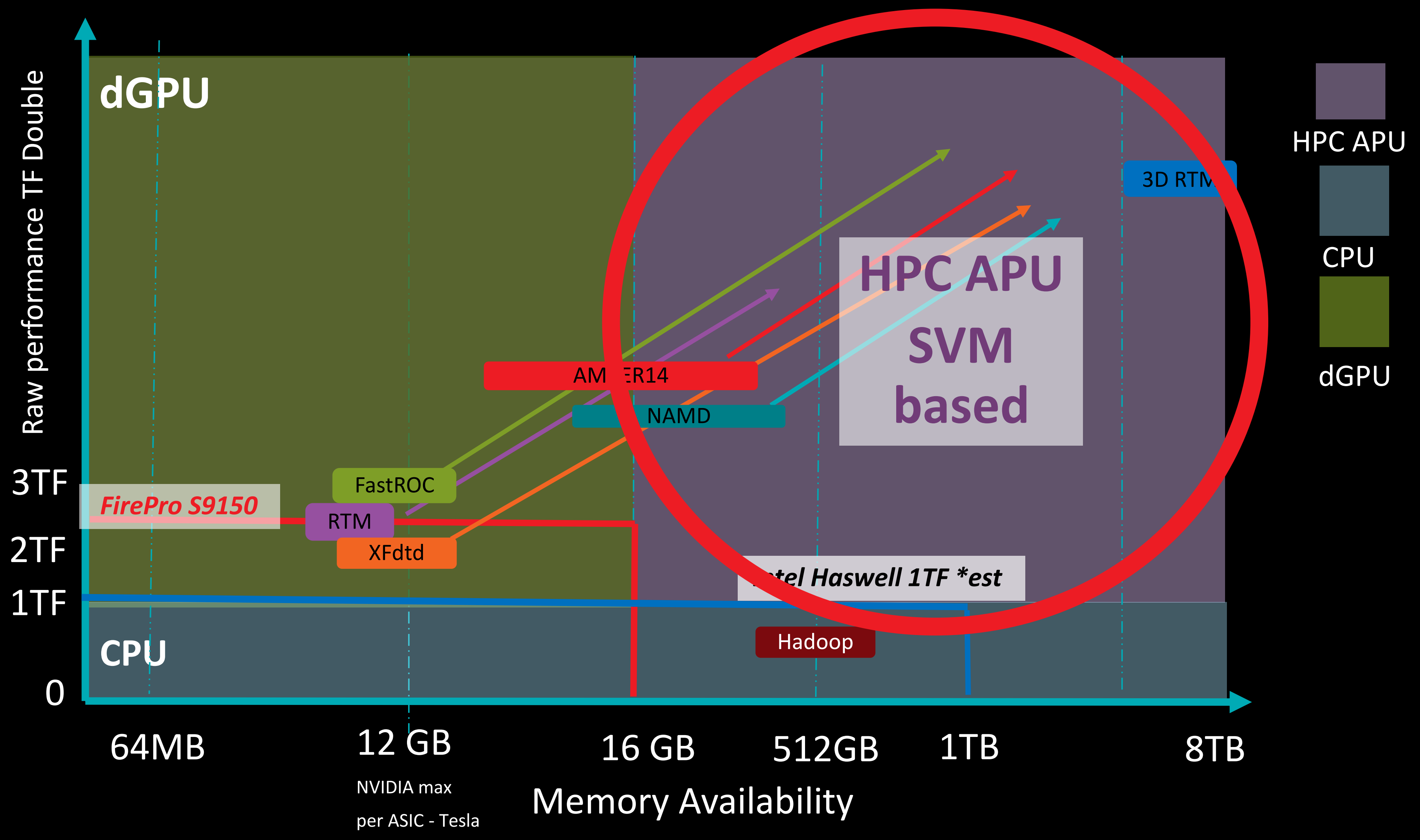
REQUIREMENT: Memory and performance



AMD HPC Roadmap Trends



AMD HPC APU delivers memory and performance



CUSTOMER CENTRIC

Smooth Transition to Heterogeneous Computing

Hardware Agnostic Open Programming Frameworks

C/C++   C++AMP   Fortran

OpenCL 2.0   OpenMP 4.0

Python   Java

Develop your code now for tomorrow's platforms

AMD

- Now — Best performance with FirePro GPU
- Summer 2015 — AMD Carrizo APU x86 64-bit laptop for code testing

THE GREEN 500

Hawaii — 44 GCN CUs

4 TF SGEMM, 2 TF DGEMM GPU

Carrizo — 4 x86 cores, 8 GCN CUs

x86 64-bit APU

HPC APU — Multi TFlops

Multi Tflops x86 64-bit APU

Code portability

OpenCL · HSA Foundation · OpenMP

Shared Virtual Memory on APU vs. PCIe data transfer on dGPU

AMD

APU SVM

```
//CL_MEM_SVM_FINE_GRAIN_BUFFER means host and device can
//concurrently access the buffer, thus no more data
//transfer….
float* Buffer = (float*)clSVMAlloc(ctx, CL_MEM_READ_WRITE |
                                   CL_MEM_SVM_FINE_GRAIN_BUFFER,
                                   1024 * sizeof(float), 0);


//fill the buffer from host, no data transfer
for (int i=0; i<1024; i++)
          Buffer[i] = ….;

// use your SVM buffer in you OpenCL kernel on device
directly
clSetKernelArgSVMPointer(my_kernel, 0, Buffer);


clEnqueueNDRangeKernel(queue, my_kernel,…)
```

dGPU PCIe

```
//create device buffer


cl_mem DeviceBuffer =  clCreateBuffer(ctx,
          CL_MEM_READ_WRITE, 1024*sizeof(float), NULL, &err
);


//create host buffer
float* hostBuffer = new float[1024];
for (int i=0; i<1024; i++)
          hostBuffer ] = ….;
//data transfer happens here
clEnqueueWriteBuffer(queue, DeviceBuffer,… , hostBuffer);



//use our device buffer on device
clSetKernelArg(my_kernel,0,sizeof(cl_mem), &DeviceBuffer );


clEnqueueNDRangeKernel(queue, my_kernel,…)
```

OpenCL 2.0 support in AMD Compute SDK 1.0

AMD

OpenCL 2.0 Core features

- Shared Virtual Memory Coarse grain, Buffer mode
- Device-side enqueue (kernels enqueuing kernels, dynamic parallelism, …)
- C11 atomics
- Generic address space
- Program scoped variables
- Pipes
- Non-uniform workgroups (flexible ND-range)
- sRGB image reads
- Create an Image2D from buffer
- New workgroup built-in functions (all, any, broadcast, reduce, scan)
- Precision for Math built-in native functions

OpenCL 2.0 Optional features APU only

- Shared Virtual Memory Fine Grain Buffer Mode
- Platform Atomics

More info in AMD Blog Series

- http://developer.amd.com/community/blog/category/opencl-2-0/

AMD CodeXL 1.7

AMD

- CodeXL helps SW developers get the best performance on AMD platforms
- Debug, Profile and Analyze applications
  - On local and remote hosts
- Power Profiling
- System level “white box” view
- AMD CPUs, GPUs and APUs
- Multiple platforms and Operating Systems
  - Standalone application for Windows® and Linux®
  - Integrated into Microsoft® Visual Studio®
- Free to download and use

http://developer.amd.com/tools-and-sdks/opencl-zone/codexl/

Debug

Profile

Build & Analyze

Start developing for OpenCL 2.0 today with latest AMD APUs and GPUs

Hawaii — 44 GCN CUs

4 TF SGEMM, 2 TF DGEMM GPU

Carrizo — 4 x86 cores, 8 GCN CUs

x86 64-bit APU

HPC APU — Multi TFlops

Multi Tflops x86 64-bit APU

Code portability

Thank you

DISCLAIMER & ATTRIBUTION

AMD

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