NSF Keeneland GPU Tutorial

Georgia Tech Hotel and Conference Center

800 Spring St. NW Atlanta, GA 30308

April 14-15, 2011

http://keeneland.gatech.edu

sponsored by NSF and the NVIDIA CCOE at Georgia Tech

Thursday 4/14

Start time	Duration	Topic	Speaker
07:30	00:30	*Registration*	Keeneland
08:00	00:15	Welcome and Introduction	Keeneland/Vetter
08:15	00:30	Heterogeneous Architectures and KID Overview	Keeneland
08:45	01:30	Introduction to CUDA	NVIDIA
10:15	00:15	*break*	
10:30	01:00	Introduction to OpenCL	NVIDIA
11:30	01:00	*Lunch (provided)*	
12:30	00:30	SDK Overview incl debugging, perf	NVIDIA
13:00	00:40	Libraries Overview: NVIDIA Libraries	NVIDIA, UTK
13:40	00:20	Libraries Overview: Magma	Keeneland/Tomov
14:00	00:45	Scalable Heterogeneous Computing	Keeneland/Roth
14:45	00:15	*break*	
15:00	00:45	Integrated Kernel Walkthroughs with SHOC	Keeneland/Spafford
15:45	00:30	GPU SW Development with Ocelot	Keeneland/Kerr
16:15	00:45	Performance analysis with Tau	Malony
17:00		*Adjourn*	

Friday 4/15

Start time	Duration	Topic	Speaker
08:00	00:05	*Announcements and Recap*	Vetter
08:05	00:30	Allinea DDT	Allinea/Maples
08:35	01:00	Advanced Topics in CUDA, OpenCL	NVIDIA
09:35	00:15	*break*	
09:50	00:40	Additional tools, compilers, and libraries	Keeneland/Roth
10:30	01:00	Application Case Studies	Keeneland/Spafford
11:30	01:00	*Lunch (provided)*	
12:30	04:10	HANDS ON WITH KEENELAND	Keeneland, NVIDIA Staff
16:40		*Adjourn*	

Logistics

- Continuous breaks
- Lunch in restaurant downstairs (11.30 sharp)
- KID Tokens available at registration desk
- Wireless information available at registration desk
- Think about what application you will analyze on Friday afternoon
- Presentations will be posted on the keeneland website
- Parking
- Evaluation
- This workshop should be useful to you!
  - Please suggest topics, ask questions, talk to presenters during the break

Background and Motivation

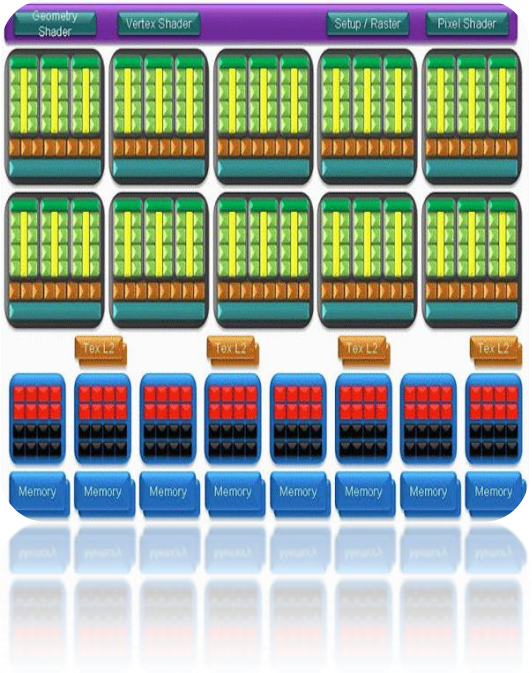
NSF Office of Cyber Infrastructure RFP

- NSF 08-573 OCI Track 2D RFP in Fall 2008
  - Data Intensive
  - Experimental Grid testbed
  - Pool of loosely coupled grid-computing resources
  - Experimental HPC System of Innovative Design

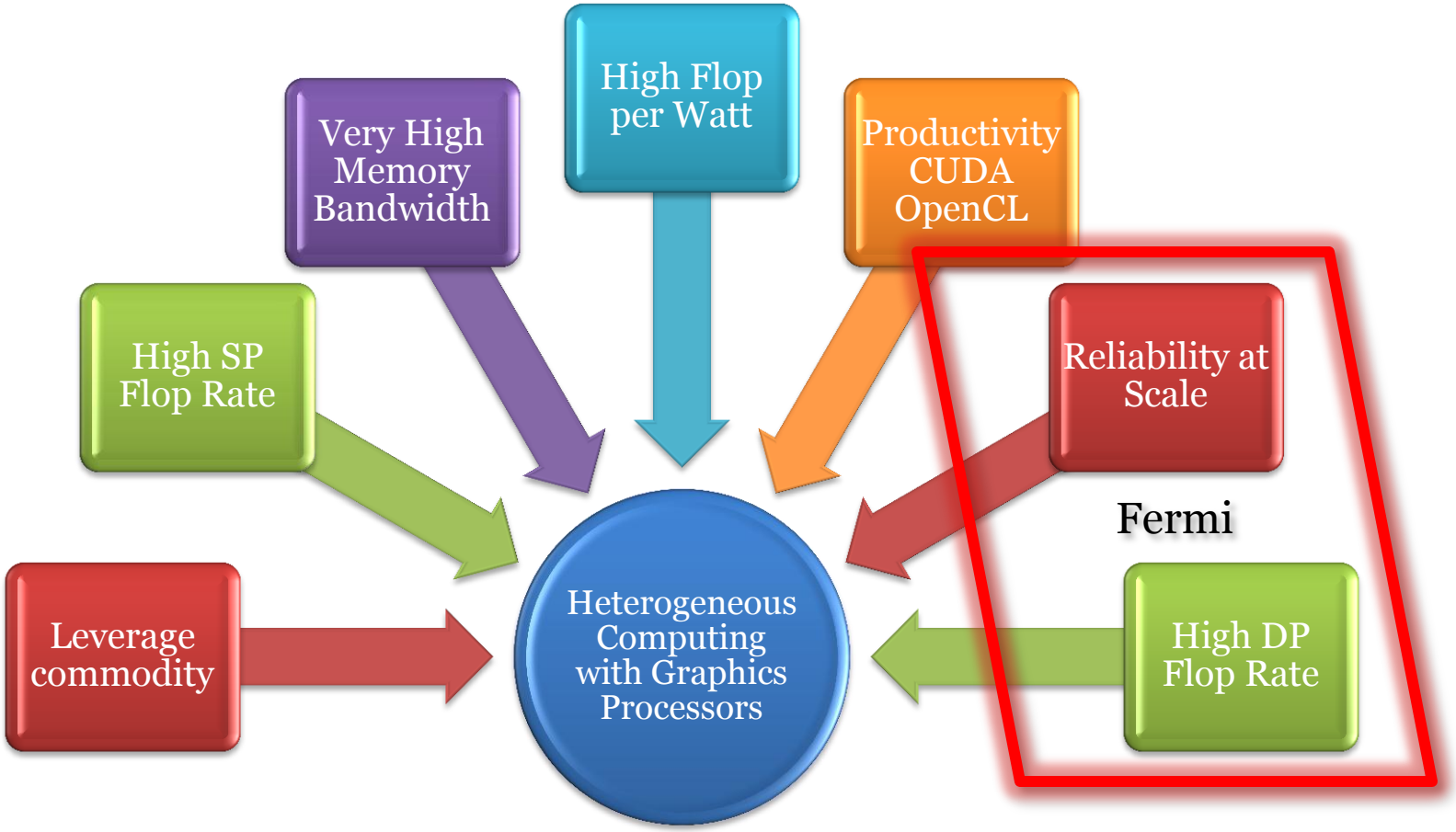
An experimental high-performance computing system of innovative design. Proposals are sought for the development and deployment of a system with an architectural design that is outside the mainstream of what is routinely available from computer vendors. Such a project may be for a duration of up to five years and for a total award size of up to $12,000,000. It is not necessary that the system be deployed early in the project; for example, a lengthy development phase might be included. Proposals should explain why such a resource will expand the range of research projects that scientists and engineers can tackle and include some examples of science and engineering questions to which the system will be applied. It is not necessary that the design of the proposed system be useful for all classes of computational science and engineering problems. When finally deployed, the system should be integrated into the TeraGrid. It is is anticipated that the system, once deployed, will be an experimental TeraGrid resource, used by a smaller number of researchers than is typical for a large TeraGrid resource. (Up to 5 years duration. Up to $12,000,000 in total budget to include development and/or acquisition, operations and maintenance, including user support. First-year budget not to exceed $4,000,000.)

Oct 2008 alternatives analysis for NSF OCI RFP concluded GPUs were a competitive solution

- Success with various applications at DOE, NSF, government, industry
  - Signal processing, image processing, etc.
  - DCA++, S3D, NAMD, many others
- Community application experiences also positive
  - Frequent workshops, tutorials, software development, university classes
  - Many apps teams are excited about using GPGPUs
- Programmability, Resilience?

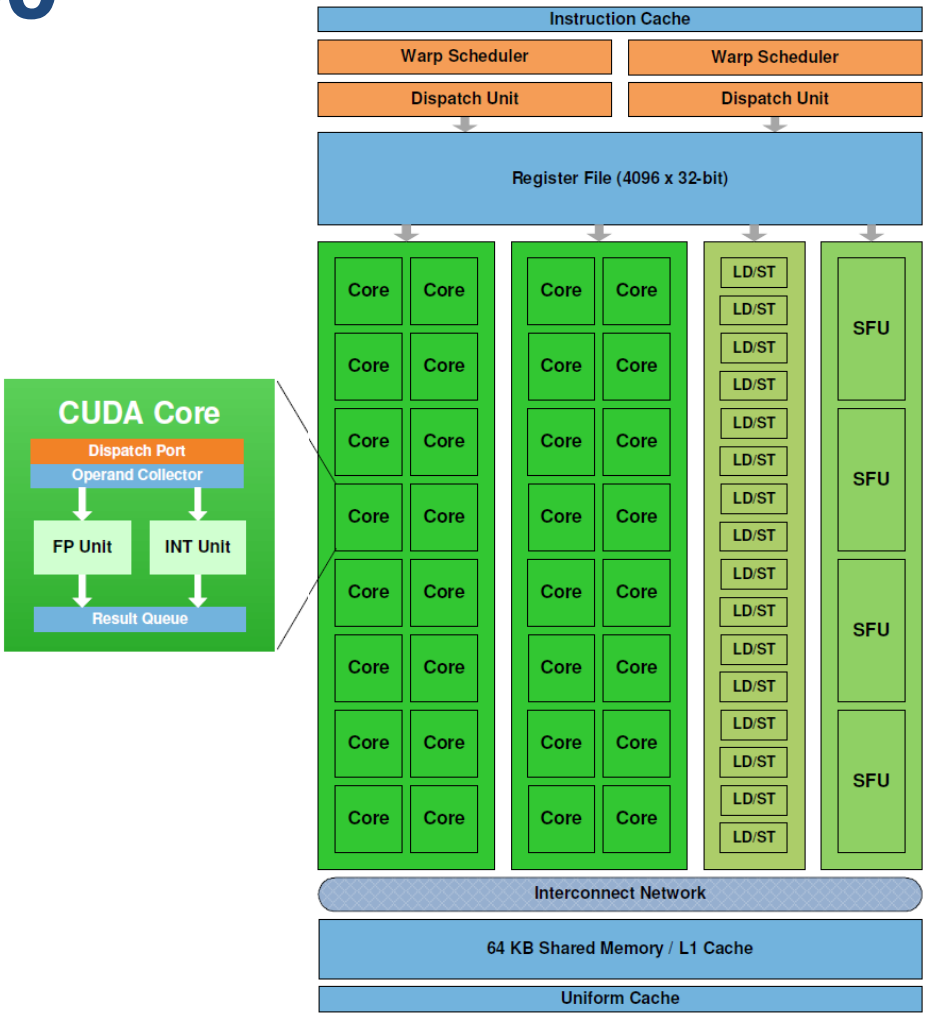


GPU Rationale – What's different now?



NVIDIA Fermi/GF100

- 3B transistors in 40nm
- Up to 512 CUDA Cores
  - New IEEE 754-2008 floating-point standard
    - FMA
    - 8× the peak double precision arithmetic performance over NVIDIA's last generation GPU
  - 32 cores per SM, 21k threads per chip
- 384b GDDR5, 6 GB capacity
  - ~120-144 GB/s memory BW
- C/M2070
  - 515 GigaFLOPS DP, 6GB
  - ECC Register files, L1/L2 caches, shared memory and DRAM



Many GPU-enabled systems blossoming worldwide….

engadget CLASSIC MOBILE HD ALT

DROID X REVIEW | IPHONE 4 REVIEW | NINTENDO 3DS | MICROSOFT AT E3

NEWS HUBS GALLERIES VIDEOS PODCASTS TOPICS REVIEWS

FILED UNDER *Desktops*

Tokyo Institute of Technology announces SSD-packing, 2.39 petaflop supercomputer

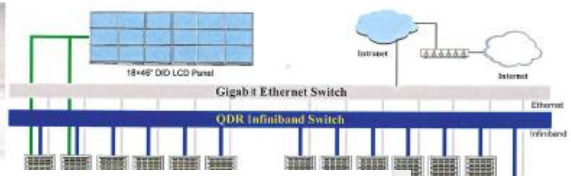
By Joseph L. Flatley posted Jun 23rd 2010 2:06PM

IBM has announced plans to start using SandForce SSDs in i
the Tokyo Institute of Technology is doing one better, work
2.0. This next-gen supercomputer will reportedly operate at
uses a new multilevel storage architecture consisting of DRA
have thirty times the computing capacity of Tsubame 1.0 (d
microprocessors and 4,224 NVIDIA Tesla M2050 GPUs), it
predecessor's. If all goes according to plan, it should be in op
(approx $35.5 million).

Development and application of a HPC system for multi-scale discrete simulation—*Mole-8.5*

Xiaowei Wang, Wei Ge, Xianfeng He, Feiguo Chen, Li Guo, Jinghai Li
Institute of Process Engineering, Chinese Academy of Sciences, Beijing, 100190

Mole-8.5 is the first GPGPU supercomputer (Rpeak of about 1100 Tflops) using NVIDIA Tesla C2050 in the world, which includes 372 nodes and is established in April 2010. It is the successor of the first supercomputer with 1.0 Petaflops peak performance in single precision in China, which was a hybrid system including four units integrating NVIDIA and AMD GPUs announced on April 20, 2009. Mole-8.5 was designed and established by Institute of Process Engineering (IPE), Chinese Academy of Sciences, one of the NVIDIA CCOEs. A designing philosophy utilizing the similarity between hardware, software and the problems to be solved is embodied, based on the multi-scale method and discrete simulation approaches developed at IPE. The whole system is connected with Gigabit Ethernet and QDR Infiniband network. *Mole-8.5* has some unique advantages over the HPC system of same performance based on CPU, for example the high performance/price ratio, the area occupied by it is only about 150 M^2. The linpack result of 320 nodes of Mole-8.5 is 2.073e+05 Gflops with a power consumption of about 480 KWatt, therefore the average power efficiency is 431 Mflop/s/Watt, manifesting an energy efficient supercomputer.



The New York Times Reprints

This copy is for your personal, noncommercial use only. You can order presentation-ready copies for distribution to your colleagues, clients or customers here or use the "Reprints" tool that appears next to any article. Visit www.nytreprints.com for samples and additional information. Order a reprint of this article now.

October 28, 2010

China Wrests Supercomputer Title From U.S.

By ASHLEE VANCE

A Chinese scientific research center has built the fastest supercomputer ever made, replacing the United States as maker of the swiftest machine, and giving China bragging rights as a technology superpower.

The computer, known as Tianhe-1A, has 1.4 times the horsepower of the current top computer, which is at a national laboratory in Tennessee, as measured by the standard test used to gauge how well the systems handle mathematical calculations, said Jack Dongarra, a University of Tennessee computer scientist who maintains the official supercomputer rankings.

Although the official list of the top 500 fastest machines, which comes out every six months, is not due to be completed by Mr. Dongarra until next week, he said the Chinese computer "blows away the existing No. 1 machine." He added, "We don't close the books until Nov. 1, but I would say it is unlikely we will see a system that is faster."

Officials from the Chinese research center, the National University of Defense Technology, are expected to reveal the computer's performance on Thursday at a conference in Beijing. The center says it is "under the dual supervision of the Ministry of National Defense and the Ministry of Education."

The race to build the fastest supercomputer has become a source of national pride as these machines are valued for their ability

d Third

ational University of
lso the World Third fastest
heoretically at its peak

号 meaning River in Sky), which later will be installed in
th "Super computing center"

erformance can reach 1.2 petaFLOPS and highest
he-I" peak performance has exceeded the JUGENE
NPACK score overrun NASA's Pleiades.

Georgia Tech

CUDA By the Numbers:

405	Universities Currently Teaching CUDA
850,000	CUDA Toolkit Downloads
125,000	Active CUDA Developers
110,000	Tesla Nodes shipped in 2010 Academia/Research

Source: NVIDIA

Keeneland Project Overview

Keeneland High Level Goals (in one slide)

- Provide a new class of computing architecture to the NSF community for science
- Acquire, deploy, and operate two GPU clusters
  - Initial delivery – Operational – KID
  - Full scale – Spring 2012
  - Operations, user support
- Software tools, application development support to ensure user productivity and success
- Technology assessment
- Education, Outreach, Training for scientists, students, industry on these new architectures

Keeneland Partners

Georgia Institute of Technology

- Project management
- Acquisition and alternatives assessment
- System software and development tools
- Education, Outreach, Training

National Institute of Computational Sciences

- Operations and TG/XD Integration
- User and Application Support
- Operational Infrastructure
- Education, Outreach, Training

Oak Ridge National Laboratory

- Applications
- Facilities
- Education, Outreach, Training

University of Tennessee, Knoxville

- Scientific Libraries
- Education, Outreach, Training

NVIDIA

- Tesla
- Applications optimizations
- Training

HP

- HPC Host System
- System integration
- Training

Acquisition Highlights

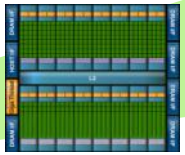
Keeneland – Initial Delivery System

- **Initial Delivery** system installed in Oct/Nov 2010
- 201 TFLOPS in 7 racks (90 sq ft incl service area)
- 677 MFLOPS per watt on HPL (#9 on Green500, Nov 2010)
- Early applications results include Gordon Bell Winner and others

Xeon 5660

67 GFLOPS

M2070



515 GFLOPS

ProLiant SL390s G7 (2CPUs, 3GPUs)

1679 GFLOPS
24/18 GB

Full PCIe X16 bandwidth to all GPUs

S6500 Chassis (4 Nodes)

6718 GFLOPS

Rack (6 Chassis)

40306 GFLOPS

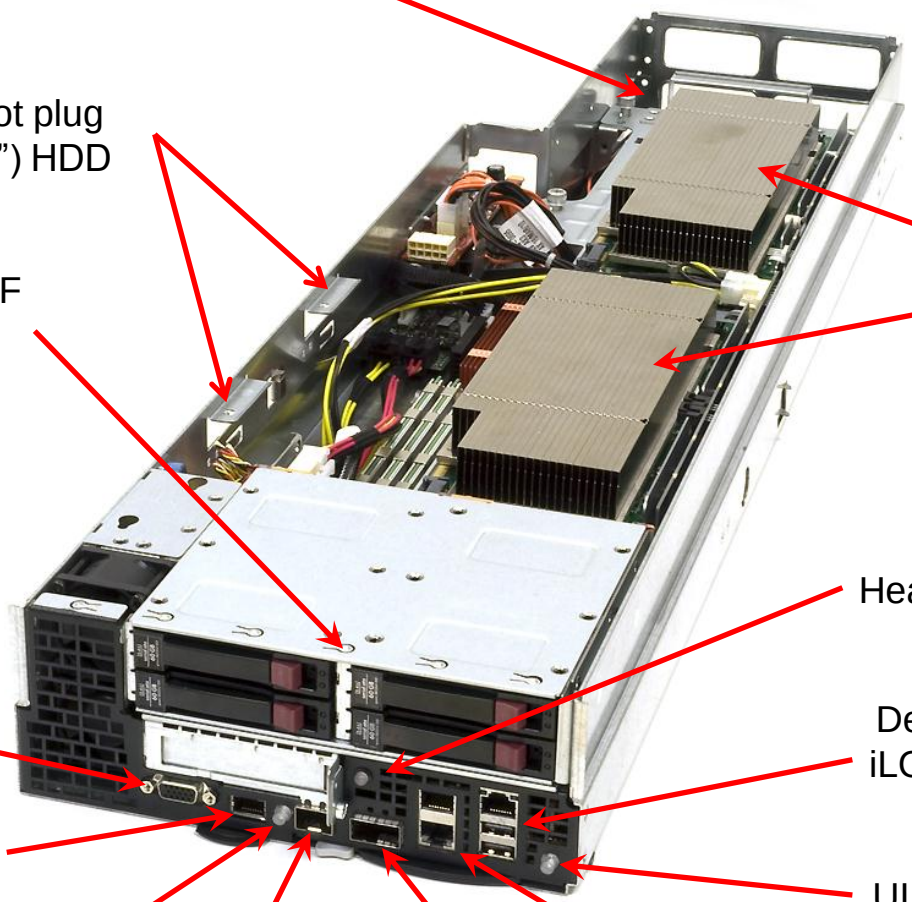
Keeneland System (7 Racks)

201528 GFLOPS

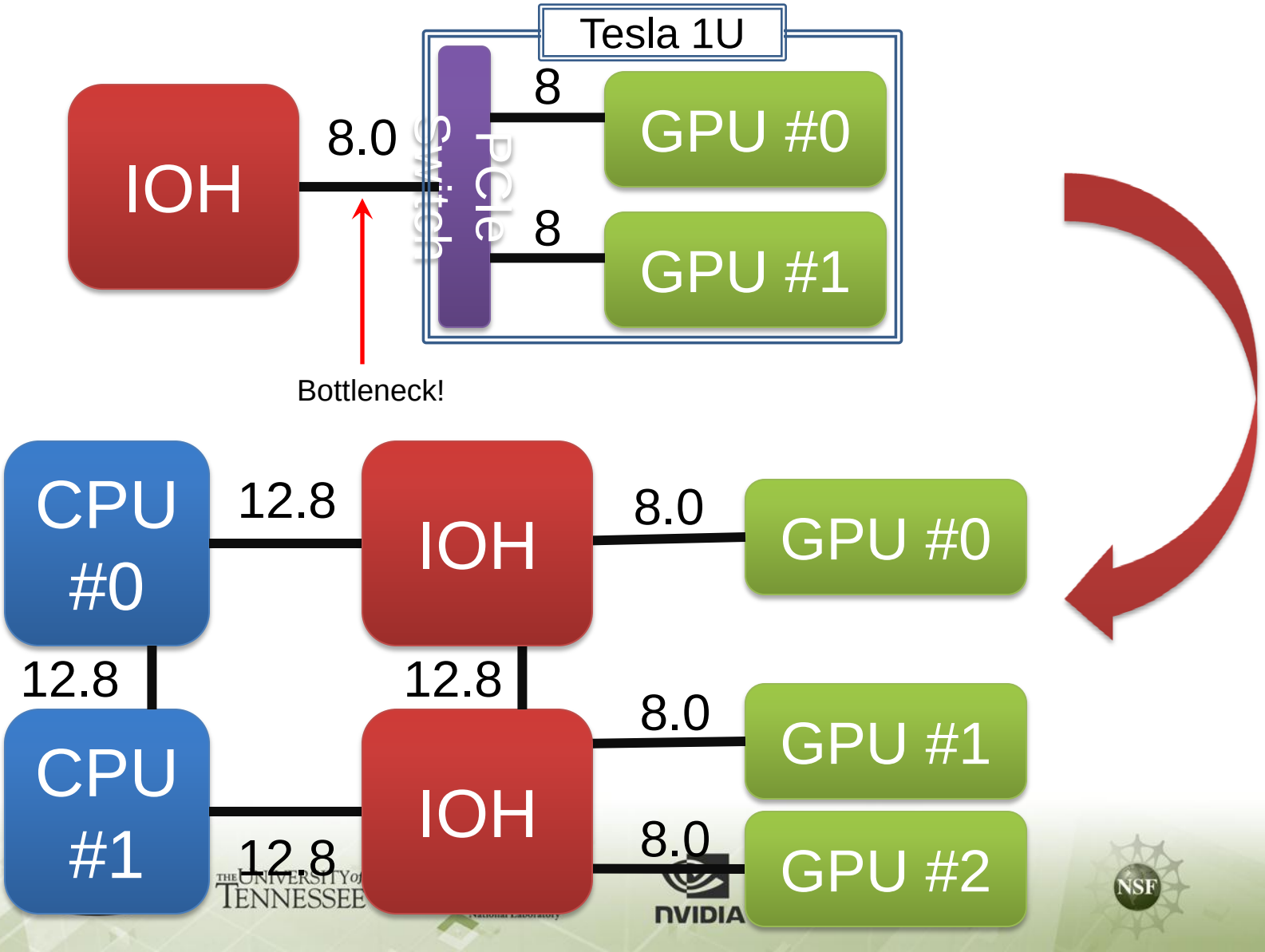
12000-Series Director Switch

Integrated with NICS Datacenter GPFS and TG

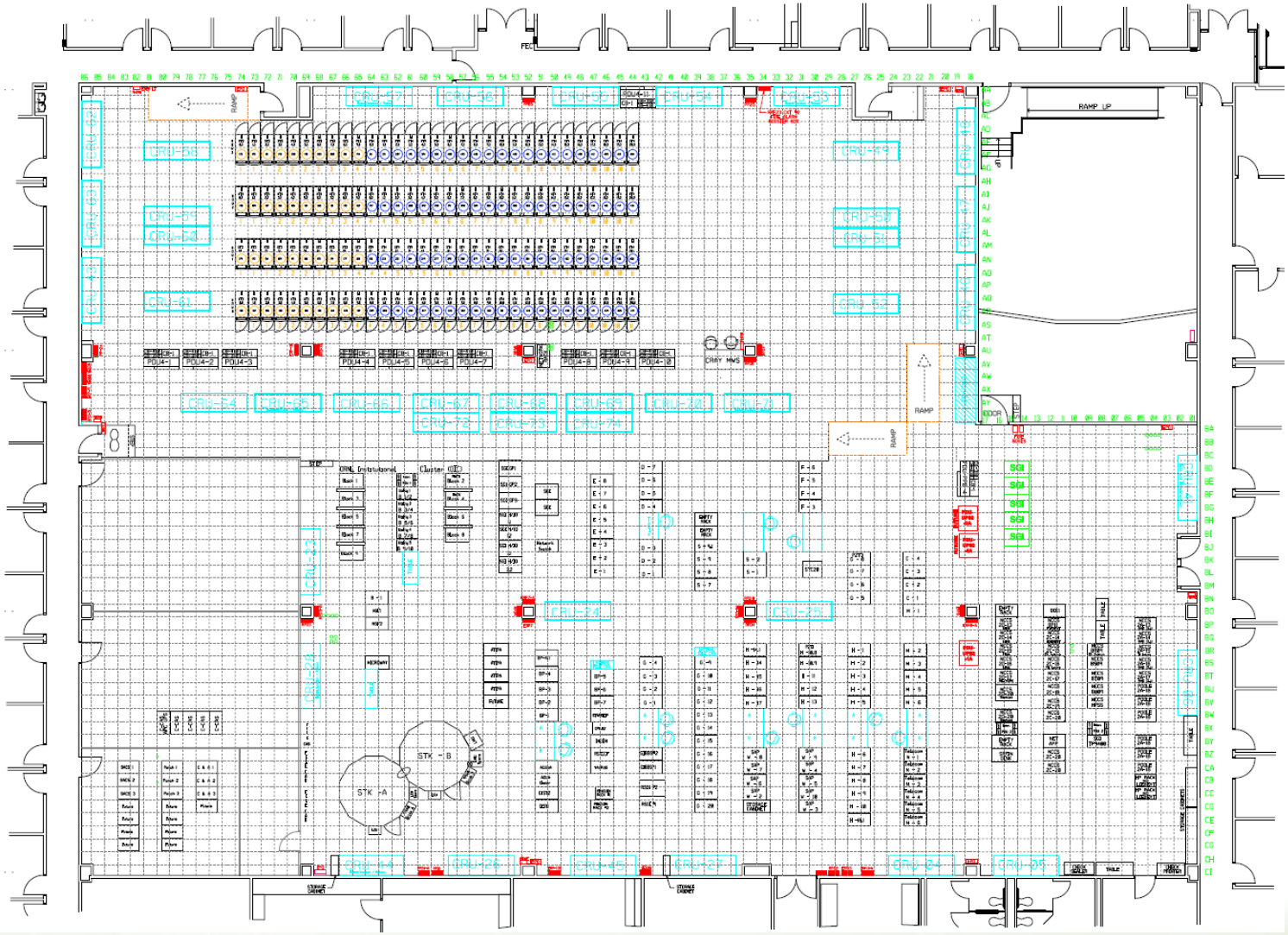
HP ProLiant SL390s G7 2U half width tray



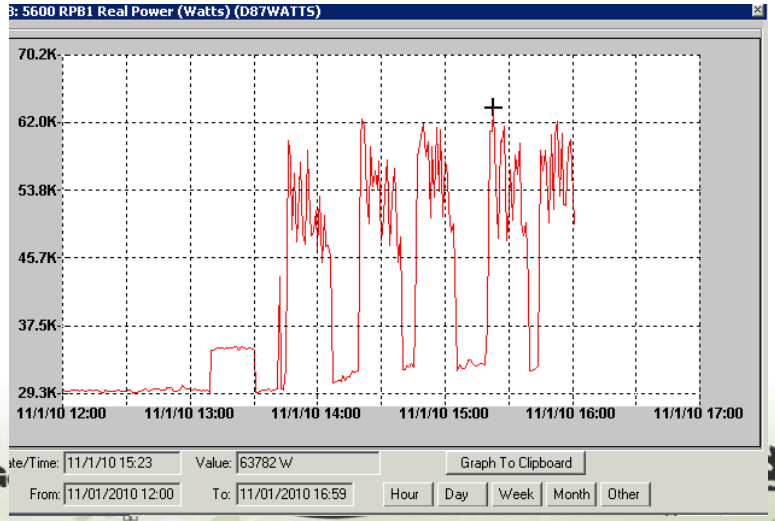
Advantages of dual I/O hub



ID Facilities



State of the Art Facilities and Staff

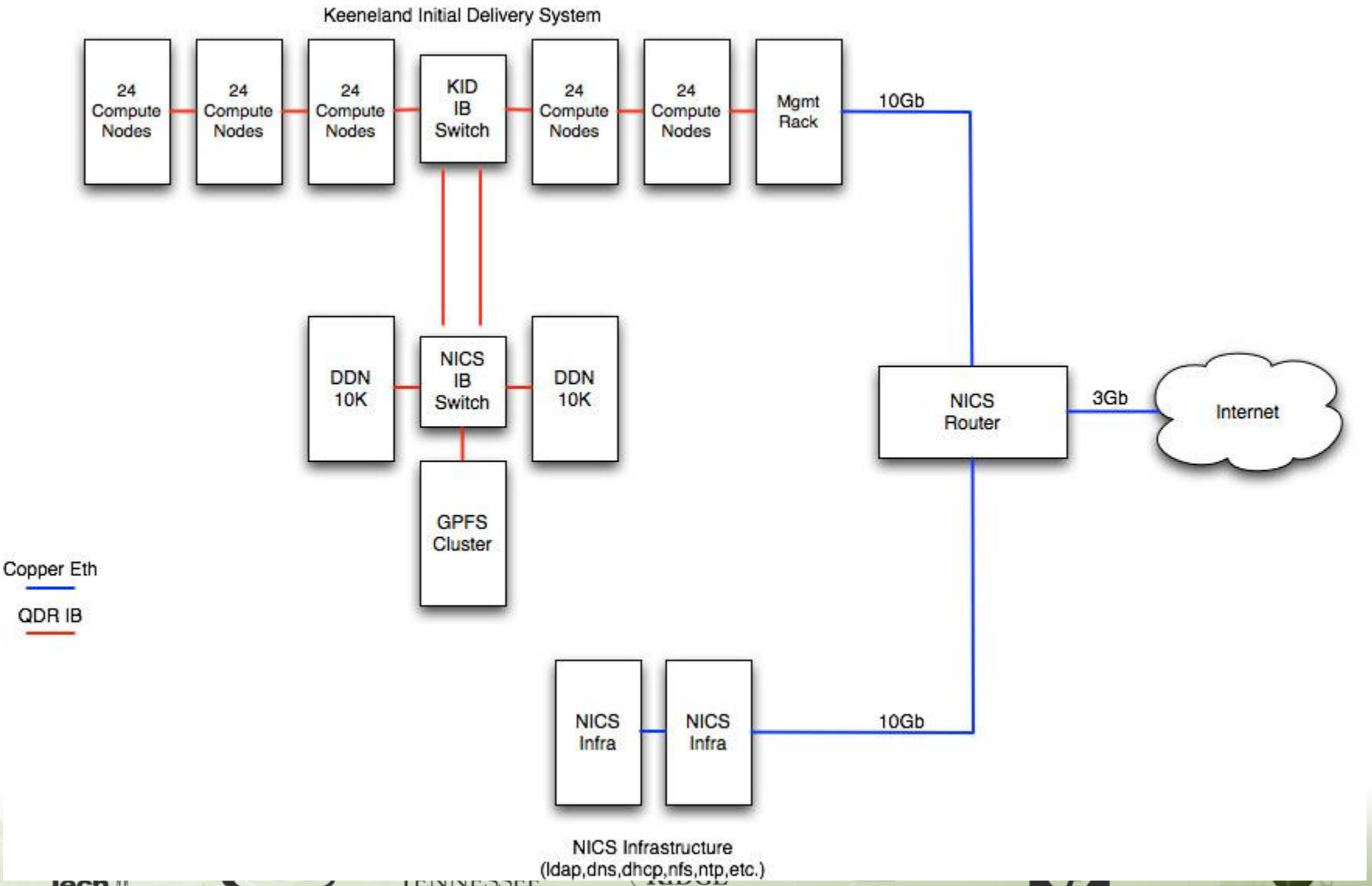


KID Installation

- From the dock to functioning system in 7 days!
  - HP Factory integration and testing prior to delivery contributed to quick uptime
- System delivered on Oct 27
- Installation completed on Oct 29
- Top500, Green500 results completed on Nov 1

Keeneland ID installation – 10/29/10

Integration with NICS Infrastructure



Benchmarks and Applications

The Scalable HeterOgeneous Computing (SHOC) Benchmark Suite

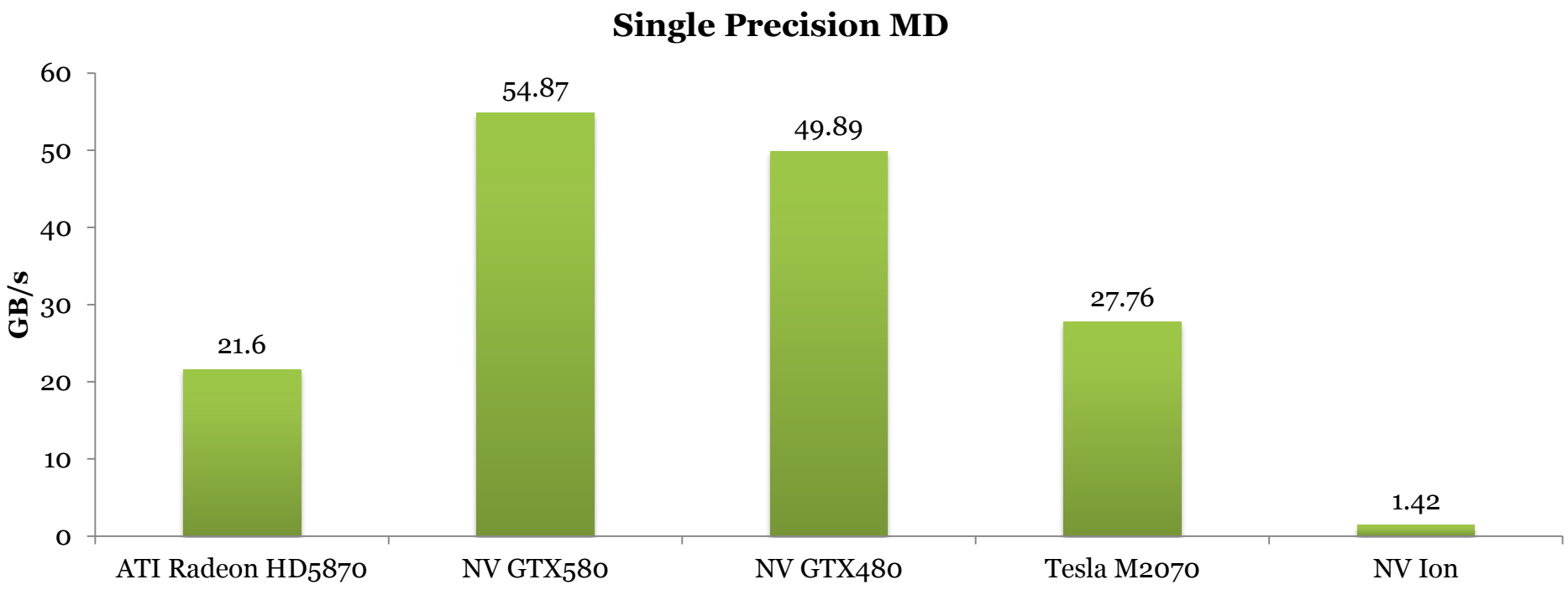
- Benchmark suite with a focus on scientific computing workloads, including common kernels like SGEMM, FFT, Stencils
- Parallelized with MPI, with support for multi-GPU and cluster scale comparisons
- Implemented in CUDA and OpenCL for a 1:1 performance comparison
- Includes stability tests

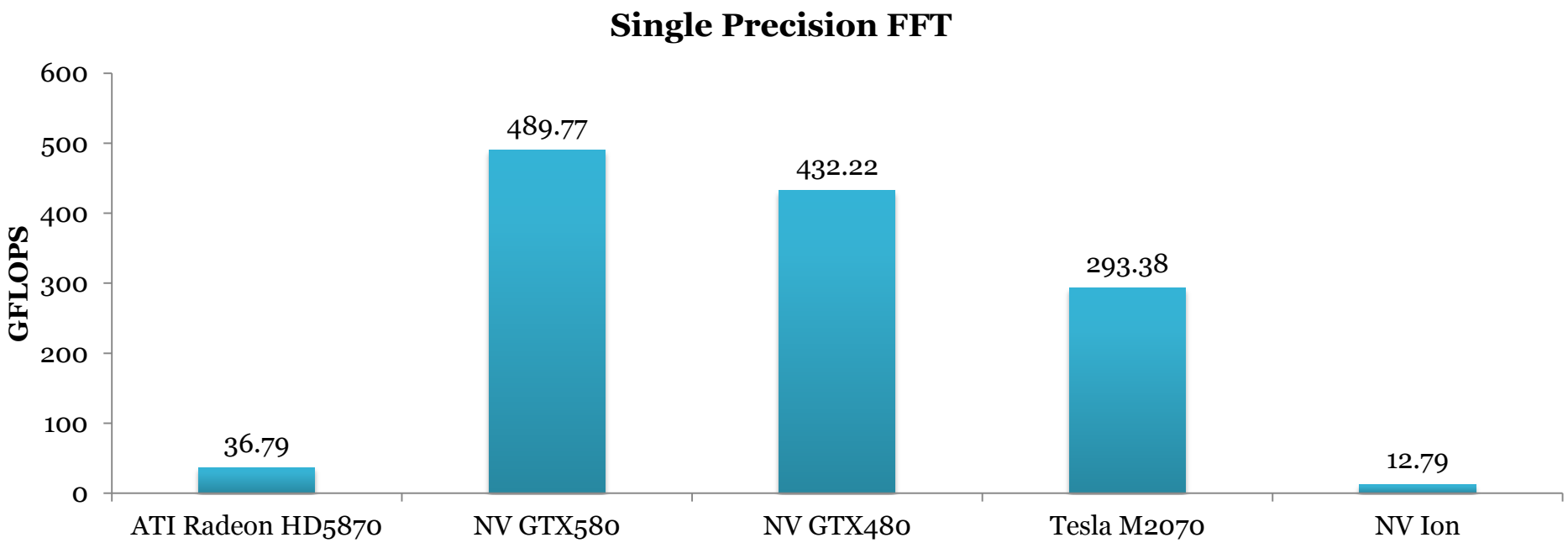
- Level 0
  - **BusSpeedDownload**: measures bandwidth of transferring data across the PCIe bus to a device.
  - **BusSpeedReadback**: measures bandwidth of reading data back from a device.
  - **DeviceMemory**: measures bandwidth of memory access to various types of device memory including global, local, and image memories.
  - **KernelCompile**: measures compile time for several OpenCL kernels, which range in complexity
  - **PeakFlops**: measures maximum achievable floating point performance using a combination of auto-generated and hand coded kernels.
  - **QueueDelay**: measures the overhead of using the OpenCL command queue.
- Level 1
  - **FFT**: forward and reverse 1D FFT.
  - **MD**: computation of the Lennard-Jones potential from molecular dynamics, a specific case of the nbody problem.
  - **Reduction**: reduction operation on an array of single precision floating point values.
  - **SGEMM**: single-precision matrix-matrix multiply.
  - **Scan**: scan (also known as parallel prefix sum) on an array of single precision floating point values.
  - **Sort**: sorts an array of key-value pairs using a radix sort algorithm
  - **Stencil2D**: a 9-point stencil operation applied to a 2D data set. In the MPI version, data is distributed across MPI processes organized in a 2D Cartesian topology, with periodic halo exchanges.
  - **Triad**: STREAM Triad operations, implemented in OpenCL.

SHOC 1.0 available
Downloaded over 1000 times

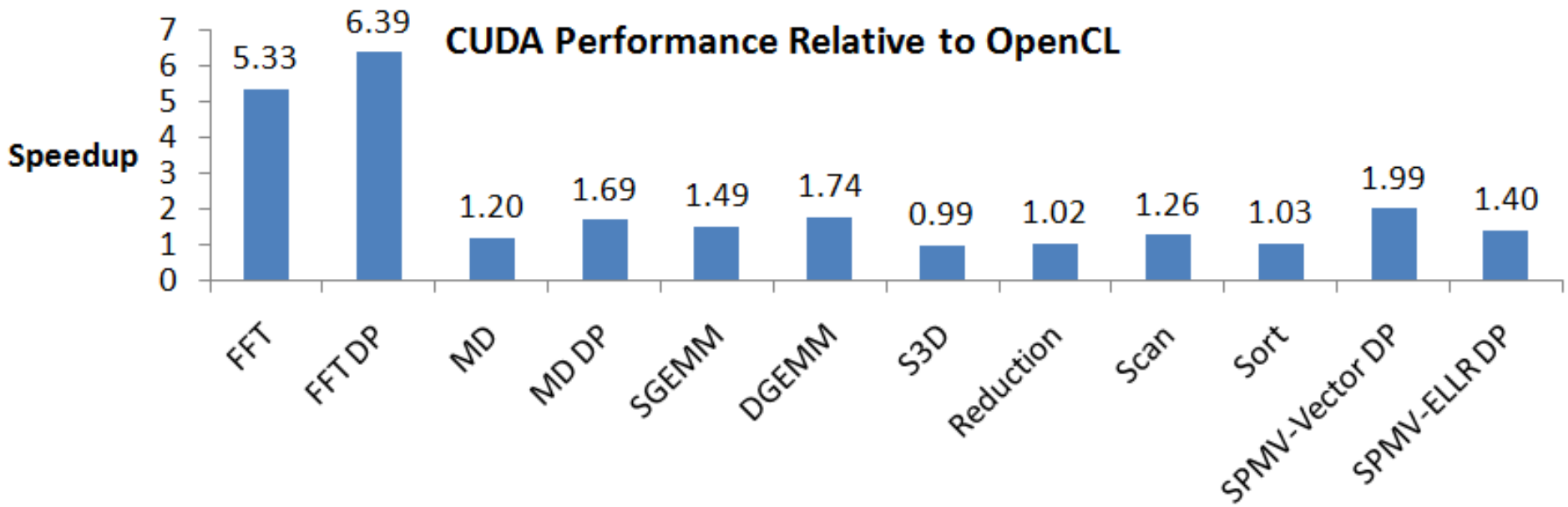
A. Danalis, G. Marin, C. McCurdy, J. Meredith, P.C. Roth, K. Spafford, V. Tipparaju, and J.S. Vetter, "The Scalable HeterOgeneous Computing (SHOC) Benchmark Suite," in Third Workshop on General-Purpose Computation on Graphics Processors (GPGPU 2010)`. Pittsburgh, 2010

Software available at http://bit.ly/shocmarx


Single Precision MD
GB/s
60
50
40
30
20
10
0
21.6
54.87
49.89
27.76
1.42
ATI Radeon HD5870
NV GTX580
NV GTX480
Tesla M2070
NV Ion


Single Precision FFT
GFLOPS
600
500
400
300
200
100
0
36.79
489.77
432.22
293.38
12.79
ATI Radeon HD5870
NV GTX580
NV GTX480
Tesla M2070
NV Ion

Comparing CUDA and OpenCL



Early Success Stories

Computational Materials

- Quantum Monte Carlo
  - High-temperature superconductivity and other materials science
  - 2008 Gordon Bell Prize
- GPU acceleration speedup of 19x in main QMC Update routine
  - Single precision for CPU and GPU: target single-precision only cards
- Full parallel app is 5x faster, start to finish, on a GPU-enabled cluster on Tesla T10

GPU study: J.S. Meredith, G. Alvarez, T.A. Maier, T.C. Schulthess, J.S. Vetter, "Accuracy and Performance of Graphics Processors: A Quantum Monte Carlo Application Case Study", *Parallel Comput., 35(3):151-63, 2009.*

Accuracy study: G. Alvarez, M.S. Summers, D.E. Maxwell, M. Eisenbach, J.S. Meredith, J. M. Larkin, J. Levesque, T. A. Maier, P.R.C. Kent, E.F. D'Azevedo, T.C. Schulthess, "New algorithm to enable 400+ TFlop/s sustained performance in simulations of disorder effects in high-Tc superconductors", SuperComputing, 2008. [*Gordon Bell Prize winner*]

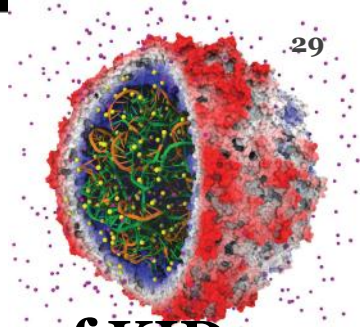
Combustion

- S3D
  - Massively parallel direct numerical solver (DNS) for the full compressible Navier-Stokes, total energy, species and mass continuity equations
  - Coupled with detailed chemistry
  - Scales to 150k cores on Jaguar
- Accelerated version of S3D's Getrates kernel in CUDA on Tesla T10
  - 31.4x SP speedup
  - 16.2x DP speedup

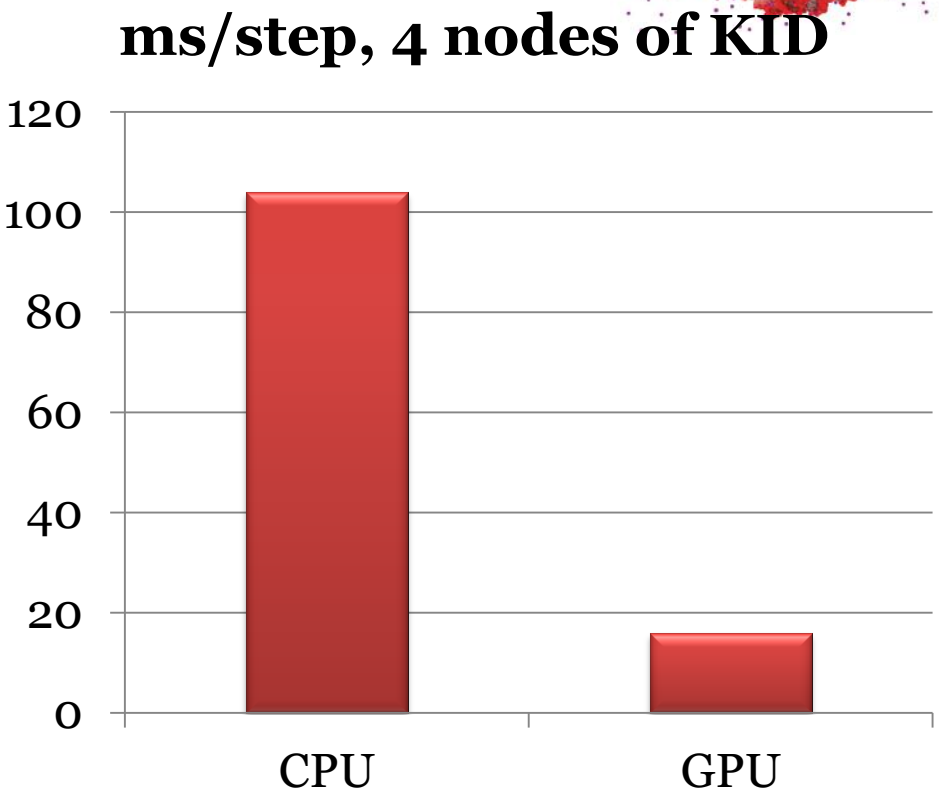
K. Spafford, J. Meredith, J. S. Vetter, J. Chen, R. Grout, and R. Sankaran. Accelerating S3D: A GPGPU Case Study. Proceedings of the Seventh International Workshop on Algorithms, Models, and Tools for Parallel Computing on Heterogeneous Platforms (HeteroPar 2009) Delft, The Netherlands.

Preliminary results from KID

NAMD



- NAMD, VMD
  - Study of the structure and function of biological molecules
- Calculation of non-bonded forces on GPUs leads to 6x on FERMI
- Framework hides most of the GPU complexity from users

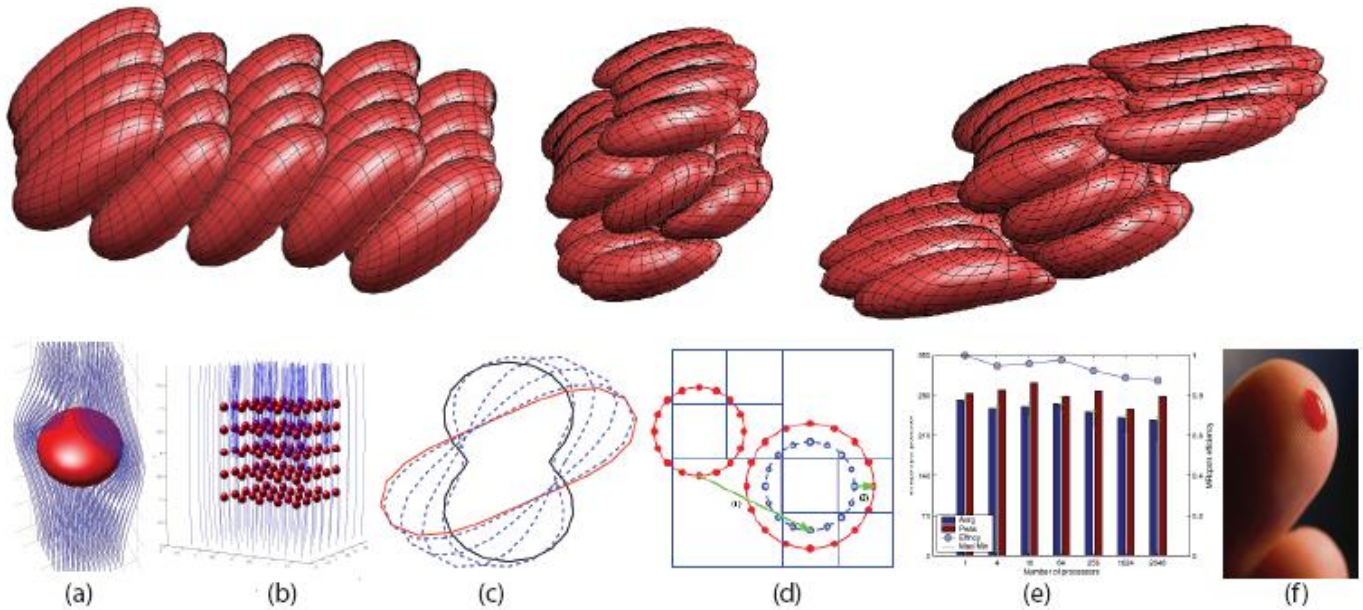


J.C. Phillips and J.E. Stone, "Probing biomolecular machines with graphics processors," *Commun. ACM, 52(10):34-41, 2009.*

Gordon Bell Award, Nov 2010

Simulating Blood Flow with FMM

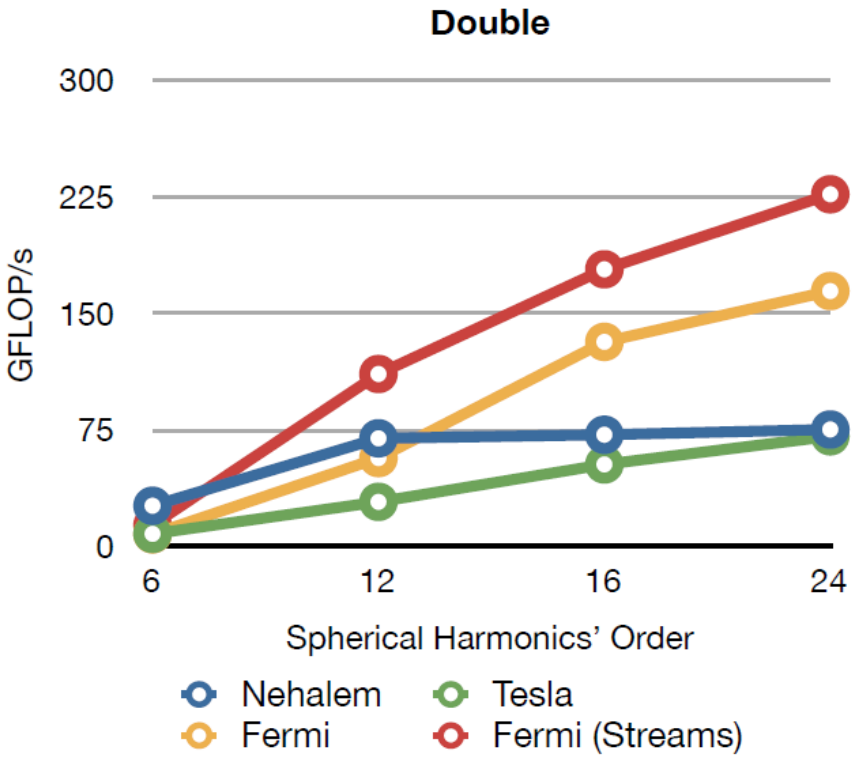
- Multiphysics particle flow of deformable cells in viscous fluid with non-uniform distribution

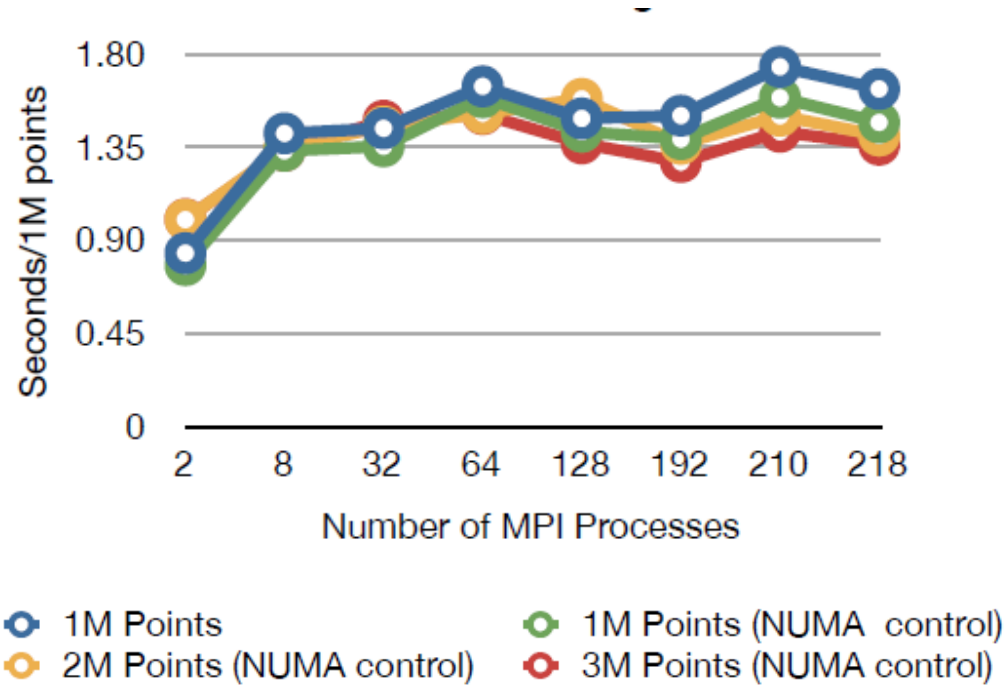


A. Rahimian, I. Lashuk, S. Veerapaneni *et al.*, *"Petascale Direct Numerical Simulation of Blood Flow on 200K Cores and Heterogeneous Architectures (Gordon Bell Winner)," in SC10: International Conf. High Performance Computing, Networking, Storage, and Analysis. New Orleans: ACM/IEEE, 2010.*

Results from KID

FMM results from KID

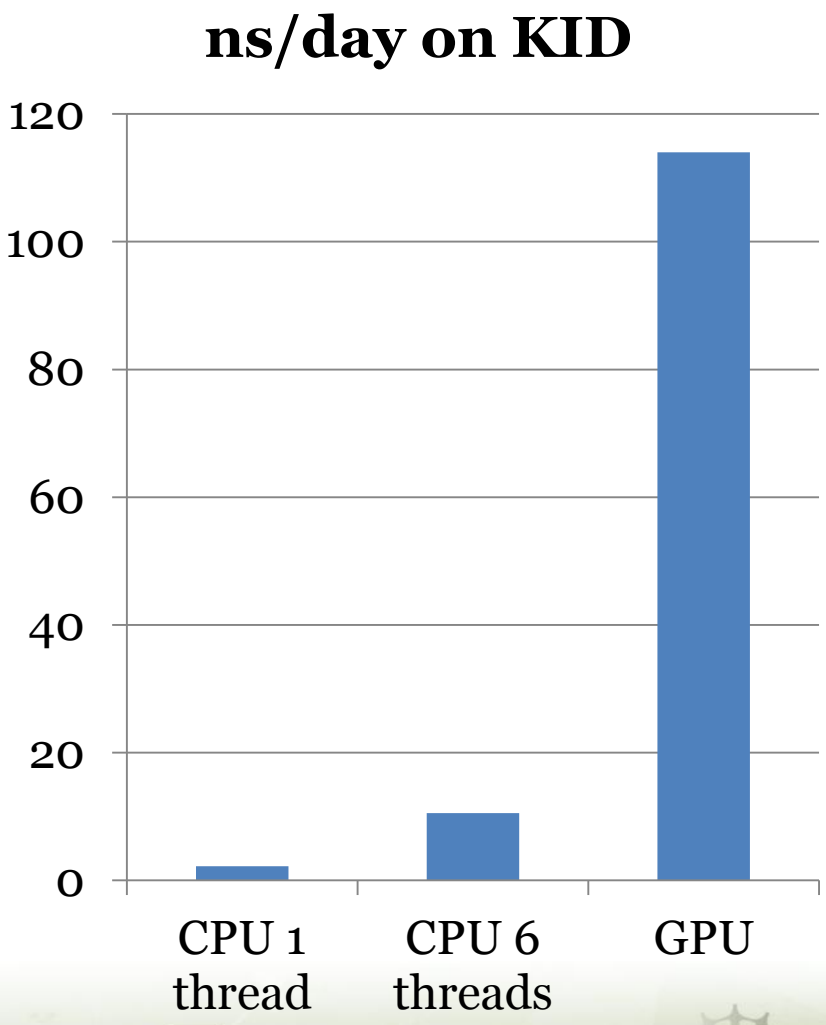




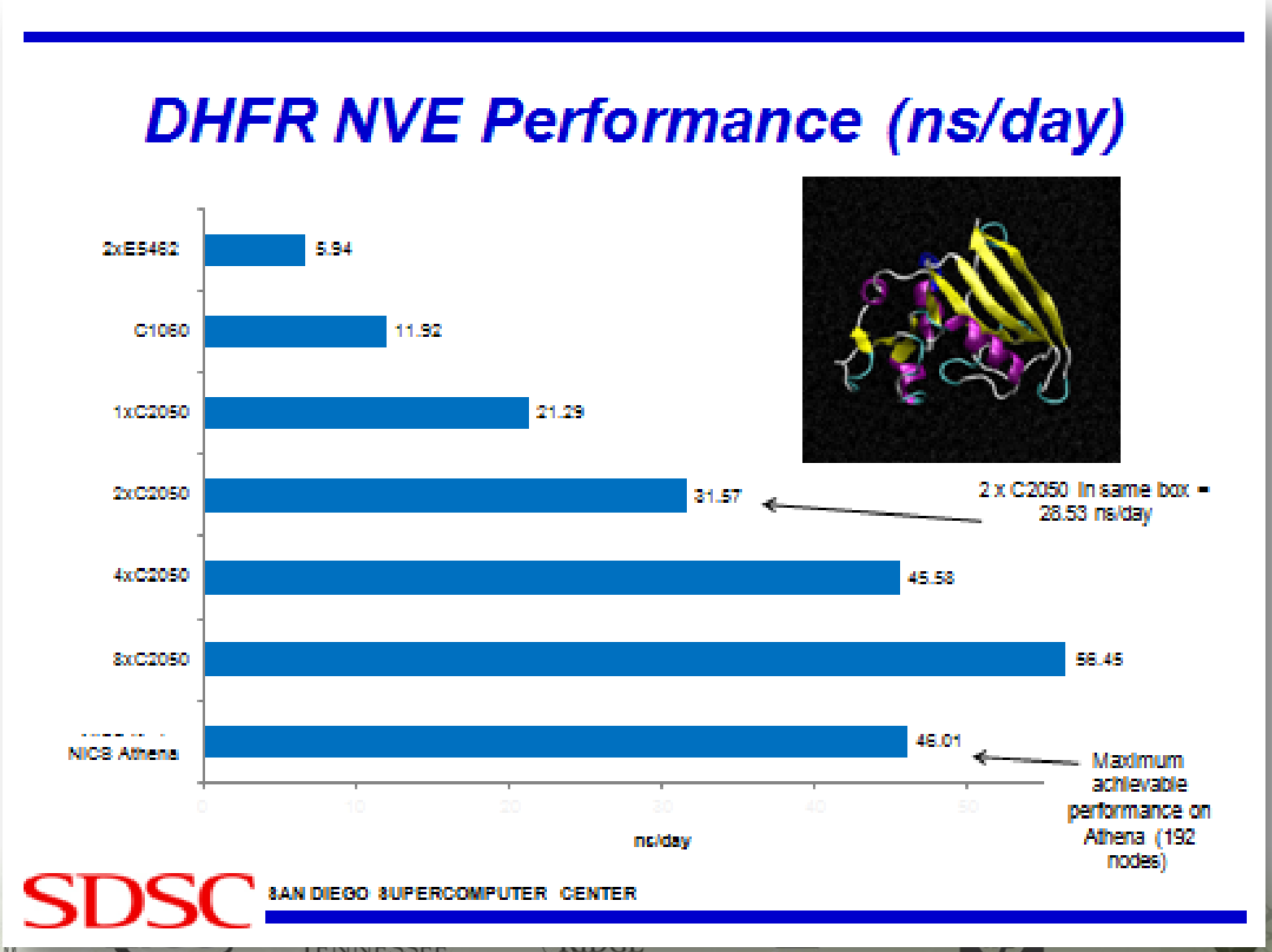
Preliminary results from KID

GROMACS

- GROMACS (GROningen MAchine for Chemical Simulations) is a molecular dynamics simulation package



AMBER on FERMI (courtesy R. Walker, D. Poole et al.)



Software Development Highlights

Keeneland Software Environment

- Integrated with NSF TeraGrid/XD
  - Including TG and NICS software stack
- Programming environments
  - NVIDIA CUDA and OpenCL SDK
  - Compilers
    - GPU-enabled
  - Scalable debuggers
    - Allinea
  - Performance tools
    - Tau
  - Libraries
    - Magma
- Tools and programming options are changing rapidly
  - HMPP, PGI, R-stream,
- Additional software activities
  - Performance and correctness tools
  - Scientific libraries
  - Virtualization

Ocelot: Dynamic Execution Infrastructure

http://code.google.com/p/gpuocelot/

Gregory Diamos, Dhuv Choudhary, Andrew Kerr, Sudhakar Yalamanchili

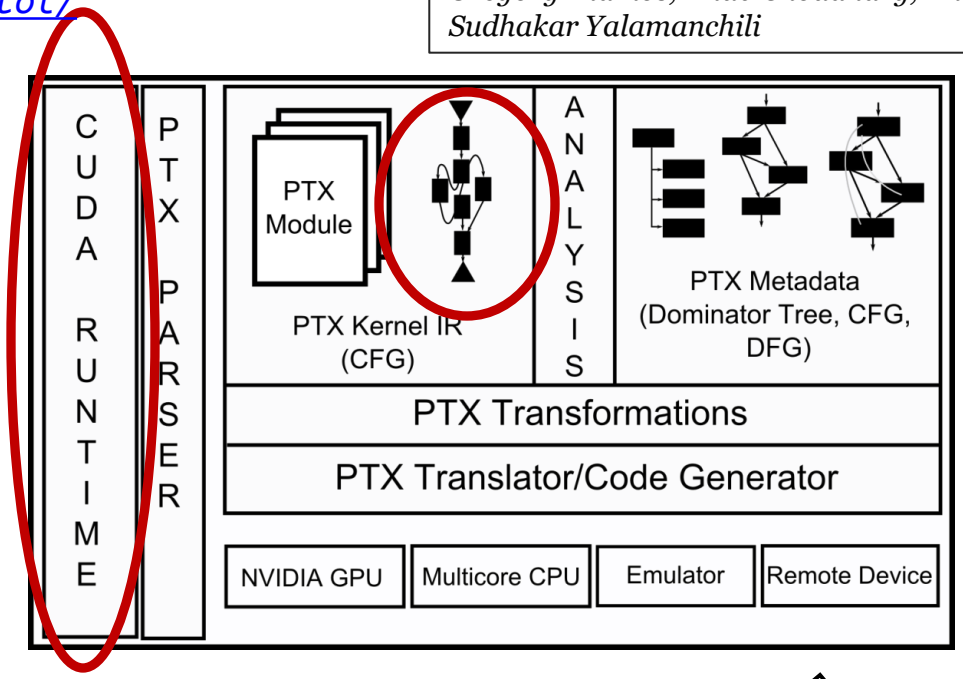
PTX Kernel

```
L_BB_1:
add.s64 %rd2, %rd1, 1
mul.s64 %rd3, %rd2, 4
mov.s64 %rd4, 256
setp.lt.s64 %p1, %rd3, %rd4
@%p1 bra L_BB_3

L_BB_2:
abs.f64 %fd1, %fd1
mov.s64 %rd5, 64
setp.lt.s64 %p2, %rd3, %rd5
@%p2 bra L_BB_4

L_BB_3:
sin.f64 %fd2, %fd1
st.f64 %fd2, [%rd0 + 4]

L_BB_4:
reconverge
reconverge
exit
```



Productivity Tools

PTX Emulator

Multiplatform Support

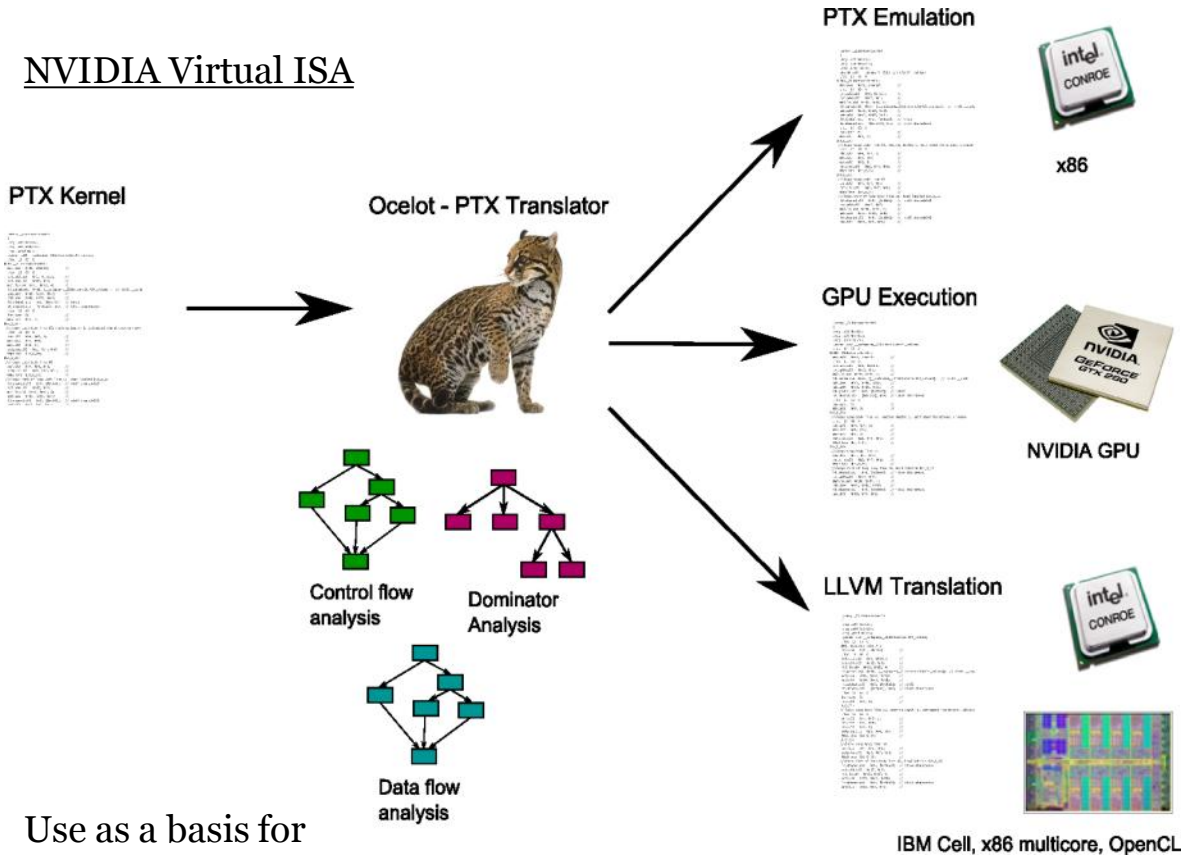
Dynamic Translation (LLVM)

Multi-GPU Support

Performance Analysis and Modeling

Ocelot: Dynamic Execution Infrastructure

NVIDIA Virtual ISA



PTX 1.4 compliant Emulation
- Validated on full CUDA SDK
- Open Source version released

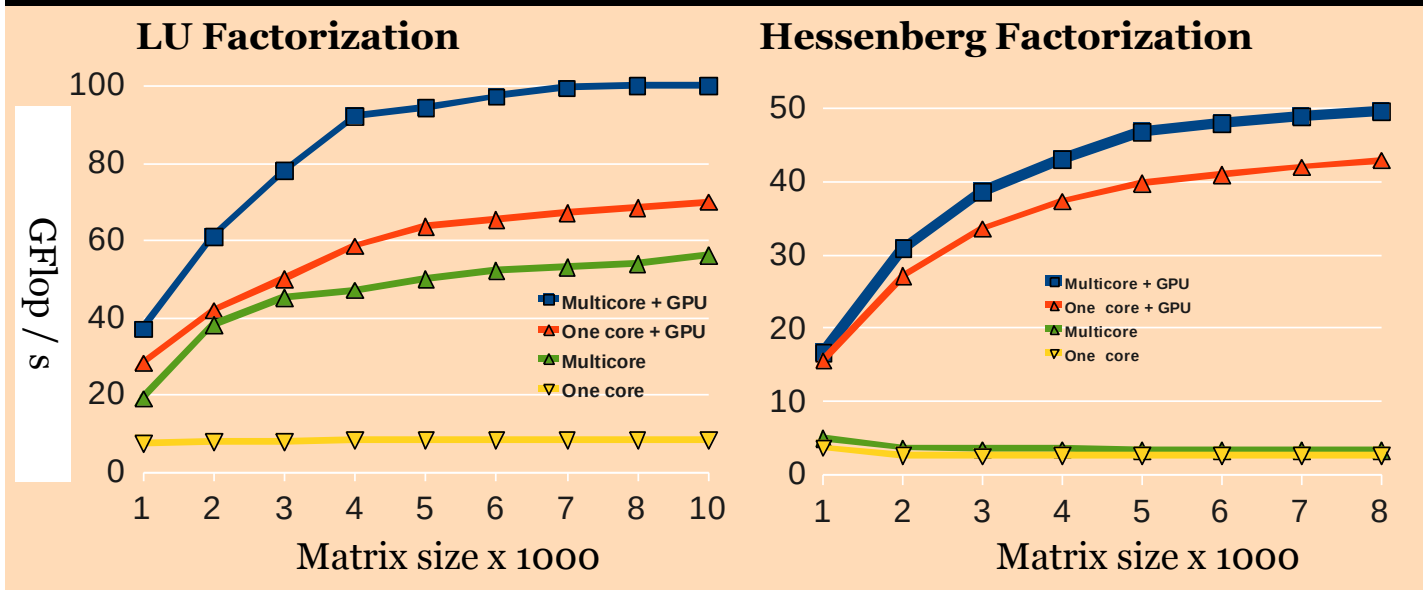
http://code.google.com/p/gpuocelot/

Use as a basis for
- Insight → workload characterization
- Performance tuning → detecting memory bank conflicts
- Debugging → illegal memory accesses, out of bounds checks, etc.

MAGMA Libraries: One and two-sided Multicore+GPU Factorizations

- These will be included in up-coming MAGMA releases
- Two-sided factorizations can not be efficiently accelerated on homogeneous x86-based multicores (above) because of memory-bound operations
  - MAGMA provided hybrid algorithms that overcome those bottlenecks (16x speedup!)

Multicore + GPU Performance in double precision



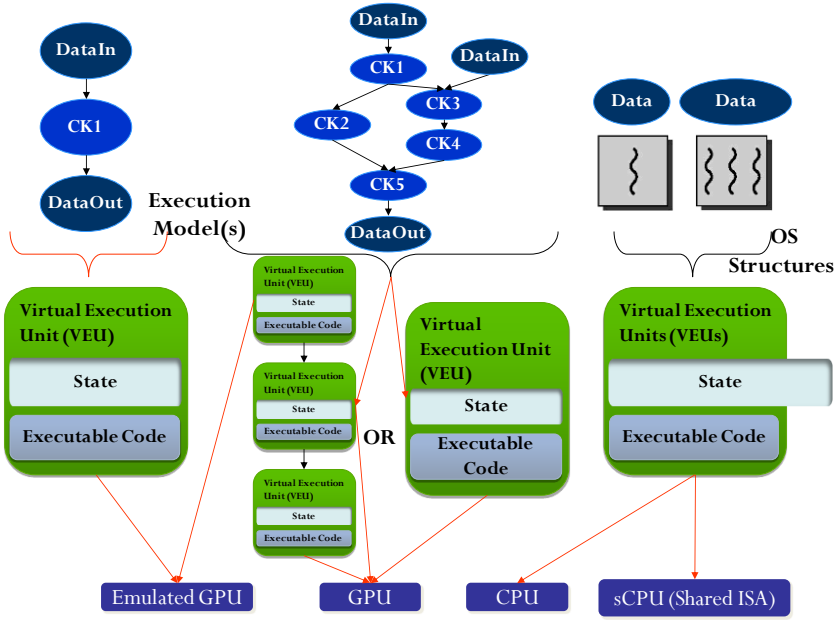
Jack Dongarra, Stan Tomov, and Rajib Nath

GPU : NVIDIA GeForce GTX 280
CPU : Intel Xeon dual socket quad-core @2.33 GHz

GPU BLAS : CUBLAS 2.2, dgemm peak: 75 GFlop/s
CPU BLAS : MKL 10.0 , dgemm peak: 65 GFlop/s

Hybrid Virtual Machine: HyVM

- Uniform runtime model for heterogeneous platforms:
- Hybrid Virtual Machines
  - Uniformity:
    - Virtual Execution Unit (VEU): hypervisor-level, uniform runtime representation for program executables
    - Heterogeneity-aware hypervisors: VMM-level management methods for improved platform utilization (incl. cache and energy) and application performance (SLAs)
    - Dynamic platform emulation: runtime CK compilation or re-writing for diverse accelerator targets (via LLVM)
  - High performance:
    - Commodity and custom VEU 'containers': Virtual Machines (VMs) – processes/threads – commodity cores; Special Execution Environments (e.g., NVIDIA) - Computational Kernels (CKs) - accelerators
    - Runtime and adaptive {CK} optimization for parallelism
    - Standards-compliant CK programming and runtime APIs (OpenCL, CUDA)
    - Compiler-based optimization techniques for {CK}



Ada Gavrilovska, Karsten Schwan, Sudha Yalamanchili, and many PhD students

DOE Vancouver: A Software Stack for Productive Heterogeneous Exascale Computing

Jeffrey Vetter, ORNL
Wen-Mei Hwu, UIUC
Allen Malony, University of Oregon
Rich Vuduc, Georgia Tech

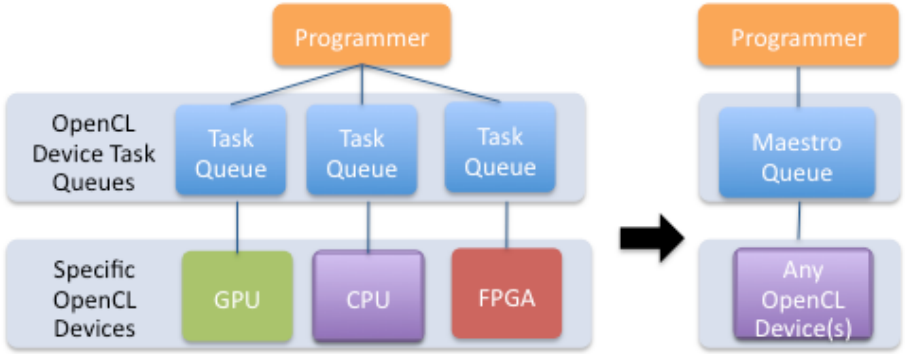
Objectives

- Enhance programmer productivity for the exascale
  - Increase code development ROI by enhancing code portability
  - Decrease barriers to entry with new programming models
- Create next-generation tools to understand the performance behavior of an exascale machine

Approach

- Programming tools
  - GAS programming model
  - Analysis, inspection, transformation
- Software libraries: autotuning
- Runtime systems: scheduling
- Performance tools
- Impact on DOE Applications

The proposed Maestro runtime simplifies programming heterogeneous systems by unifying OpenCL task queues into a single high-level queue.

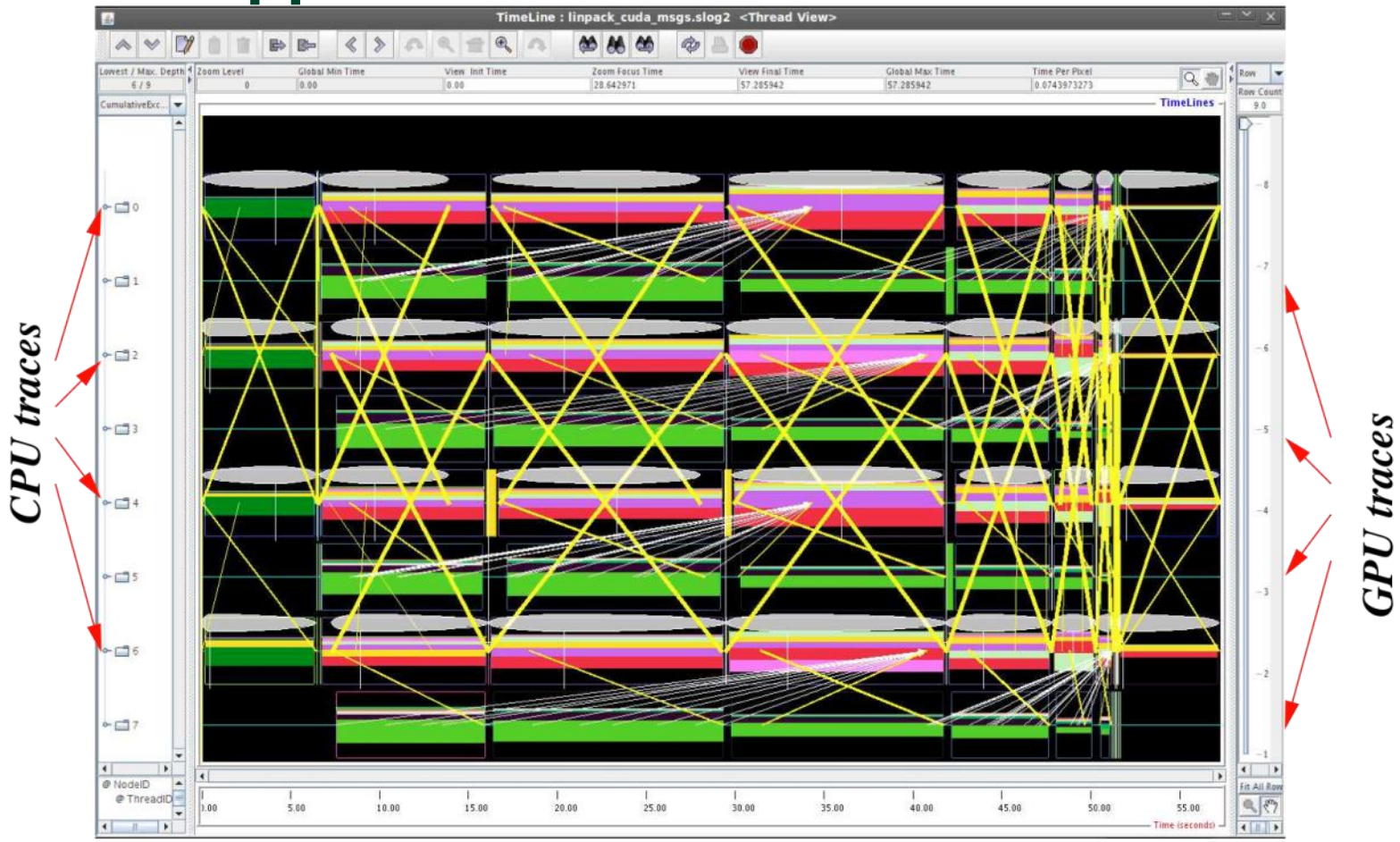


Impact

- Reduced application development time
- Ease of porting applications to heterogeneous systems
- Increased utilization of hardware resources and code portability

Georgia Tech College of Computing Computational Science and Engineering
ILLINOIS
UNIVERSITY OF OREGON

DOE Vancouver: Performance Analysis of MPI/GPU Applications at Scale

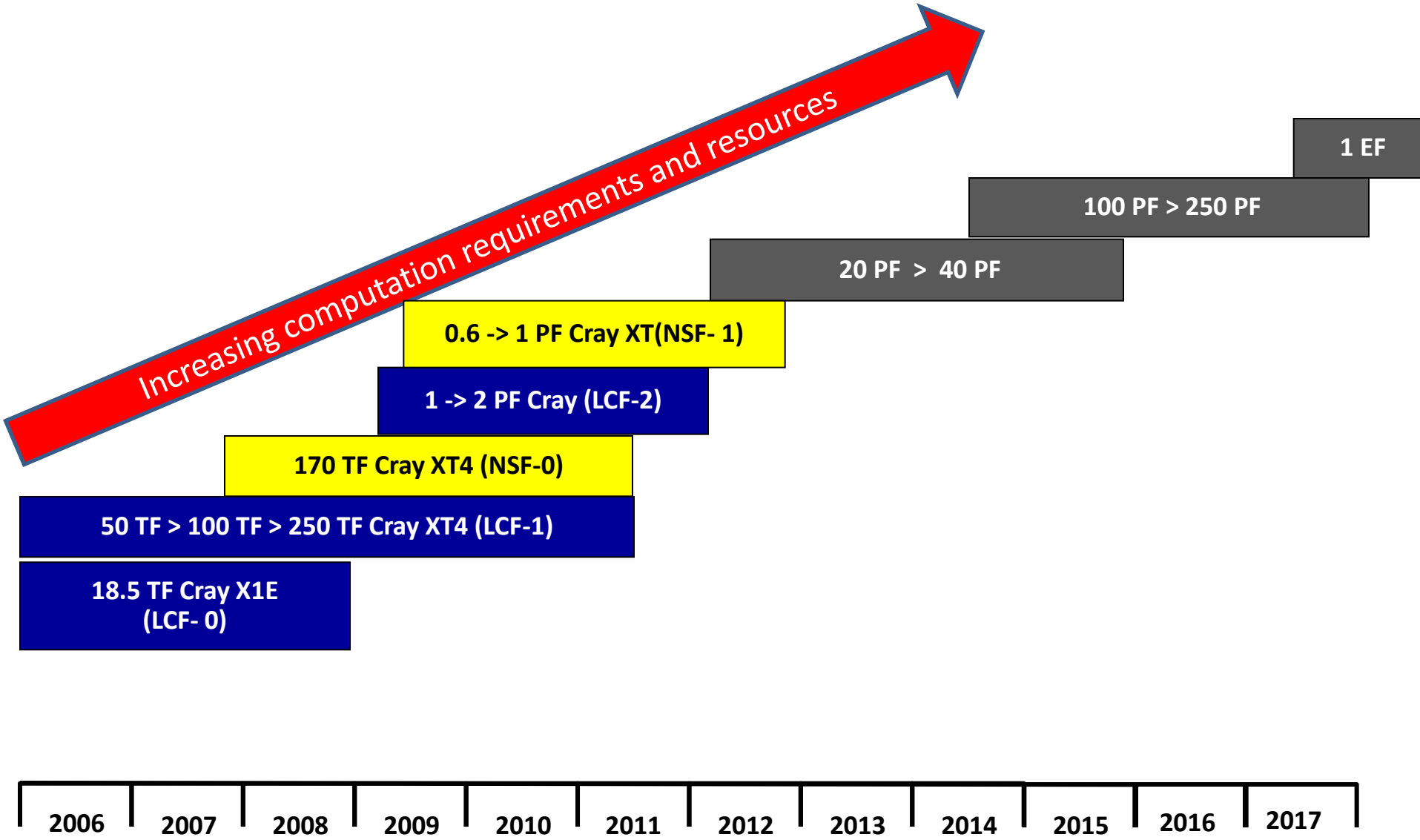


CUDA memory transfer (white) MPI communication (yellow)

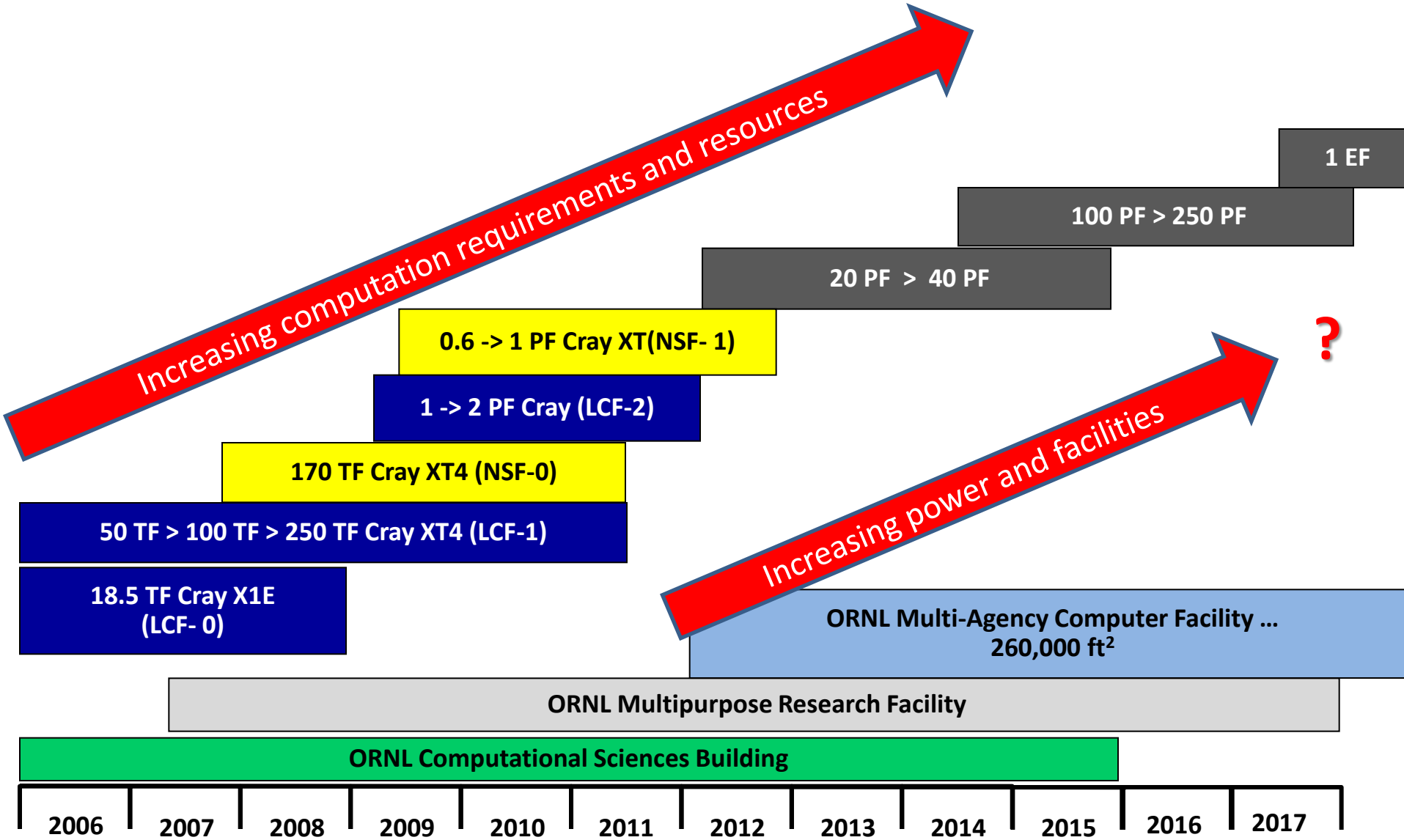
A.D. Malony, S. Biersdorff, W. Spear, and S. Mayanglambam, "An experimental approach to performance measurement of heterogeneous parallel applications using CUDA," in *Proc 24th ACM International Conferen Supercomputing., 2010.*

The Road Ahead

ORNL Roadmap to Exascale



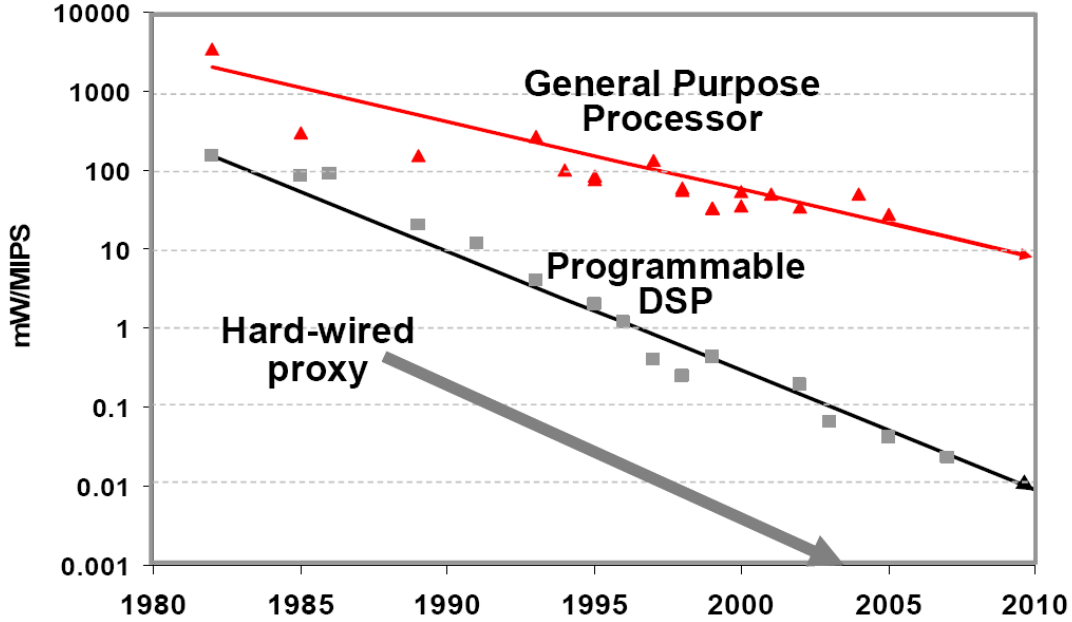
ORNL Roadmap to Exascale



Facilities and Power … Not just ORNL

Mobile and Embedded Systems Community has Experiences

No single architecture solves all power problems



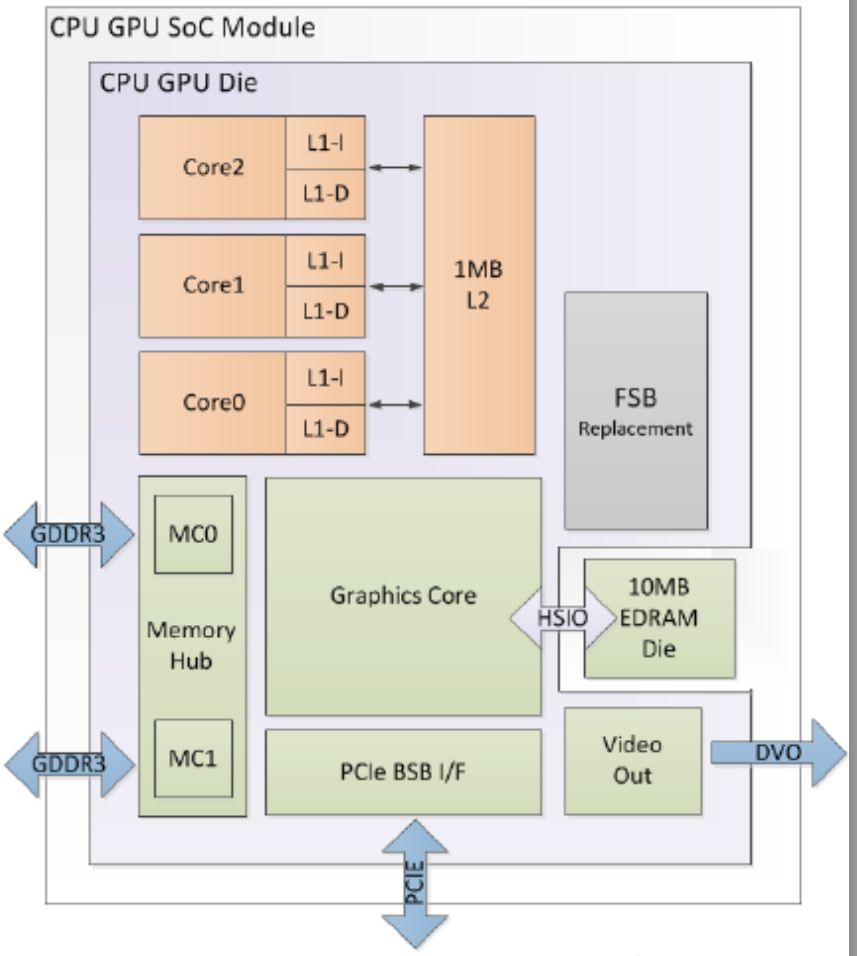
- Industry has debated merits of each architecture for decades…
- Combination of all approaches optimizes power and performance

©2010 IEEE International Solid-State Circuits Conference ©2010 IEEE

Source: Delagi, ISSCC 2010

CPU GPU SoC: Features & Block Diagram

- **CPU**
  - Three 3.2 GHz PowerPC® cores
  - Shared 1MB L2 cache
  - Per Core:
    - Dual Thread Execution
    - 32K L1 I-cache, 32K L1 D-cache
    - 2-issue per cycle
    - Branch, Integer, Load/Store Units
    - VMX128 Units enhanced for games
- **GPU**
  - 48 parallel unified shaders
  - 24 billion shader instructions per second
  - 4 billion pixels/sec pixel fill rate
  - 500 million triangles/sec geometry rate
  - High Speed IO interface to 10 MB EDRAM
- **Compatibility**
  - Functional and Performance equivalent to prior Xbox 360 GPU/CPU
  - FSB Latency and BW match prior FSB



XBOX 360

Microsoft | IBM

Source: Microsoft, HotChips22

Systems on the Horizon

Delivery	System	Location	Comp	Comm	Peak (PF)	Power (MW)
2010	Tianhe-1A	National Supercomputing Center in Tianjin, China	Intel + NVIDIA	Proprietary	4.5	4
2010	Tsubame 2	TiTech	Intel + NVIDIA	Infiniband	2.4	
~2012	Cray 'Titan'	ORNL	AMD + NVIDIA	Gemini	20	
~2012	BlueWaters	NCSA/UIUC	POWER7	IBM Hub	10	
~2012	K Computer	Kobe	SPARC64 VIIIfx	Tofu	10	20
~2012	BlueGeneQ	ANL	SoC	IBM	10	
~2012	BlueGeneQ	LLNL	SoC	IBM	20	
Others…						

Notional System Architecture Targets

System attributes	2010	"2015"		"2018"	
System peak	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	6 MW	15 MW		20 MW	
System memory	0.3 PB	5 PB		32-64 PB	
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW	25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW	1.5 GB/s	150 GB/sec	1 TB/sec	250 GB/sec	2 TB/sec
MTTI	day	O(1 day)		O(1 day)	

Echelon: Extreme-scale Compute Hierarchies with Efficient Locality-Optimized Nodes

Main Objectives

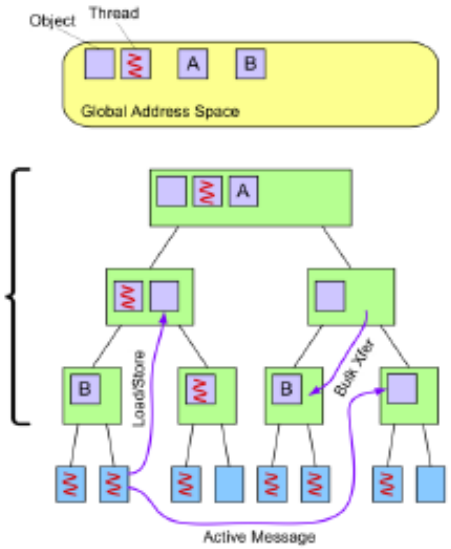
- Two orders of magnitude increase in application execution energy efficiency over today's CPU systems.
- Improve programmer productivity so that the time required to write a parallel program achieving a large fraction of peak efficiency is comparable to the time required to write a serial program today.
- Strong scaling for many applications to tens of millions of threads in UHPC system.
- High application mean-time to interrupt (AMTTI) with low overhead; matched to application needs.
- Machines resilient to attack; enables reliable software.

Key Innovations

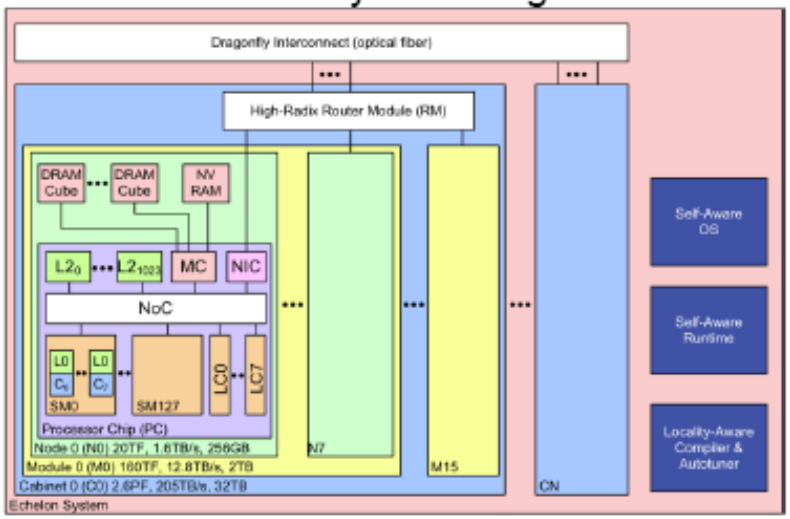
- Programming systems that express concurrency/locality abstractly; autotuning for hardware mapping.
- Self-aware runtime reacts to changes in environment, workload (load-balance), fault states.
- Fine-grained, energy-optimized, multithreaded throughput cores + latency-optimized cores.
- Software selective memory hierarchy configuration; selective coherence for non-critical data.
- HW/SW cooperative resilience for energy- and performance-efficient fault protection.
- Guarded pointers for memory safety.
- Low-power, high speed communication circuits.

Echelon Execution Model

- Programmability: global address space, abstract memory hierarchy, autotuning; runtime task placement/scheduling.
- Efficiency: Active messages, bulk transfer.
- Dependability: software selective redundancy, hardware accelerated guarded pointers.

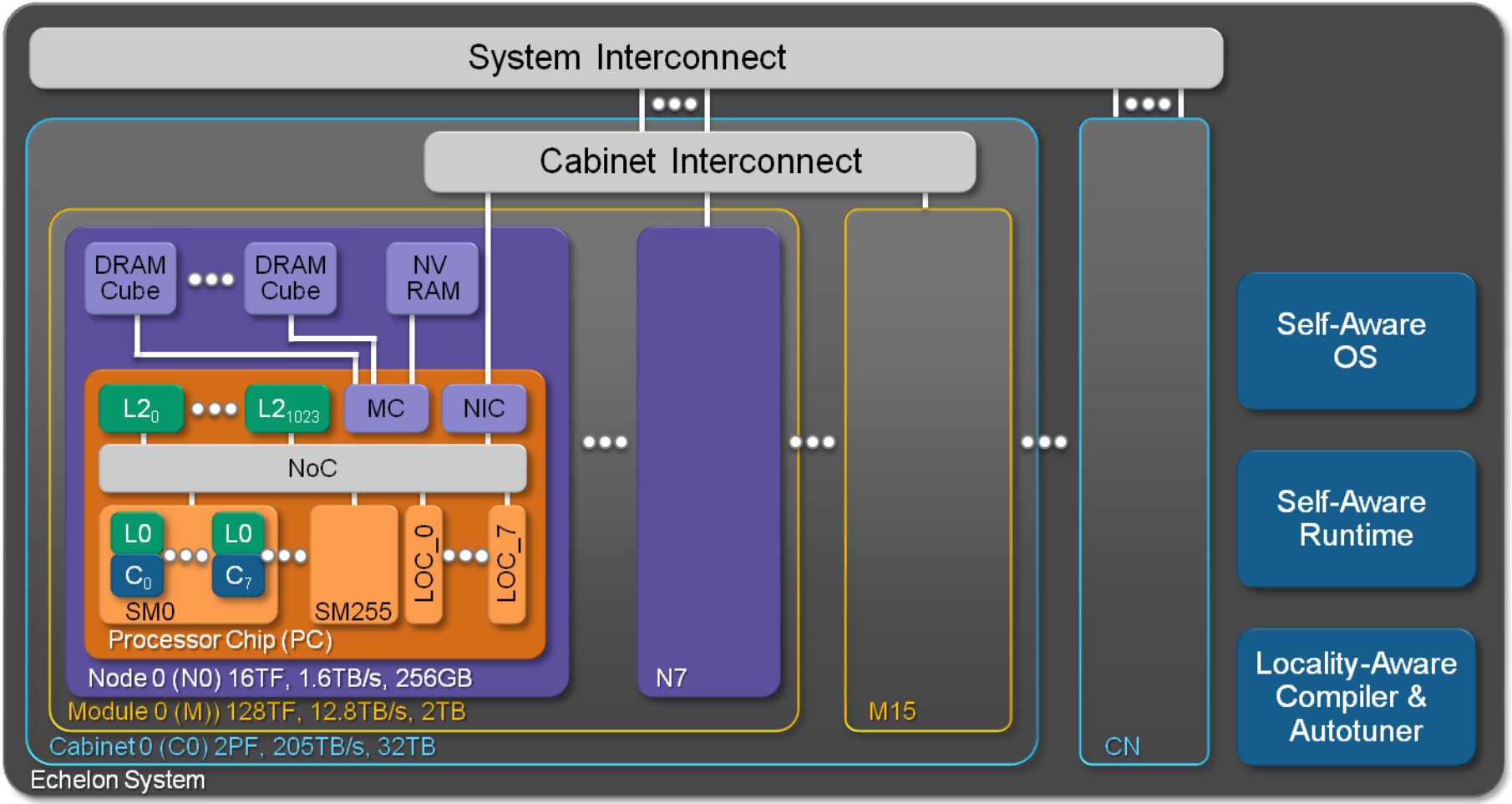


Echelon System Diagram



Georgia Tech Stanford UC-Berkeley UPenn UT-Austin U. Utah Tennessee Lockheed Martin

NVIDIA Echelon System Sketch



Source: NVIDIA Echelon Proj

Bonus

Members of Our CCOE

- Georgia Tech
  - Jeffrey Vetter; David Bader; George Biros; Edmond Chow; Tom Conte; Richard Fujimoto; Richard Glassbrook; Alex Gray; Bo Hong; Seung Soon Jang; Hyesoon Kim; Pablo Laguna; Aaron Lanterman; Hsien-Hsin Lee; Lew Lefton; Blair MacIntyre; Santosh Pande; Chris Rorden; Jarek Rossignac; Ryan Russell; Karsten Schwan; David Sherrill; Deirdre Shoemaker; Jeff Skolnick; Rich Vuduc; Yan Wang; Sudha Yalamanchili; PK Yeung; Haomin Zhou
- GTRI
  - Richard Boyd; Dan Campbell; Daniel Faircloth
- CDC
  - Christopher Lynberg; Beth Neuhaus
- ORNL
  - Jeremy Meredith; Philip Roth
- Accelereyes
  - John Melonakos; Dave Gibson

Kickoff workshop had over 100 participants

CCOE Activities Span the Range of Science

Applications

- Biology
- Chemistry
- Security
- Materials
- Aerospace Eng
- Mechanical Eng
- Knowledge Discovery
- Advanced Graphics
- Augmented Reality
- Graph Theory
- Industrial Eng - Optimization

Software

- Libraries, Autotuning
- Compilers and Runtime
- Tools for correctness and performance
- Scheduling
- Virtualization

Architecture

- Facilities
  - NSF Keeneland
  - DOE OLCF3
- Simulation and Performance Prediction
- Benchmarking
- Future Architectures
  - DARPA Echelon

New ProLiant SL6500 series

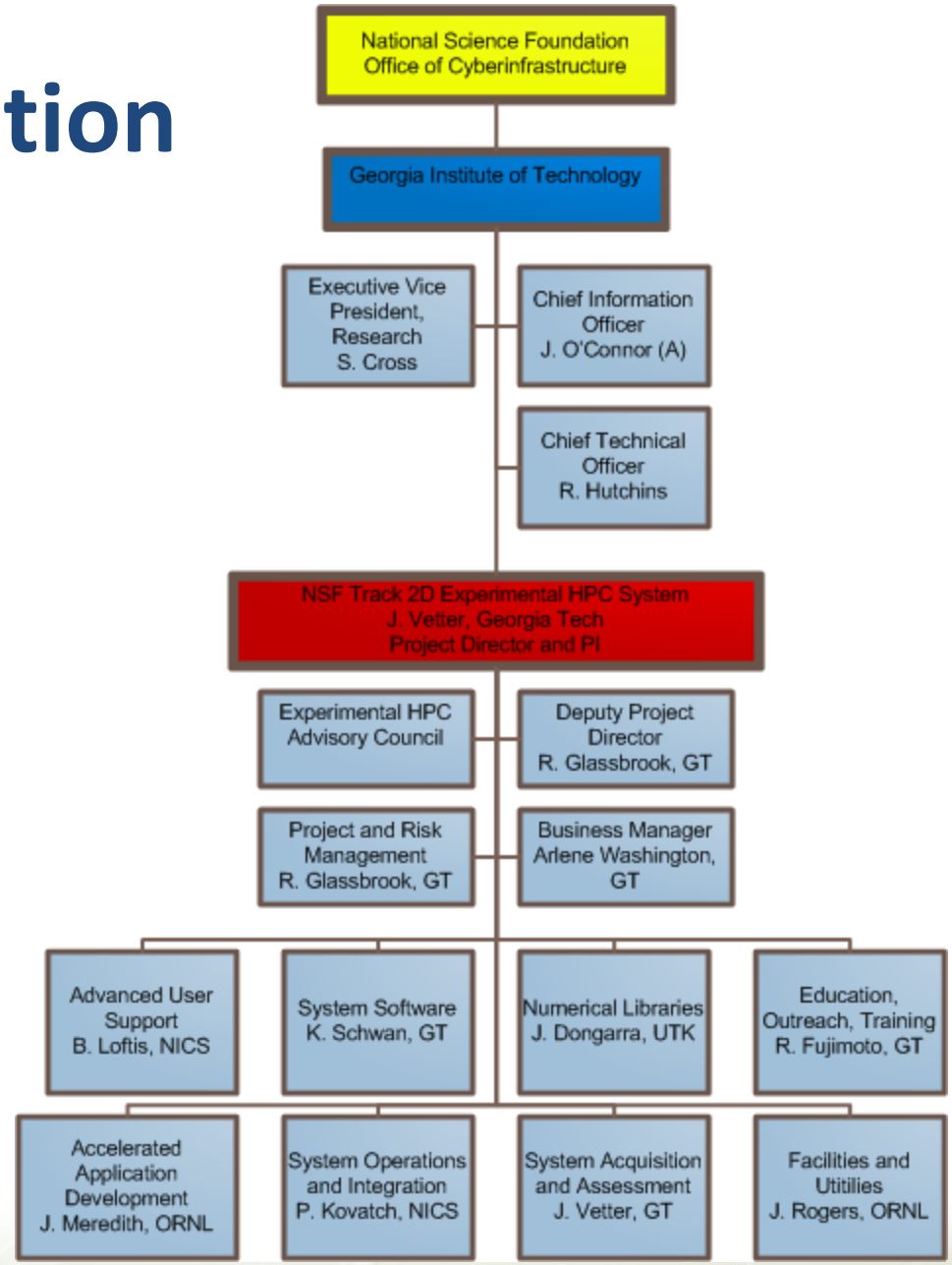
Highly Flexible s6500 Chassis

Multinode, Shared Power and Cooling Architecture

- Shared power and fans
- Optional hot-plug redundant PSU
- Energy efficient hot-plug fans
- 3-phase load balancing
- 94% platinum common slot power supplies
- N+1 capable power supplies (up to 4)

Benefits: Low Cost, High Efficiency Chassis

- 4U chassis for deployment flexibility
- Standard 19” racks, with front I/O cabling
- Unrestricted airflow (no mid-plane or I/O connectors)
  - Reduced weight
    - <u>Individually serviceable nodes</u>
      - Variety of optimized node modules

- SL Advanced Power Manager support
  - Power monitoring
  - Node level power off/on

Organization