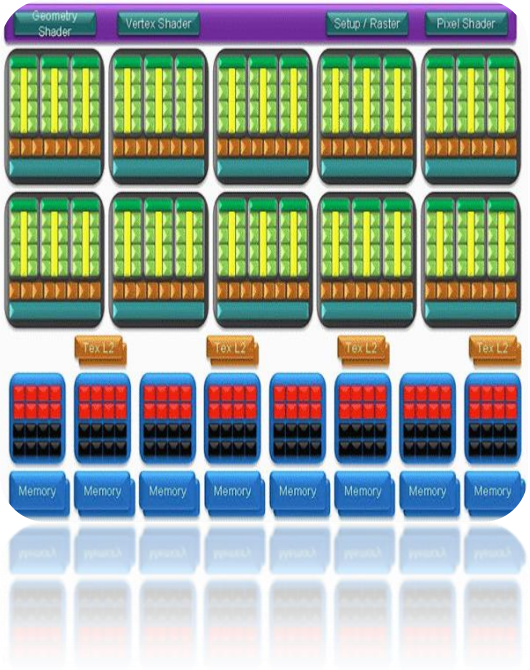
Jeffrey Vetter, Dick Glassbrook, Jack Dongarra, Richard Fujimoto, Thomas Schulthess, Karsten Schwan, Sudha Yalamanchili, Kathlyn Boudwin, Jim Ferguson, Patricia Kovatch, Bruce Loftis, Stephen McNally, Jeremy Meredith, Jim Rogers, Philip Roth, Kyle Spafford, Arlene Washington, Don Reed, Tracy Rafferty, Ursula Henderson, Terry Moore, and many others

KEENELAND - ENABLING HETEROGENEOUS COMPUTING FOR THE OPEN SCIENCE COMMUNITY

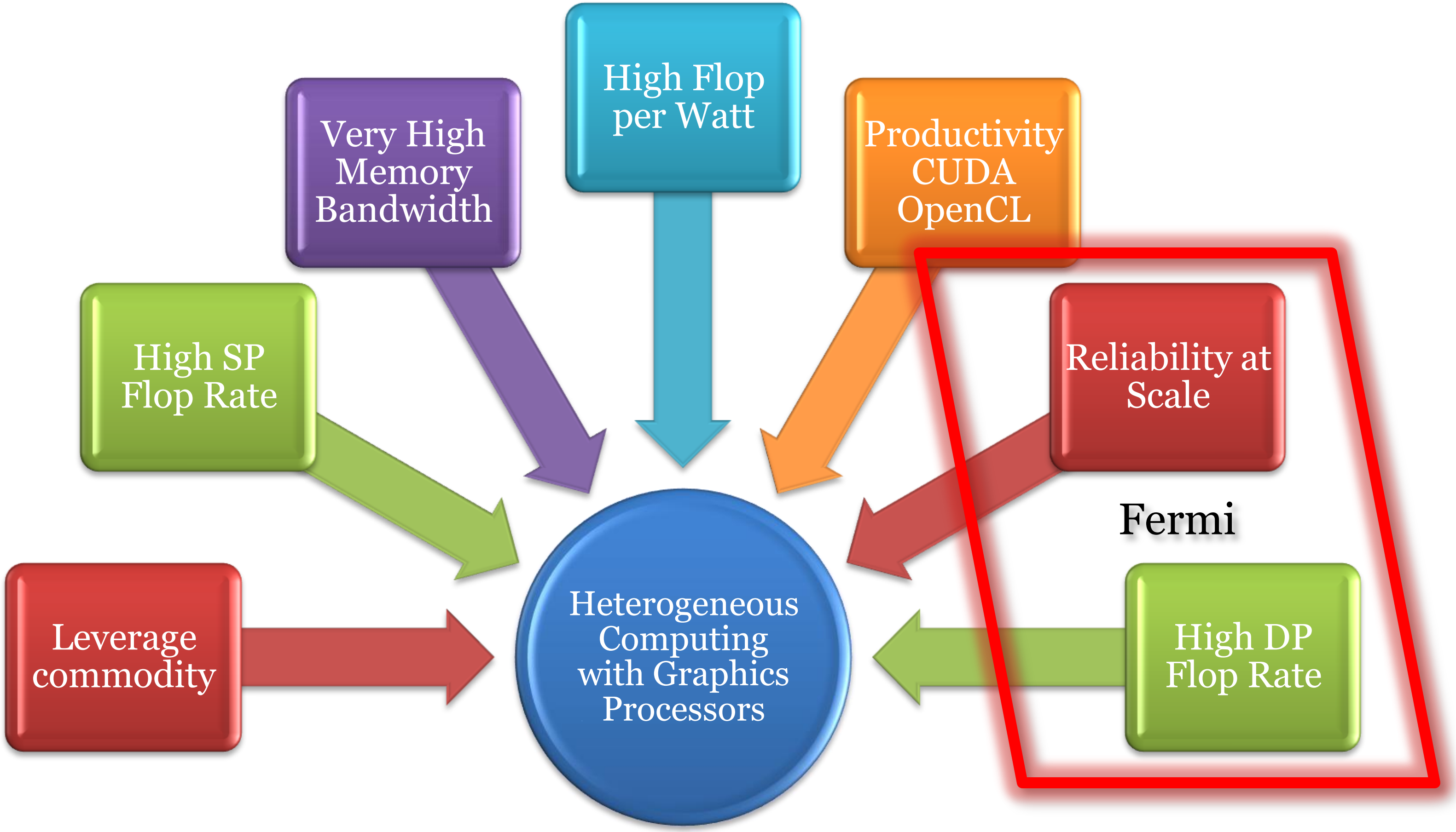
BACKGROUND – HOW DID WE GET HERE?

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?



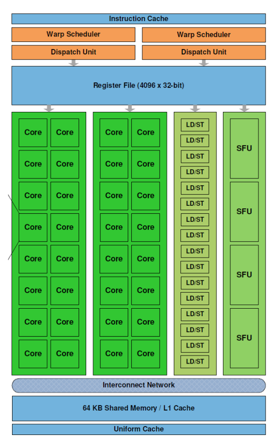
U.S. DEPARTMENT OF ENERGY

Notional System Architecture Targets and "swim lanes"

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)	

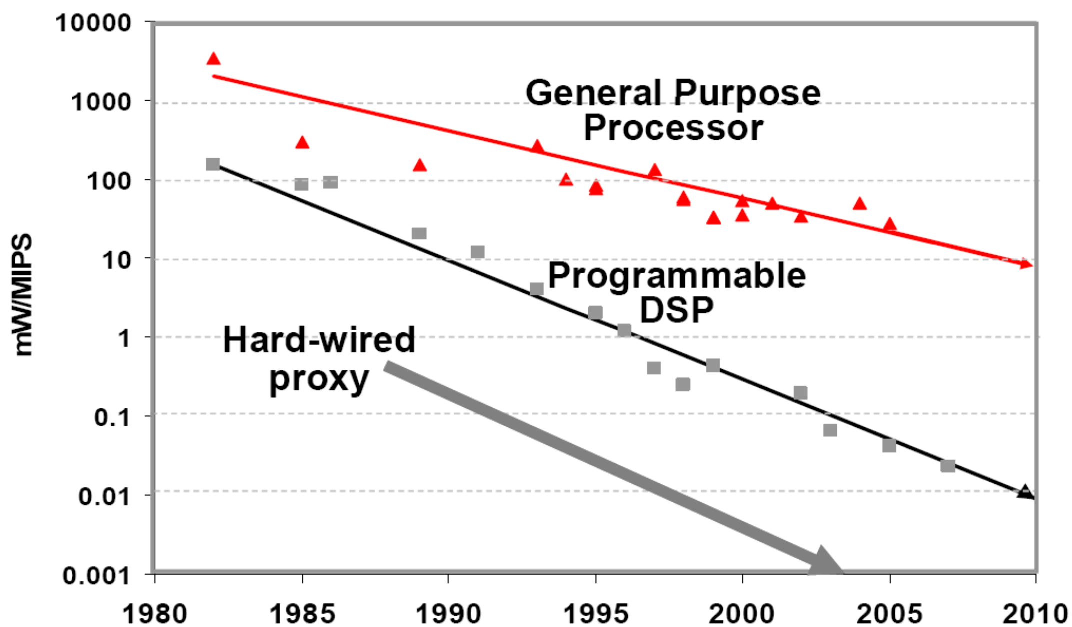
Exascale computing will require tough decisions and/or innovative technologies

- **Build bigger buildings and plan to pay $$$ for ops**
- **Improve efficiencies**
  > PUE
  > Power distribution
  > Workload scheduling
  > Software
- **Use architectures that 'match' your workload**
  > GPUs, FPGAs
- **Design new underlying technologies**
  > Optical networks
  > 3D stacking
  > MRAM, PCM, R-RAM



Heterogeneous architectures can offer better performance, power

No single architecture solves all power problems



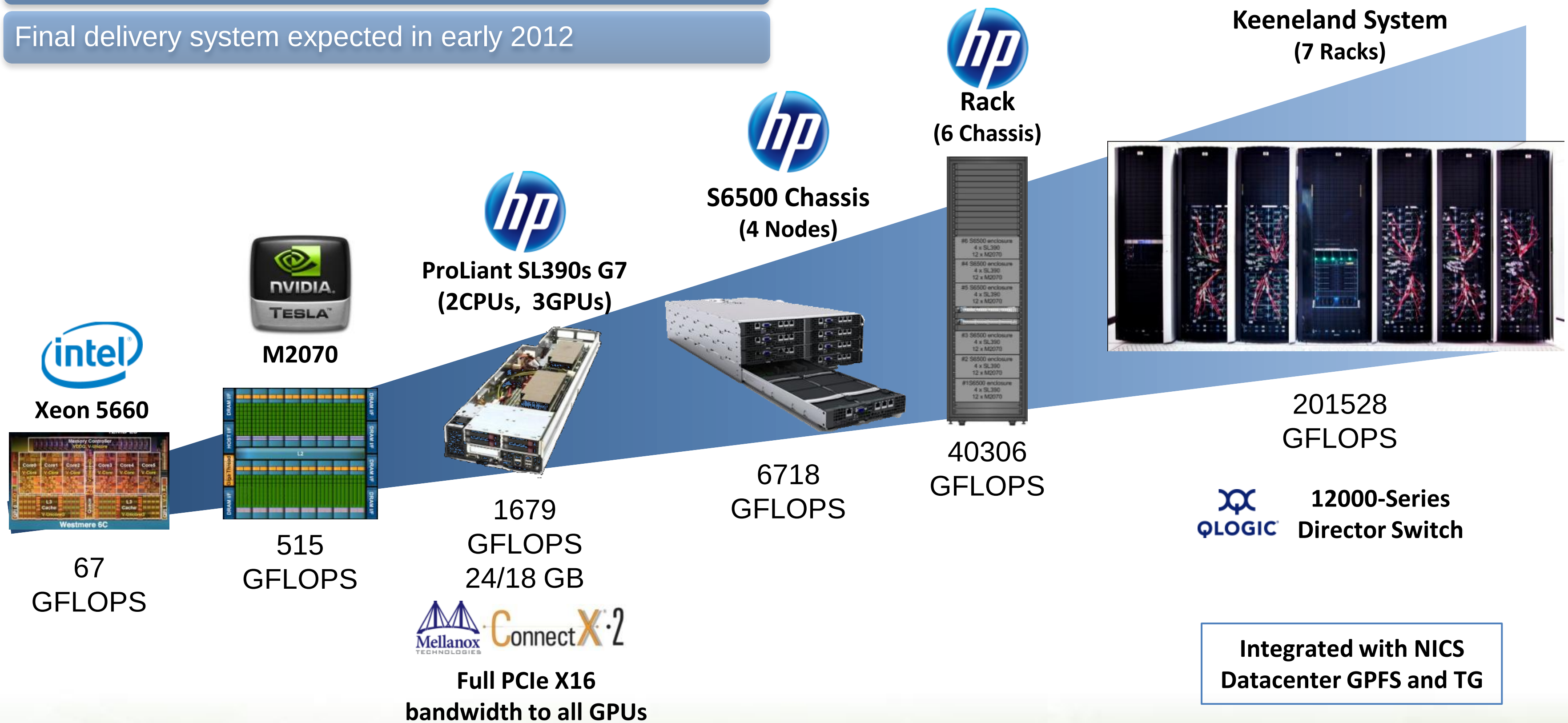
- Industry has debated merits of each architecture for decades…
- Combination of all approaches optimizes power and performance

Source: Delagi, ISSCC 2010

KIID ARCHITECTURE

Keeneland – Initial Delivery System Architecture

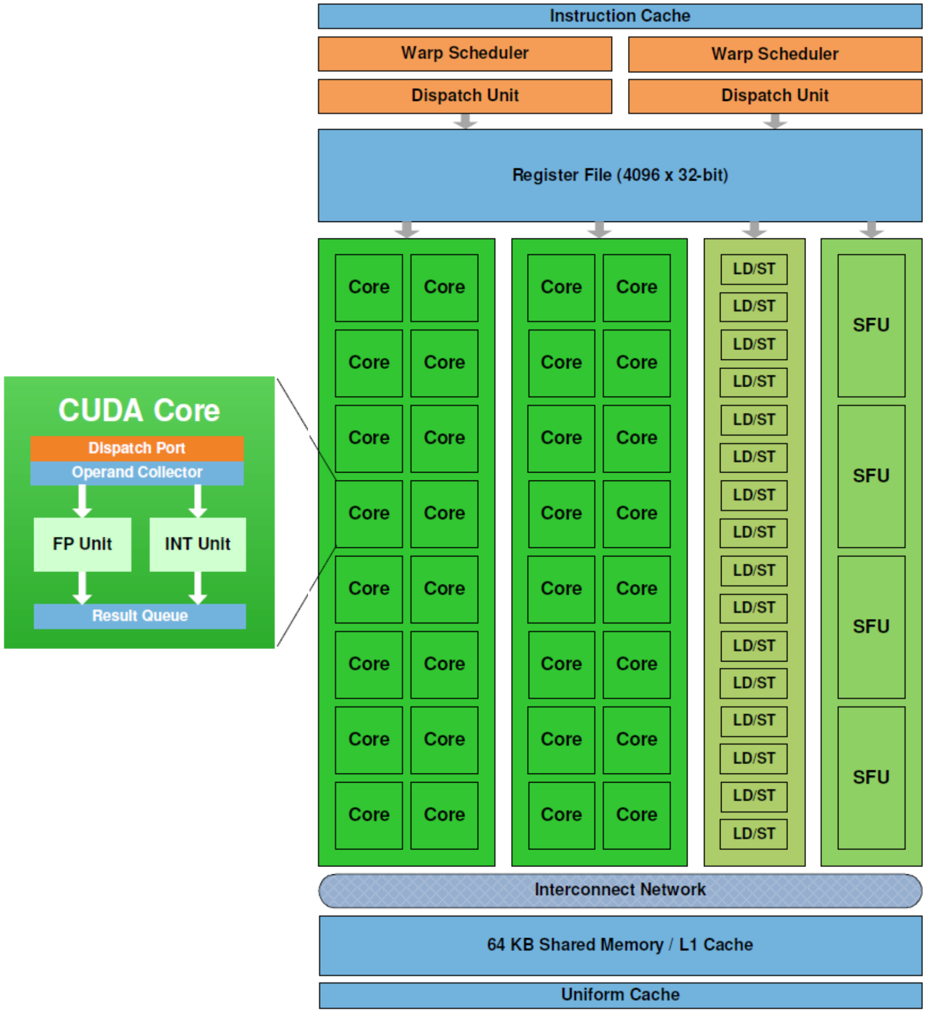
- **Initial Delivery** system procured and installed in Oct 2010
- 201 TFLOPS in 7 racks (90 sq ft incl service area)
- 677 MFLOPS per watt on HPL
- Final delivery system expected in early 2012



Integrated with NICS Datacenter GPFS and TG

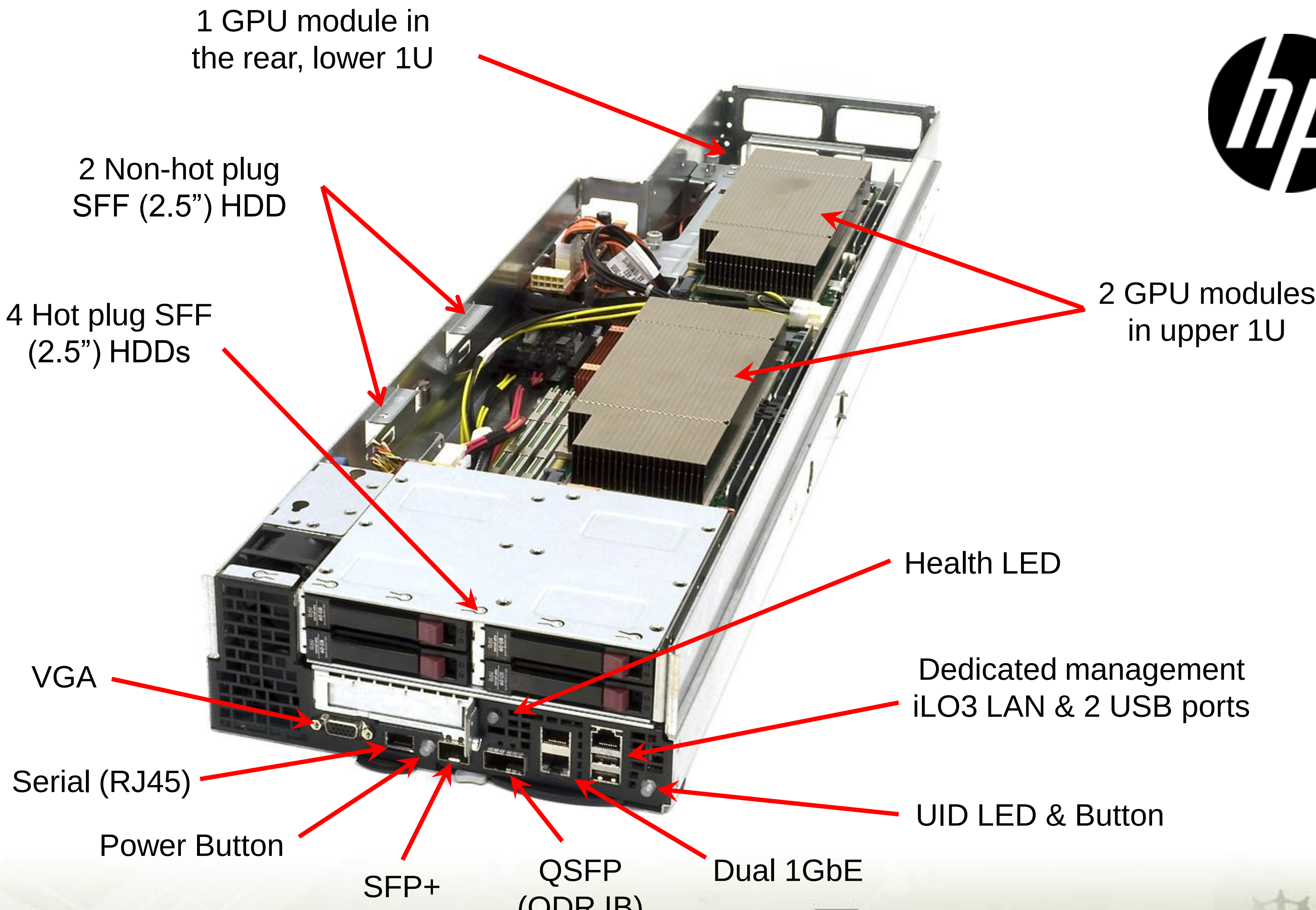
NVIDIA Fermi

- 3B transistors!
- Error correction
- 448 CUDA Cores featuring the new IEEE 754-2008 floating-point standard
  - 8× the peak double precision arithmetic performance over NVIDIA's last generation GPU
  - 515 DP GF
  - 1030 SP GF
  - 32 cores per SM, 21k threads per chip
- 120-144 GB/s memory BW
- NVIDIA Parallel DataCache
- NVIDIA GigaThread Engine
- Debuggers, language support

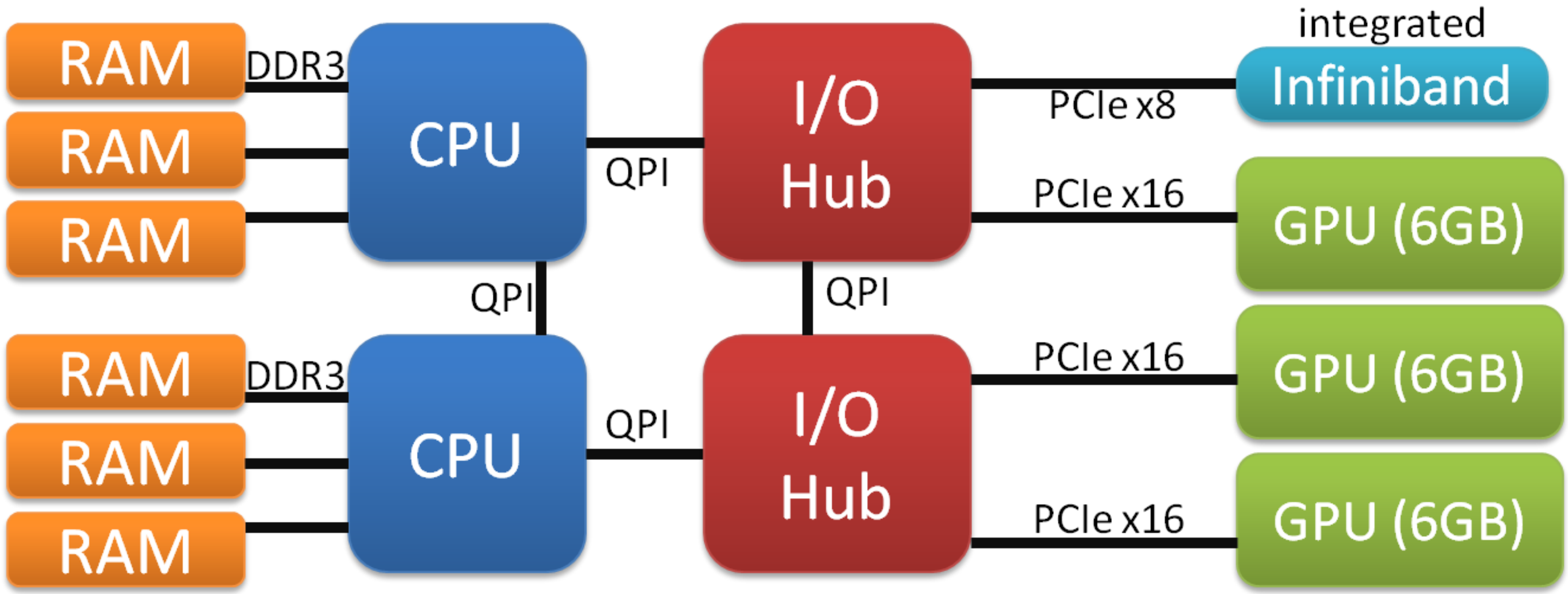


Georgia Tech NICS THE UNIVERSITY of TENNESSEE OAK RIDGE National Laboratory NVIDIA hp NSF

HP ProLiant SL390s G7 2U half width tray



Keeneland Node Architecture SL390



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
    - Individually serviceable nodes
      - Variety of optimized node modules

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

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
- Finishing acceptance testing this week

Keeneland ID installation – 10/29/10

Installation Team has worked long hours






Thanks to many at HP, NVIDIA, Qlogic: Paul Salerno, Glen Lupton, etc

Clockwise from upper right: Stephen McNally, Kyle Spafford, Philip Roth, Jeremy Meredith, Dave Holton (HP), Jeffrey Vetter, Dale Southard (NVIDIA).

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

Status

- Finish acceptance testing on KID
- Enter early science operation
- KID goals
  - Connected to TG/XD
  - Resource for applications teams with GPU codes
  - Resource for GPU software and tool development
- Larger, final delivery system planned for mid 2012

APPLICATIONS

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

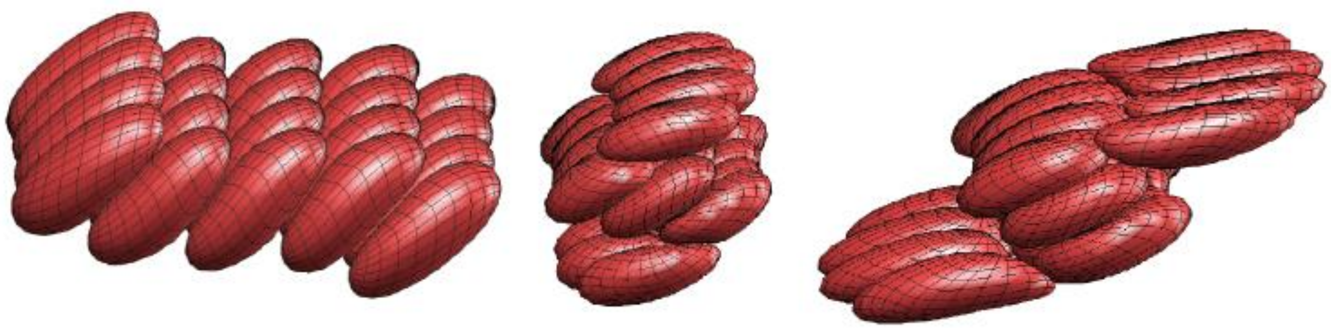
- S3D
  - Massively parallel direct numerical solver (DNS) for the full compressible Navier-Stokes, total energy, species and mass continuity equations
  - Coupled with detailed chemistry
  - Scales to 150k cores on Jaguar
- Accelerated version of S3D’s Getrates kernel in CUDA on Tesla T10
  - 31.4x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

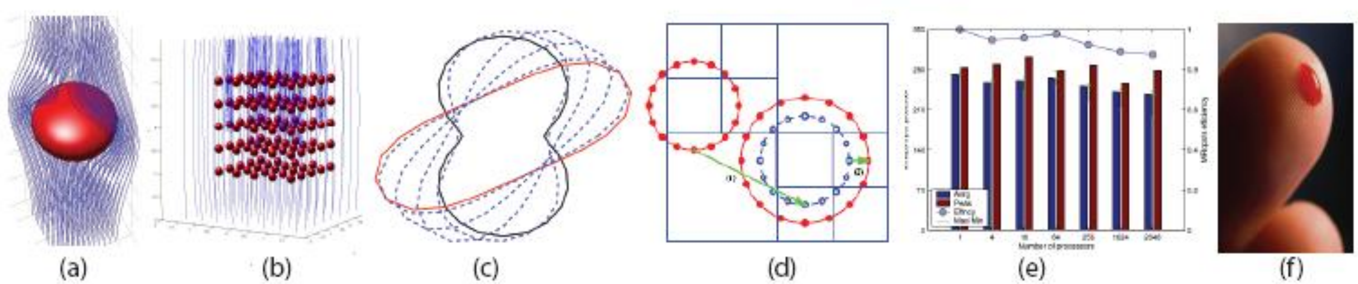
K. Spafford, J. Meredith, J. S. Vetter, J. Chen, R. Grout, and R. Sankaran. Accelerating S3D: A GPGPU Case Study. Proceedings of the Seventh International Workshop on Algorithms, Models, and Tools for Parallel Computing on Heterogeneous Platforms (HeteroPar 2009) Delft, The Netherlands.

Gordon Bell Presentation today at 1.30

Simulating Blood Flow with FMM

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

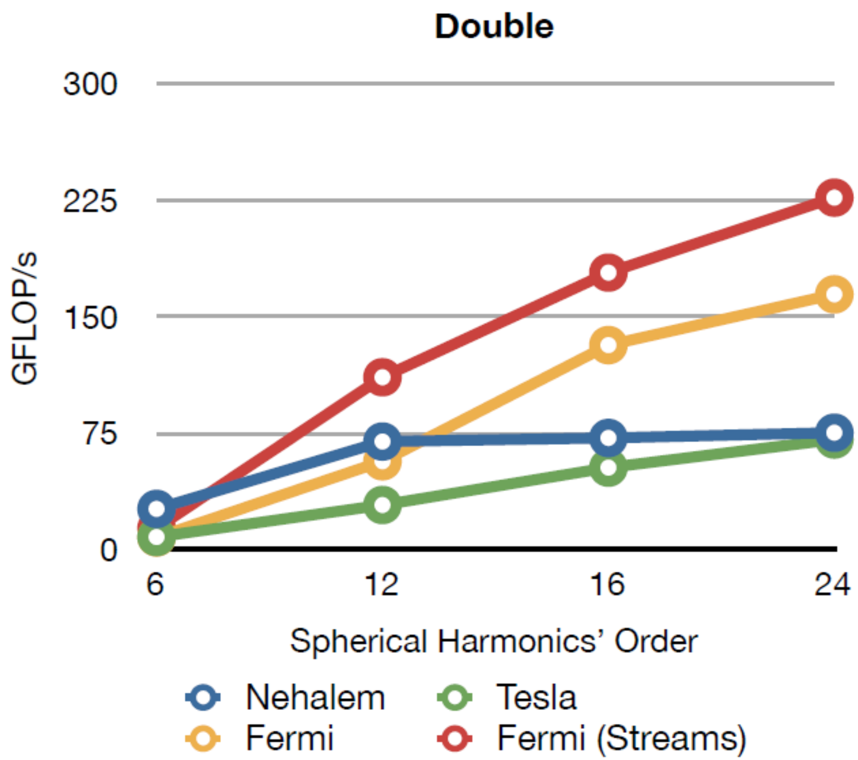


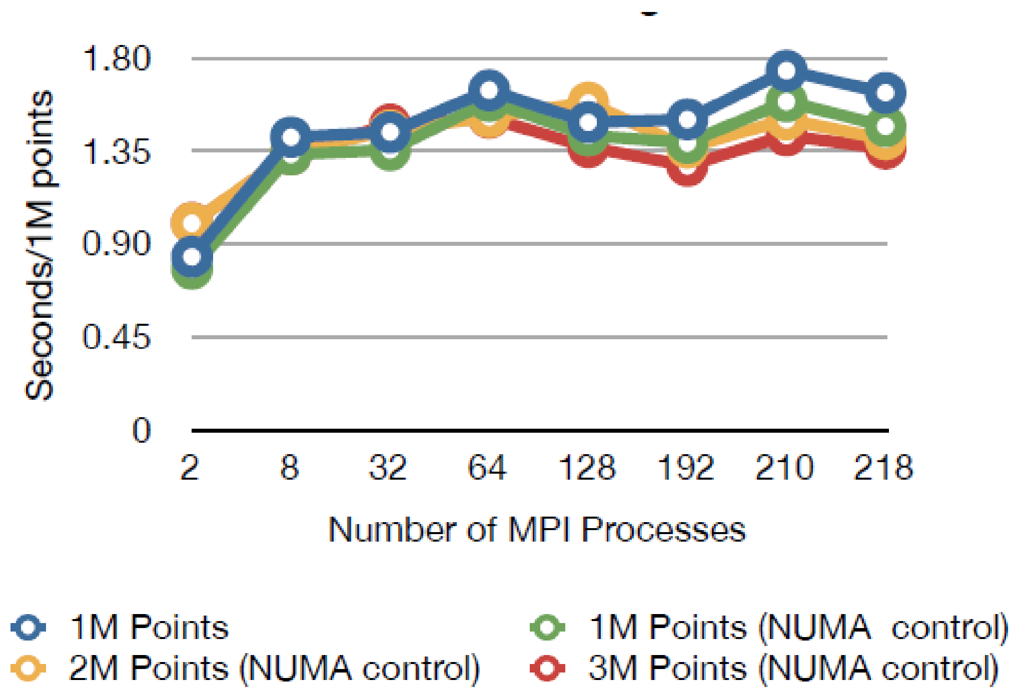


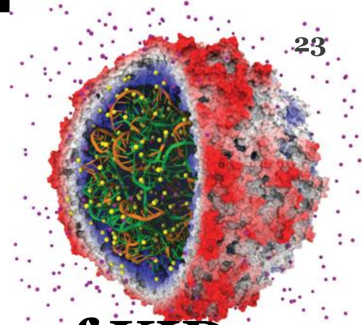
A. Rahimian, I. Lashuk, S. Veerapaneni *et al.*, *"Petascale Direct Numerical Simulation of Blood Flow on 200K Cores and Heterogeneous Architectures (Gordon Bell Finalist)," in SC10: International Conf. High Performance Computing, Networking, Storage, and Analysis. New Orleans: ACM/IEEE, 2010.*

Preliminary results from KID

FMM results from KID



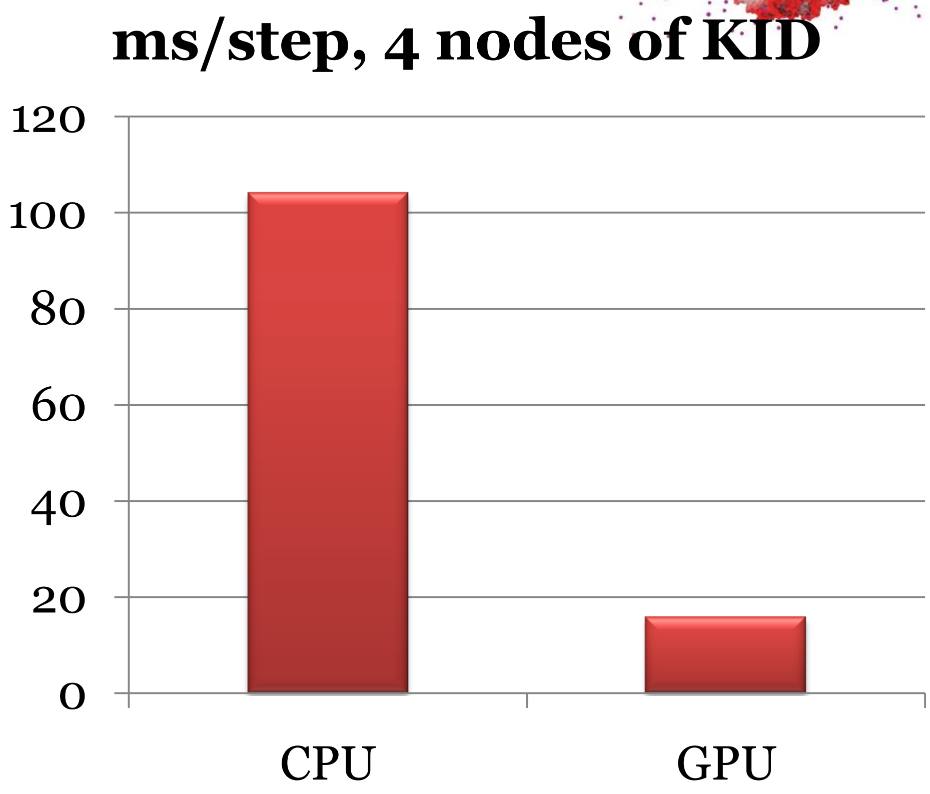




Preliminary results from KID

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

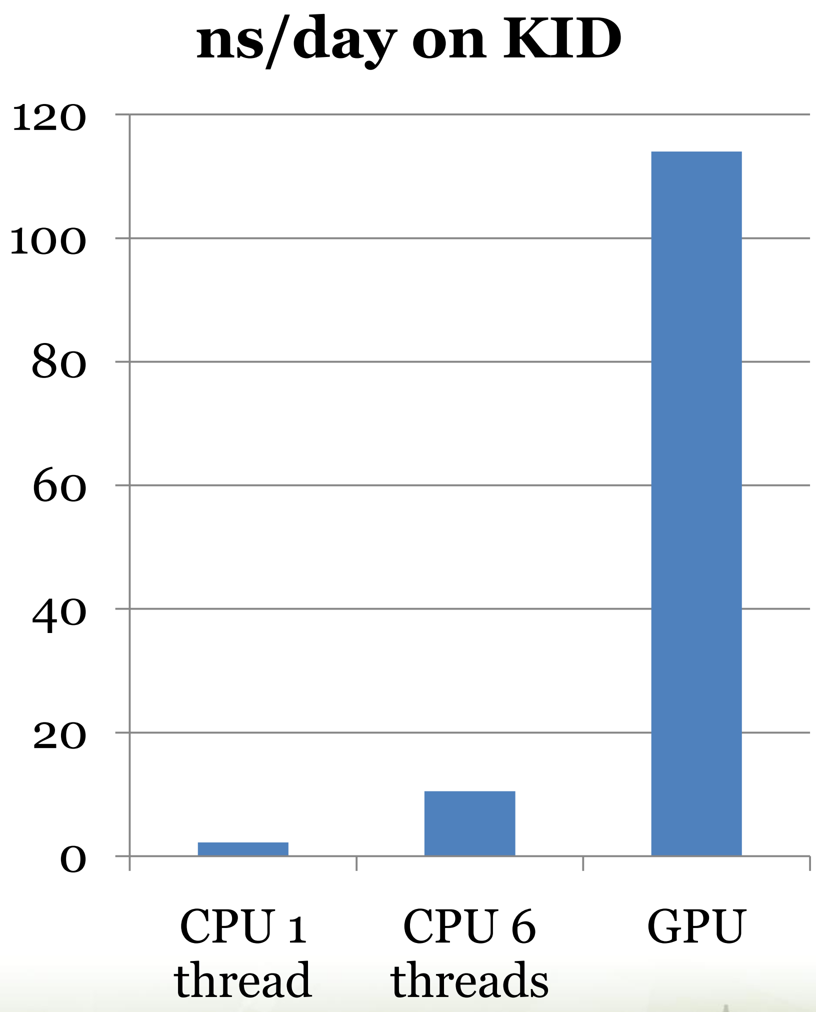


J.C. Phillips and J.E. Stone, "Probing biomolecular machines with graphics processors," *Commun. ACM, 52(10):34-41, 2009.*

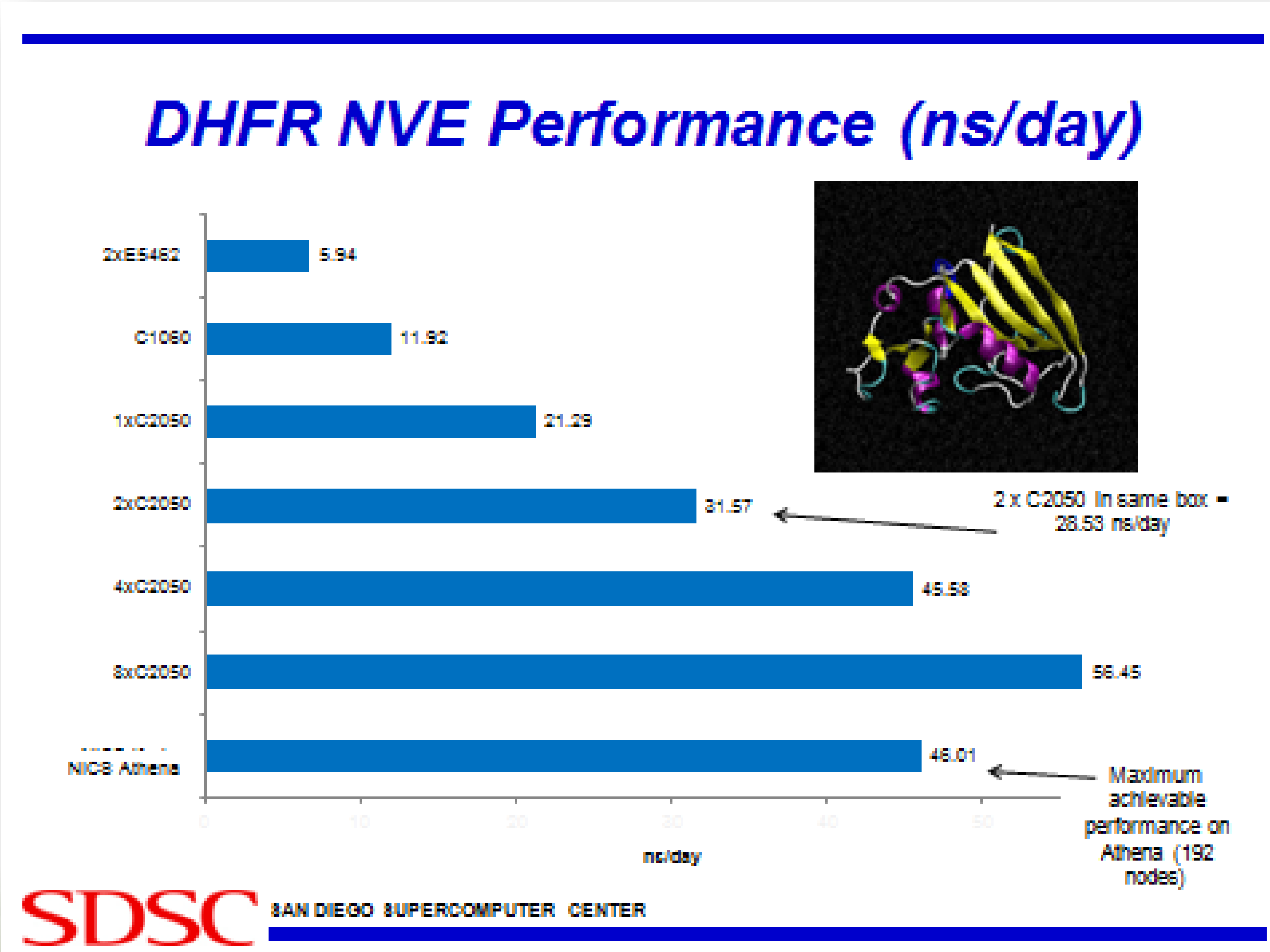
Preliminary results from KID

GROMACS

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



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



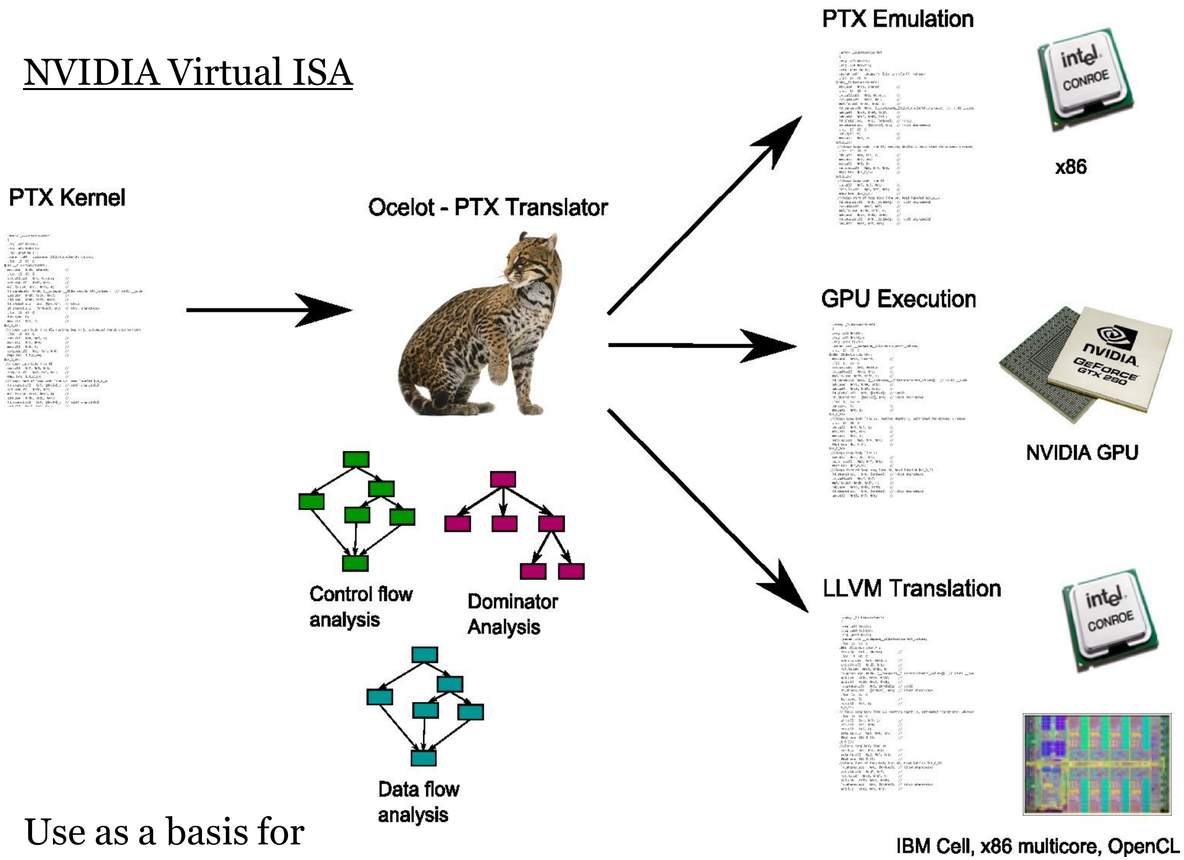
SOFTWARE

Keeneland Software Environment

- Integrated with NSF TeraGrid/XD
  - Including TG and NICS software stack
- Programming environments
  - CUDA
  - OpenCL
  - Compilers
    - PGI
      - Accelerate, CUDA Fortran
    - OpenMP 3.0
  - Scalable debuggers
  - Performance tools
- Additional software activities
  - Benchmarks
  - Performance and correctness tools
  - Scientific libraries
  - Virtualization

Ocelot: Dynamic Execution Infrastructure

<u>NVIDIA Virtual ISA</u>



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.

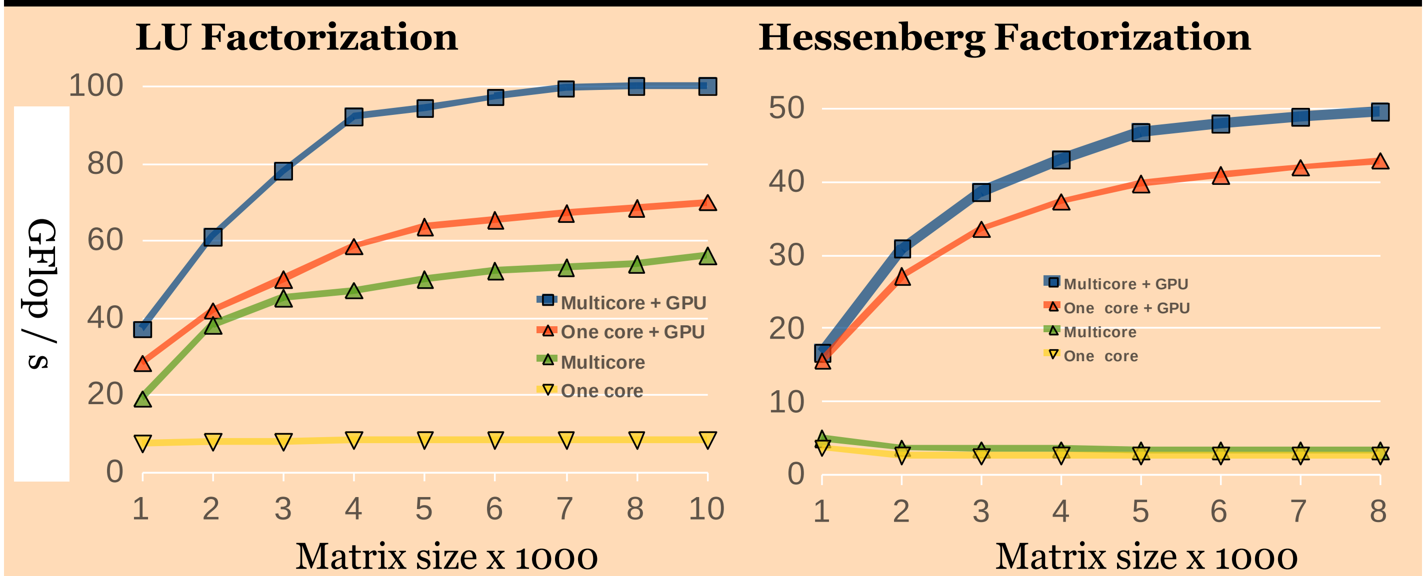
Gregory Diamos, Dhuw Choudhary, Andrew Kerr, Sudhakar Yalamanchili

Georgia Tech NICS THE UNIVERSITY of TENNESSEE OAK RIDGE National Laboratory NVIDIA hp NSF

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

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

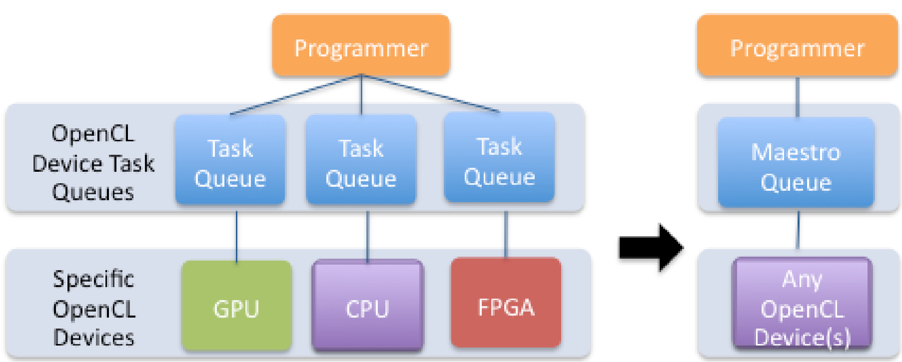
Objectives

- 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

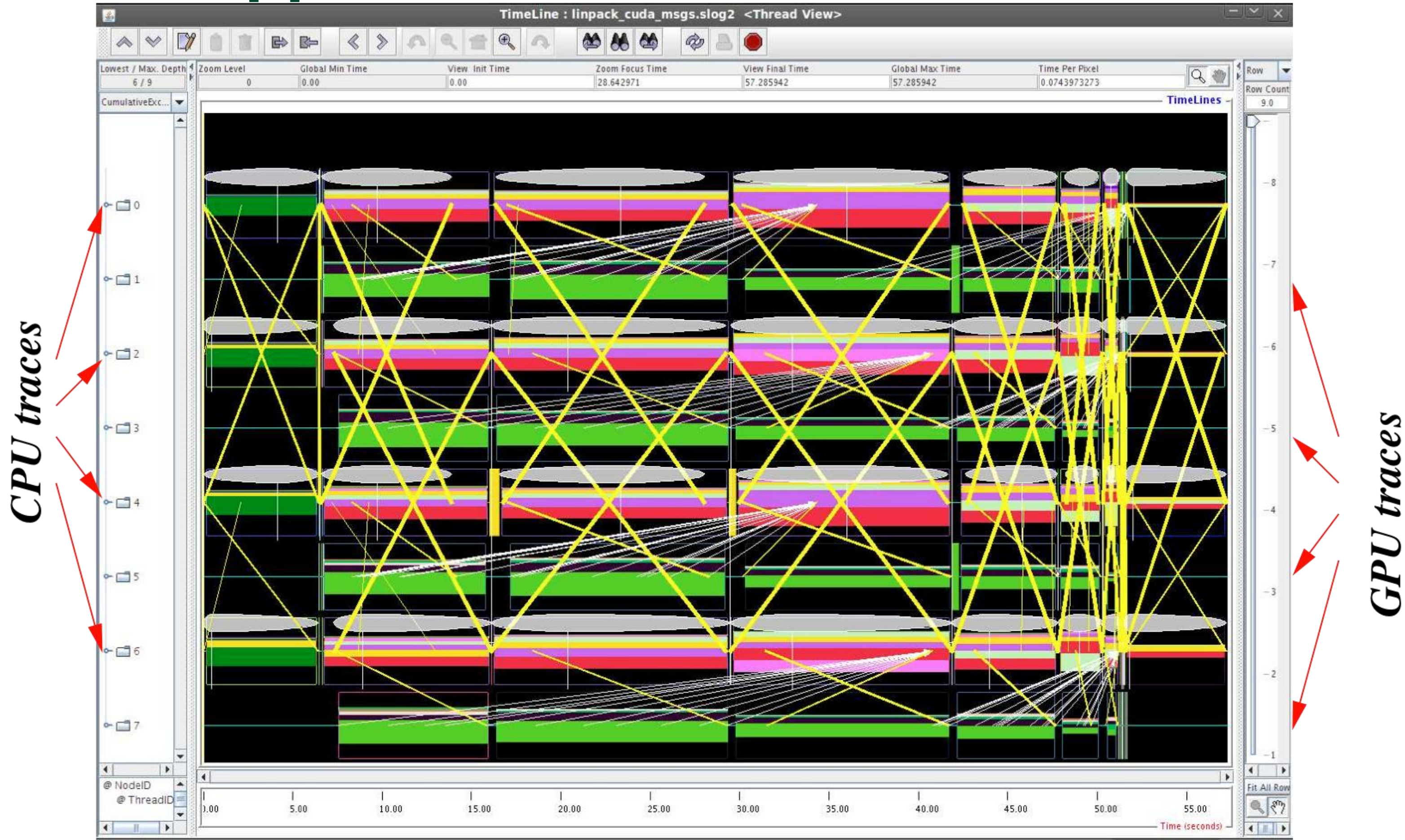
The proposed Maestro runtime simplifies programming heterogeneous systems by unifying OpenCL task queues into a single high-level queue.



Impact

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

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 Conference Supercomputing., 2010.*

BENCHMARKS

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

SHOC 1.0 available

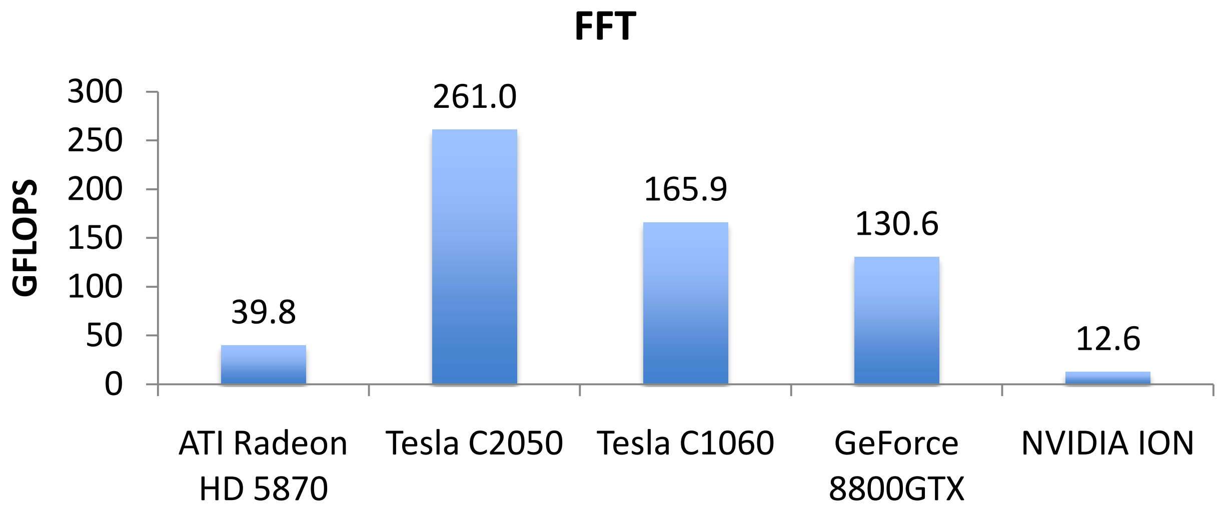
- Level 0
  - **BusSpeedDownload**: measures bandwidth of transferring data across the PCIe bus to a device.
  - **BusSpeedReadback**: measures bandwidth of reading data back from a device.
  - **DeviceMemory**: measures bandwidth of memory accesses to various types of device memory including global, 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 n-body 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.

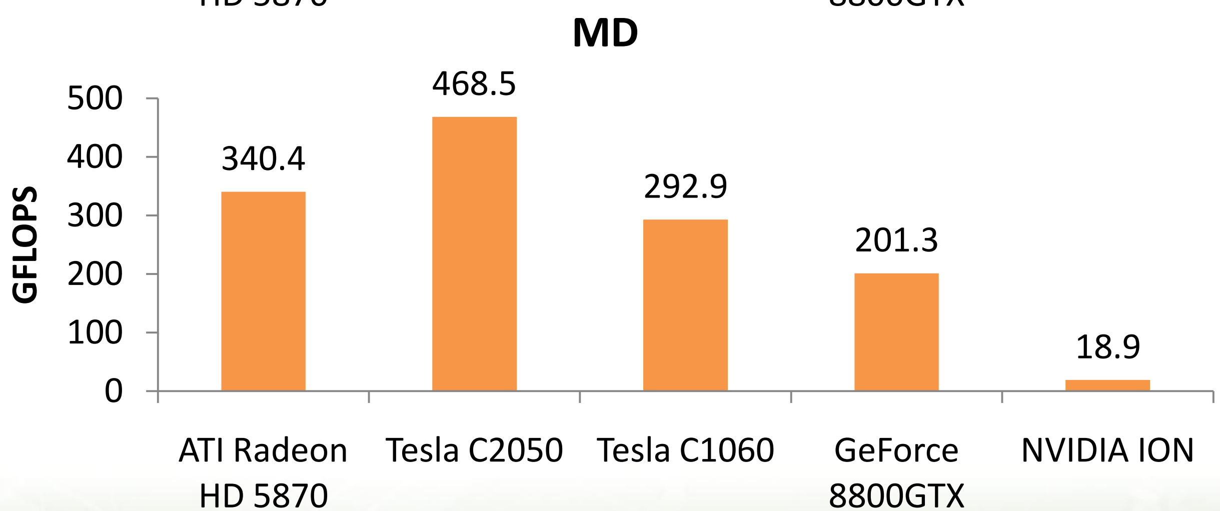
A. Danalis, G. Marin, C. McCurdy, J. Meredith, P.C. Roth, K. Spafford, V. Tipparaju, and J.S. Vetter, "The Scalable HeterOgeneous Computing (SHOC) Benchmark Suite," in Third Workshop on General-Purpose Computation on Graphics Processors (GPGPU 2010)`. Pittsburgh, 2010

Paper also includes energy and CUDA comparisons.

Beta software available at http://bit.ly/shocmarx

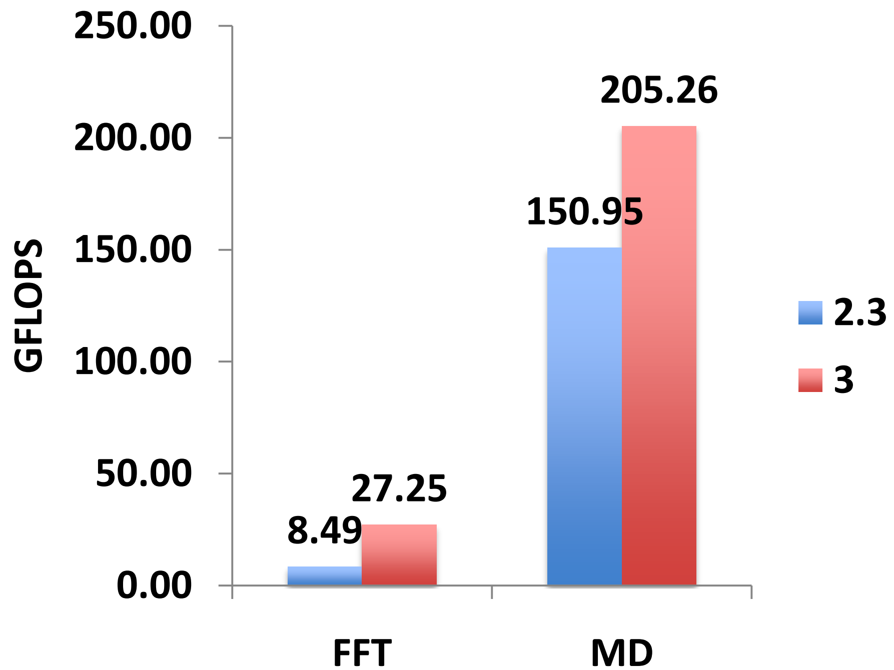
Compare Different GPUs





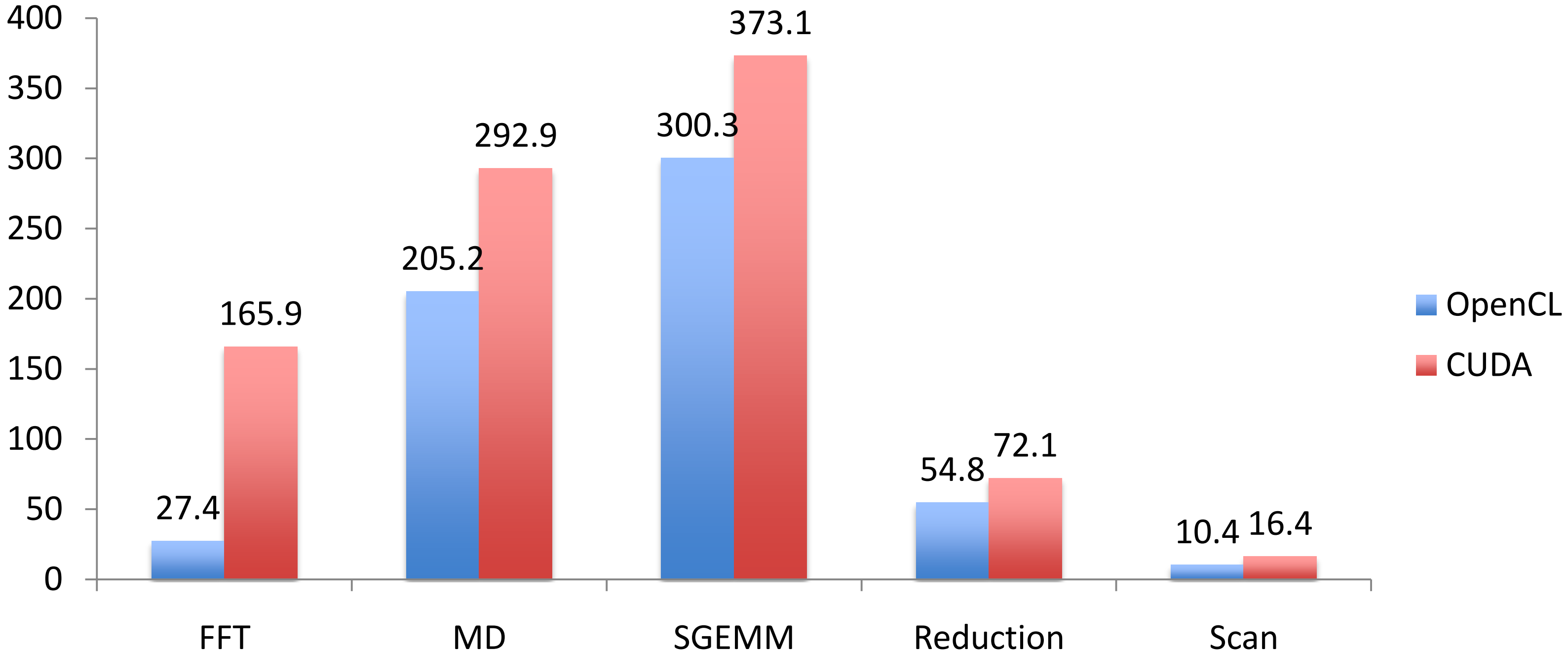
- Single Precision
- ECC On (for Tesla C2050)
- Radeon HD 5870: AMD OpenCL v2.1
- Tesla C2050 CUDA 3.1b
- Others CUDA 3.0

Longitudinal OpenCL Performance



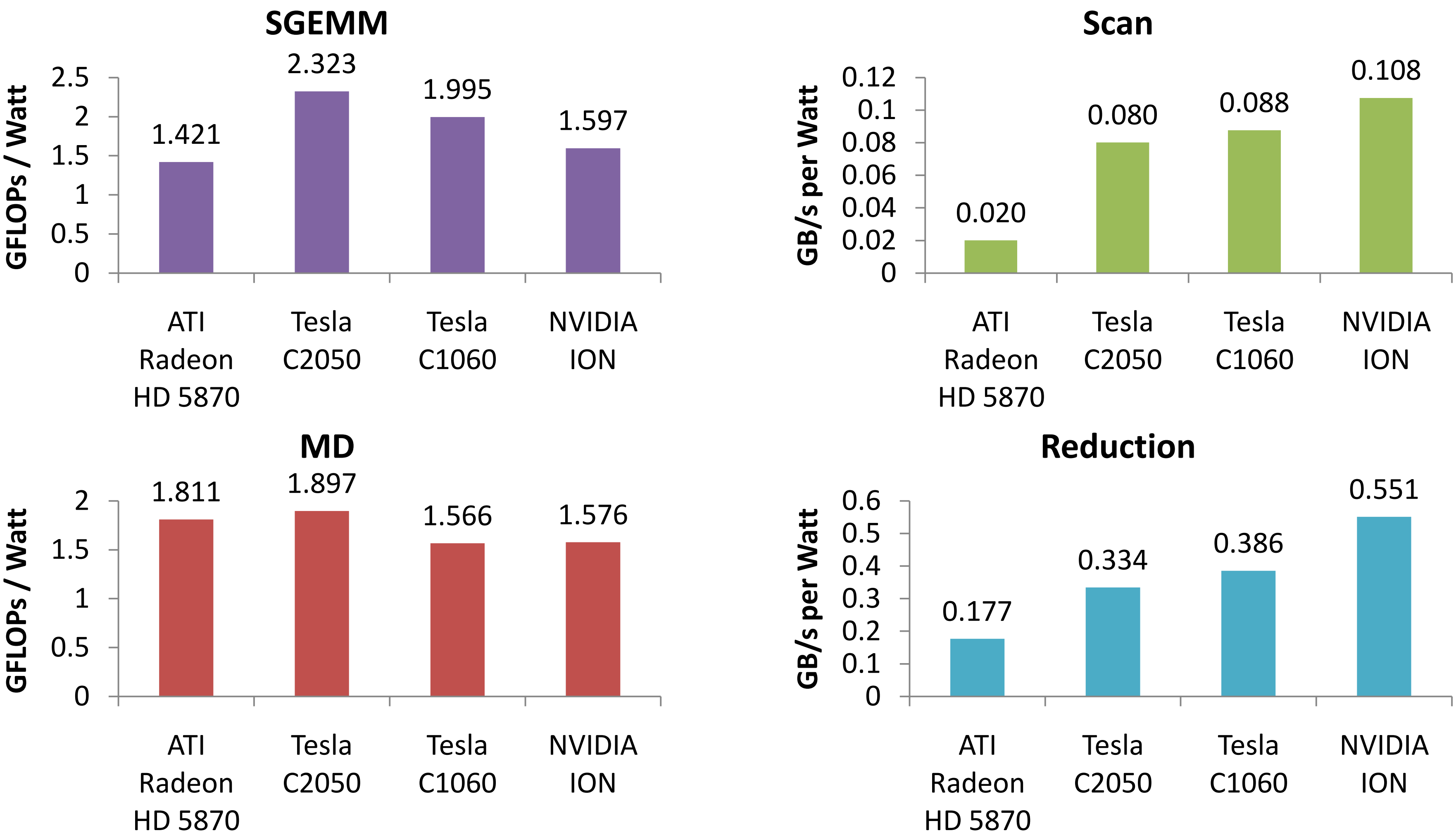
- Single precision, Tesla C1060 GPU
- Comparing NVIDIA OpenCL implementation from 2.3 and 3.0 GPU Computing SDK

Compare OpenCL and CUDA



- OpenCL improving, but still trailing CUDA
- Tesla C1060, Single Precision, CUDA and OpenCL 3.0
- FFT/MD/SGEMM – GFLOPS, Reduction/Scan – GB/s

Energy Efficiency



- Single precision, calculated using vendor’s TDP – Ion very efficient for bandwidth bound problems

FUTURE SYSTEMS

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

Main Objectives

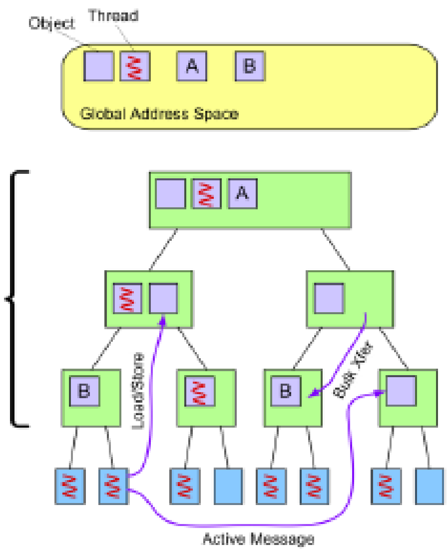
- Two orders of magnitude increase in application execution energy efficiency over today's CPU systems.
- Improve programmer productivity so that the time required to write a parallel program achieving a large fraction of peak efficiency is comparable to the time required to write a serial program today.
- 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

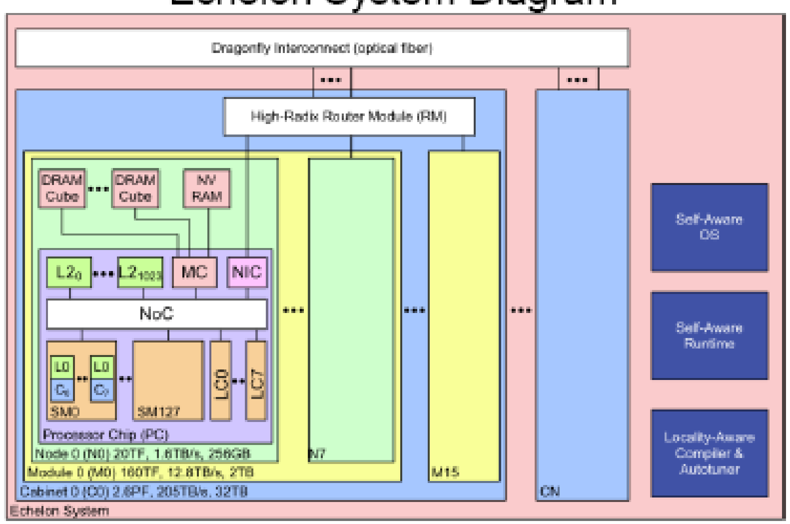
- Programming systems that express concurrency/locality abstractly; autotuning for hardware mapping.
- Self-aware runtime reacts to changes in environment, workload (load-balance), fault states.
- Fine-grained, energy-optimized, multithreaded throughput cores + latency-optimized cores.
- Software selective 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

Recap

- The HPC community has several (new) constraints
  - Power, Performance, Facilities, Cost
- Emerging technologies will play a critical role
- Heterogeneous computing with GPUs offers some opportunities and challenges
  - High performance; good performance per watt
  - Programmability; limited applicability
- KID is up and running
- Keeneland Final System coming in 2012

For more information:

vetter@computer.org

http://keeneland.gatech.edu

http://ft.ornl.gov

http://www.cse.gatech.edu

http://www.cercs.gatech.edu

http://icl.cs.utk.edu

http://www.nics.tennessee.edu/

http://nsf.gov/dir/index.jsp?org=OCI