A UT/ORNL PARTNERSHIP
NATIONAL INSTITUTE FOR COMPUTATIONAL SCIENCES
NICS

Keenland—Enabling Heterogeneous Computing for the Open Science Community

Jeffrey Vetter, Jack Dongarra, Richard Fujimoto, Thomas Schulthess, Karsten Schwan, Sudha Yalamanchili, Richard Glassbrook, Jim Ferguson, Patricia Kovatch, Stephen McNally, Bruce Loftis, Jeremy Meredith, Philip Roth, Kyle Spafford, Jim Rogers, Arlene Washington, and others

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.)

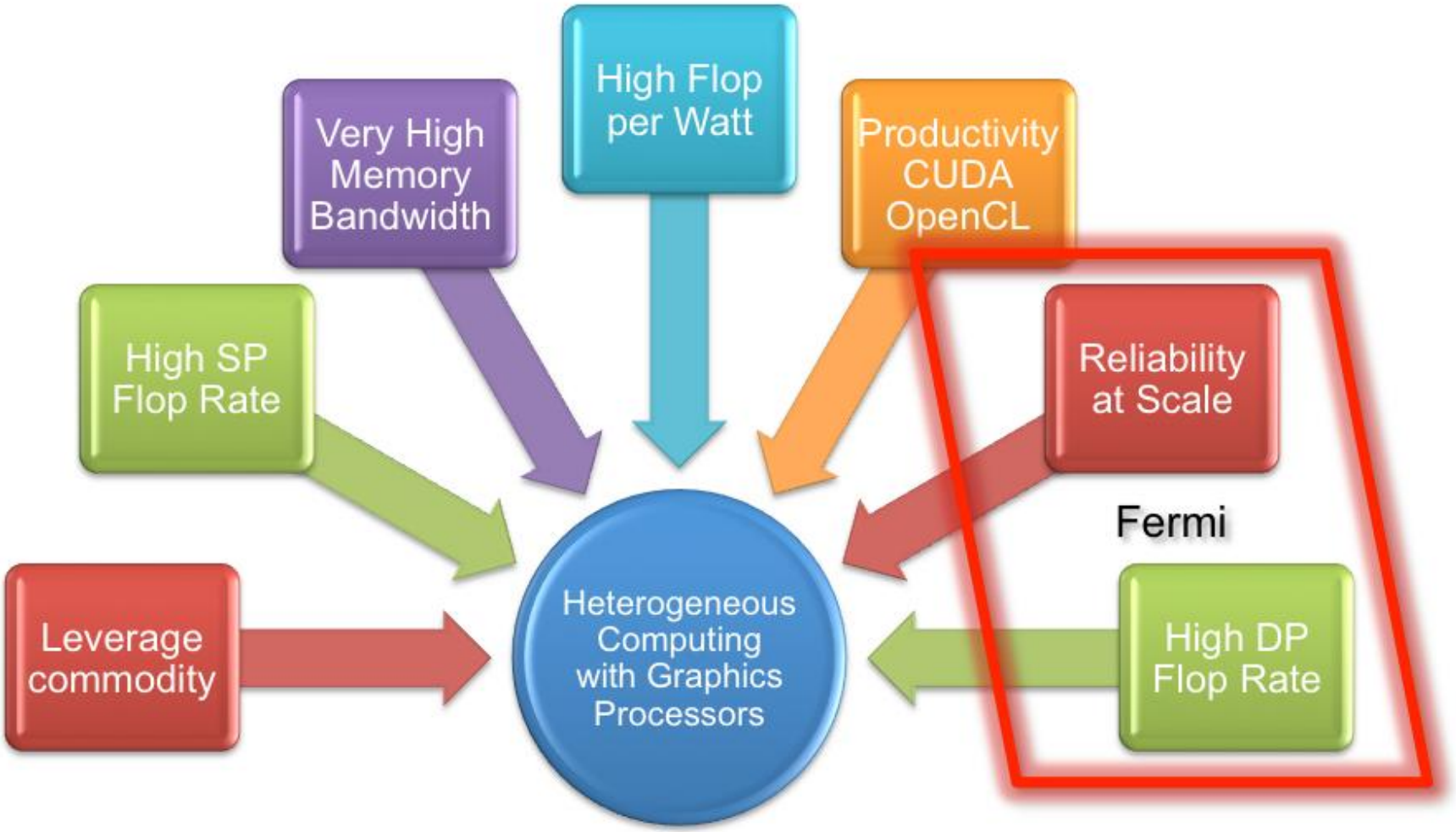
NICS

Keeneland is a NSF-funded partnership to enable large-scale computational science on heterogeneous architectures using GPUs

- Deploy and operate a large-scale heterogeneous computer
  - Keeneland Initial Delivery System (KIDS)
    - October 2010 – NOW OPERATIONAL
  - Full-Scale (FS) system – Spring 2012
- Operations, user support as a TG/XD resource
- Productivity tools and applications readiness
- Education, outreach, and training for scientists, students, industry on heterogeneous platforms

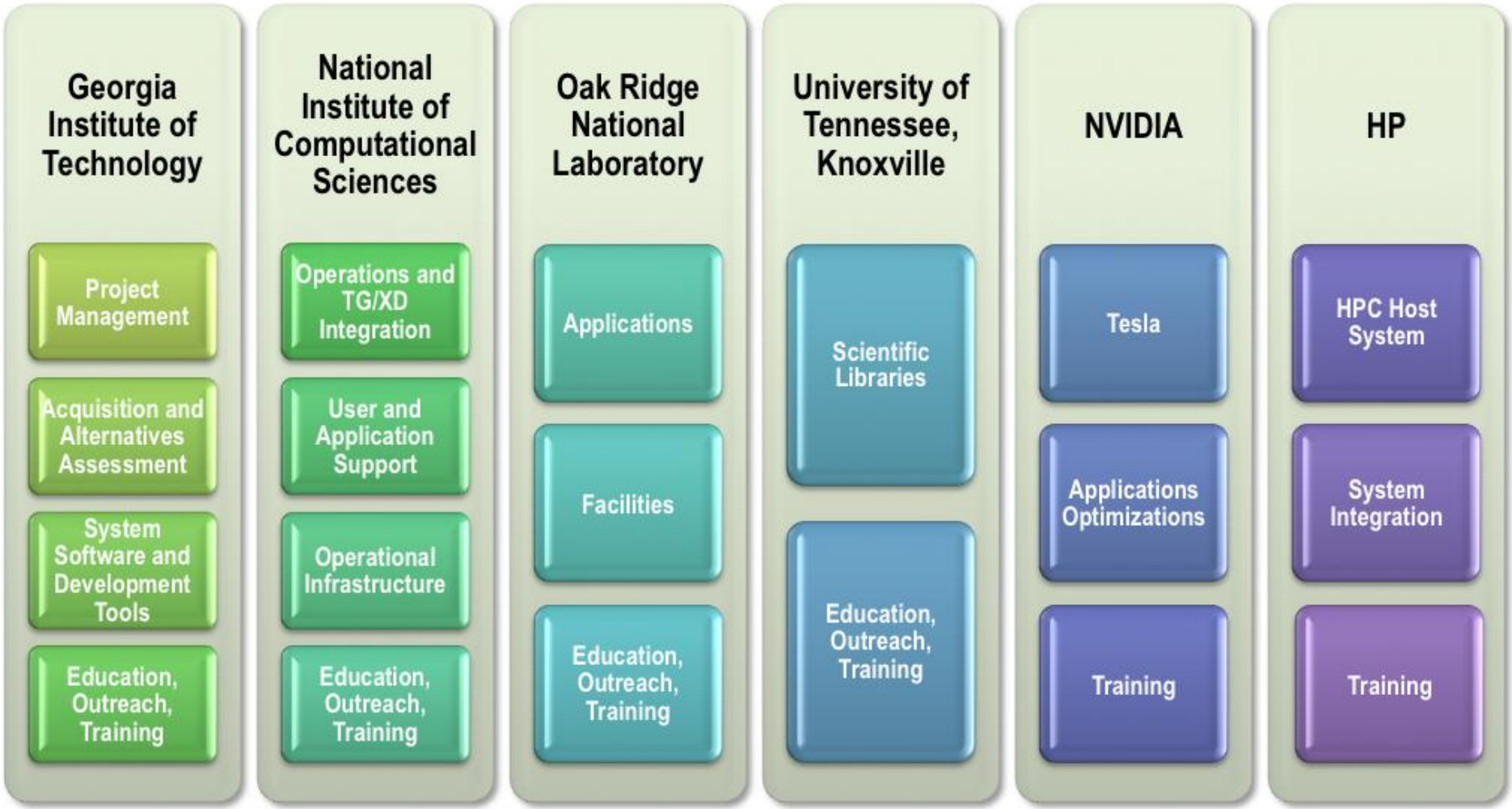
NICS

GPU rationale: What's different now?



NICS

Keeneland partners



NICS

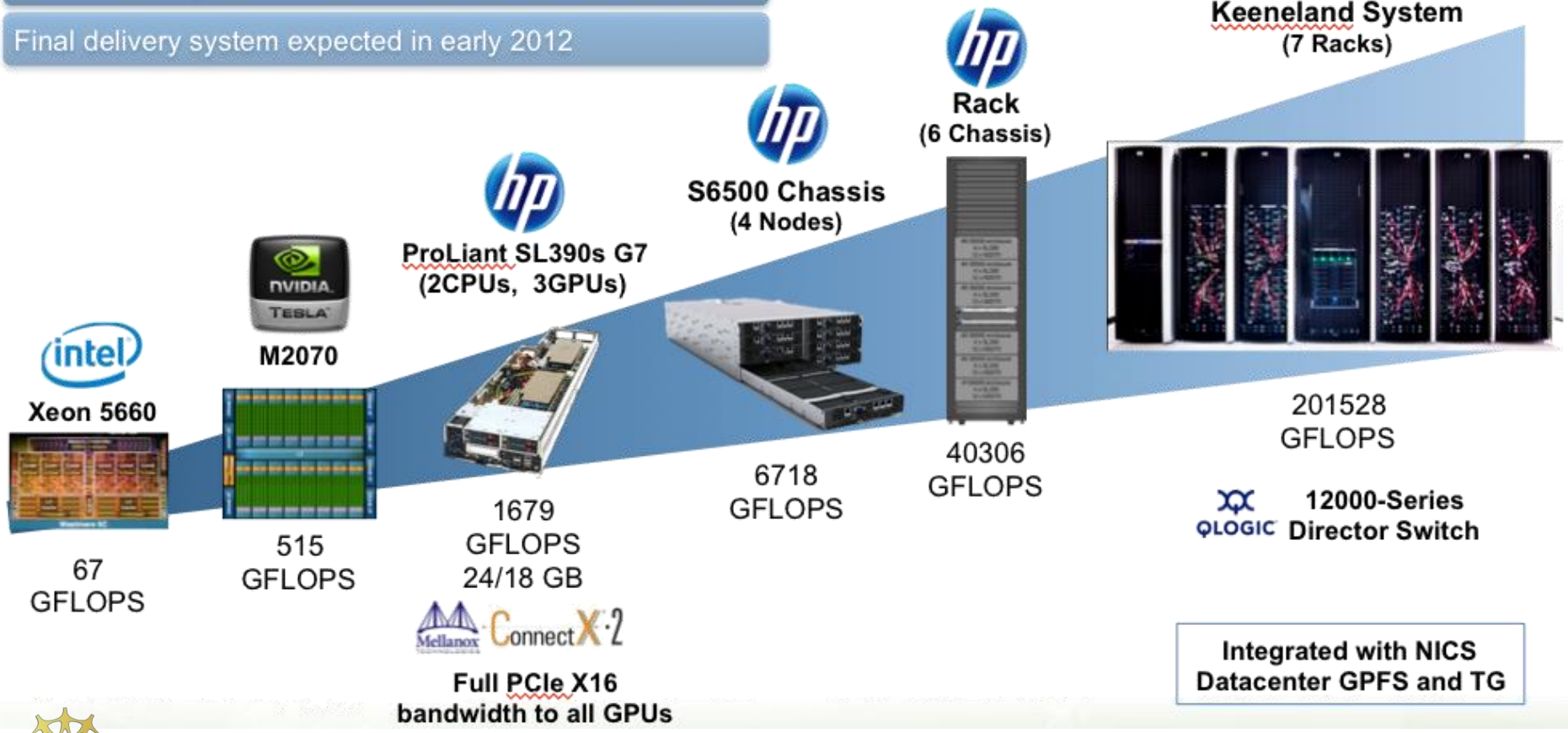
Keeneland – enabling heterogeneous computing for the open science community

Initial Delivery system procured and installed in Oct 2010

201 TFLOPS in 7 racks (90 sq ft incl service area)

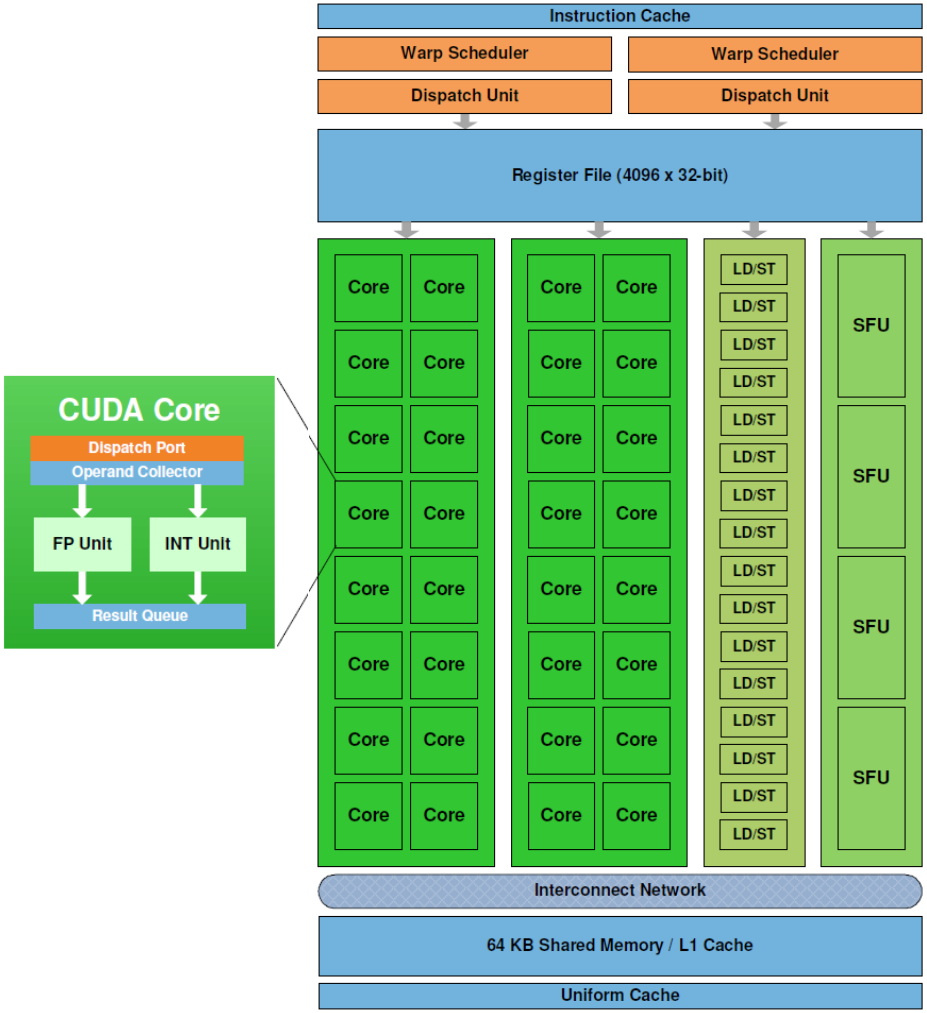
650 MFLOPS per watt on HPL

Final delivery system expected in early 2012



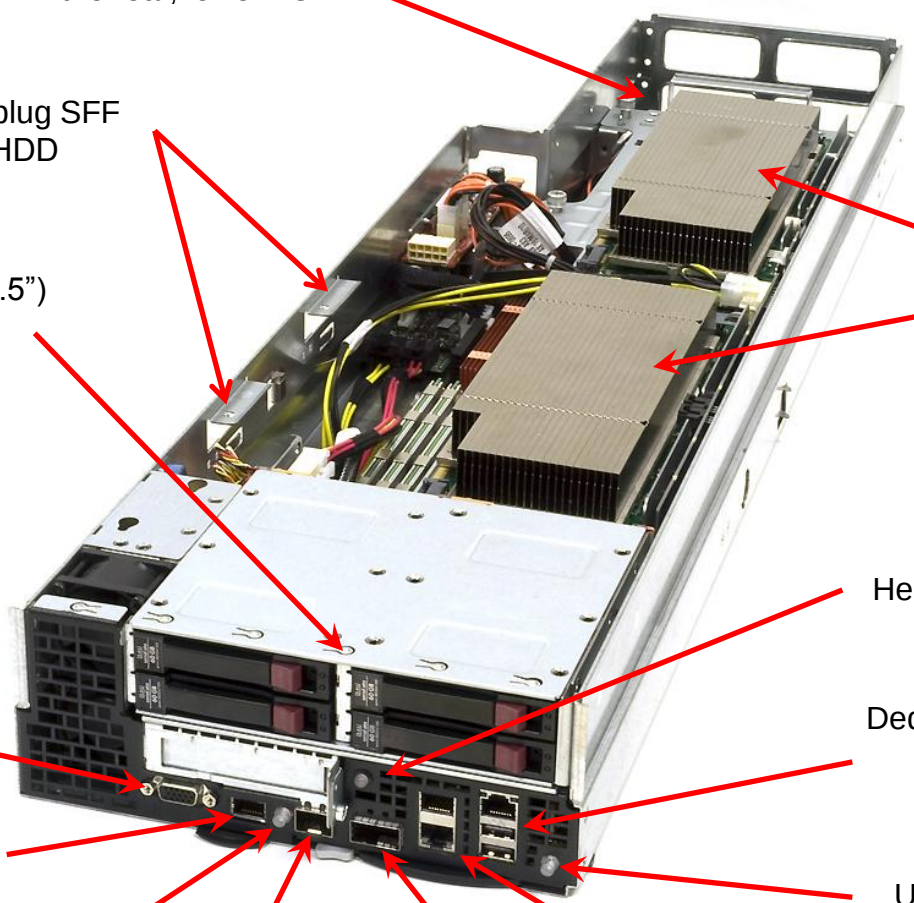
NVIDIA Fermi

- 3B transistors
- ECC
- 8x the peak double precision arithmetic performance over NVIDIA's last generation GPU
- 448 CUDA Cores featuring the new IEEE 754-2008 floating-point standard
- NVIDIA Parallel DataCache
- NVIDIA GigaThread Engine
- Debuggers, language support



NICS

HP ProLiant SL390s G7 2U half width tray



1 GPU module in the rear, lower 1U

2 Non-hot plug SFF (2.5”) HDD

4 Hot plug SFF (2.5”) HDDs

2 GPU modules in upper 1U

Health LED

VGA

Dedicated management iLO3 LAN & 2 USB ports

Serial (RJ45)

UID LED & Button

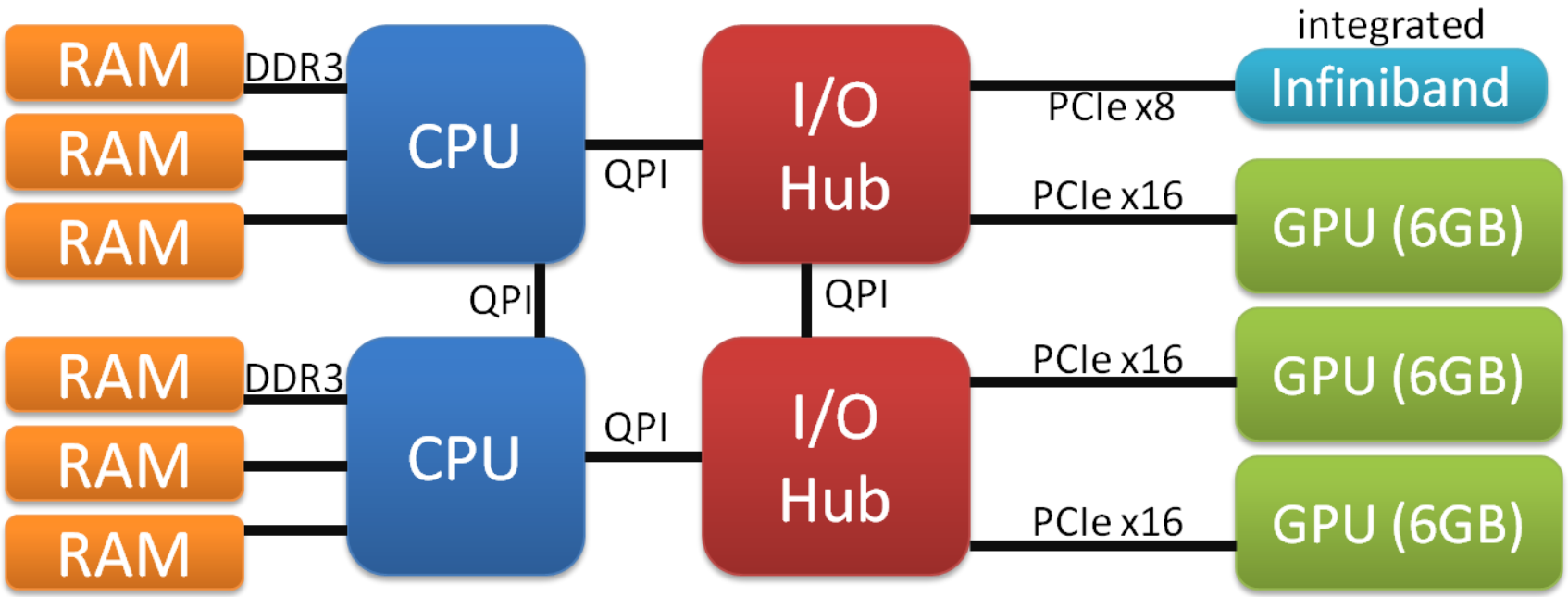
Power Button

SFP+

QSFP (QDR IB)

Dual 1GbE

Keeneland node architecture SL390



NICS

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

NICS

Keeneland ID installation – 10/29/10

NICS

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
- Acceptance tests completed on ??

NICS

Productivity tools are key!

- XHPC system has a number of differences that warrant risk mitigation efforts
  - Scientific libraries
  - Software tools
  - Runtime software support
- The rapid pace of change in GPU design requires ongoing application evaluation
  - Fermi's cache and DP improvements may allow new applications to take advantage of GPUs
  - Identify promising applications and application metrics
- Application acceleration
  - Migrate selected applications to Keeneland
- Keeneland is working on
  - Performance and correctness tools
  - Scientific libraries
  - Virtualization
  - Benchmarks

NICS

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 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 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.

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

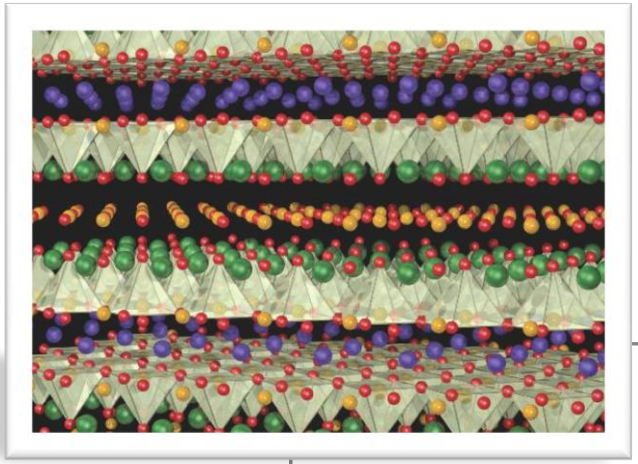
Paper also includes energy and CUDA comparisons.

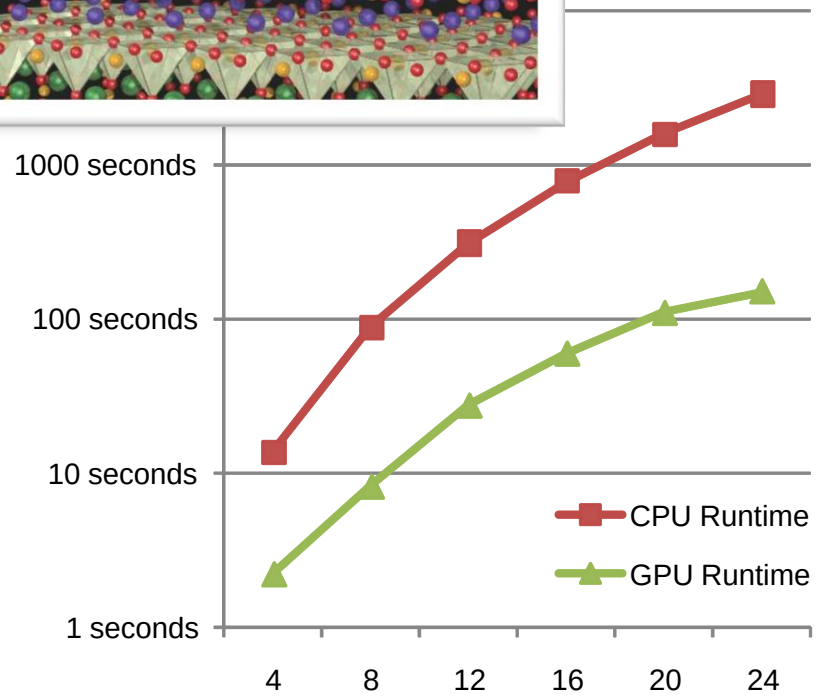
Beta software available at http://ft.ornl.gov/doku/shoc/start

NICS

Computational materials: Case study

- **Quantum Monte Carlo simulation**
  - **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**
  - **Required detailed accuracy study and mixed precision port of app**
- **Full parallel app is 5x faster, start to finish, on a GPU-enabled cluster**





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*]

NICS

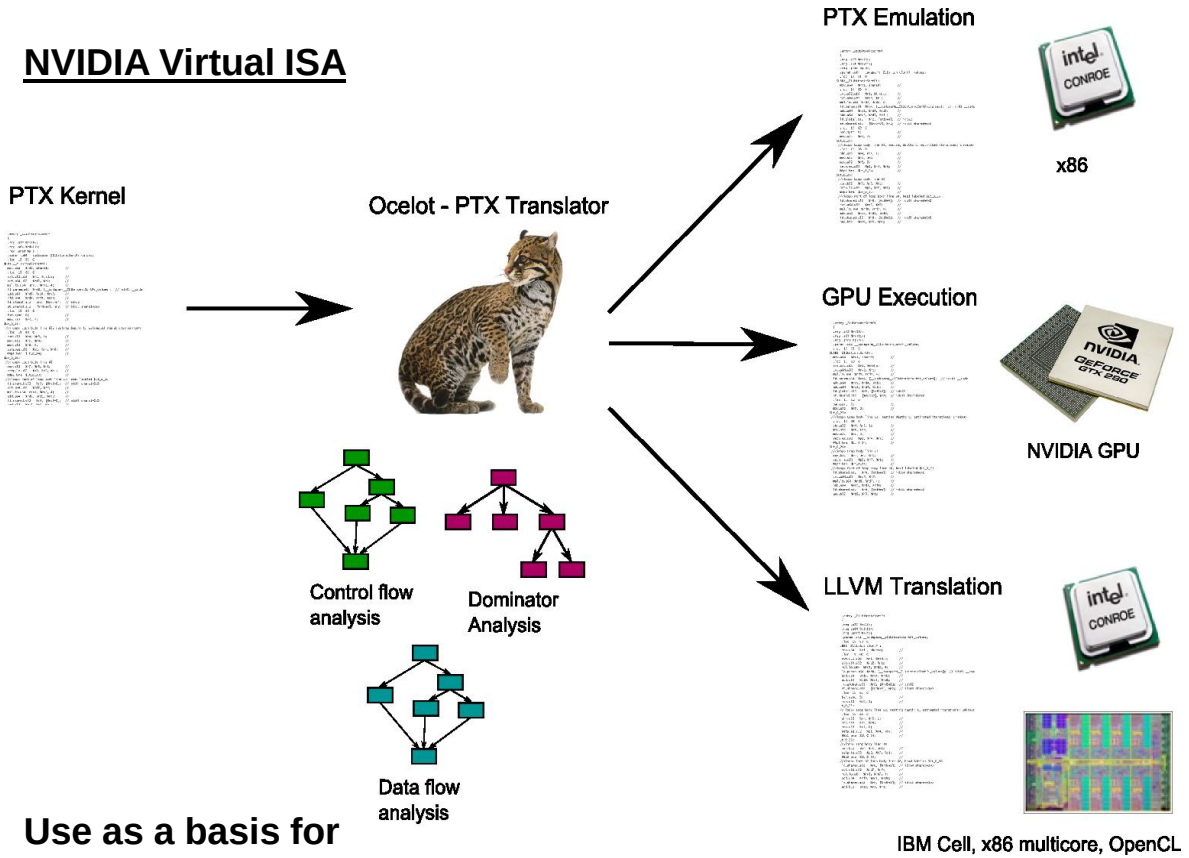
Combustion with S3D: Case study

- Application for 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
  - 14.3x SP speedup
  - 9.32x 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

NICS

Ocelot: Dynamic execution infrastructure



PTX 1.4 compliant emulation
- **Validated on full CUDA SDK**
- **Open source version released**

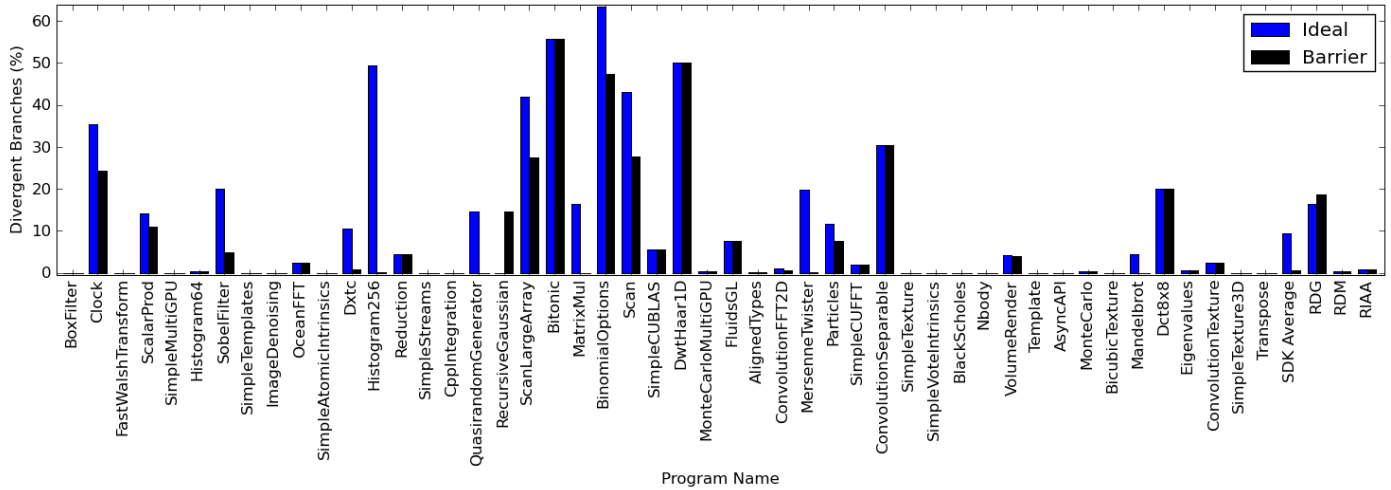
http://code.google.com/p/gpuocelot/

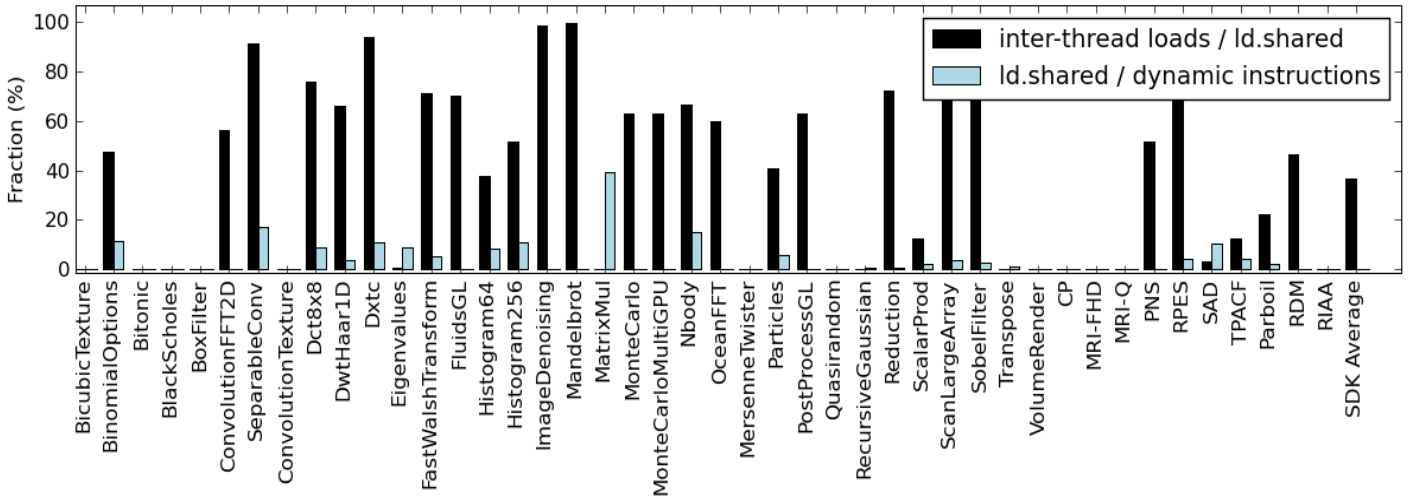
Use as a basis for
- **Insight → workload characterization**
- **Performance tuning → detecting memory bank conflicts**
- **Debugging → illegal memory accesses, out of bounds checks, etc.**

Gregory Diamos, Dhuv Choudhary, Andrew Kerr, Sudhakar Yalamanchili

NICS

Workload analysis: Examples





Branch Divergence

- Study of control flow behavior
- Motivate synchronization support

Inter-thread Data Flow

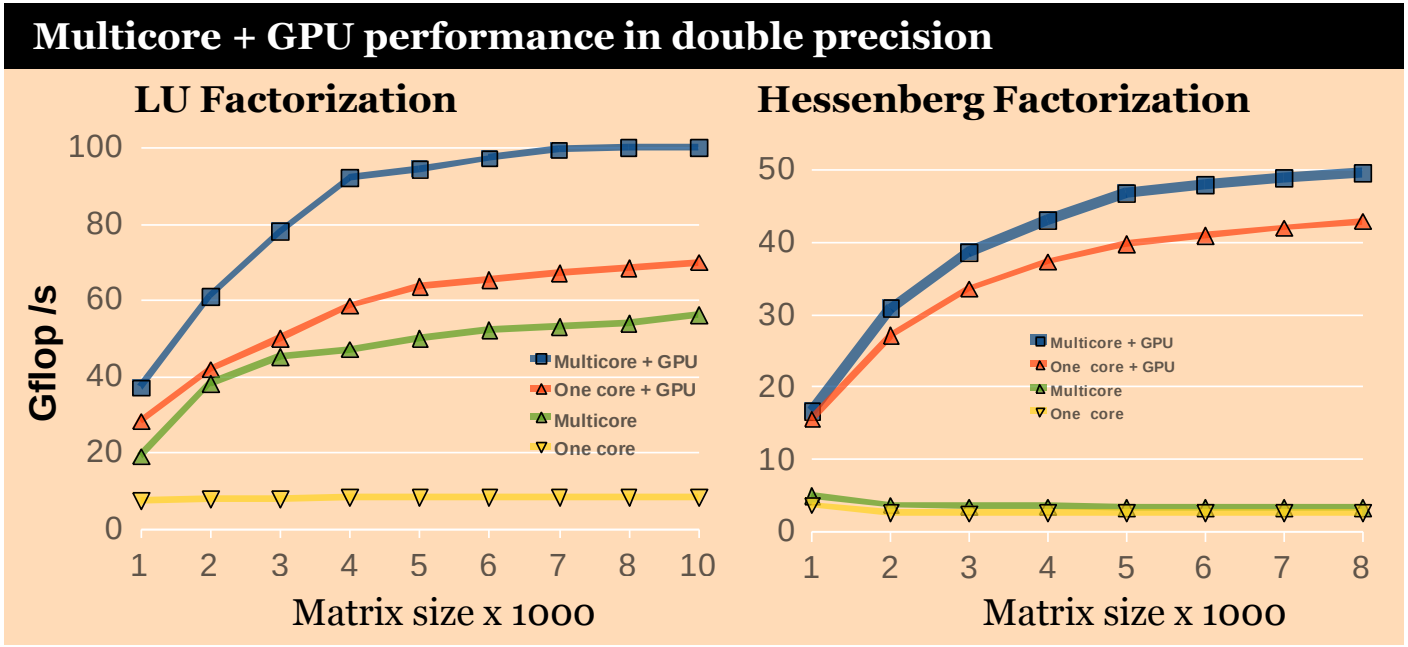
- Study of data sharing patterns
- Motivate architectural support

Gregory Diamos, Dhuv Choudhary, Andrew Kerr, Sudhakar Yalamanchili

NICS

One- and two-sided multicore+GPU factorizations

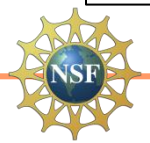
- These will be included in upcoming MAGMA releases
- Two-sided factorizations cannot be efficiently accelerated on homogeneous x86-based multicores (above) because of memory-bound operations
  - MAGMA provided hybrid algorithms that overcome those bottlenecks (16x speedup!)



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



NICS

Contact

vetter@computer.org

http://keeneland.gatech.edu

http://www.cse.gatech.edu

http://www.cercs.gatech.edu

http://icl.cs.utk.edu

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

http://ft.ornl.gov

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