Keeneland - Enabling Heterogeneous Computing for the Open Science Community

Jeffrey Vetter, Dick Glassbrook, Jack Dongarra, Karsten Schwan, Sudha Yalamanchili, Bruce Loftis, Stephen McNally, Jeremy Meredith, Patti Reed, Jim Rogers, Philip Roth, Kyle Spafford, Kevin Sharkey, and many others

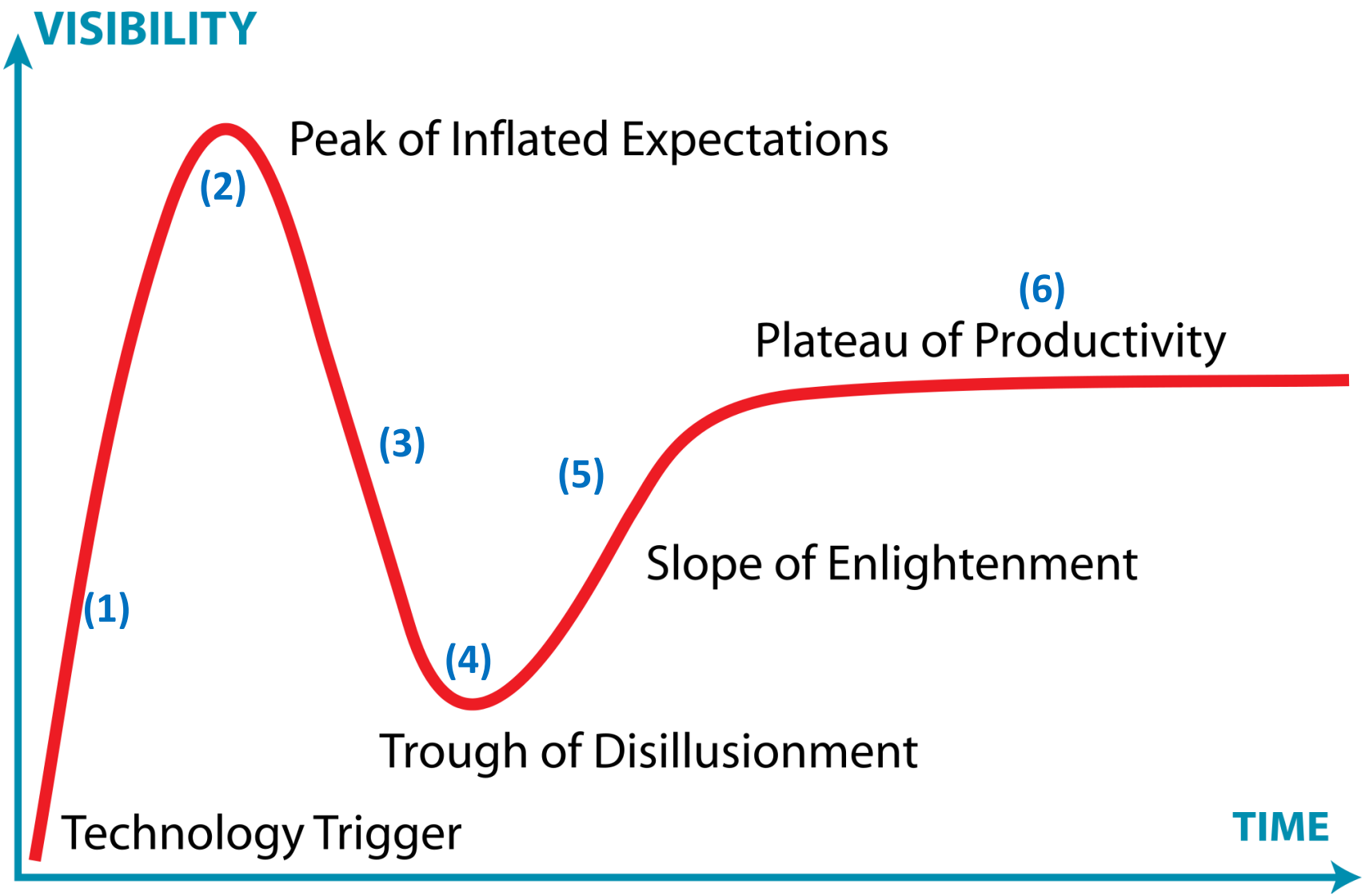
Presented to the

Programming weather, climate, and earth-system models on heterogeneous multi-core platforms workshop

September 7, 2011
National Center for Atmospheric Research in Boulder, Colorado

http://keeneland.gatech.edu

Morning Quiz

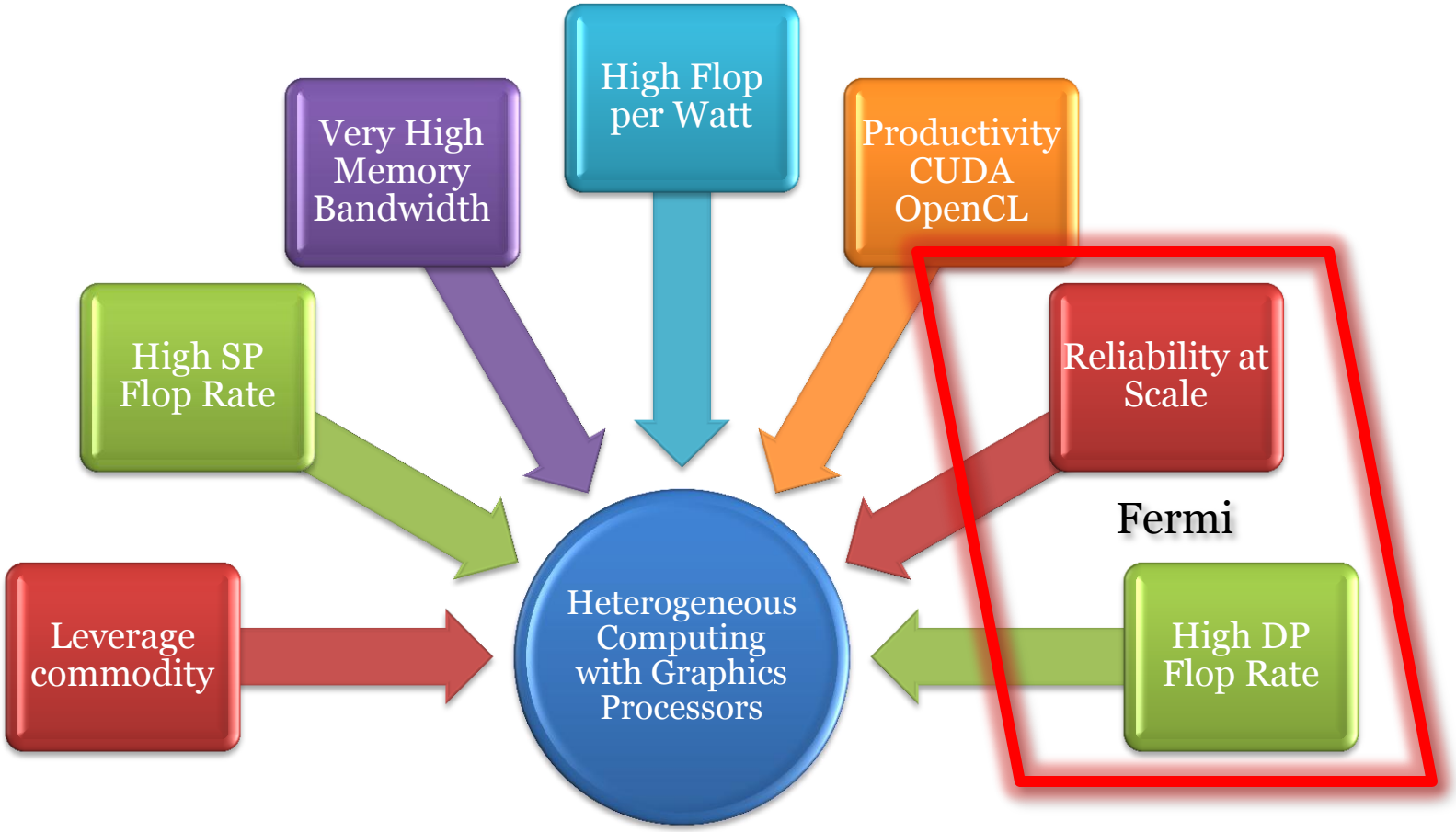


NSF Office of Cyber Infrastructure RFP

- 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.)

GPU Rationale – What's different now?



Keeneland High Level Goals (in one slide)

- Provide a new, innovative class of computing architecture to the NSF community for science
- Acquire, deploy, and operate two GPU clusters
  - Initial delivery - Operational
  - Full scale – Spring 2012
  - Operations, user support
- Ensure software tools, application development support for user productivity and success
- Perform technology assessment to track fast moving hardware and software
- Perform education, Outreach, Training for scientists, students, industry on these new architectures

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

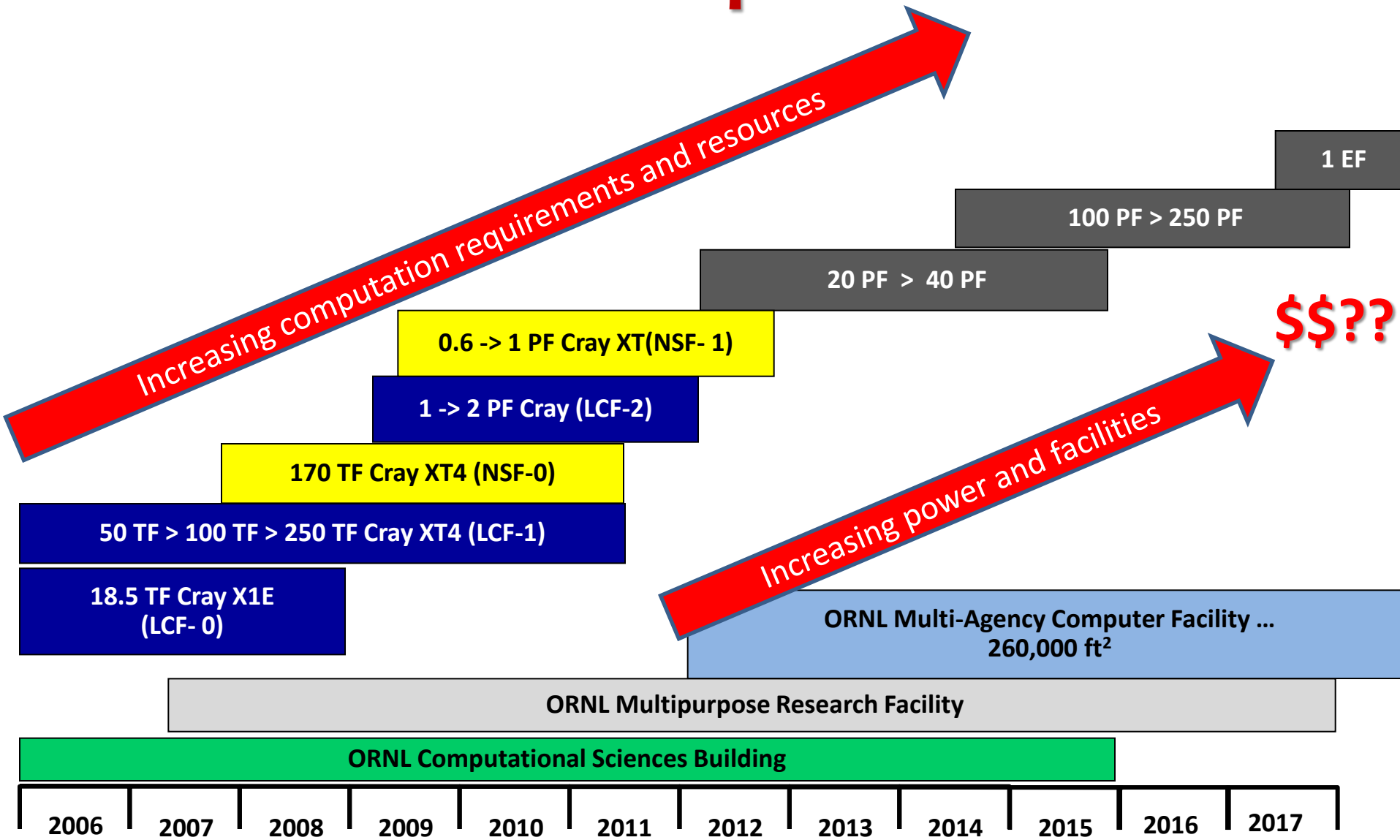
- Tesla
- Applications optimizations
- Training

HP

- HPC Host System
- System integration
- Training

Technology Trends

ORNL Roadmap to Exascale

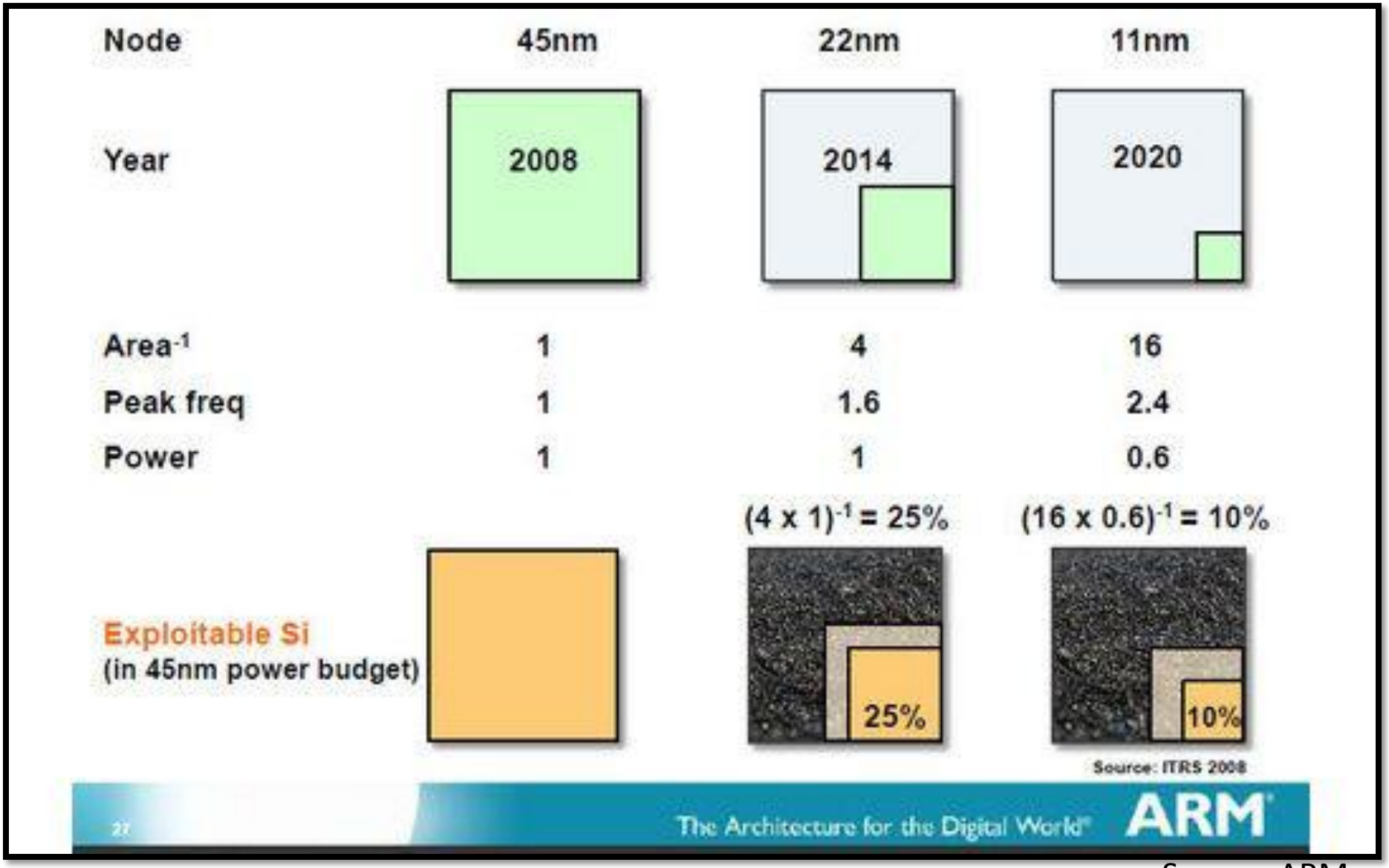


If energy costs ~$1/MW/yr, then how much is the energy cost for an exascale system?!?!

Contemporary Systems

Date	System	Location	Comp	Comm	Peak (PF)	Power (MW)
2010	Tianhe-1A	NSC in Tianjin	Intel + NVIDIA	Proprietary	4.7	4.0
2010	Nebulae	NSC In Shenzhen	Intel + NVIDIA	IB	2.9	2.6
2010	Tsubame 2	TiTech	Intel + NVIDIA	IB	2.4	1.4
2011	K Computer (612 cabinets)	Kobe	SPARC64 VIIIfx	Tofu	8.7	9.8
~2012	Cray ‘Titan’	ORNL	AMD + NVIDIA	Gemini	20?	7?
~2012	BlueGeneQ	ANL	SoC	IBM	10?	?
~2012	BlueGeneQ	LLNL	SoC	IBM	20?	?
	Others…					

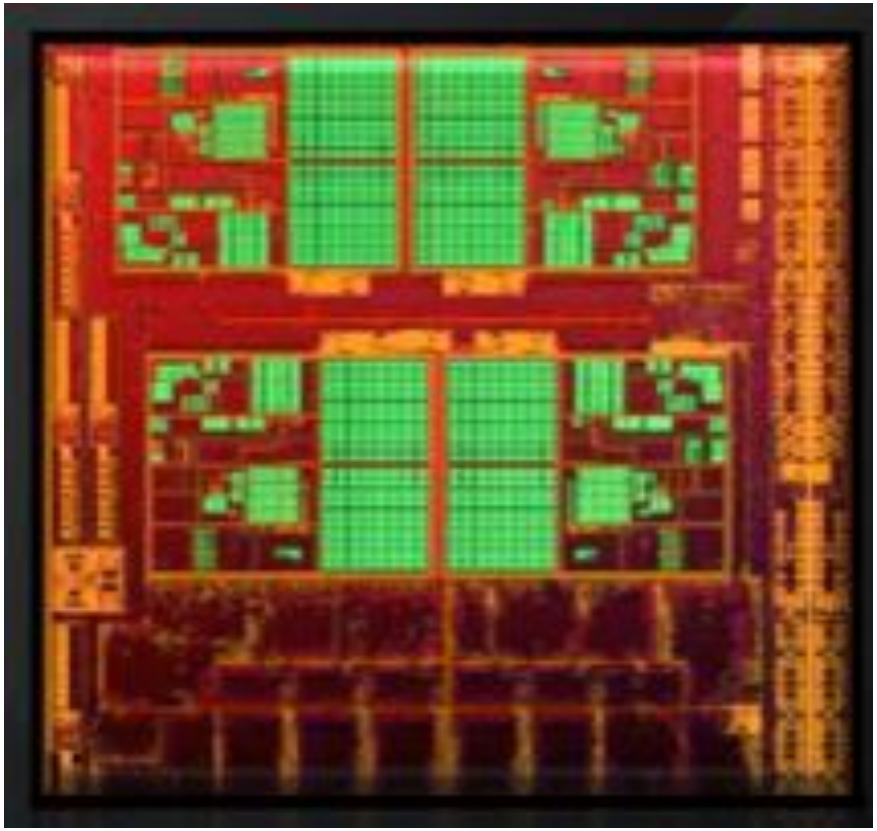
Dark Silicon



Source: ARM

AMD's Llano: A-Series APU

- Combines
  - 4 x86 cores
  - Array of Radeon cores
  - Multimedia accelerators
  - Dual channel DDR3
- 32nm
- Up to 29 GB/s memory bandwidth
- Up to 500 Gflops SP
- 45W TDP



Source: AMD

Keeneland ID Architecture

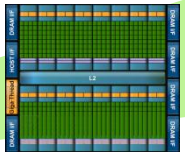
Keeneland – Initial Delivery System Architecture

- **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

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)

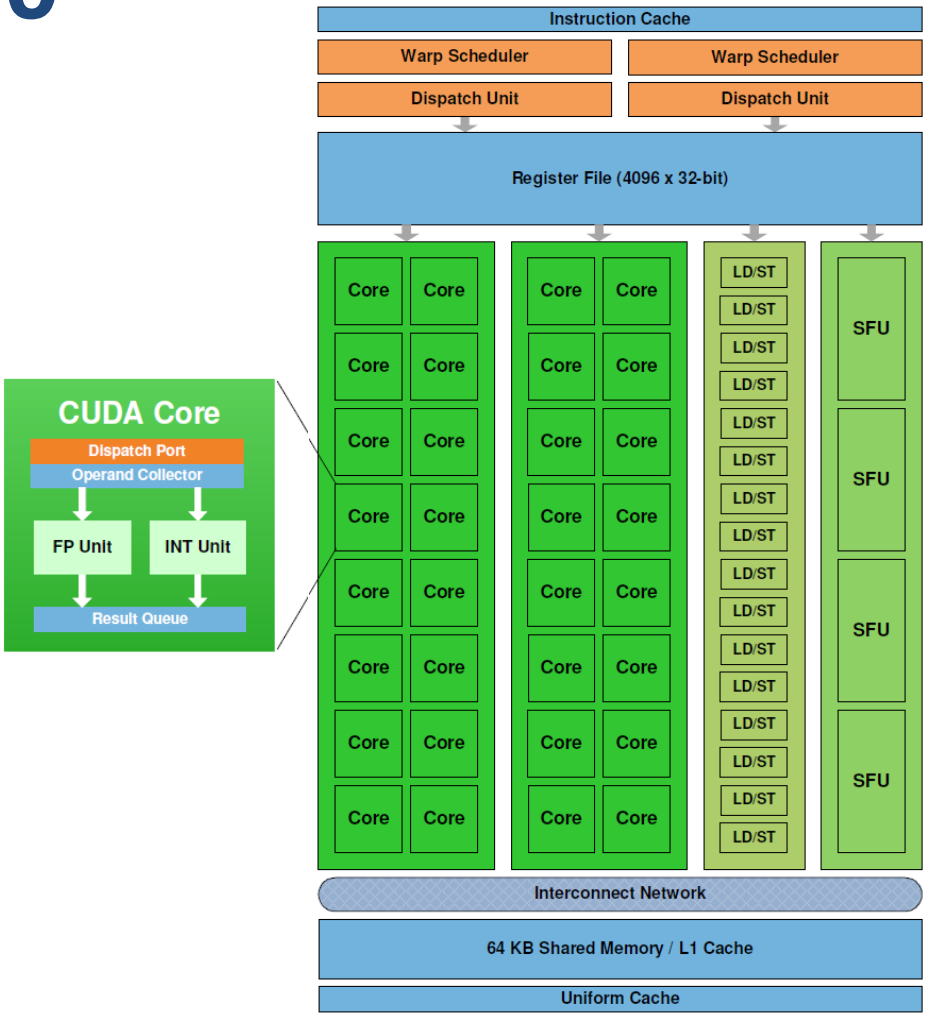
201528 GFLOPS

12000-Series Director Switch

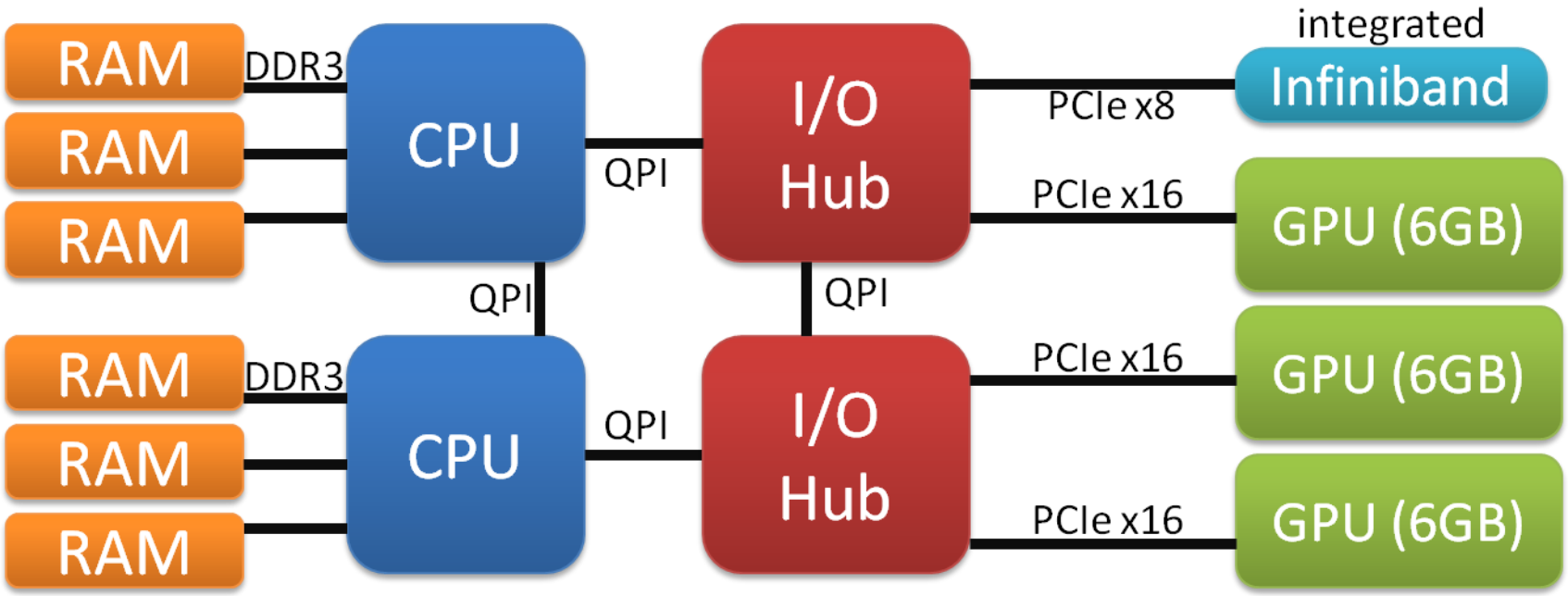
Integrated with NICS Datacenter GPFS and TG

NVIDIA Fermi/GF100

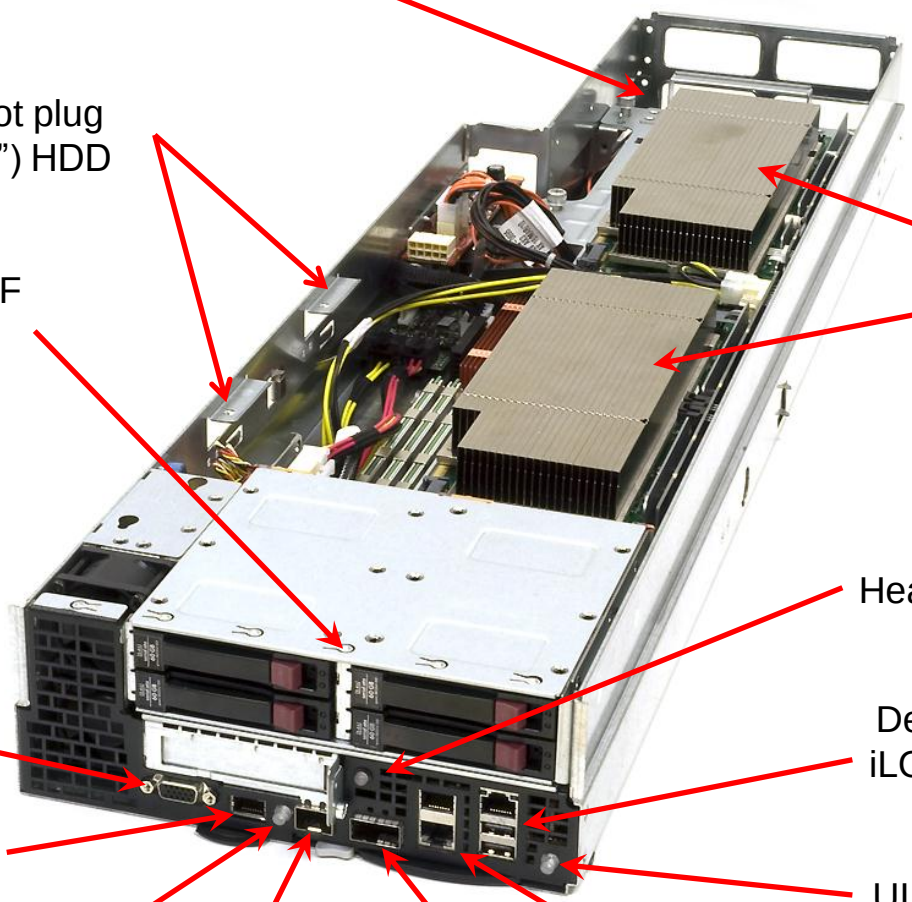
- 3B transistors in 40nm
- Up to 512 CUDA Cores
  - New IEEE 754-2008 floating-point standard
    - FMA
    - 8× the peak double precision arithmetic performance over NVIDIA's last generation GPU
  - 32 cores per SM, 21k threads per chip
- 384b GDDR5, 6 GB capacity
  - ~120-144 GB/s memory BW
- C/M2070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



Keeneland Node Architecture SL390



HP ProLiant SL390s G7 2U half width tray



Sampling of Applications

Early (Co-design) 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

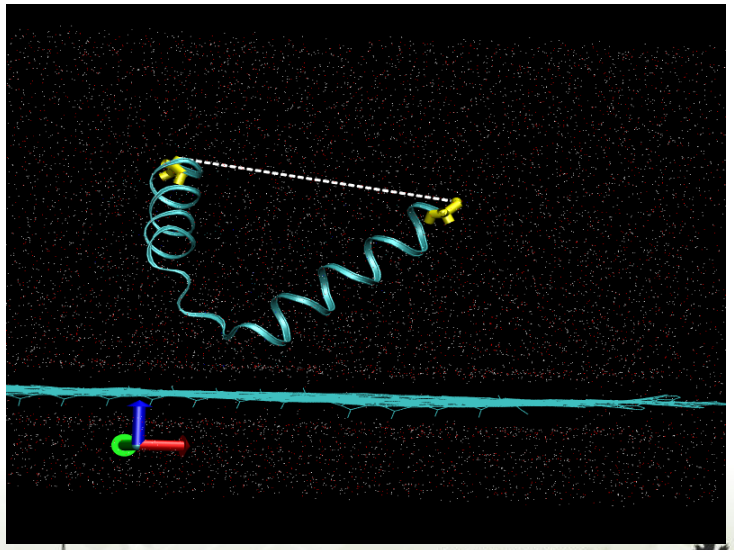
K. Spafford, J. Meredith, J. S. Vetter, J. Chen, R. Grout, and R. Sankaran. Accelerating S3D: A GPGPU Case Study. Proceedings of the Seventh International Workshop on Algorithms, Models, and Tools for Parallel Computing on Heterogeneous Platforms (HeteroPar 2009) Delft, The Netherlands.

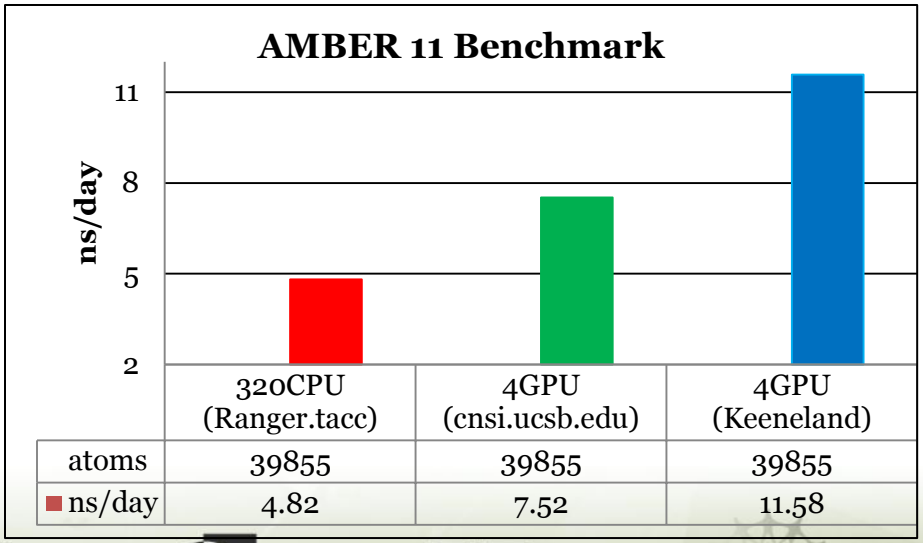
Joan-Emma Shea at UCSB

Peptide folding on surfaces

- Peptide folding on a hydrophobic surface
  - www.chem.ucsb.edu/~sheagroup
- Surfaces can modulate the folding and aggregation pathways of proteins. Here, we investigate the folding of a small helical peptide in the presence of a hydrophobic surface of graphite. Simulations are performed using explicit solvent and a fully atomic representation of the peptide and the surface.

- Benefits of running on a GPU cluster:
  - Reduction in the the number of computing nodes needed: one GPU is at least faster than 8 CPUs in GPU-accelerated AMBER Molecular Dynamics.
  - The large simulations that we are currently running would be prohibitive using CPUs. The efficiency of the CPU parallelization becomes poorer with increasing number of CPUs.
  - It can also decrease consumption of memory and network bandwidth in simulations with large number of atoms.

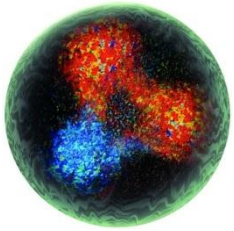


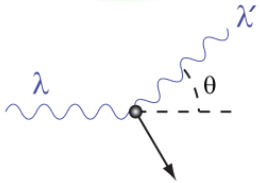


Hadron Polarizability in Lattice QCD

Andrei Alexandru
The George Washington University

Understanding the structure of subnuclear particles represents the main challenge for today's nuclear physics. Photons are used to probe this structure in experiments carried out at laboratories around the world. To interpret the results of these experiments we need to understand how electromagnetic field interacts with subnuclear particles. Theoretically, the structure of subnuclear particles is described by Quantum Chromodynamics (QCD). Lattice QCD is a 4-dimensional discretized version of this theory that can be solved numerically. The focus of our project is to understand how the electric field deforms neutrons and protons by computing the polarizability using lattice QCD techniques.

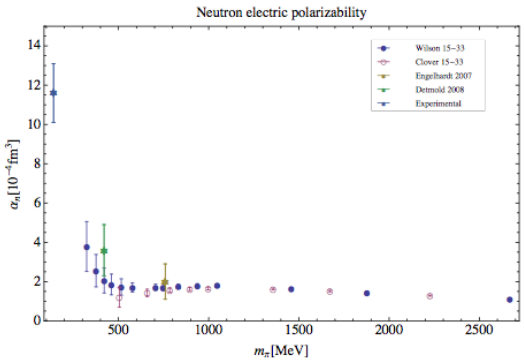


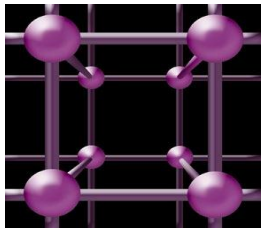


http://samurai.phys.gwu.edu/wiki/index.php/Hadron_polarizability

Why GPUs?

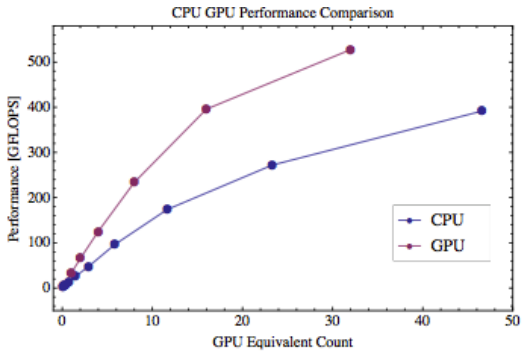
- Lattice QCD simulations require very large bandwidth to run efficiently. GPUs have 10–15 times larger memory bandwidth compared to CPUs.
- Lattice QCD simulations can be efficiently parallelized.
  - Bulk of calculation spent on one kernel.
  - The kernel requires only nearest neighbor information.
  - Cut the lattice into equal sub-lattices. Effectively use single instruction multiple-data (SIMD) paradigm.





Experimental and current values for neutron electric polarizability in lattice QCD.

Alexandru and F. X. Lee, [arXiv:0810.2833]



Performance comparison between Keeneland's GPU cluster and Kraken's Cray XT-5 machine. The CPU core count is translated to GPU equivalent count by dividing the total number of CPUs by 22, which is the number of CPU cores equivalent to a single-GPU performance.

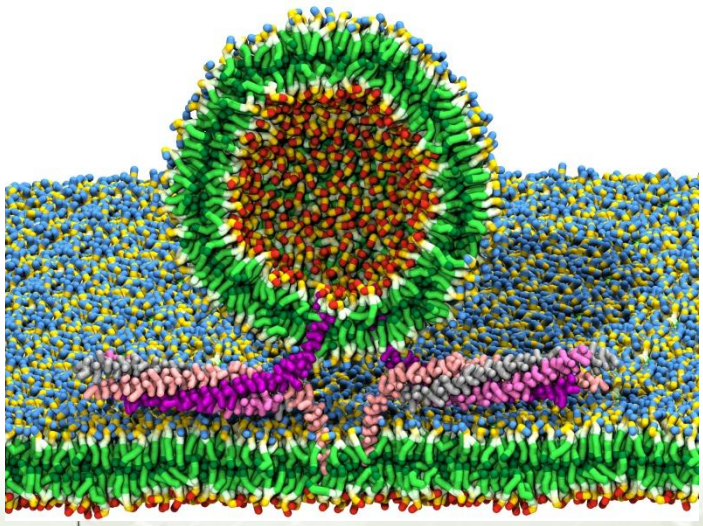
A. Alexandru. *et. al*, [arXiv:1103.5103]

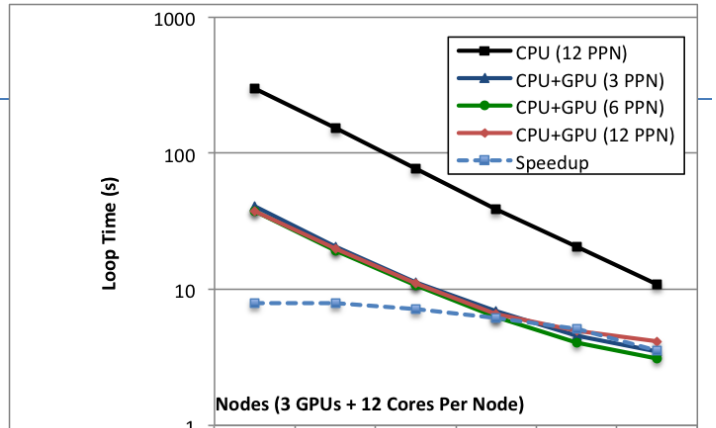
LAMMPS with GPUs

PI: Axel Kohlmeyer,
Temple University

- Parallel Molecular Dynamics
  - http://lammps.sandia.gov
  - Classical Molecular Dynamics
  - Atomic models, Polymers, Metals, Bio-simulations, Coarse-grain (picture), Ellipsoids, etc.
  - Already good strong and weak scaling on CPUs via MPI

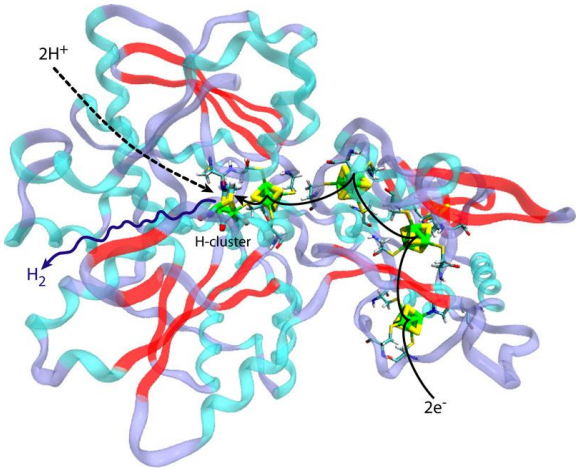
- Better performance on fewer nodes => larger problems faster
- Neighbor, non-bonded force, and long-range GPU acceleration
- Allows for CPU/GPU concurrency
- Implementation and benchmarks by W. Michael Brown, NCCS, ORNL





	1	2	4	8	16	32
CPU (12 PPN)	297.796	151.528	76.7366	38.7134	20.4366	10.83
CPU+GPU (3 PPN)	40.68	20.56	11.29	6.86	4.57	3.49
CPU+GPU (6 PPN)	37.44	19.16	10.66	6.31	4.02	3.08
CPU+GPU (12 PPN)	37.67	19.88	10.96	6.62	4.93	4.15
Speedup	7.95	7.91	7.20	6.14	5.09	3.52

NAMD



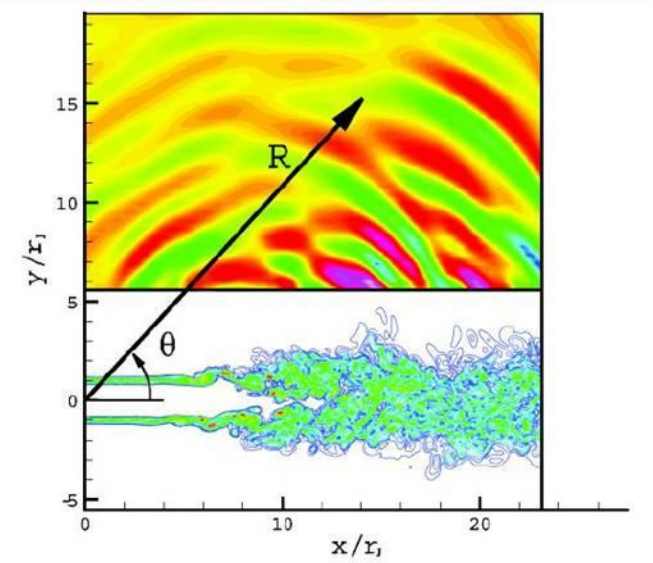
- Biomolecular dynamics
- Public 2.7 Beta 4 pre-release
  - configuration flags to enable GPU support
  - minor benchmark input file modifications
- Single-node performance:
  - 4x faster than CPU on small benchmark, 9x faster on large

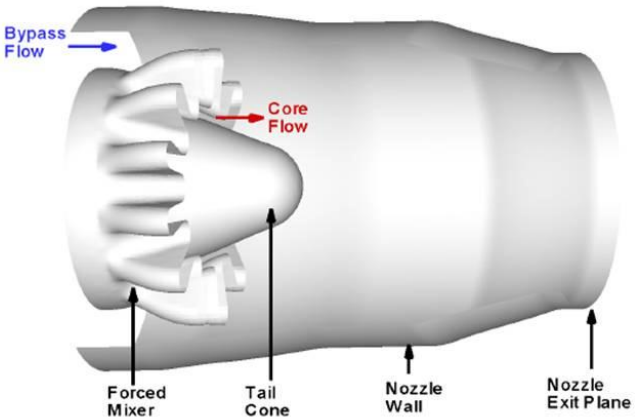
Benchmark	CPU Runtime	GPU Runtime
dhfr	41.1 sec	11.5 sec
apoa1	580.6 sec	67.56 sec

Jet Engine Noise

User: Gregory Blaisdell, Purdue

- Large eddy simulation, computational aeroacoustics
- NSF PetaApps project
- Applications team:
  - function offload model results in communication bottlenecks
  - interested in pursuing other approaches to minimize communications



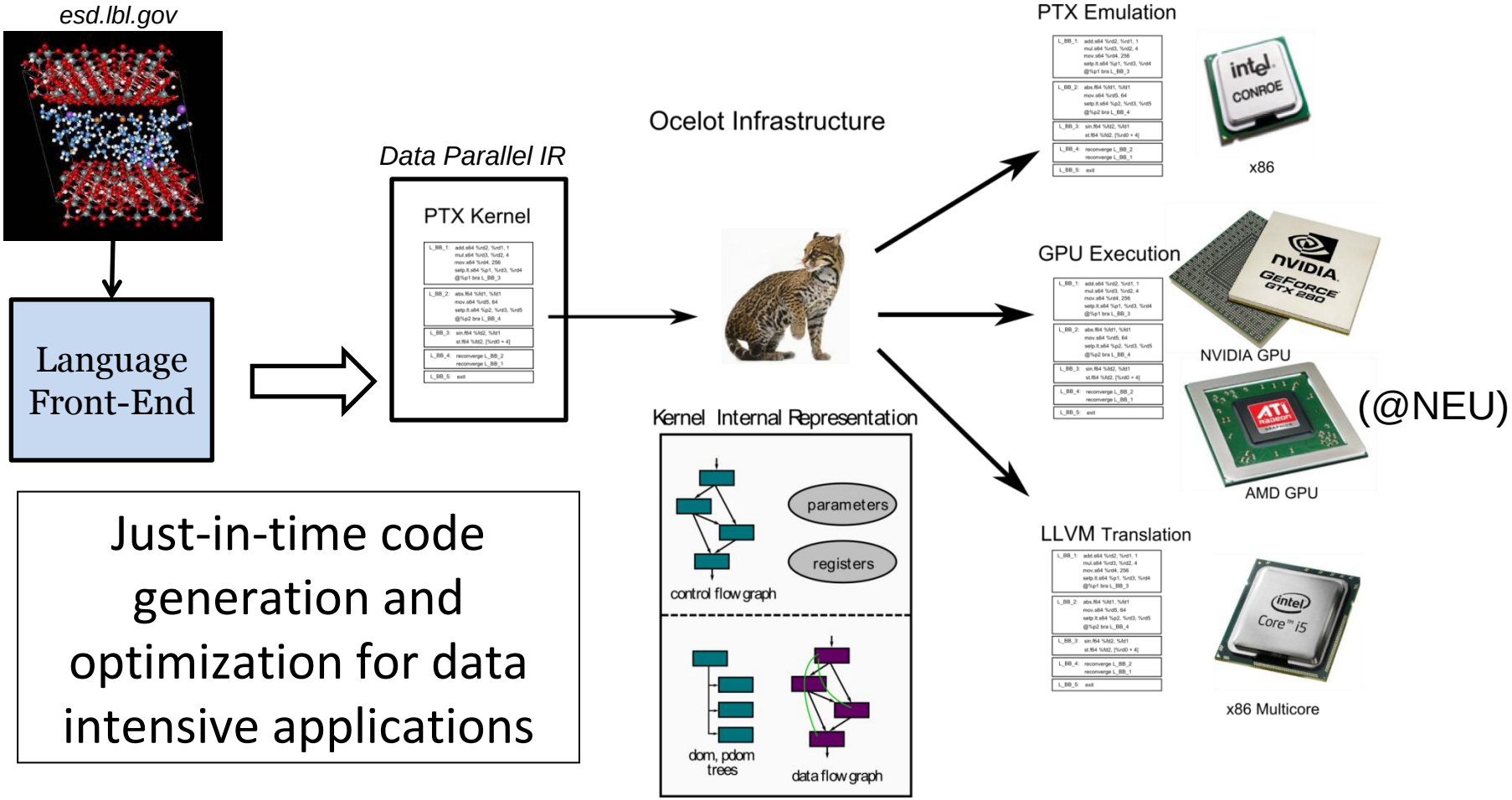


Software Development

Keeneland Software Environment

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

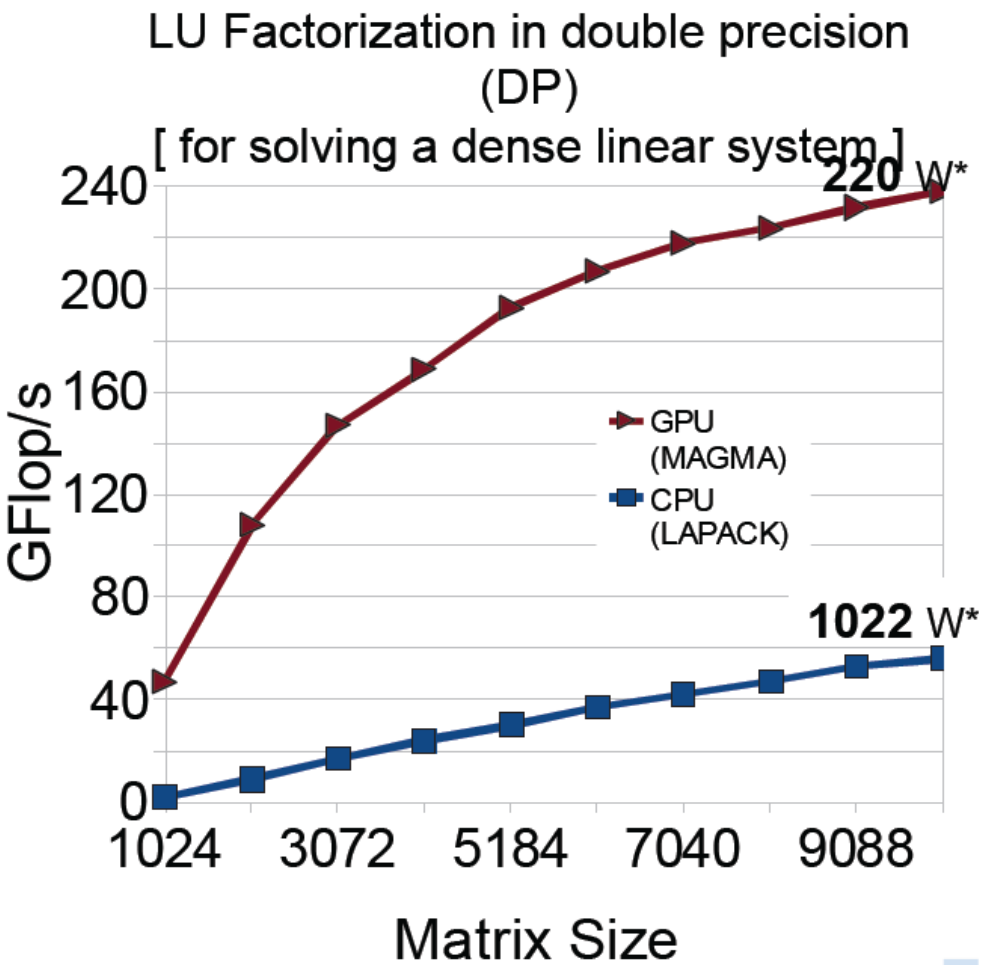
Review: Ocelot Vision

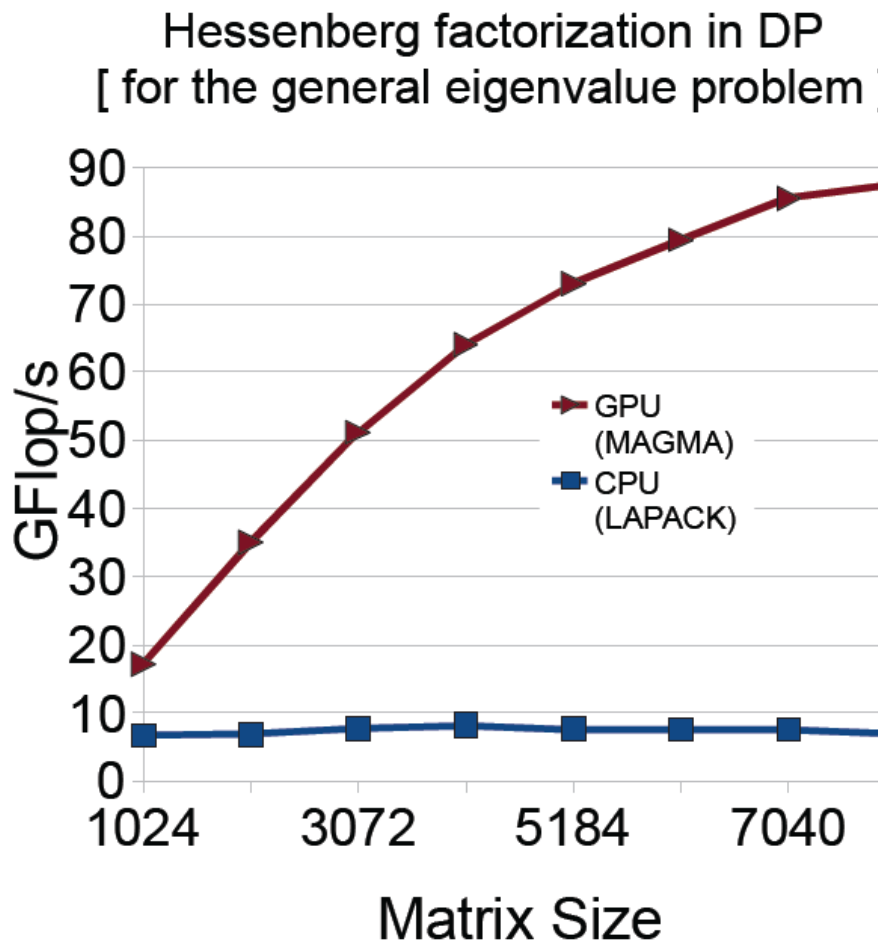


Just-in-time code generation and optimization for data intensive applications

- Environment for i) compiler research, ii) architecture research, and iii) productivity tools

Accelerating Dense Linear Algebra with GPUs

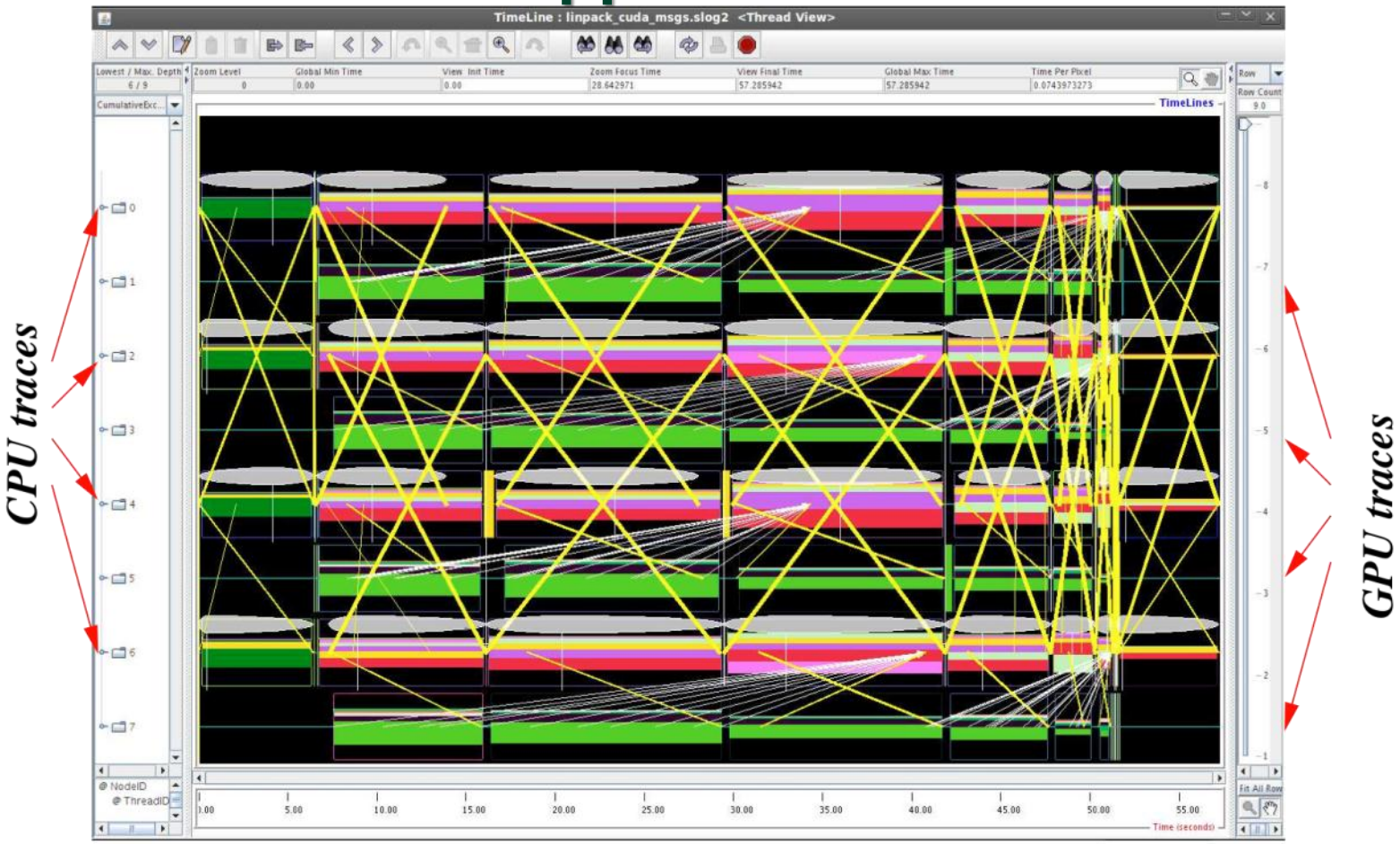




GPU	Fermi C2050 [448 CUDA Cores @ 1.15 GHz] + Intel Q9300 [4 cores @ 2.50 GHz]	**CPU**	AMD ISTANBUL [8 sockets x 6 cores (48 cores) @2.8GHz]
DP peak	**515 + 40** GFlop/s	DP peak	**538** GFlop/s
System cost	~ $**3,000**	System cost	~ $**30,000**
Power *	~ **220** W	Power *	~ **1,022** W

* Computation consumed power rate (total system rate minus idle rate), measured with *KILL A WATT PS, Model P430*

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.*

Technical Assessment

The SHOC Benchmark Suite

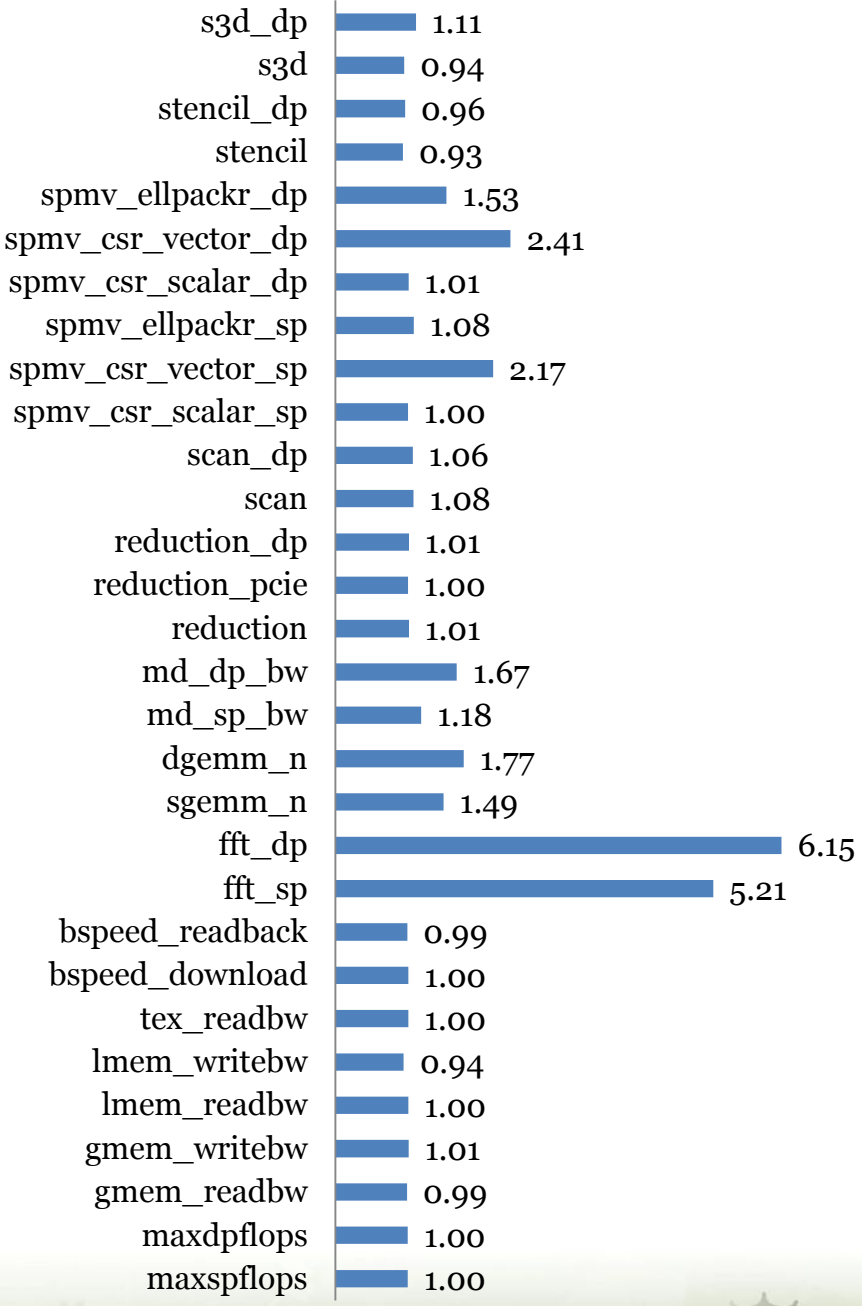
- Benchmark suite with a focus on scientific computing workloads, including common kernels like SGEMM, FFT, Stencils
- Parallelized with MPI, with support for multi-GPU and cluster scale comparisons
- Implemented in CUDA and OpenCL for a 1:1 performance comparison
- Includes stability tests

- 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

CUDA and OpenCL

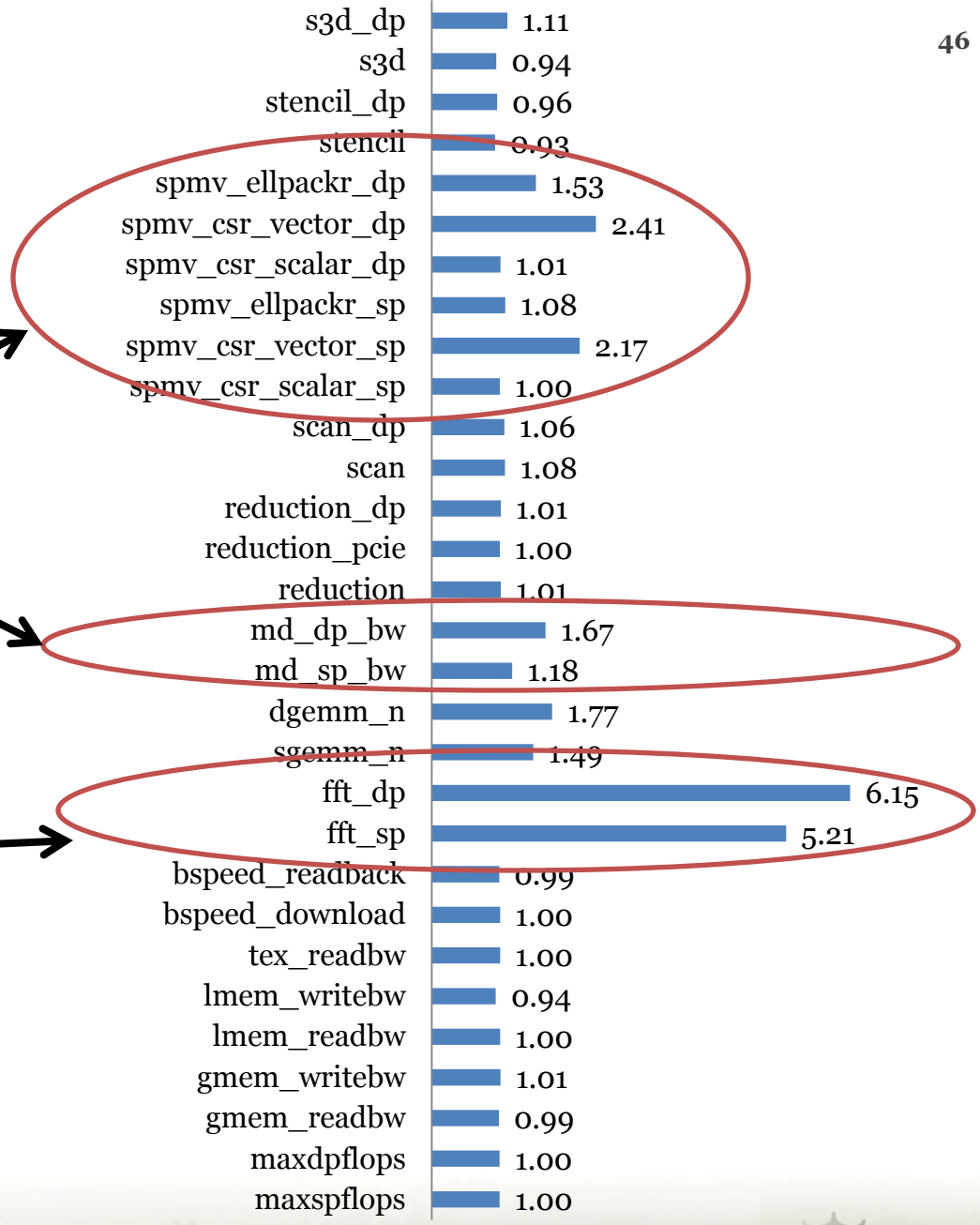
- What does performance look like today?
- This chart shows the speedup of CUDA over OpenCL on a single Tesla M2070 on KIDS (CUDA 4.0, May 2011)
- Note that performance is (in most cases, close to equivalent)
- Cases where it's not tend to be related to texture memory or transcendental intrinsics



A second look

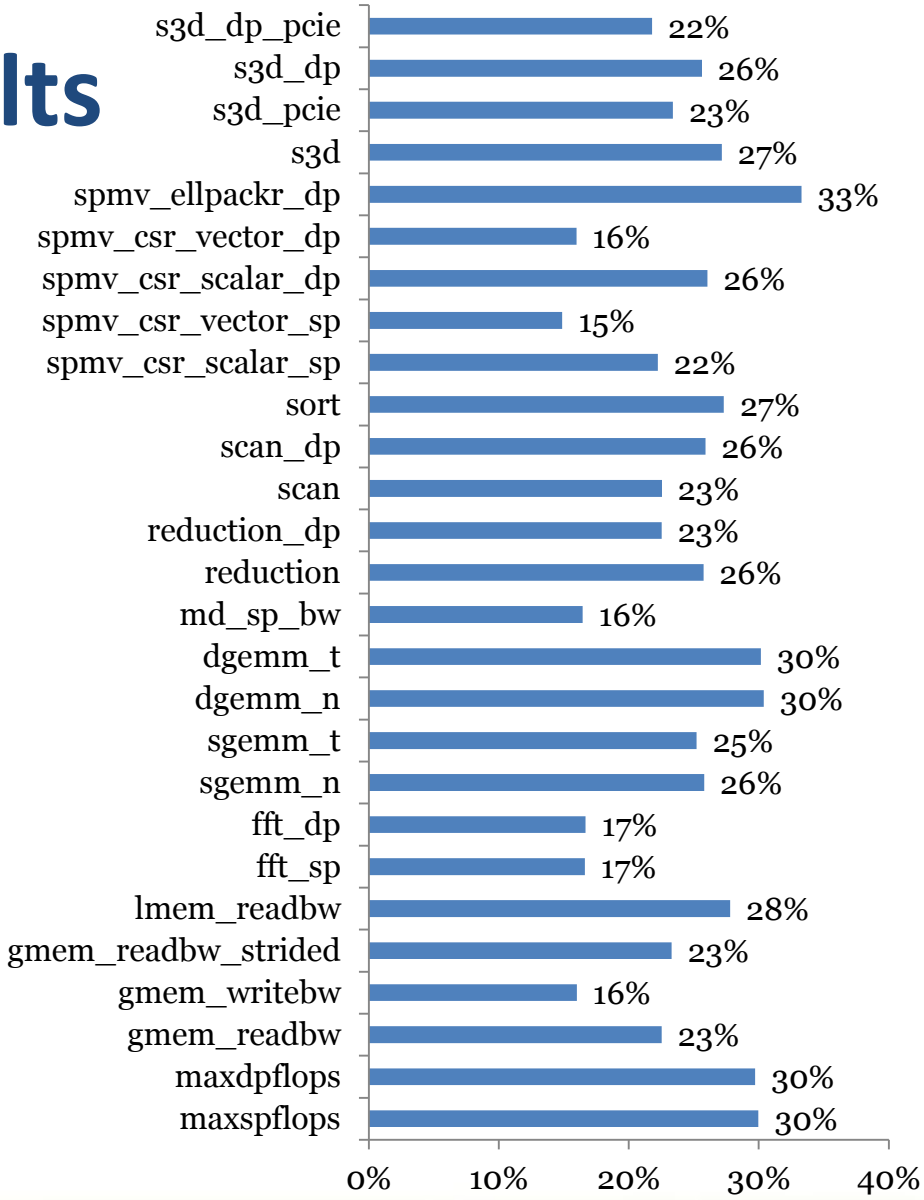
Reliant on texture memory

Dependent on transcendentals



M2090 SHOC Results

- This chart shows improvement compared to the M2070 in CUDA 4.0
- Performance improvements commensurate with expectation.



Assessing other technologies as well

- Compilers
- Debuggers
- New libraries and runtimes
  - E.g., GPU-Direct

- Alternative hardware
  - MIC
  - AMD

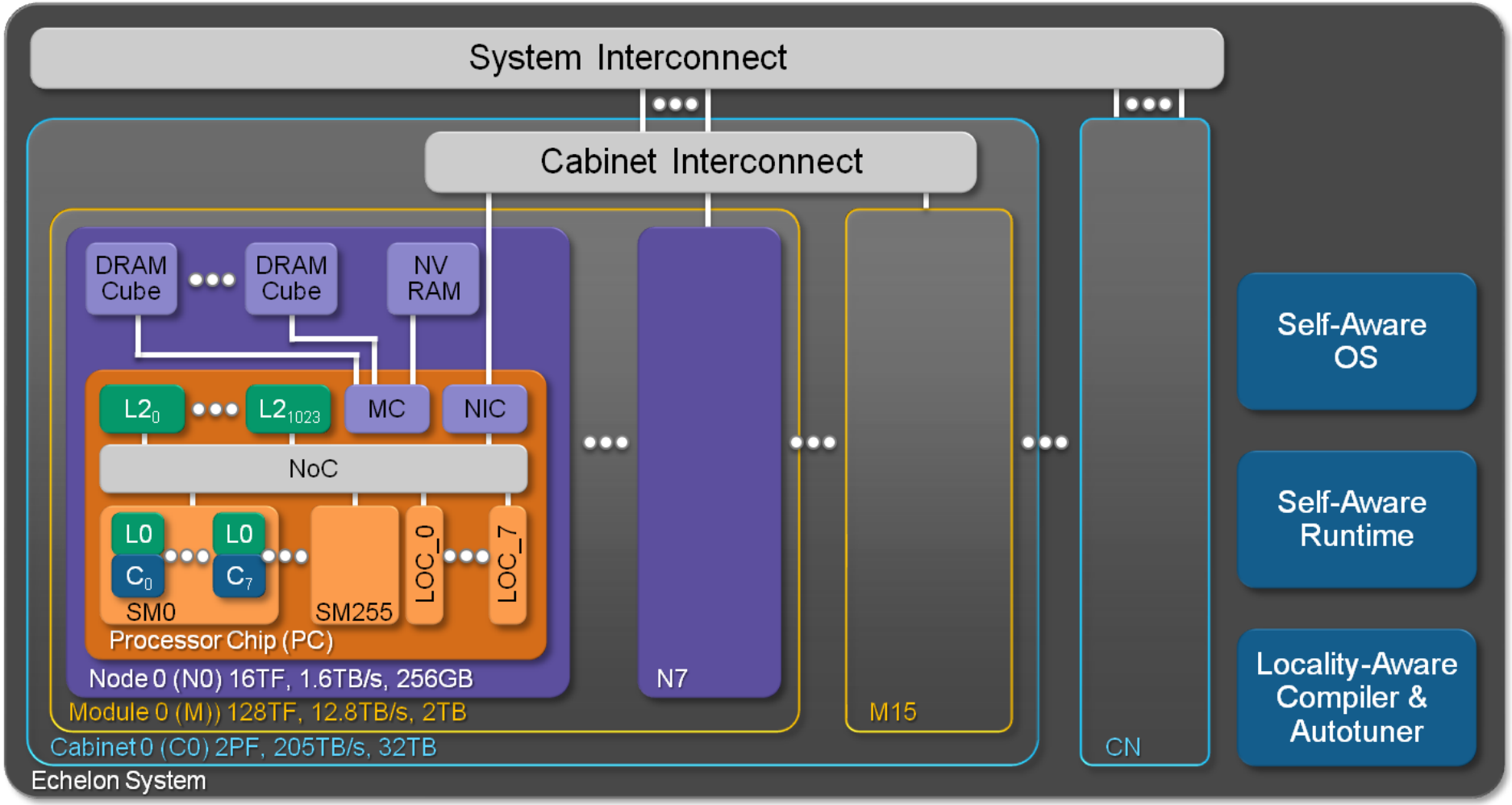
Exascale Trends and Challenges

Notional Exascale Architecture Targets

Exascale Arch Report 2009

System attributes	2001	2010	"2015"		"2018"	
System peak	10 Tera	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	~0.8 MW	6 MW	15 MW		20 MW	
System memory	0.006 PB	0.3 PB	5 PB		32-64 PB	
Node performance	0.024 TF	0.125 TF	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	16	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	416	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)	

NVIDIA Echelon System Sketch



DARPA Echelon team: NVIDIA, ORNL, Micron, Cray, Georgia Tech, Stanford, UC-Berkeley, U Penn, Utah, Tennessee, Lockheed Martin

TG + SC11 tutorials

- Co-organized Track 2D introductions at OGF and TG11 in July in SLC
  - Overview
  - How to get access

- SC11 Tutorial
- Hands on with KID
  - Overview of scalable heterogeneous systems, and refreshers in MPI and CUDA/OpenCL;
  - Successful, demonstrated methodologies for migrating existing applications to scalable heterogeneous systems;
  - Performance and Correctness tools for integrated MPI and CUDA/OpenCL analysis
  - Integrated MPI and CUDA/OpenCL optimization techniques

Summary

- Keeneland is bringing heterogeneous computing to NSF
- Learn more: http://keeneland.gatech.edu
  - Apply for an account
  - Access tutorial materials
  - Download software like MAGMA, SHOC
- SC11 Tutorial – Hands on

- Technical papers
  - http://ft.ornl.gov
  - J.S. Vetter, R. Glassbrook *et al.*, *"Keeneland: Bringing heterogeneous GPU computing to the computational science community," IEEE Computing in Science and Engineering, 13(5):90-5, 2011, http://dx.doi.org/10.1109/MCSE.2011.83*.