Titan: A New Leadership Computer for Science

Presented to:

DOE Advanced Scientific Computing Advisory Committee

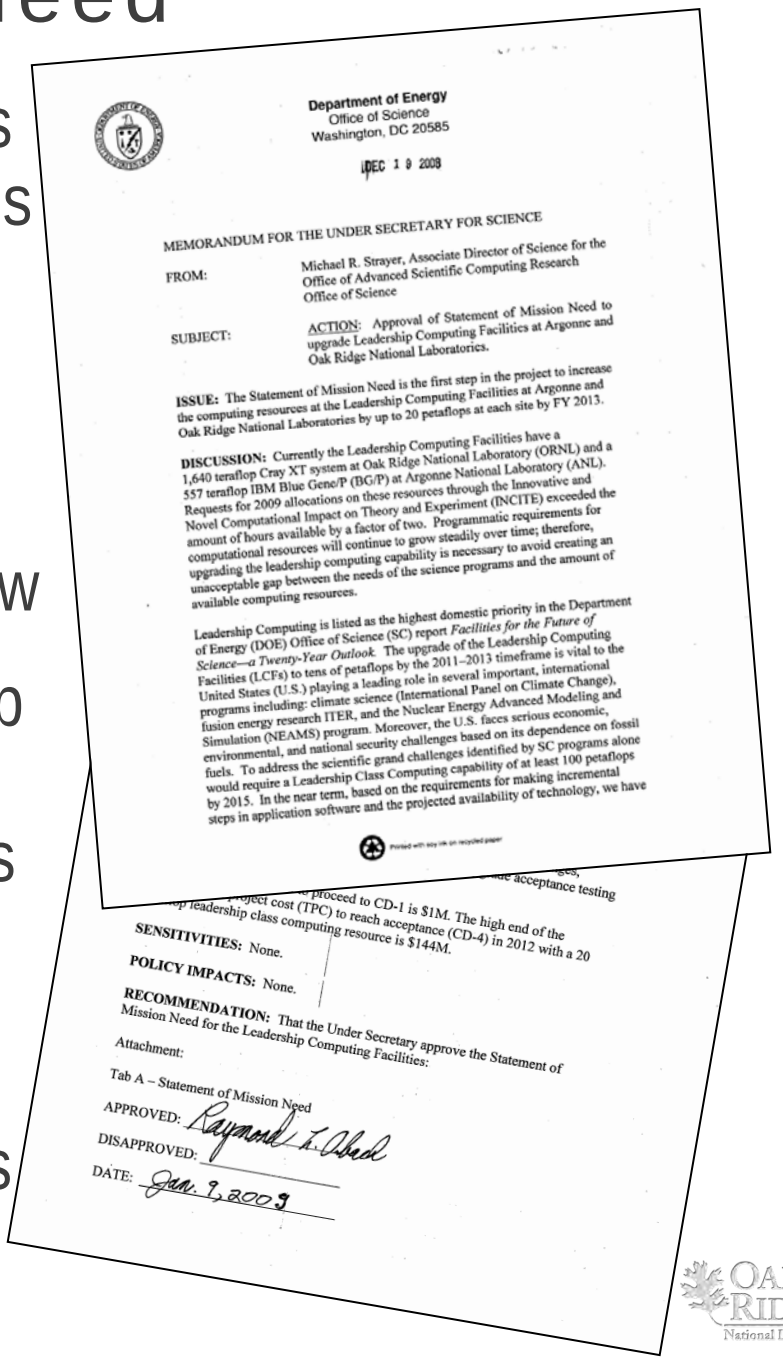
November 1, 2011

Arthur S. Bland

OLCF Project Director

Statement of Mission Need

- Increase the computational resources of the Leadership Computing Facilities by 20-40 petaflops
- INCITE program is oversubscribed
- Programmatic requirements for leadership computing continue to grow
- Needed to avoid an unacceptable gap between the needs of the science programs and the available resources
- Approved: Raymond Orbach
  January 9, 2009
- The OLCF-3 project comes out of this requirement



Department of Energy
Office of Science
Washington, DC 20585

DEC 19 2008

MEMORANDUM FOR THE UNDER SECRETARY FOR SCIENCE

FROM: Michael R. Strayer, Associate Director of Science for the Office of Advanced Scientific Computing Research
Office of Science

SUBJECT: ACTION: Approval of Statement of Mission Need to upgrade Leadership Computing Facilities at Argonne and Oak Ridge National Laboratories.

ISSUE: The Statement of Mission Need is the first step in the project to increase the computing resources at the Leadership Computing Facilities at Argonne and Oak Ridge National Laboratories by up to 20 petaflops at each site by FY 2013.

DISCUSSION: Currently the Leadership Computing Facilities have a 1,640 teraflop Cray XT system at Oak Ridge National Laboratory (ORNL) and a 557 teraflop IBM Blue Gene/P (BG/P) at Argonne National Laboratory (ANL). Requests for 2009 allocations on these resources through the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) exceeded the amount of hours available by a factor of two. Programmatic requirements for computational resources will continue to grow steadily over time; therefore, upgrading the leadership computing capability is necessary to avoid creating an unacceptable gap between the needs of the science programs and the amount of available computing resources.

Leadership Computing is listed as the highest domestic priority in the Department of Energy (DOE) Office of Science (SC) report *Facilities for the Future of Science—a Twenty-Year Outlook*. The upgrade of the Leadership Computing Facilities (LCFs) to tens of petaflops by the 2011–2013 timeframe is vital to the United States (U.S.) playing a leading role in several important, international programs including: climate science (International Panel on Climate Change), fusion energy research ITER, and the Nuclear Energy Advanced Modeling and Simulation (NEAMS) program. Moreover, the U.S. faces serious economic, environmental, and national security challenges based on its dependence on fossil fuels. To address the scientific grand challenges identified by SC programs alone would require a Leadership Class Computing capability of at least 100 petaflops by 2015. In the near term, based on the requirements for making incremental steps in application software and the projected availability of technology, we have

Printed with soy ink on recycled paper

...proceed to CD-1 is $1M. The high end of the ...ject cost (TPC) to reach acceptance (CD-4) in 2012 with a 20 ...leadership class computing resource is $144M.

SENSITIVITIES: None.

POLICY IMPACTS: None.

RECOMMENDATION: That the Under Secretary approve the Statement of Mission Need for the Leadership Computing Facilities:

Attachment:

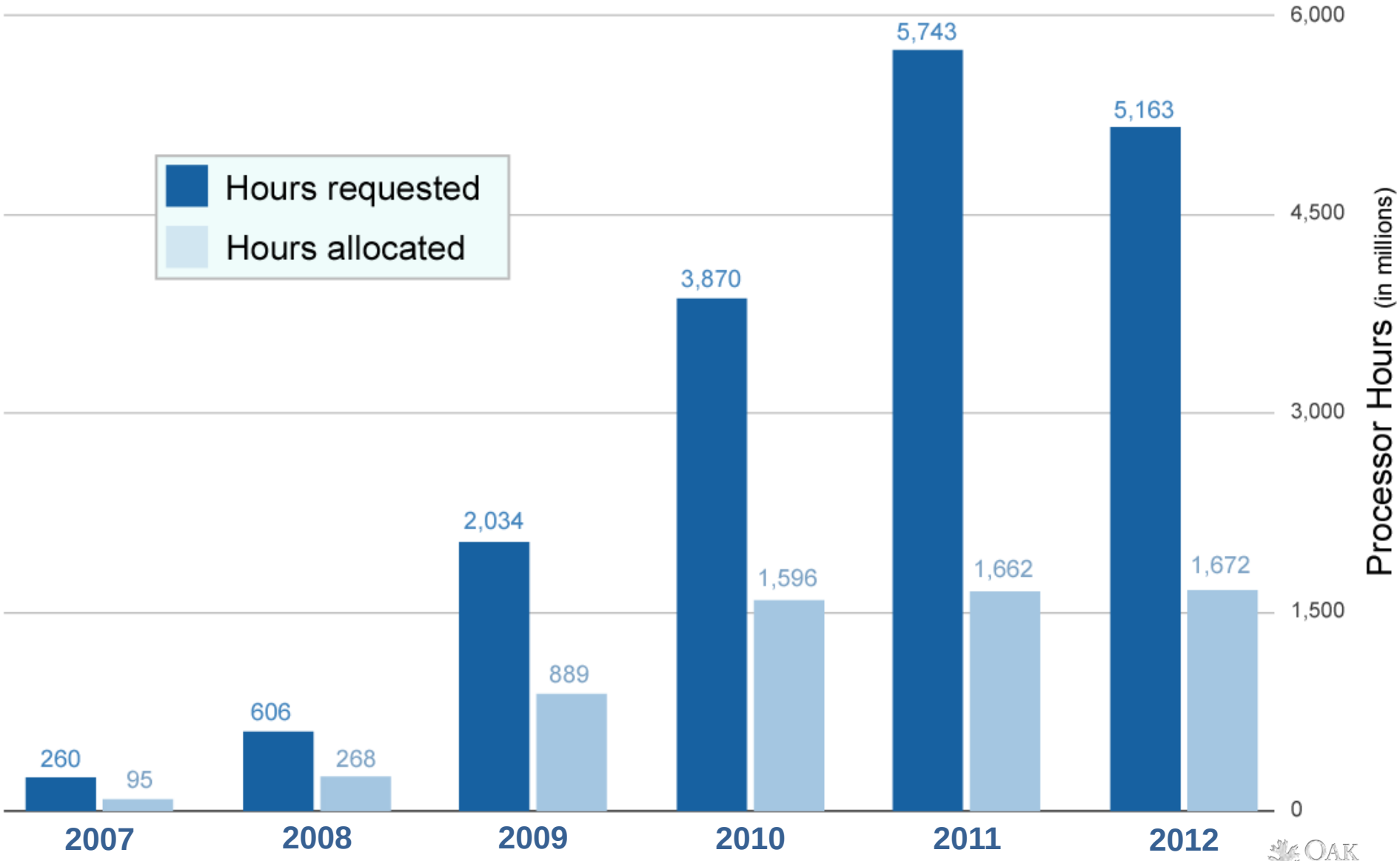
Tab A – Statement of Mission Need

APPROVED: Raymond L. Orbach

DISAPPROVED:

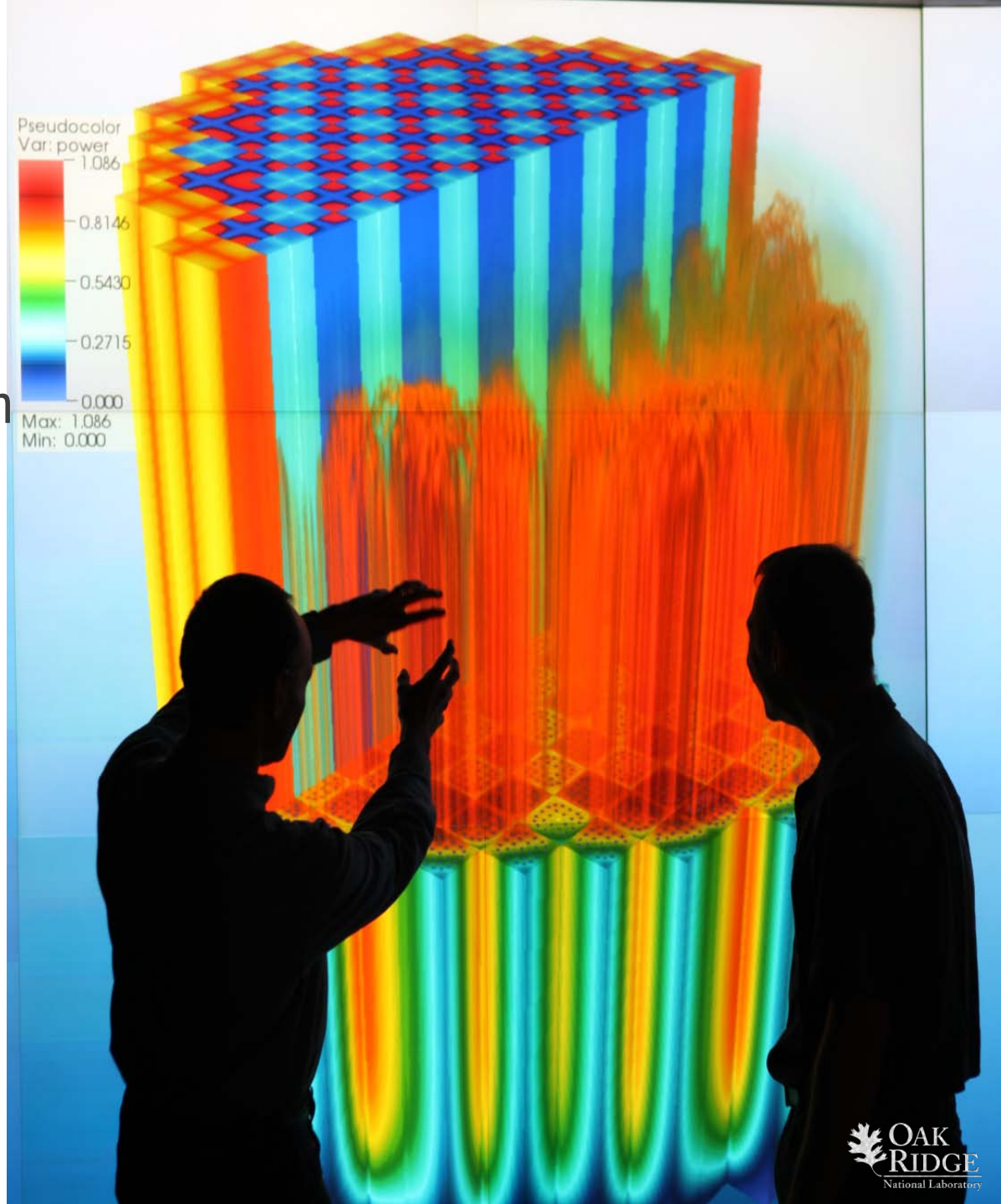
DATE: Jan. 9, 2009

INCITE is 2.5 to 3.5 times oversubscribed



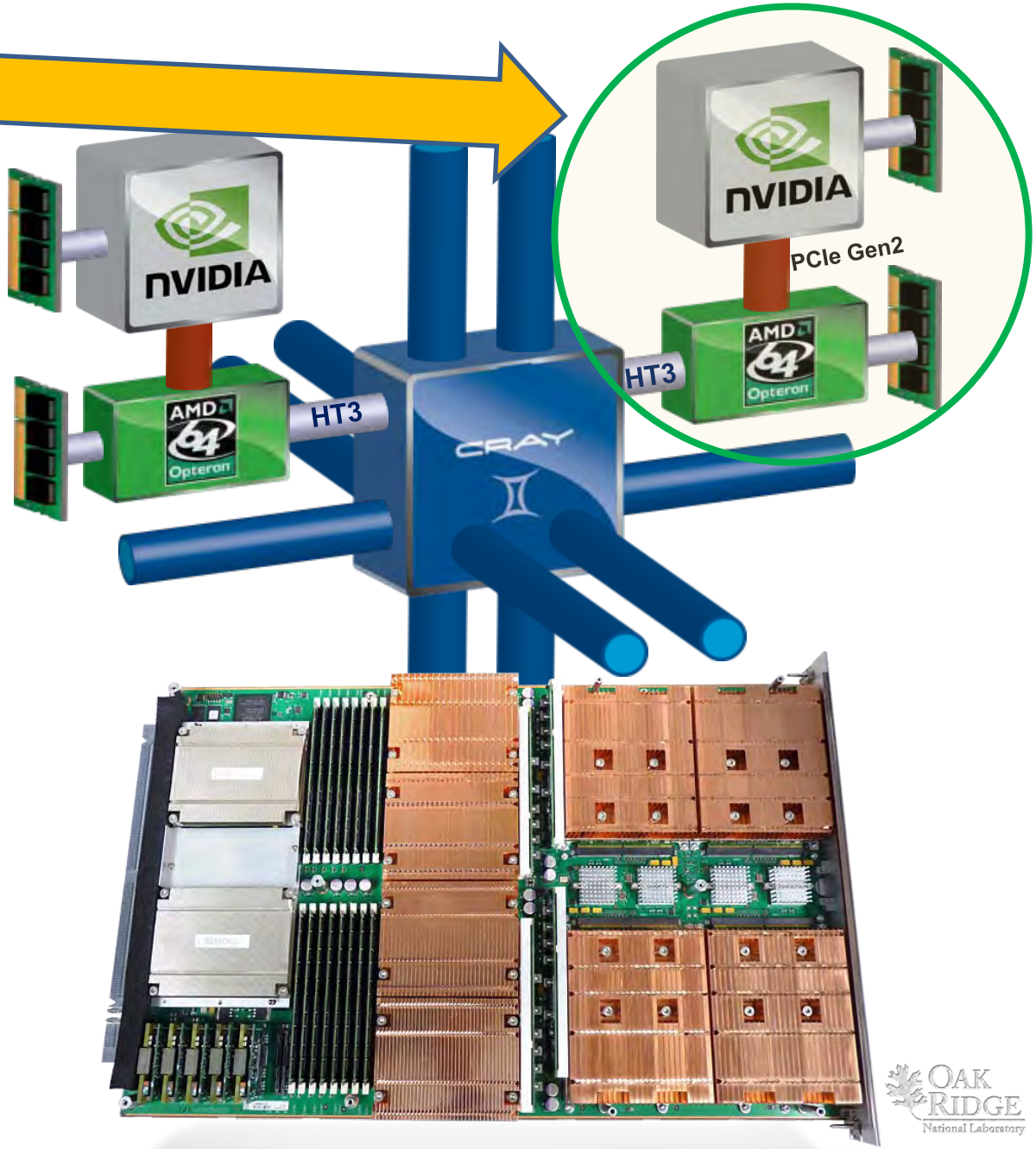
What is OLCF-3

- The next phase of the Leadership Computing Facility program at ORNL
- An upgrade of Jaguar from 2.3 Petaflops (peak) today to between 10 and 20 PF by the end of 2012 with operations in 2013
- Built with Cray's newest XK6 compute blades
- When completed, the new system will be called Titan



Cray XK6 Compute Node

XK6 Compute Node Characteristics
AMD Opteron 6200 “Interlagos” 16 core processor @ 2.2GHz
Tesla M2090 “Fermi” @ 665 GF with 6GB GDDR5 memory
Host Memory 32GB 1600 MHz DDR3
Gemini High Speed Interconnect
Upgradeable to NVIDIA’s next generation “Kepler” processor in 2012
Four compute nodes per XK6 blade. 24 blades per rack



ORNL's "Titan" System

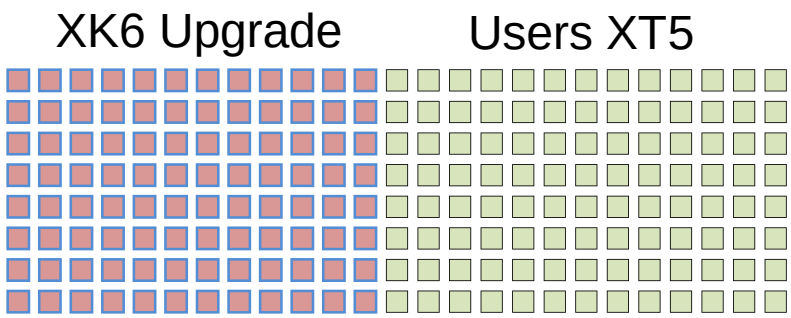
- Upgrade of existing Jaguar Cray XT5
- Cray Linux Environment operating system
- Gemini interconnect
  - 3-D Torus
  - Globally addressable memory
  - Advanced synchronization features
- AMD Opteron 6200 processor (Interlagos)
- New accelerated node design using NVIDIA multi-core accelerators
  - 2011: 960 NVIDIA M2090 "Fermi" GPUs
  - 2012: 10-20 PF NVIDIA "Kepler" GPUs
- 10-20 PFlops peak performance
  - Performance based on available funds
- 600 TB DDR3 memory (2x that of Jaguar)

Titan Specs	
Compute Nodes	18,688
Login & I/O Nodes	512
Memory per node	32 GB + 6 GB
NVIDIA "Fermi" (2011)	665 GFlops
# of Fermi chips	960
NVIDIA "Kepler" (2012)	>1 TFlops
Opteron	2.2 GHz
Opteron performance	141 GFlops
Total Opteron Flops	2.6 PFlops
Disk Bandwidth	~ 1 TB/s

2011 Upgrade from XT5 to XK6

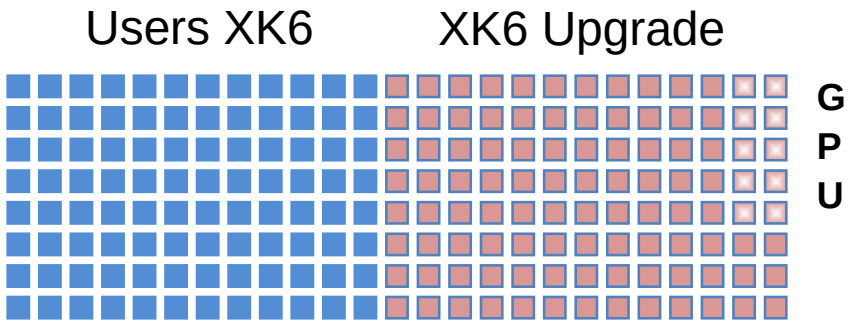
Oct: Segment system into 104 cabinets of existing Jaguar for users (1.2 PF)

- Upgrade 96 cabinets to XK6 (16-core) nodes and test



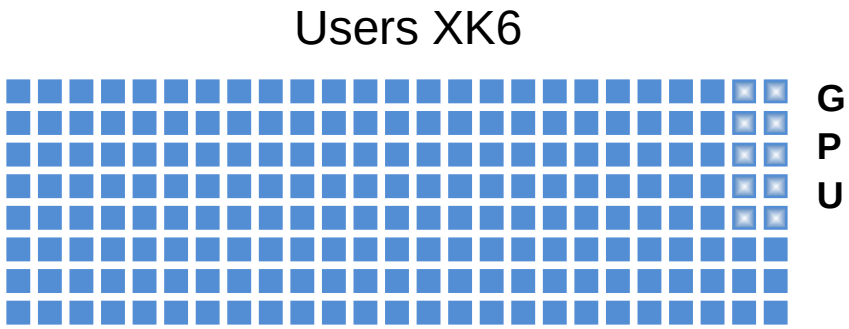
Nov: Move users to upgraded 96 cabinets (1.3 PF)

- Upgrade and test 104 cabinets to XK6 nodes including 10 cabinets with GPUs



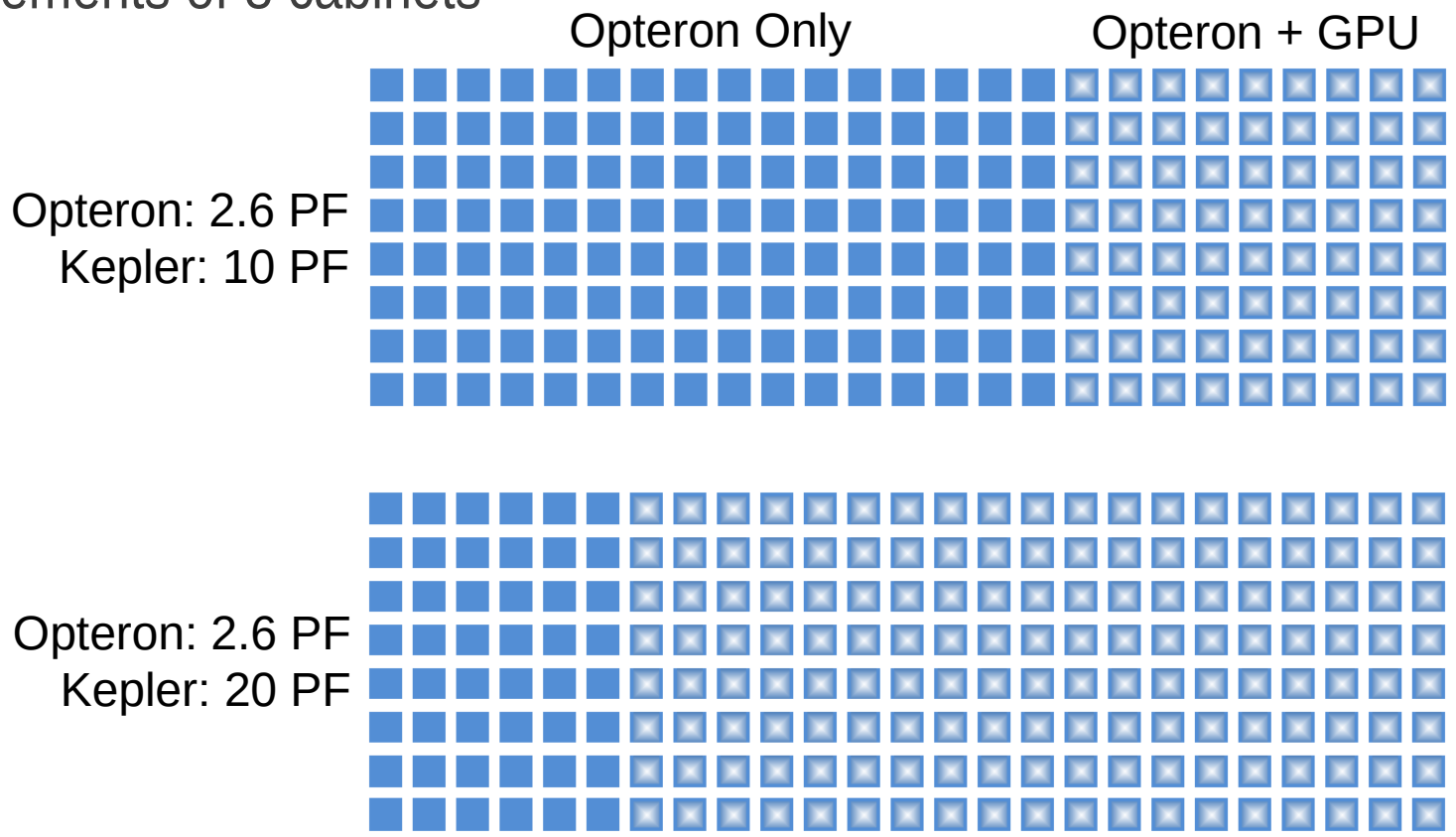
Dec: Combine two halves and run acceptance test (3.3 PF)

- Users will be running during acceptance test period



2012 Upgrade: Add Kepler GPUs

- When NVIDIA Kepler GPUs are available we will replace Fermi GPUs and increase system to 10 to 20 petaflops based on available funding
- The contract with Cray has options to add GPUs in increments of 8 cabinets



Processor Architecture: Power vs. Single Thread Performance

- Multi-core architectures are a good first response to power issues
  - Performance through parallelism, not frequency
  - Exploit on-chip locality
- However, conventional processor architectures are optimized for single thread performance rather than energy efficiency
  - Fast clock rate with latency(performance)-optimized memory structures
  - Wide superscalar instruction issue with dynamic conflict detection
  - Heavy use of speculative execution and replay traps
  - Large structures supporting various types of predictions
  - Relatively little energy spent on actual ALU operations
- Could be much more energy efficient with multiple simple processors, exploiting vector/SIMD parallelism and a slower clock rate
- But serial thread performance is really important (Amdahl's Law):
  - If you get great parallel speedup, but hurt serial performance, then you end up with a niche processor (less generally applicable, harder to program)

Courtesy of Steve Scott
presented at SciDAC 2010

Exascale Conclusion: Heterogeneous Computing

- To achieve scale and sustained performance per {$,watt}, must adopt:
  - …a *heterogeneous* node architecture
    - fast serial threads coupled to many efficient parallel threads
  - …a deep, explicitly managed memory hierarchy
    - to better exploit locality, improve predictability, and reduce overhead
  - …a microarchitecture to exploit parallelism at all levels of a code
    - distributed memory, shared memory, vector/SIMD, multithreaded
    - (related to the "concurrency" challenge—leave no parallelism untapped)

- **This sounds a lot like GPU accelerators…**
- NVIDIA Fermi™ has made GPUs feasible for HPC
  - Robust error protection and strong DP FP, plus programming enhancements
- Expect GPUs to make continued and significant inroads into HPC
  - Compelling technical reasons + high volume market
- **Programmability remains primary barrier to adoption**
  - Cray is focusing on compilers, tools and libraries to make GPUs easier to use
  - There are also some structural issues that limit applicability of current designs…
- Technical direction for Exascale:
  - Unified node with "CPU" and "accelerator" on chip sharing common memory
  - Very interesting processor roadmaps coming from Intel, AMD and NVIDIA….

Courtesy of Steve Scott
presented at SciDAC 2010

Application Programmability Challenges for Titan

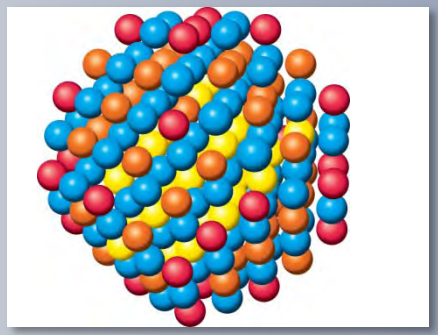
- Node level concurrency
  - Finding sufficient concurrency through vectorization and threads to amortize or hide data movement costs
    - Jaguar today has 224,256 Opteron cores
    - If we fill all the Titan cabinets with Kepler GPUs, it will have almost 10 *million* threads of execution!
  - Thread creation and management
- Data locality
  - Managing data locality
  - Minimize data movement between CPU and GPU
  - PCIe connection between CPU and GPU limits bandwidth to 8 GB/s
- Programming models
  - Must support an evolving code base
  - Maintaining code portability and performance
- Availability of compilers and tools for hybrid systems

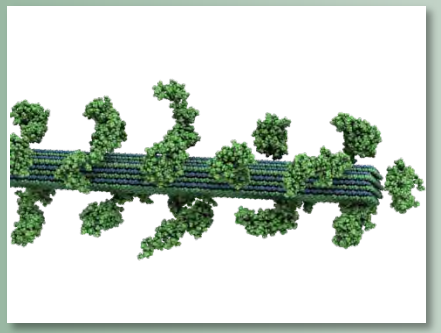
OLCF-3 Application Selection Considerations

- Science (highest priority is mission critical science)
  - Science results, impact, timeliness
  - Alignment with DOE and US science mission (CD-0)
  - Broad coverage of science domains
- Implementation (mix of models, algorithms, software)
  - Broad coverage of relevant programming models, environment, languages, implementations
  - Broad coverage of relevant algorithms and data structures (motifs)
  - Broad coverage of scientific library requirements
- Good representation of current and anticipated INCITE workloads
  - Goal is to identify best practices and templates for how to expose parallelism
- Practical Considerations
  - Mix of straightforward and difficult to port applications (as evaluated by the actual app developers)
  - Availability of liaisons and key code development personnel to engage in and guide readiness activities
  - Limited resources can only support a few of these efforts

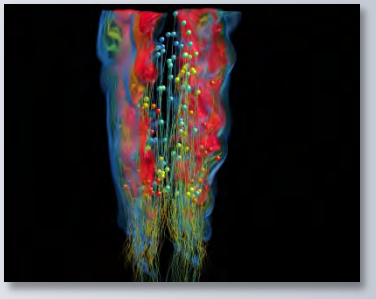
Titan: Early Science Applications

WL-LSMS
Role of material disorder, statistics, and fluctuations in nanoscale materials and systems.





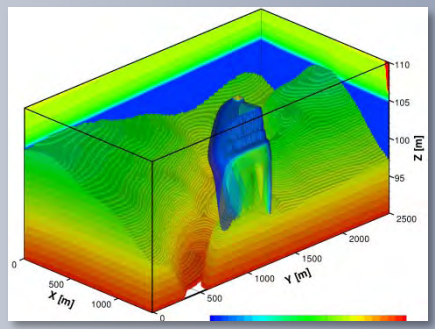
LAMMPS
A parallel particle simulator that can simulate soft materials (biomolecules, polymers), solid-state materials (metals, semiconductors) and coarse-grained or mesoscopic systems

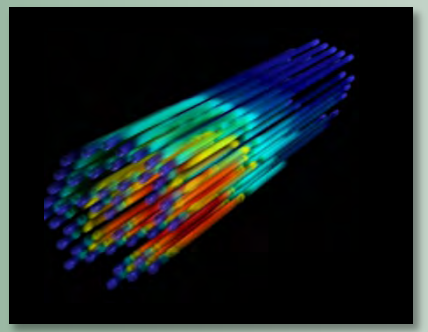


S3D
How are going to efficiently burn next generation diesel/bio fuels?
.

CAM / HOMME
Answer questions about specific climate change adaptation and mitigation scenarios; realistically represent features like precipitation patterns/statistics and tropical storms

PFLOTRAN
Stability and viability of large scale CO_2 sequestration; predictive containment groundwater transport





Denovo
Unprecedented high-fidelity radiation transport calculations that can be used in a variety of nuclear energy and technology applications.

S3D

- Description
  - Scientific software for direct numerical simulation (DNS) to study fundamental turbulence-chemistry interaction
  - Compressible flow solver for structured mesh with detailed models for chemical kinetics and transport
- Science problem
  - 3D DNS of auto-ignition in a high-pressure dimethyl-ether/air mixture
  - 2 Billion cells and 60 chemical species
- Progress
  - Key kernels of S3D have been ported to GPU using CUDA and compiler directives
  - The kernel performance numbers indicate that the S3D application will perform 4X faster on the hybrid architecture compared to Jaguar

WL-LSMS

- Description
  - Classical statistical physics using Wang-Landau Monte-Carlo & First principles Density Functional Theory solved in real space using multiple scattering theory
- Science problem
  - Calculate thermodynamic properties of magnetic materials:
    - Curie Temperature, Magnetization, Susceptibility, Specific Heat
    - Iron, Steel, Magneto-Caloric Materials, Permanent Magnets, …
  - Acceptance Test: 1024 Fe atoms & fixed number of iterations
  - Early Science: Acceptance test problem iterated to convergence
- Progress
  - Ported main compute kernel [zblock_lu] (95% time on CPU) to GPU
    - 25x speedup of kernel execution on Fermi vs. one CPU core (JaguarPF)
  - Ongoing: porting full WL-LSMS to hybrid MPI/OpenMP/GPU model

CAM-SE

- Description
  - Community Atmosphere Model – Spectral Element
  - Climate-scale atmospheric simulation using spectral element method
- Science problem
  - New dynamical core and new atmospheric chemistry required to make reliable, long-term climate predictions.
  - CAM-SE: 240x240 elements / panel, 4x4 moments / elements, 26 vertical levels
    - Using MOZART chemistry package introducing ~100 advected constituents
- Progress
  - All kernels completed. Now merging into newer code base
  - Vertical remap improved by over 2x even on CPU
  - Isolated data that remains on-node. Now optimizing to keep on GPU
    - Significant reduction in PCI-e transfers and packs / unpacks
    - Will provide a boost CPU performance as well
  - Improved kernel-level concurrency by pulling another loop inside

PFLOTRAN

- Description
  - Massively parallel reactive groundwater code written in F90 interfaced with C++ adaptive mesh refinement (AMR) framework (SAMRAI)
  - Different modes: single phase variably saturated flow, reactive transport and CO_2 sequestration; AMR enabled for first two modes, in progress for third
- Science problem
  - Simulating CO_2 sequestration in subsurface geologic formations containing saline aquifers
- Progress
  - Initial profiling on Jaguar done, AMR for coupled single phase variably saturated flow and reactive transport modes demonstrated, conversion to SAMRAI v3 in progress
  - > 4x speedup of host+Fermi C2090 versus a 4-core i7 920 (2.67GHz) processor socket for stand alone kernel experiments

LAMMPS

- Description
  - LAMMPS (http://lammps.sandia.gov) is a parallel particle simulator that can simulate soft materials (biomolecules, polymers), solid-state materials (metals, semiconductors) and coarse-grained or mesoscopic systems.
- Science problem
  - Biomass recalcitrance is a major impediment for ethanol production
  - Molecular dynamics simulations of lipid bilayer fusion
    - A charge system of ~850,000,000 coarse grained beads of approximate system dimension 32,000Å x 32,000Å x 33 Å
- Progress
  - LAMMPS is currently 2x to 5x faster with GPU acceleration using all 16 cores on the XK6 using existing LAMMPS algorithms
  - 95% complete with implementing a new parallel linear scaling electrostatic solver (MSM) algorithm

Denovo

- Description
  - Solves the radiation transport problem for nuclear reactor design
  - Computes solutions to the six-dimensional linear Boltzmann equation
- Science problem
  - Neutron transport is the most compute-intensive component of nuclear reactor simulation.
  - Neutron transport simulation with 128 million spatial zones, 16 energy groups and 256 discrete ordinates to resolve the radiation field
- Progress
  - Code has been ported to NVIDIA Fermi GPUs
  - The 3-D sweep kernel, which consumes 90% of the runtime, runs 40X faster on Fermi compared to an Opteron core
  - The new GPU-aware sweeper also runs 2X faster on CPUs compared to the previous CPU-based sweeper due to performance optimizations

Titan's Programming Environment
(Ease of Use and Application Portability)

Goals:

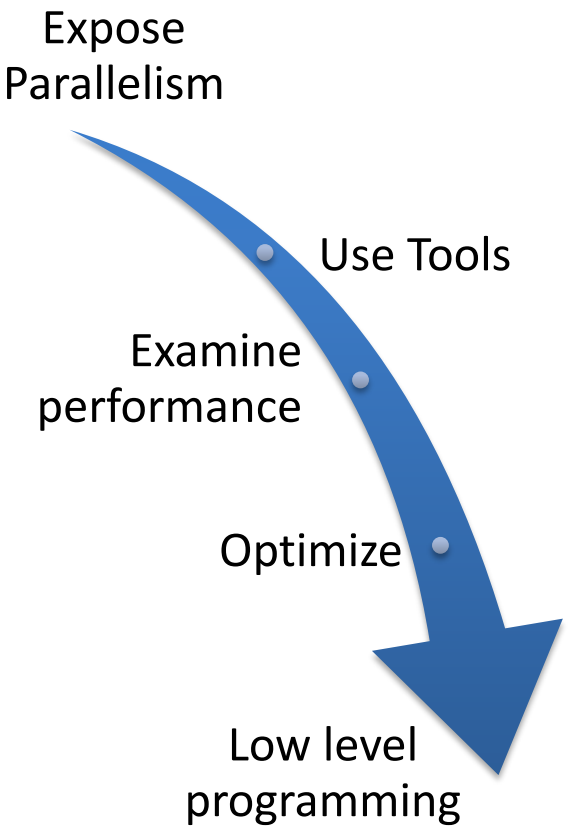
- Full fledged programming environment for hybrid CPU / GPU nodes
  - Compilers
  - Debuggers
  - Performance Analysis tools
  - Mathematical Libraries
- Hardware agnostic programming model - portable
  - Code based directives:
    - Describe execution parallelism - expose (hierarchical) parallelism
    - Describe data layout
    - Being standardized through the OpenMP Architecture Review Board and other industry initiatives (see more at SC'11)
- Leverage existing software capabilities
  - Extend existing tools to add GPU support

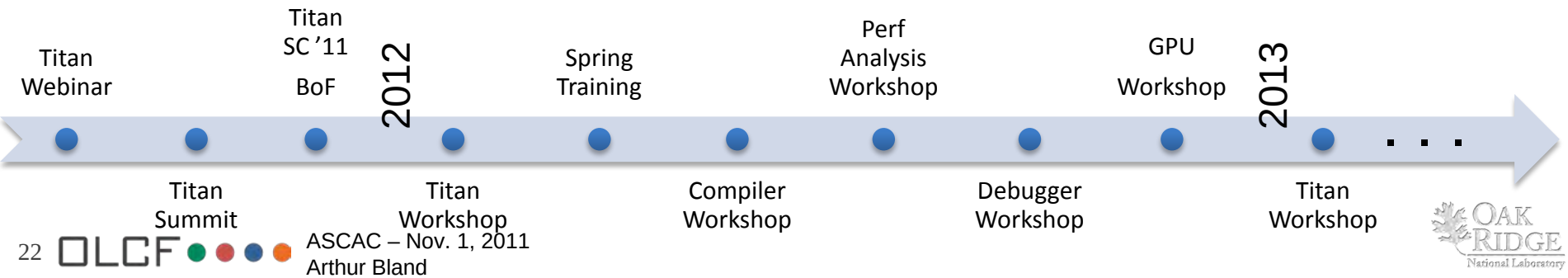
Programming Environment Components
(Multiple Options Improves Performance & Reduces Risk)

- Compilers
  - Cray Compiler Environment – add GPU support
  - HMPP from CAPS – add C++, Fortran, and additional GPU support
  - PGI – new GPU support
- Debuggers
  - DDT – Add GPU support
  - Scalability already being addressed outside of the project
- Performance Analysis tools
  - Cray – add GPU support
  - Vampir suite – add GPU support, and increase scalability
  - Supply other third party tools (HPCToolkit, TAU)
- Math libraries
  - Cray – add GPU support
  - Third party – CULA, MAGMA

Titan Training Program

- **Goal:** Enable break-through science through education
- **Strategy:** Provide conferences, workshops, tutorials, case studies, and lessons learned on tools and techniques for realizing hybrid architecture benefits. Provide content via traditional venues, online, and pre-recorded sessions.
- **Objective:** Users will be able to expose hierarchical parallelism, use compiler directive-based tools, analyze / optimize / debug codes, and use low-level programming techniques if required





OLCF-3 Project Timeline

Date	Event	Description
Jan 2009	CD-0	Statement of Mission Need for LCF program upgrade signed
July 2009	Lehman Review	Presented plan for hybrid system
Jan 2010	CD-1	Alternative selection and cost baseline approved
Aug 2010	Application Review	Present plan for application performance and portability across many platforms
Dec 2010	Lehman Review	Presented baseline and acquisition plan
Feb 2011	CD-3a	Approve long lead time acquisitions
Aug 2011	Lehman Review & CD-2/3b	Approved performance baseline and system acquisition
Oct 2011	Begin Upgrade	Jaguar's 200 cabinets upgraded to new Cray XK6 blades
2H 2012	Kepler GPUs	Add in the 10-20 PF of NVIDIA Kepler GPUs
June 2013	Complete Acceptance Testing	Finish all acceptance testing

Questions?

The research and activities described in this presentation were performed using the resources of the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC0500OR22725.