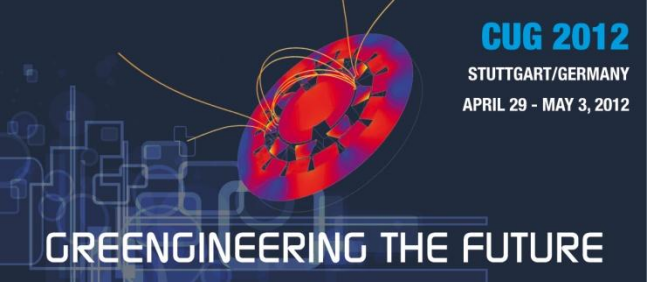


Titan: Early experience with the Cray XK6 at Oak Ridge National Laboratory

Buddy Bland
Jack Wells
Bronson Messer
Oscar Hernandez
Jim Rogers

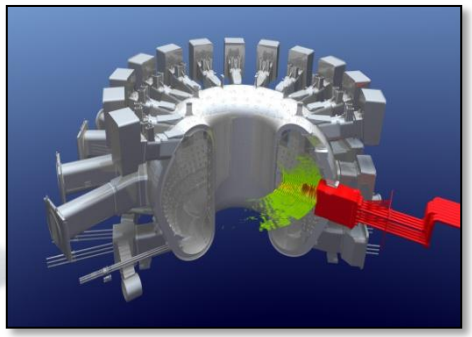
May 2, 2012

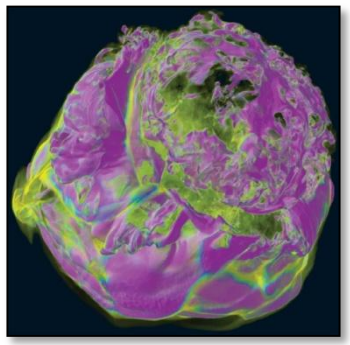
Outline

- Goals of Titan project
- XK6 nodes
- Gemini Interconnect
- Titan system
- Upgrade process
- Programming model
- Tools
- Results

Titan System Goals: Deliver breakthrough science for DOE/SC, industry, and the nation

Geosciences	Energy	Fundamental Science
Understanding our earth and the processes that impact it – Sea level rise – Regional climate change – Geologic carbon sequestration – Biofuels – Earthquakes and Tsunamis	**Reducing U.S. reliance on foreign energy & reducing carbon footprint of production** – Carbon free energy production from fusion, fission, solar, wind, and geothermal sources – Improving the efficiency of combustion energy sources	**Understanding the physical processes from the scale of subatomic particles to the universe** – Understanding the makeup of atoms to supernovae – Developing advanced materials for applications such as photovoltaics & electronic components

Accomplishing these missions requires the power of Titan

Titan System Goals: Promote application development for highly scalable architectures through the Center for Accelerated Application Readiness (CAAR)

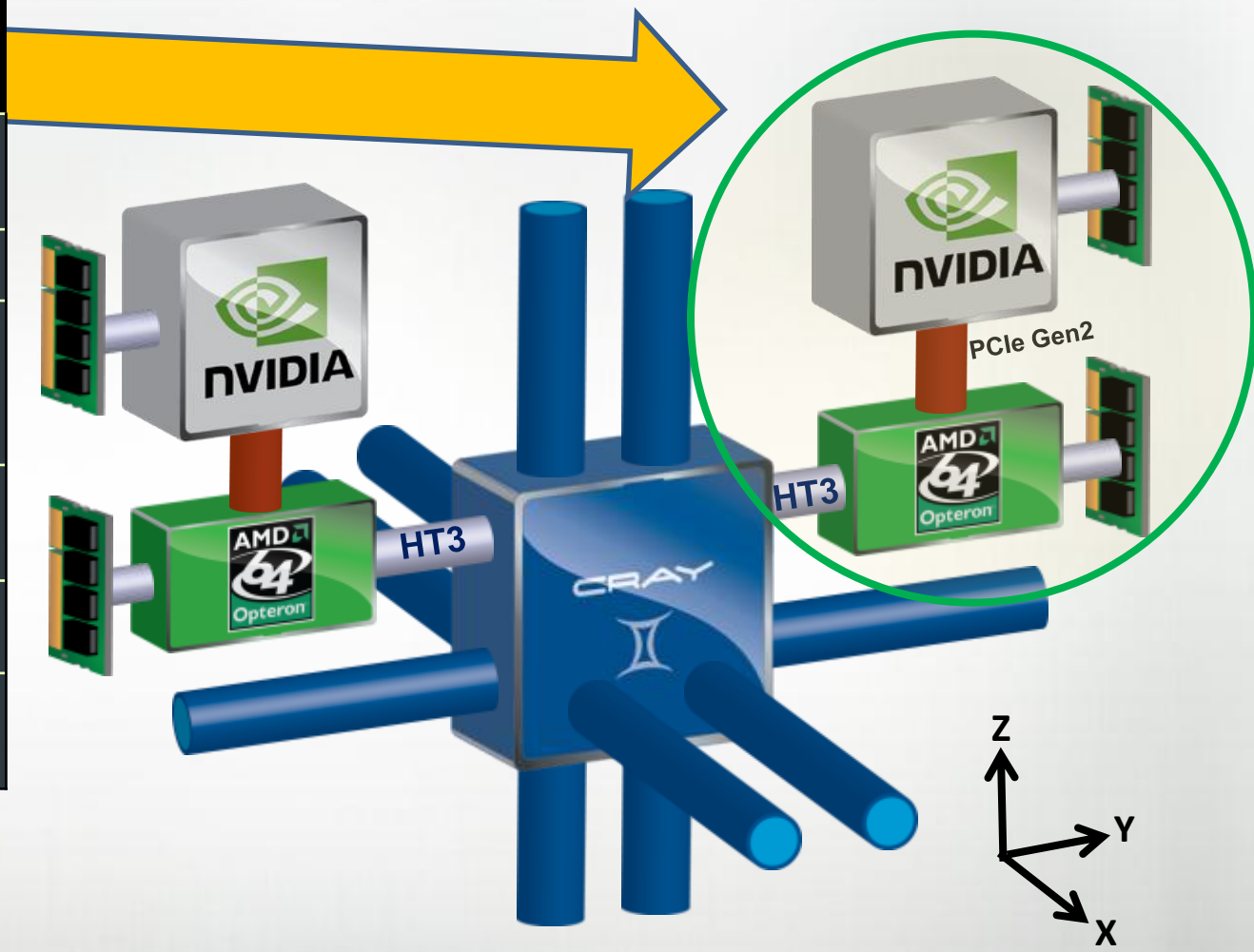
Using six representative apps to explore techniques to effectively use highly scalable architectures

- **CAM-SE** – Atmospheric model
- **Denovo** – Nuclear reactor neutron transport
- **wl-LSMS** - First principles statistical mechanics of magnetic materials
- **S3D** – Combustion model
- **LAMMPS** – Molelcular dynamics
- **NRDF** – Adaptive mesh refinement

- Data locality
- Explicit data management
- Hierarchical parallelism
- Exposing more parallelism through code refactoring and source code directives
- Highly parallel I/O
- Heterogeneous multi-core processor architecture

Cray XK6 Compute Node

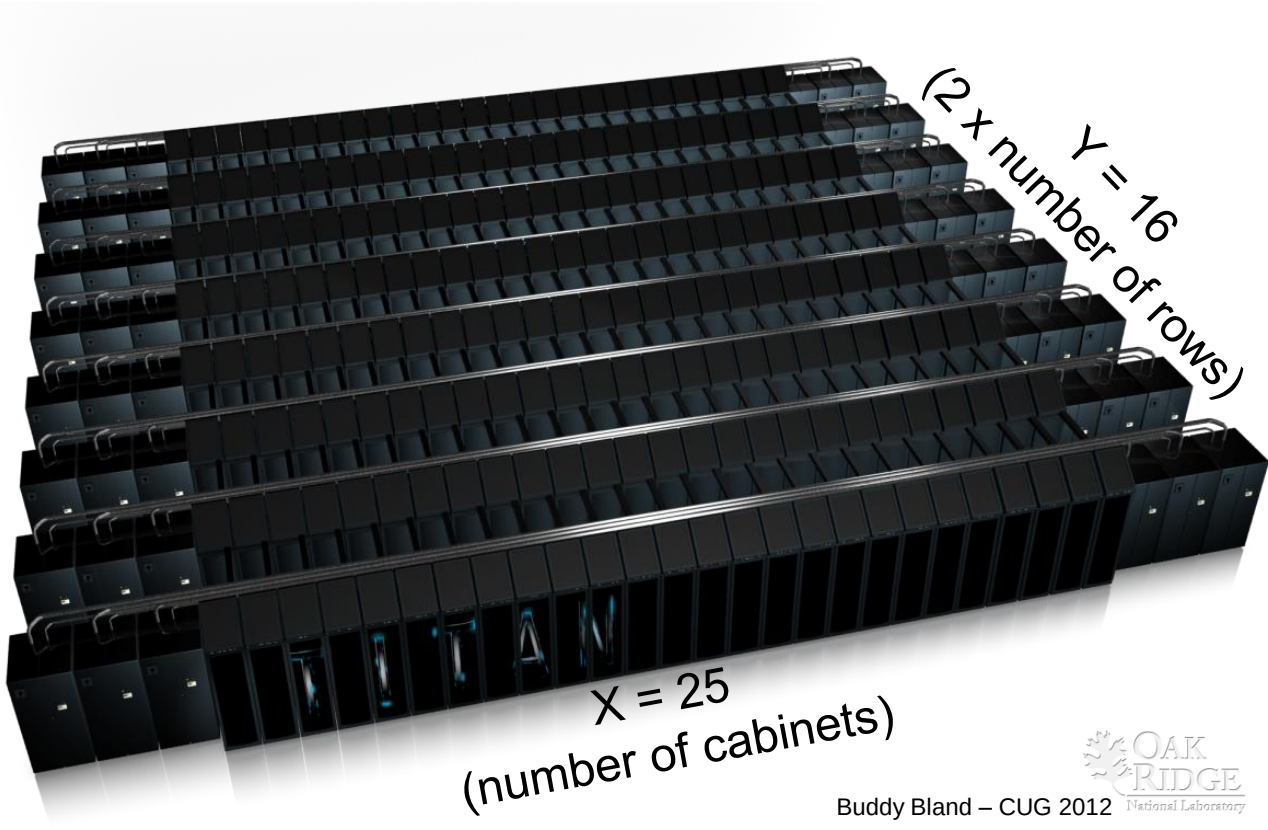
XK6 Compute Node Characteristics
AMD Opteron 6200 Interlagos 16 core processor
Tesla X2090 @ 665 GF
Host Memory 16 or 32GB 1600 MHz DDR3
Tesla X090 Memory 6GB GDDR5 capacity
Gemini High Speed Interconnect
Upgradeable to NVIDIA's KEPLER many-core processor



Slide courtesy of Cray, Inc.

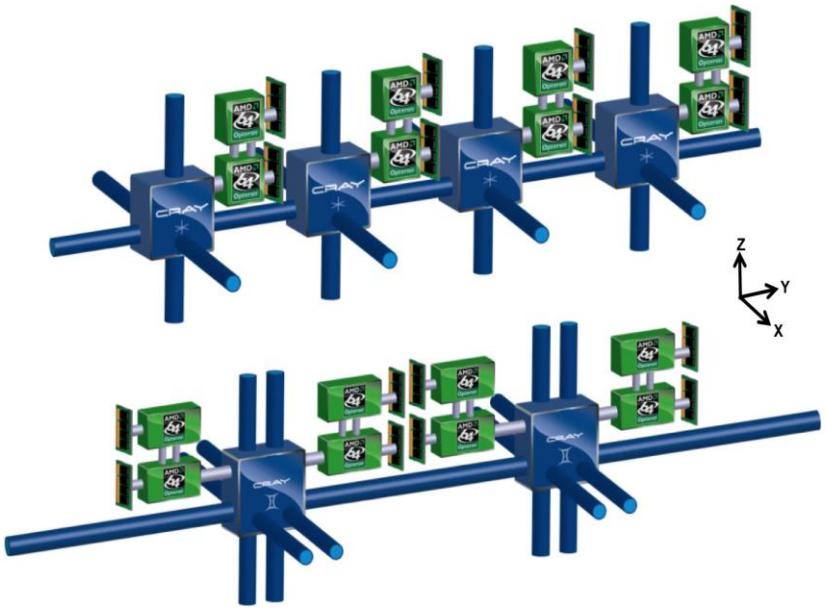
How does Gemini differ from SeaStar?

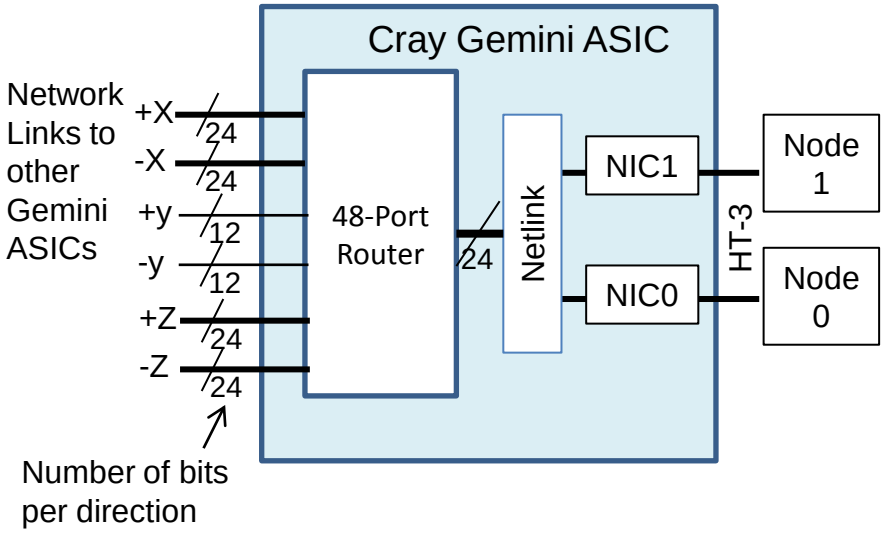
- Torus Layout
  - Each Gemini connects to 2 nodes
  - Result is half as many torus vertices in the Y dimension
  - But since the Gemini uses the same backplane and cables as SeaStar, each vertex has two cables in the X and Z dimensions
  - Jaguar: 25x32x24
  - Titan: 25x16x24



Gemini Torus Bandwidth

Link Type	Bits per direction	Clock Rate (Gbits/s)	Bandwidth per Direction (Gbits/s)
X & Z cables	24	3.125	75
Y cables	12	3.125	37.5
Y mezzanine traces	12	6.25	75
Z backplane traces	24	5.0	120





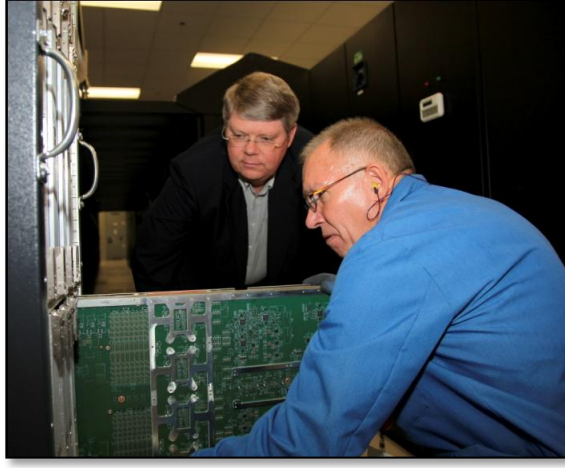
Images courtesy of Cray Inc.

ORNL's "Titan" System

- Upgrade of Jaguar from Cray XT5 to XK6
- Cray Linux Environment operating system
- Gemini interconnect
  - 3-D Torus
  - Globally addressable memory
  - Advanced synchronization features
- AMD Opteron 6274 processors (Interlagos)
- New accelerated node design using NVIDIA multi-core accelerators
  - 2011: 960 NVIDIA x2090 "Fermi" GPUs
  - 2012: 14,592 NVIDIA "Kepler" GPUs
- 20+ PFlops peak system performance
- 600 TB DDR3 mem. + 88 TB GDDR5 mem

Titan Specs	
Compute Nodes	18,688
Login & I/O Nodes	512
Memory per node	32 GB + 6 GB
# of Fermi chips (2012)	960
# of NVIDIA "Kepler" (2013)	14,592
Total System Memory	688 TB
Total System Peak Performance	20+ Petaflops
Cross Section Bandwidths	X=14.4 TB/s Y=11.3 TB/s Z=24.0 TB/s

Two Phase Upgrade Process



- Phase 1: XT5 to XK6 without GPUs
  - Remove all XT5 nodes and replace with XK6 and XIO nodes
  - 16-core processors, 32 GB/node, Gemini
  - 960 nodes (10 cabinets) have NVIDIA Fermi GPUs
  - Users ran on half of system while other half was upgraded
- Add NVIDIA Kepler GPUs
  - Cabinet Mechanical and Electrical upgrades
    - New air plenum bolts on to cabinet to support air flow needed by GPUs
    - Larger fan
    - Additional power supply
    - New doors ☺
  - Rolling upgrade of node boards
    - Pull board, add 4 Kepler GPUs modules, replace board, test, repeat 3,647 times!
    - Keep most of the system available for users during upgrade
  - Acceptance test of system

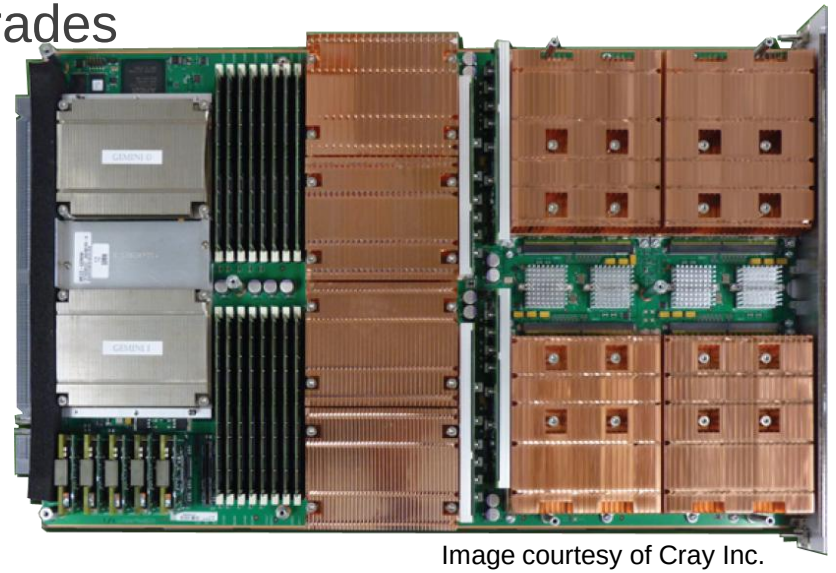


Image courtesy of Cray Inc.

Hybrid Programming Model

- On Jaguar today with 299,008 cores, we are seeing the limits of a single level of MPI scaling for most applications
- To take advantage of the vastly large parallelism in Titan, users need to use hierarchical parallelism in their codes
  - Distributed memory: MPI, Shmem, PGAS
  - Node Local: OpenMP, Pthreads, local MPI communicators
  - Within threads: Vector constructs on GPU, libraries, CPU SIMD
- ***These are the same types of constructs needed on all multi-PFLOPS computers to scale to the full size of the systems!***

How do you program these nodes?

- **Compilers**
  - OpenACC is a set of compiler directives that allows the user to express hierarchical parallelism in the source code so that the compiler can generate parallel code for the target platform, be it GPU, MIC, or vector SIMD on CPU
  - Cray compiler supports XK6 nodes and is OpenACC compatible
  - CAPS HMPP compiler supports C, C++ and Fortran compilation for heterogeneous nodes and is adding OpenACC support
  - PGI compiler supports OpenACC and CUDA Fortran
- **Tools**
  - Allinea DDT debugger scales to full system size and with ORNL support will be able to debug heterogeneous (x86/GPU) apps
  - ORNL has worked with the Vampir team at TUD to add support for profiling codes on heterogeneous nodes
  - CrayPAT and Cray Apprentice support XK6 programming

Titan Tool Suite

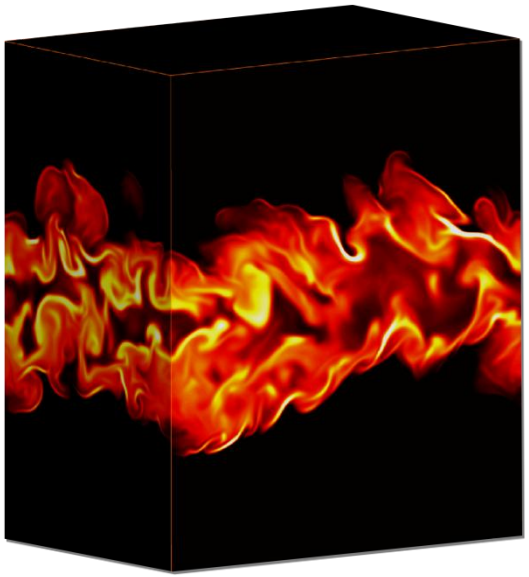
Compilers	Performance Tools	GPU Libraries	Debuggers	Source Code
Cray PGI CAP-HMPP Pathscale NVIDIA CUDA GNU Intel	CrayPAT Apprentice Vampir VampirTrace TAU HPCToolkit CUDA Profiler	MAGMA CULA Trillinos libSCI	DDT NVIDIA Gdb	HMPP Wizard

S3D

Direct Numerical Simulation of Turbulent Combustion

Code Description

- Compressible Navier-Stokes equations
- 3D Cartesian grid, 8th-order finite difference
- Explicit 4th-order Runge-Kutta integration
- Fortran, 3D Domain decomposition, non-blocking MPI



DNS provides unique fundamental insight into the chemistry-turbulence interaction

Porting Strategy

- Hybrid MPI/OpenMP/OpenACC application
- All intensive calculations can be on the accelerator
- Redesign message passing to overlap communication and computation

Early Performance Results on XK6

- **Refactored code was 2x faster on Cray XT5**
- OpenACC acceleration with minimal overhead
- XK6 outperforms XE6 by 1.4x

Science Target (20 PF Titan)

Increased chemical complexity for combustion:

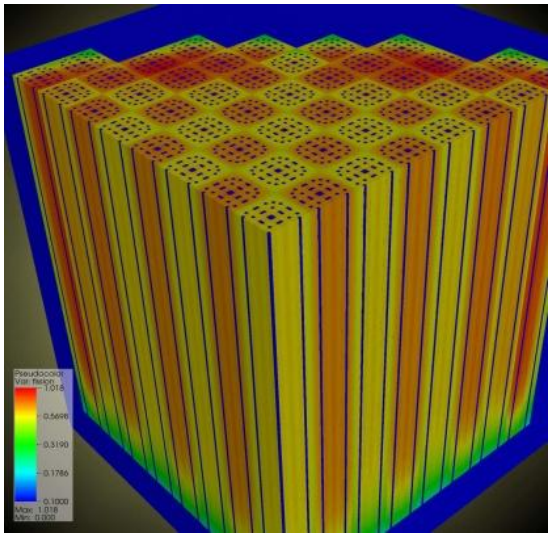
- Jaguar: 9-22 species (H_2, syngas, ethylene)
- Titan: 60-100 species (n-heptane, iso-octane, biofuels)

DENOVO

3D Neutron Transport for Nuclear Reactor Design

Code Description

- Linear Boltzmann radiation transport
- Discrete ordinates method
- Iterative eigenvalue solution
- Multigrid, preconditioned linear solves
- C++ with F95 kernels



DENOVO is a component of the DOE CASL Hub, necessary to achieve CASL challenge problems

Porting Strategy

- SWEEP kernel re-written in C++ & CUDA, runs on CPU or GPU
- Scaling to over 200K cores with opportunities for increased parallelism on GPUs
- Reintegrate SWEEP into DENOVO

Early Performance Results on XK6

- **Refactored code was 2x faster on Cray XT5**
- XK6 performance exceeds XE6 by 3.3x

Science Target (20PF Titan)

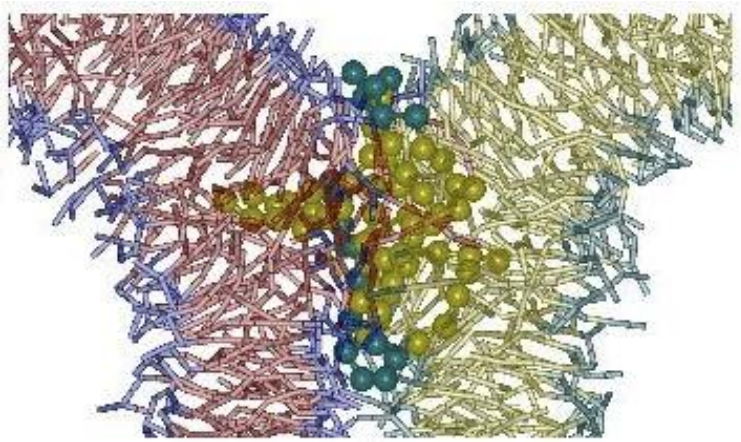
- 3-D, Full reactor radiation transport for CASL challenge problems within 12 wall clock hours. This is the CASL stretch goal problem!

LAMMPS

Large-scale, massively parallel molecular dynamics

Code Description

- Classical N-body problem of atomistic modeling
- Force fields available for chemical, biological, and materials applications
- Long-range electrostatics evaluated using a "particle-particle, particle-mesh" (PPPM) solver.
- 3D FFT in particle-mesh solver limits scaling



Insights into the molecular mechanism of membrane fusion from simulation. Stevens et al., *PRL* **91** (2003)

Porting Strategy

- For PPPM solver, replace 3-D FFT with grid-based algorithms that reduce inter-process communication
- Parallelism through domain decomposition of particle-mesh grid
- Accelerated code builds with OpenCL or CUDA

Early Performance Results on XK6:

- XK6 outperforms XE6 by 3.2x
- XK6 outperforms XK6 w/o GPU by 6.5x

Science Target (20PF Titan)

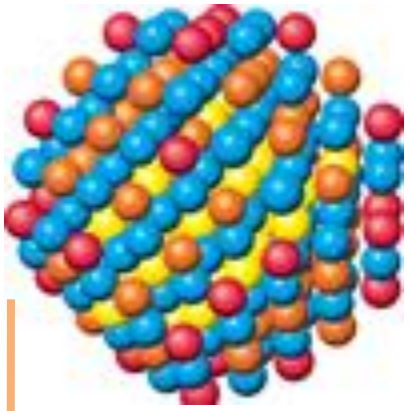
- Simulate biological membrane fusion in coarse-grained MD within 5 wall clock days

Wang-Landau LSMS

First principles, statistical mechanics of magnetic materials

Code Description

- Combines classical statistical mechanics (W-L)for atomic magnetic moment distributions with first-principles calculations (LSMS) of the associated energies.
- Main computational effort is dense linear algebra for complex numbers
- F77 with some F90 and C++ for the statistical mechanics driver.



Compute the magnetic structure and thermodynamics of low-dimensional magnetic structures

Porting Strategy

- Leverage accelerated linear algebra libraries, e.g., cuBLAS + CULA, LibSci_acc
- Parallelization over (1) W-L Monte-Carlo walkers, (2) over atoms through MPI process, (3) OpenMP on CPU sections.
- Restructure communications: moved outside energy loop

Early Performance Results on XK6:

- XK6 outperforms XE6 by 1.6x
- XK6 outperforms XK6 w/o GPU by 3.1x

Science Target (20PF Titan)

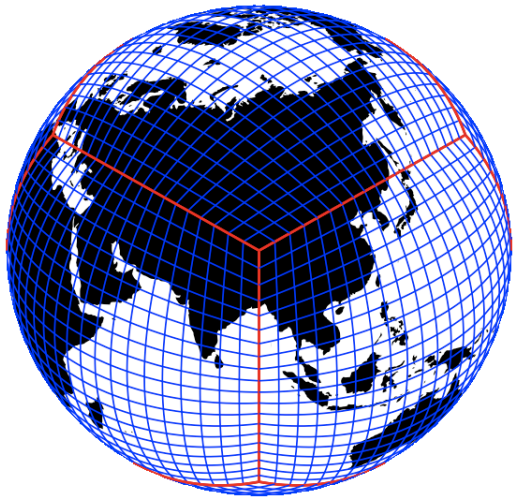
- Calculate both the magnetization and the free-energy for magnetic materials.

CAM-SE

Community Atmosphere Model – Spectral Elements

Code Description

- Employs equal-angle, cubed-sphere grid and terrain-following coordinate system.
- Scaled to 172,800 cores on XT5
- Exactly conserves dry mass without the need for *ad hoc* fixes.
- Original baseline code achieves parallelism through domain decomposition using one MPI task per element



Cubed-sphere grid of CAM spectral element model. Each cube panel is divided into elements.

http://www-personal.umich.edu/~paullric/A_CubedSphere.png

Porting Strategy

- Using realistic "Mozart" chemical tracer network, tracer transport (i.e., advection) dominates the run time.
- Use hybrid MPI/OpenMP parallelism
- Intensive kernels are coded in CUDA Fortran
- Migration in future to OpenACC

Early Performance Results on XK6:

- **Refactored code was 1.7x faster on Cray XT5**
- XK6 outperforms XE6 by 1.5x
- XK6 outperforms XK6 w/o GPU by 2.6x

Science Target (20PF Titan)

- CAM simulation using Mozart tropospheric chemistry with 106 constituents at 14 km horizontal grid resolution

How Effective are GPUs on Scalable Applications?
OLCF-3 Early Science Codes -- Current performance measurements on TitanDev

	XK6 (w/ GPU) vs. XK6 (w/o GPU)	XK6 (w/ GPU) vs. XE6	Cray XK6: Fermi GPU plus Interlagos CPU Cray XE6: Dual Interlagos and no GPU
Application	**Performance Ratio**	**Performance Ratio**	**Comment**
S3D	1.5	1.4	• Turbulent combustion • **6% of Jaguar workload**
Denovo	3.5	3.3	• 3D neutron transport for nuclear reactors • **2% of Jaguar workload**
LAMMPS	6.5	3.2	• High-performance molecular dynamics • **1% of Jaguar workload**
WL-LSMS	3.1	1.6	• Statistical mechanics of magnetic materials • **2% of Jaguar workload** • 2009 Gordon Bell Winner
CAM-SE	2.6	1.5	• Community atmosphere model • **1% of Jaguar workload**

Additional Applications from Community Efforts
Current performance measurements on TitanDev

	XK6 (w/ GPU) vs. XK6 (w/o GPU)	XK6 (w/ GPU) vs. XE6	Cray XK6: Fermi GPU plus Interlagos CPU Cray XE6: Dual Interlagos and no GPU
Application	**Performance Ratio**	**Performance Ratio**	**Comment**
NAMD	2.6	1.4	• High-performance molecular dynamics • **2% of Jaguar workload**
Chroma	8.8	6.1	• High-energy nuclear physics • **2% of Jaguar workload**
QMCPACK	3.8	3.0	• Electronic structure of materials • New to OLCF, Common to
SPECFEM-3D	4.7	2.5	• Seismology • 2008 Gordon Bell Finalist
GTC	2.5	1.6	• Plasma physics for fusion-energy • **2% of Jaguar workload**
CP2K	2.8	1.5	• Chemical physics • **1% of Jaguar workload**

Access to Titan via INCITE

U.S. DEPARTMENT OF ENERGY
INCITE
LEADERSHIP COMPUTING

INCITE seeks computationally intensive, large-scale research projects with the potential to significantly advance key areas in science and engineering.

1 **Impact criterion**

High-impact science and engineering

2 **Computational leadership criterion**

Computationally intensive runs that cannot be done anywhere else

3 **Eligibility criterion**

- INCITE grants allocations regardless of funding source (ex. DOE, NSF, private, etc)
- Researchers at non-US institutions may apply

Call for 2013 proposals open now through June 27, 2012

The INCITE program seeks proposals for high-impact science and technology research challenges that require the power of the leadership-class systems

In 2013 over 4 billion core-hours will be available through the INCITE program

Email: INCITE@DOEleadershipcomputing.org
Web: http://hpc.science.doe.gov

Questions?

Buddy Bland
Email: BlandAS@ORNL.Gov

Want to join our team? ORNL is hiring. Contact any of the Oak Ridge team.

The research and activities described in this presentation were performed using the resources of the National Center for Computational Sciences 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.

OAK RIDGE National Laboratory