# Overview of ORNL Leadership Computing Facility and Usage

20 Years of Excellence in Computational Science
OLCF
OAK RIDGE LEADERSHIP COMPUTING FACILITY
1992–2012

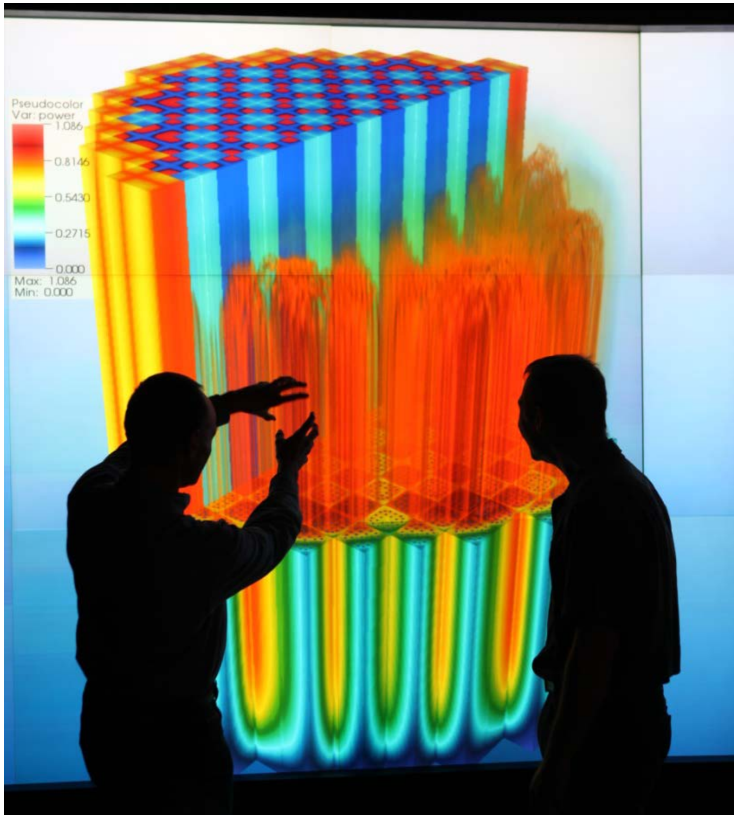
CScADS Summer Workshop
Leadership Computing Platforms, Extreme-scale Applications, and Performance Strategies
July 23, 2012

By Hai Ah Nam
Scientific Computing Group
Oak Ridge Leadership Computing Facility
National Center for Computational Sciences Division

# Oak Ridge Leadership Computing Facility Mission

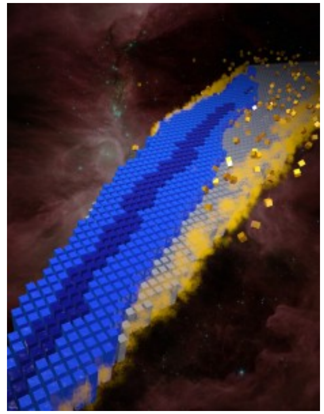

The OLCF is a DOE Office of Science National User Facility whose mission is to enable breakthrough science by:

- Fielding the most powerful capability computers for scientific research,
- Building the required infrastructure to facilitate user access to these computers,
- Selecting a few time-sensitive problems of national importance that can take advantage of these systems,
- And partnering with these teams to deliver breakthrough science.

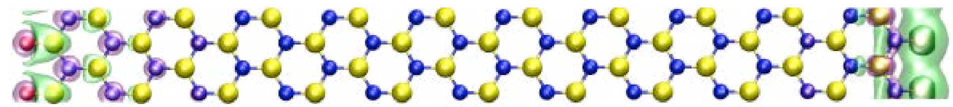


# Breakthrough Science at Every Scale

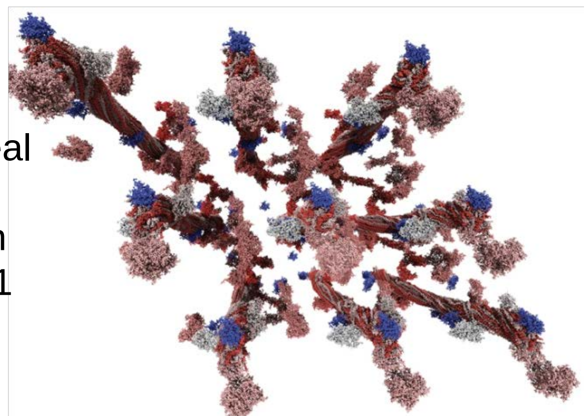
**Nuclear Physics**
Nazarewicz et al., map the nuclear driplines that mark the borders of nuclear existence, predicting ~7000 bound nuclei, though only ~3000 have been observed. Nature 2012

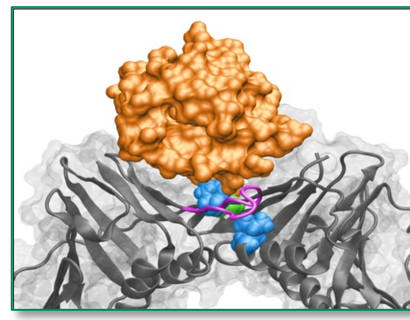


**New Materials**
Lopez-Bezanilla et al., discover that boron-nitride monolayers are an ideal dielectric substrate for future nanoelectronic devices constructed with graphene as the active layer

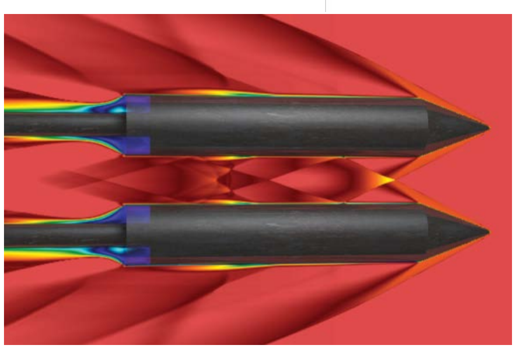


**Biofuels**
Smith et al., reveal the surface structure of lignin clumps down to 1 angstrom



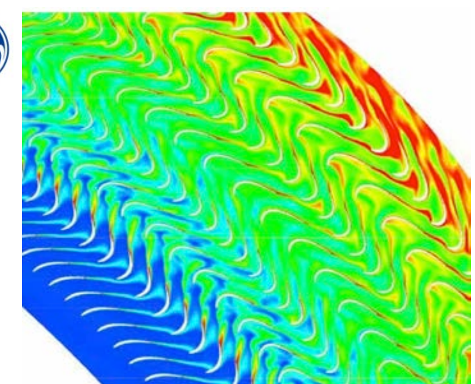


**Biochemistry**
Ivanov et al., illuminate how DNA replication continues past a damaged site so a lesion can be repaired later



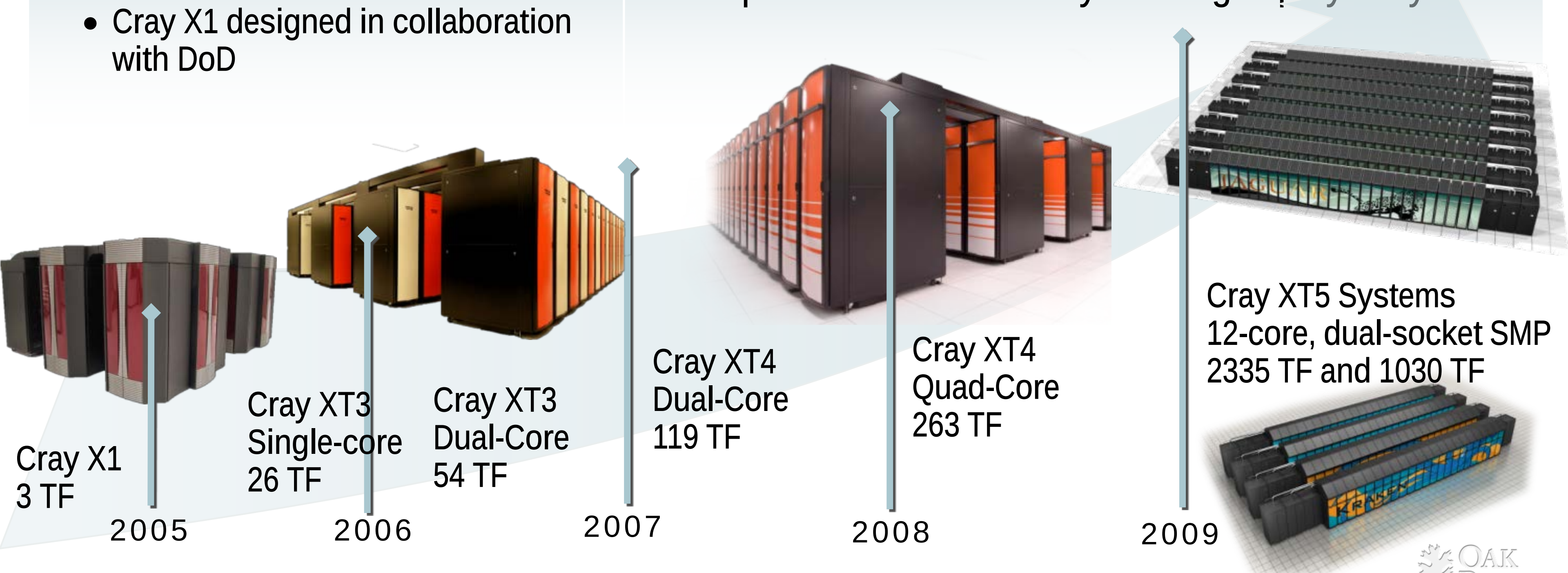
**Design Innovation**
Ramgen Power Systems accelerates their design of shock wave turbo compressors for carbon capture and sequestration

**Turbo Machinery Efficiency**
General Electric, for the first time, simulated unsteady flow in turbo machinery, opening new opportunities for design innovation and efficiency improvements.



# We have increased system performance by 1,000 times since 2004

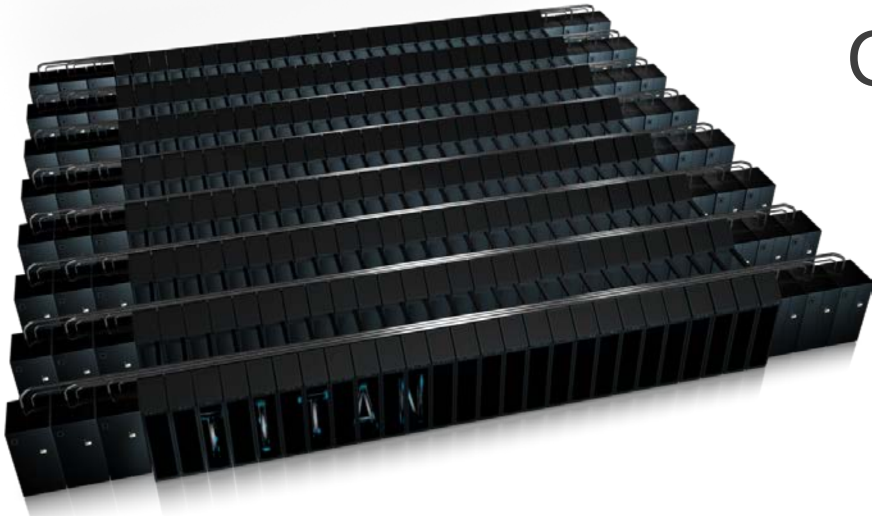
| Hardware scaled from single-core through dual-core to quad-core and dual-socket , 12-core SMP nodes | Scaling applications and system software is the biggest challenge |
|---|---|
| • NNSA and DoD have funded much of the basic system architecture research<br>  • Cray XT based on Sandia Red Storm<br>  • IBM BG designed with Livermore<br>  • Cray X1 designed in collaboration with DoD | • DOE SciDAC and NSF PetaApps programs are funding scalable application work, advancing many apps<br>• DOE-SC and NSF have funded much of the library and applied math as well as tools<br>• Computational Liaisons key to using deployed systems |



# Hierarchical Parallelism

- Parallelism on multiple levels
  - MPI parallelism between nodes (or PGAS)
  - On-node, SMP-like parallelism via threads
  - Vector parallelism
    - SSE/AVX on CPUs
    - GPU threaded parallelism
- It doesn't matter if you use GPU-based machines or not
  - GPUs [CUDA, OpenCL, directives]
  - FPUs on Power [xlf, etc.]
  - Cell [SPE]
  - SSE/AVX; MIC (Knights Ferry, Knights Corner)[?]

**Increasing node-level parallelism and data locality are universally needed**




# ORNL's "Titan" System

## Upgrade of Jaguar from Cray XT5 to XK6

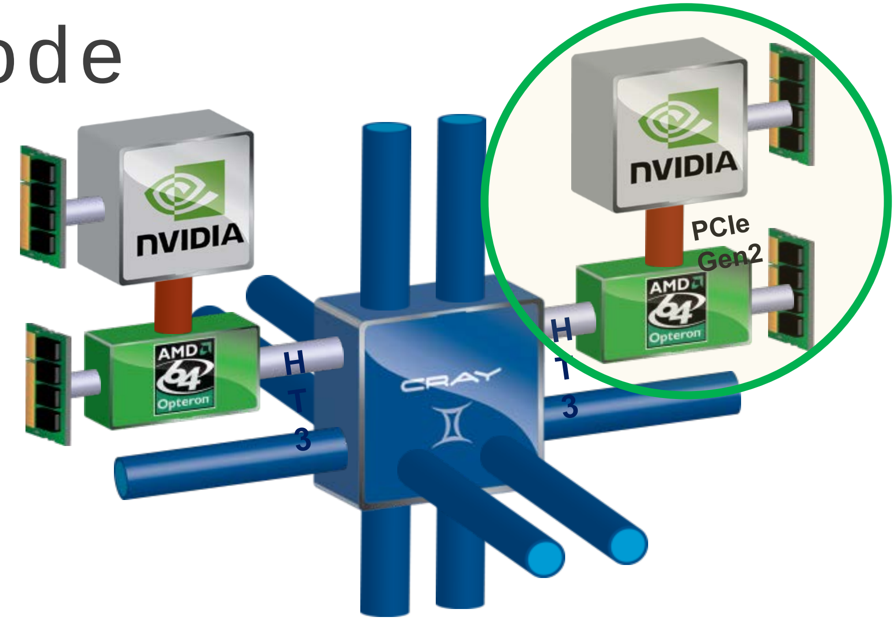
- AMD Opteron 6274 processors (Interlagos)
- New accelerated node design using NVIDIA multi-core accelerators
  - 2011: 960 NVIDIA x2090 "Fermi" GPUs
  - 2012: 14,592 NVIDIA "Kepler" GPUs
- Gemini interconnect
  - 3-D Torus, Globally addressable memory
  - Advanced synchronization features

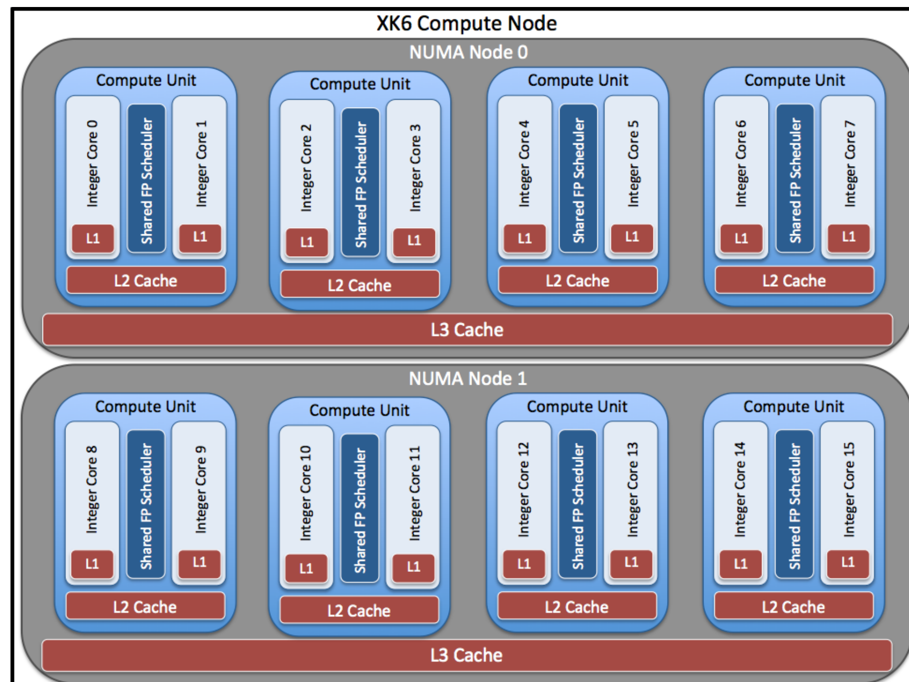
| Titan Specs | |
|---|---|
| Compute Nodes | 18,688 |
| Cores (CPU) | 299,008 |
| Login & I/O Nodes | 512 |
| Memory per node | 32 GB + 6 GB |
| # of Fermi chips (2012) | 960 |
| # of NVIDIA "Kepler" (2013) | 14,592 |
| Total System Memory | 688 TB |
| Total System Peak Performance | 20+ Petaflops |
| Cross Section Bandwidths | X=14.4 TB/s<br>Y=11.3 TB/s<br>Z=24.0 TB/s |

20+ PFlops peak system performance | 600 TB DDR3 + 88 TB GDDR5 mem

# Cray XK6 Compute Node

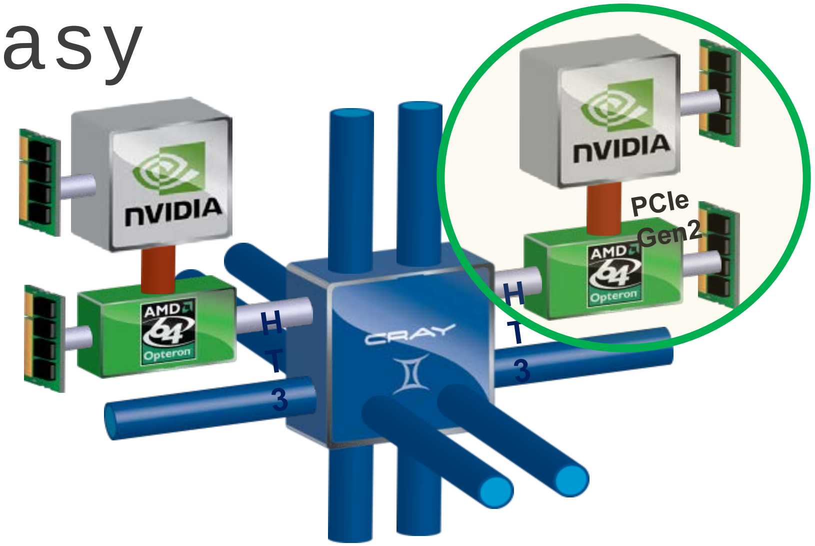
| XK6 Compute Node Characteristics |
|---|
| AMD Opteron 6200 "Interlagos" 16 core processor @ 2.2GHz |
| Tesla M2090 "Fermi" @ 665 GF with 6GB GDDR5 memory |
| Host Memory, 32GB 1600 MHz DDR3 |
| Gemini High Speed Interconnect |
| Upgradeable to NVIDIA's next generation "Kepler" processor in 2012 |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

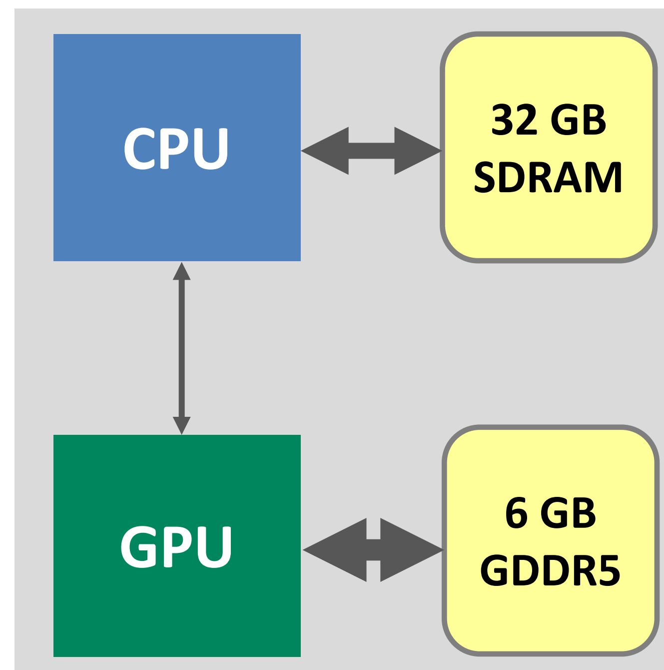


# In principle, GPUs are easy

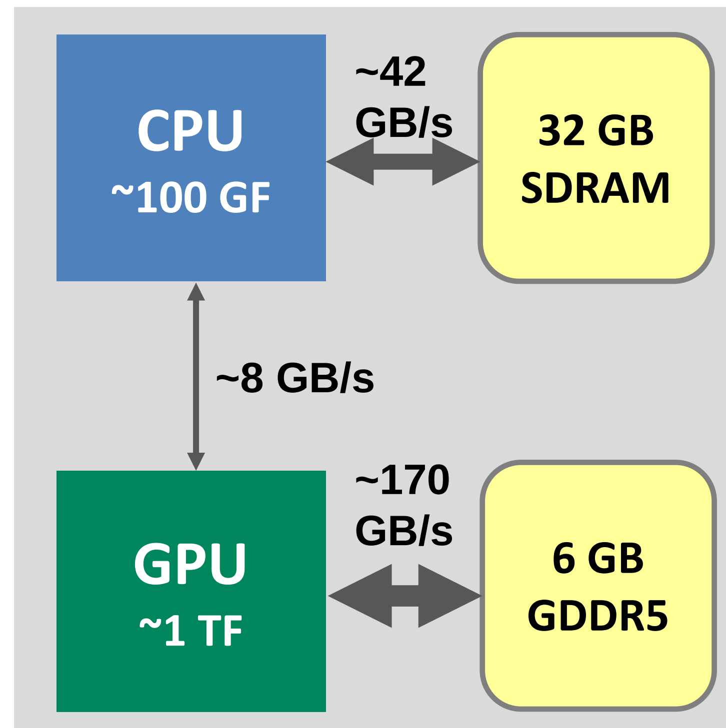
- Identify opportunities for acceleration (loops/high flops)
  - Allocate arrays on GPU
  - Move data from host to GPU
  - Launch computer kernel on GPU
  - Move results from GPU to host
  - Deallocate arrays on GPUs



# In practice, it's tricky

- Optimization problem!
  - Exploiting strengths (FLOPS), avoiding weakness (DATA MOVEMENT)
- Identifying acceleration opportunities is not obvious
  - New algorithm
    - Minimize data, flops → minimize data movement
    - Multiple levels of parallelism
  - Revisit good coding practices and vector parallelism



# Titan (GPU) programming tools (for now)

| Compiler Directives (OpenACC) | | |
|---|---|---|
| Cray | PGI | HMPP Toolkit* (CAPS) |

| Low-level GPU Languages | | |
|---|---|---|
| OpenCL (agnostic) | CUDA C (NVIDIA) | CUDA Fortran (PGI) |

Intended for portability (GPU, MIC, APU, etc.)
OpenACC: Standard for directives to designate areas for GPU kernels
* OpenCL/CUDA converted source provided

| Accelerated Libraries** | | | | | |
|---|---|---|---|---|---|
| Libsci_acc (Cray) | MAGMA (ICL/UT) (GNU) | CULA (EM Photonics) | cuBLAS/ cuSparse (NVIDIA) | Trilinos | Etc, etc. |

** Libraries are based on CUDA

| Performance Tools | | | | |
|---|---|---|---|---|
| CrayPAT / Apprentice | Vampir / VampirTrace | TAU | HPCToolkit | CUDA Profiler |

| Debuggers | | |
|---|---|---|
| Allinea DDT | NVIDIA | gdb |

# CAAR @ the OLCF

*Center for Accelerated Application Readiness*

- Titan System Goals: Promote application development for highly scalable architectures

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

**CAM-SE**
- **Community Atmospheric Model**

**Denovo**
- **3D neutron transport for nuclear reactors**

**wl-LSMS**
- **First principles statistical mechanics of magnetic materials**

**S3D**
- **Turbulent Combustion model**

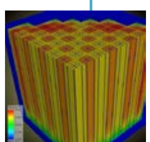
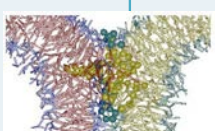
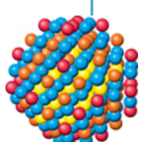
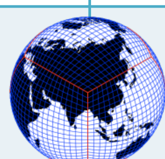
**LAMMPS**
- **Molecular Dynamics**

**NRDF**
- **Adaptive mesh refinement**

**Addressing:**
- 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

# GPUs on Scalable Applications
## OLCF-3 Early Science Codes

Current performance (ratio) measurements on TitanDev (XK6) vs. XE6

| Application | XK6 (w/ GPU) vs. XK6 (w/o GPU) | XK6 (w/ GPU) vs. XE6 | Comment |
|---|---|---|---|
| **S3D**  | **1.5** | **1.4** | • Hybrid MPI/OpenMP/OpenACC<br>• Redesign message passing – overlap<br>• 6% of Jaguar workload |
| **Denovo**  | **3.5** | **3.3** | • SWEEP kernel rewritten in C++ & CUDA<br>• 2% of Jaguar workload |
| **LAMMPS**  | **6.5** | **3.2** | • Builds with either OpenCL or CUDA<br>• 1% of Jaguar workload |
| **WL-LSMS**  | **3.1** | **1.6** | • Accelerated linear algebra libraries<br>• 2% of Jaguar workload<br>• 2009 Gordon Bell Winner |
| **CAM-SE**  | **2.6** | **1.5** | • Hybrid MPI/OpenMP/CUDA<br>• 1% of Jaguar workload |

**Cray XK6: Fermi GPU plus Interlagos CPU; Cray XE6: Dual Interlagos and no GPU**

# Community Efforts

Current performance (ratio) measurements on TitanDev (XK6) vs. XE6

| Application | XK6 (w/ GPU) vs. XK6 (w/o GPU) | XK6 (w/ GPU) vs. XE6 | Comment |
|---|---|---|---|
| **NAMD** | **2.6** | **1.4** | • High-performance molecular dynamics<br>• **2% of Jaguar workload** |
| **Chroma** | **8.8** | **6.1** | • High-energy nuclear physics<br>• **2% of Jaguar workload** |
| **QMCPACK** | **3.8** | **3.0** | • Electronic structure of materials<br>• New to OLCF, Common to |
| **SPECFEM-3D** | **4.7** | **2.5** | • Seismology<br>• 2008 Gordon Bell Finalist |
| **GTC** | **2.5** | **1.6** | • Plasma physics for fusion-energy<br>• **2% of Jaguar workload** |
| **CP2K** | **2.8** | **1.5** | • Chemical physics<br>• **1% of Jaguar workload** |

**Cray XK6: Fermi GPU plus Interlagos CPU; Cray XE6: Dual Interlagos and no GPU**

# 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

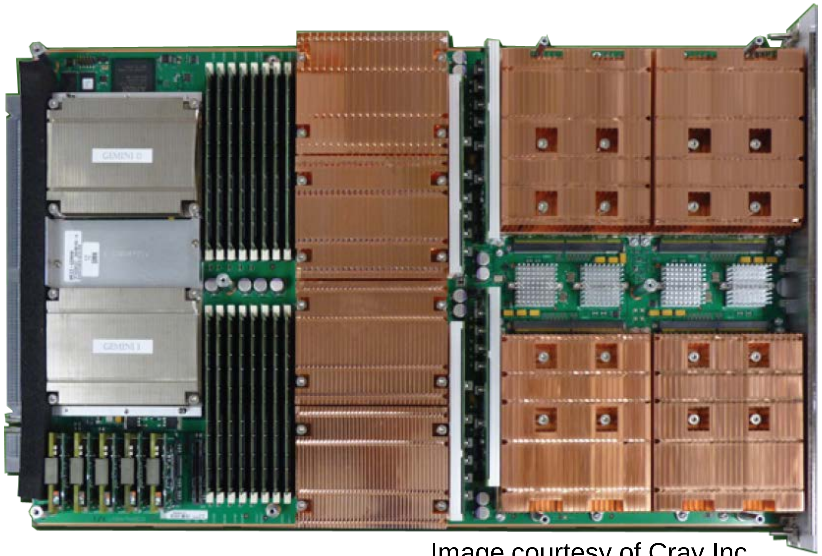
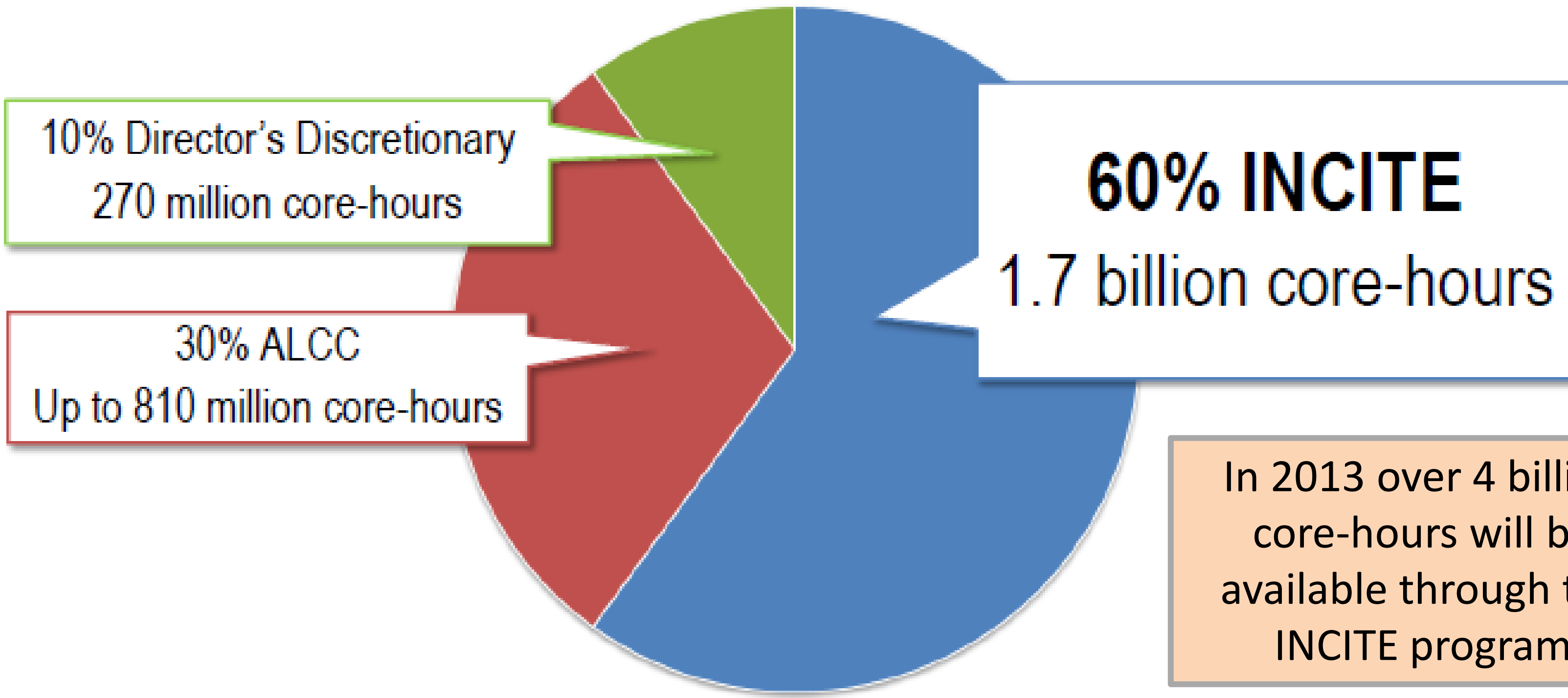


Image courtesy of Cray Inc.

# Access to Titan

## Provide access to LCFs:
## More than 2.7 billion core hours awarded in 2012




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

INCITE Webinar: http://www.doeleadershipcomputing.org/faqs

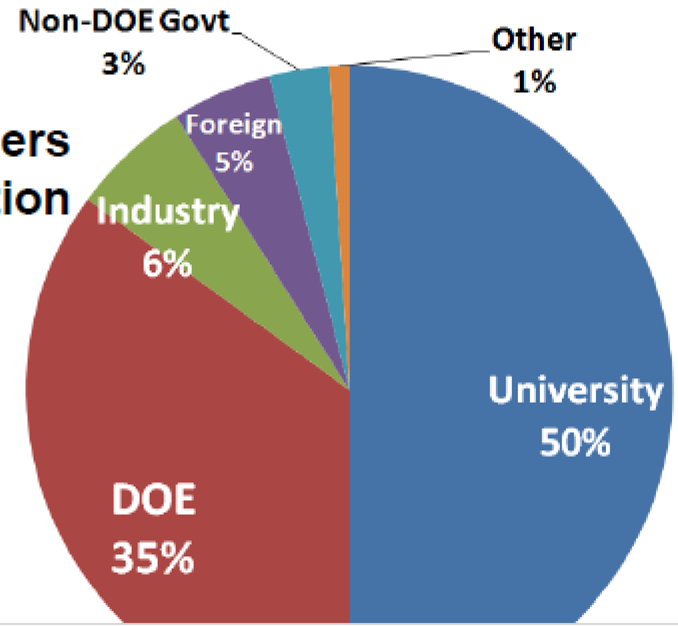
# OLCF Allocation Programs
## *Selecting applications of national importance*

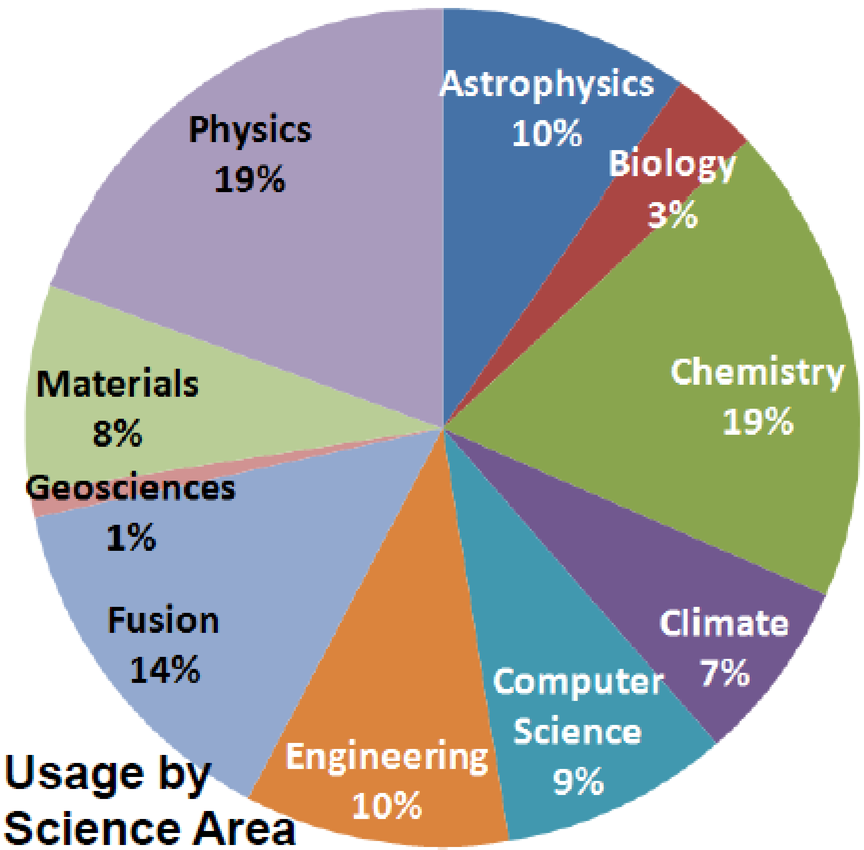
| | INCITE (60%) | | ALCC (30%) | | Director's Discretionary (10%) | |
|---|---|---|---|---|---|---|
| **Mission** | High-risk, high-payoff science that requires LCF-scale resources* | | High-risk, high-payoff science aligned with DOE mission | | Strategic LCF goals | |
| **Call** | 1x/year – (Closes June) | | 1x/year – (Closes February) | | Rolling | |
| **Duration** | 1-3 years, yearly renewal | | 1 year | | 3m,6m,1 year | |
| **Typical Size** | 30 - 40 projects | 10M - 100M core-hours/yr. | 5 - 10 projects | 1M – 75M core-hours/yr. | 100s of projects | 10K – 1M core-hours |
| **Review Process** | Scientific Peer-Review | Computational Readiness | Scientific Peer-Review | Computational Readiness | Strategic impact and feasibility | |
| **Managed By** | INCITE management committee (ALCF & OLCF) | | DOE Office of Science | | LCF management | |
| **Availability** | Open to all scientific researchers and organizations<br>***Capability >20% of cores*** | | | | | |

# OLCF User Demographics

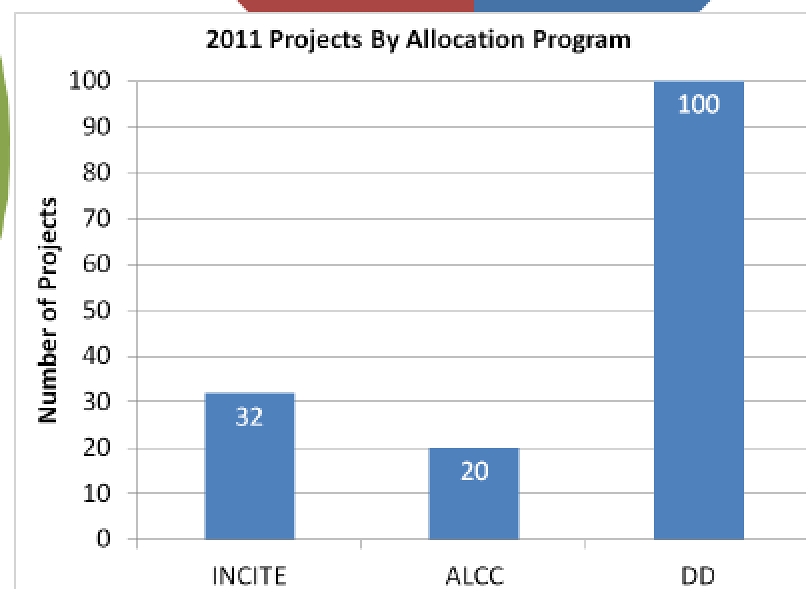
- Number of Users: 857

Percentage of Users by Affiliation




Usage by Science Area



# Leadership Metric & Scheduling

- As a DOE Leadership Computing Facility, the OLCF has a mandate to be used for large, leadership-class (aka capability) jobs.
- To that end, the OLCF implements queue policies that enable large jobs to run in a timely fashion.

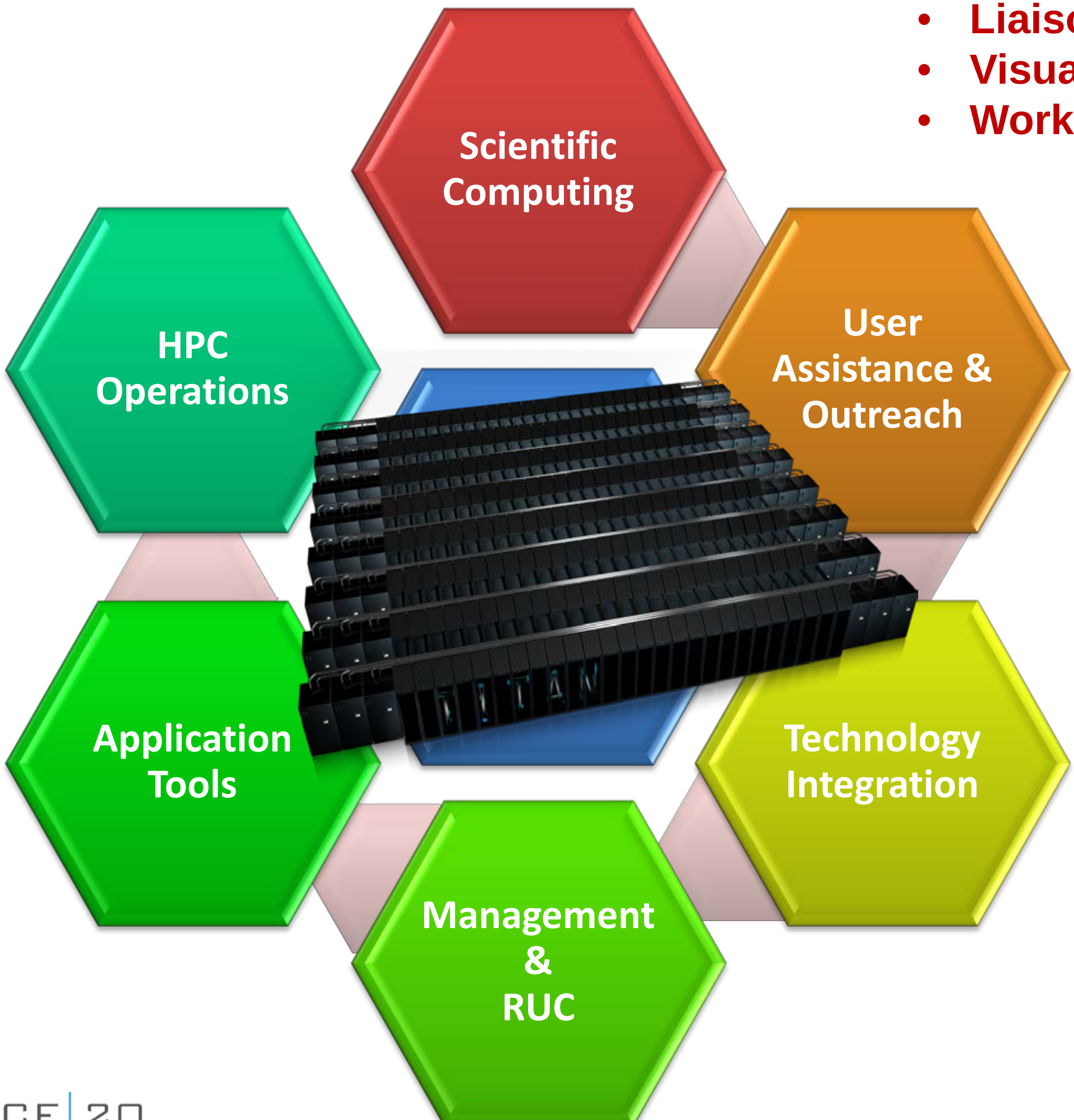
| Bin | Min Cores | Max Cores | Max Walltime (hours) | Aging Boost (Days) |
|---|---|---|---|---|
| 5 | 180,000 | -- | 24.0 | 15 |
| 4 | 60,000 | 179,999 | 24.0 | 5 |
| 3 | 5,008 | 59,999 | 12.0 | 0 |
| 2 | 2,004 | 5,007 | 6.0 | 0 |
| 1 | 1 | 2,003 | 2.0 | 0 |

www.olcf.ornl.gov/support/user-guides-policies/jaguar-xk6-user-guide

**Leadership Usage Metric:**

35% of the CPU time used on the system will be accumulated by jobs using 20% or more of the available processors (60,000 cores)

# The OLCF




- Liaisons
- Visualization
- Work Flow (end-to-end)

- Advocate
- Porting, Scaling, Performance, & Debugging Assistance
- Project Collaboration

# OLCF Training Programs

- 2012
  - January: Titan Workshop
  - February: Electronic Structure Calculation Methods on Accelerators
  - March: Performance Analysis Tools
  - April: OLCF Spring Training & User's Meeting
  - May: GPU Technology Conference, San Jose
  - June: Crash Course in Supercomputing
  - June - August: HPC Fundamentals Series Summer
  - October 9: Cray technical Workshop on XK6 Programming
- www.olcf.ornl.gov | help@nccs.gov
- Presentations, webinars available

# Questions

Hai Ah Nam
namha@ornl.gov

The 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.