Accelerating Science on Titan: PanDA-Related HEP and NP Payloads

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

Kenneth Read
Oak Ridge National Laboratory/
University of Tennessee

PanDA Workshop
University of Texas, Arlington, TX
September 4, 2013

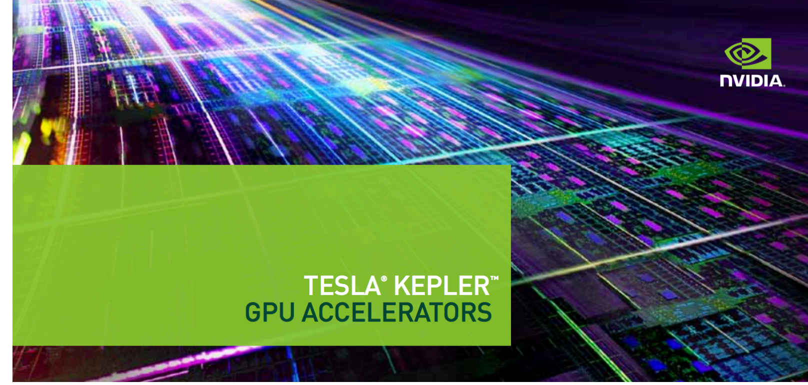
What is the Leadership Computing Facility?

- Collaborative DOE Office of Science program at Oak Ridge and Argonne National Laboratories
- Mission: Provide the computational and data science resources required to solve the most important scientific & engineering problems in the world.
- Highly competitive user allocation programs (INCITE, ALCC).
- Projects receive 10x to 100x more resources than at other generally available centers.
- LCF centers partner with users to enable science & engineering breakthroughs.

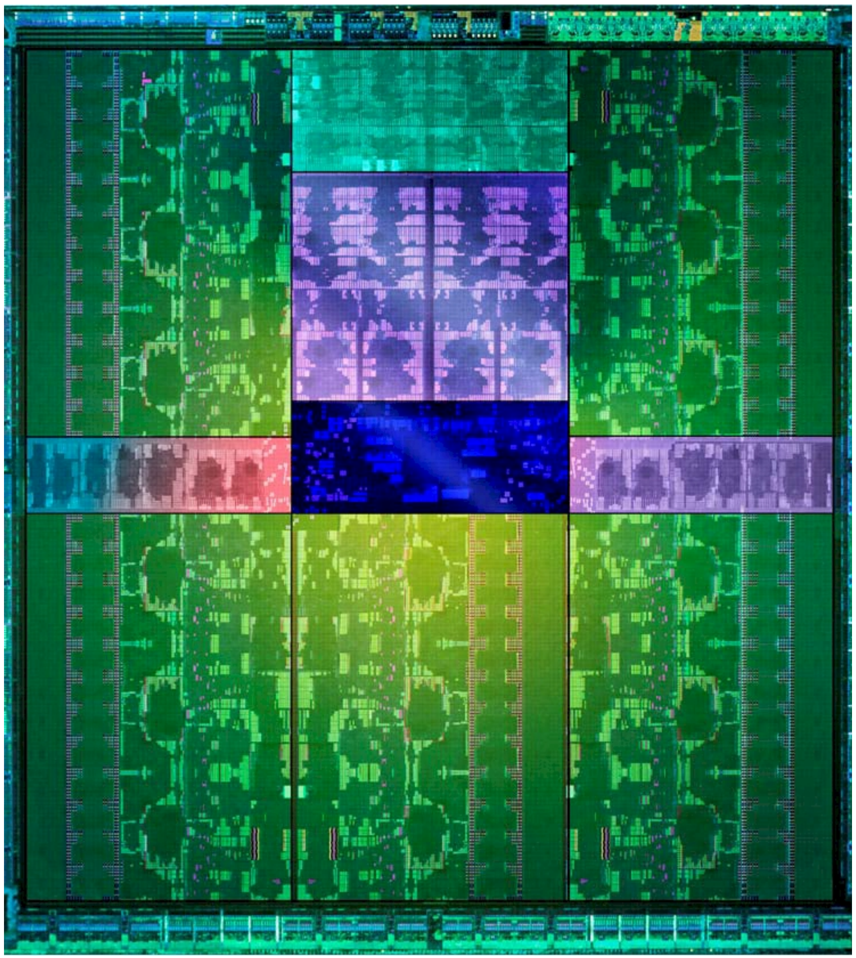
Oak Ridge National Laboratory

Argonne National Laboratory

Kepler GK110 GPU



Most complex semiconductor device ever.
Delivers 1.3 TFlop peak double precision.



2.3 cm

NVIDIA Tesla Kepler K20X

NVIDIA GeForce GTX Titan – On Sale Now

Science breakthroughs at the LCF:

A few of the many science and engineering advances through the INCITE program

Hours requested vs. allocated: ***~2X per year*** ***~3X per year***

Hours allocated	4.9M	6.5M	18.2M	95M	268M	889M	1.6B	1.7B	1.7B	***5B***
Projects	3	3	15	45	55	66	69	57	60	61

2004 2005 2006 2007 2008 2009 2010 2011 2012 2013

Researchers solved the 2D Hubbard model and presented evidence that it predicts HTSC behavior, ***Phys. Rev. Lett*** *(2005)* .

Modeling of molecular basis of Parkinson's disease named #1 computational accomplishment, ***Breakthroughs*** *(2008).*

Largest simulation of a galaxy's worth of dark matter, showed for the first time the fractal-like appearance of dark matter substructures, ***Nature*** (2008), ***Science*** *(2009).*

World's first continuous simulation of 21,000 years of Earth's climate history, ***Science*** *(2009).*

Largest-ever LES of a full-sized commercial combustion chamber used in an existing helicopter turbine, **Compte Rendus de Mecanique** (2009).

Unprecedented simulation of magnitude-8 earthquake over 125-square miles, ***Proceedings SC10.***

NIST proposes new standard reference materials from LCF concrete simulations, ***Eur Phys J E Soft Matter*** *(2012).*

Calculation of the number of bound nuclei in nature, ***Nature*** *(2012).*

New method to rapidly determine protein structure, with limited experimental data, ***Science*** *(2010),* ***Nature*** *(2011).*

OMEN breaks the petascale barrier using more than 220,000 cores, ***Proceedings SC10.***

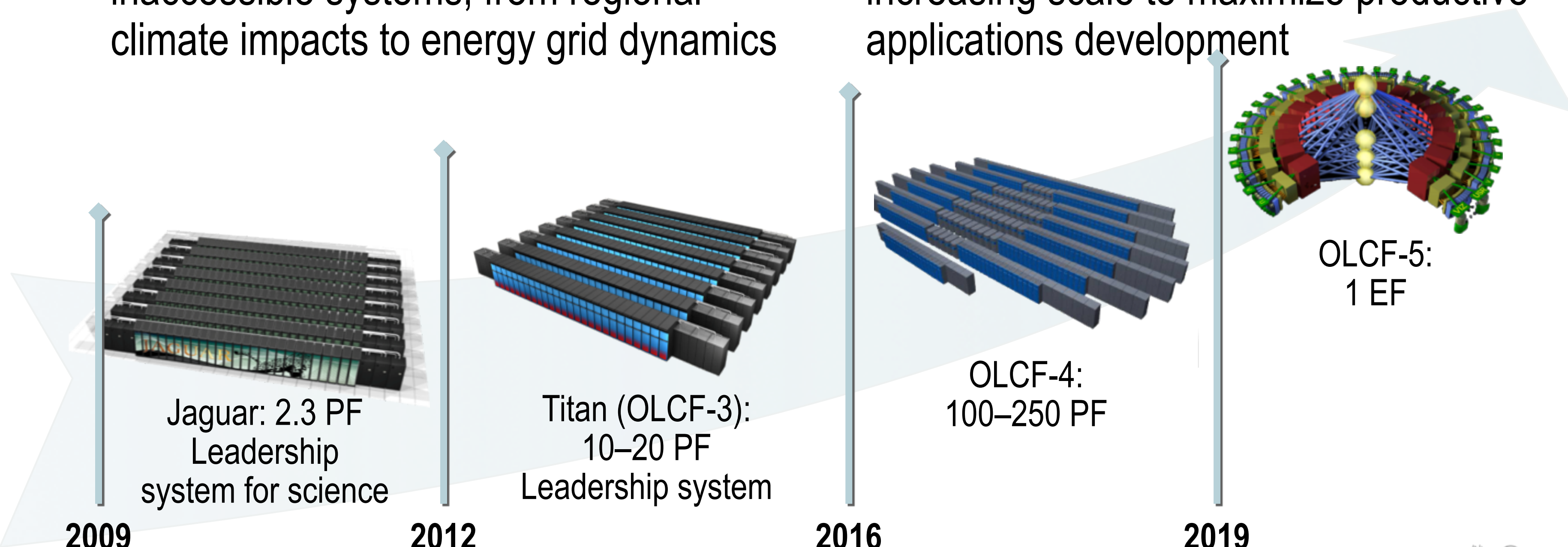
Science requires exascale capability in this decade

Mission: Deploy and operate the computational resources required to tackle global challenges

Vision: Maximize scientific productivity and progress on largest scale computational problems

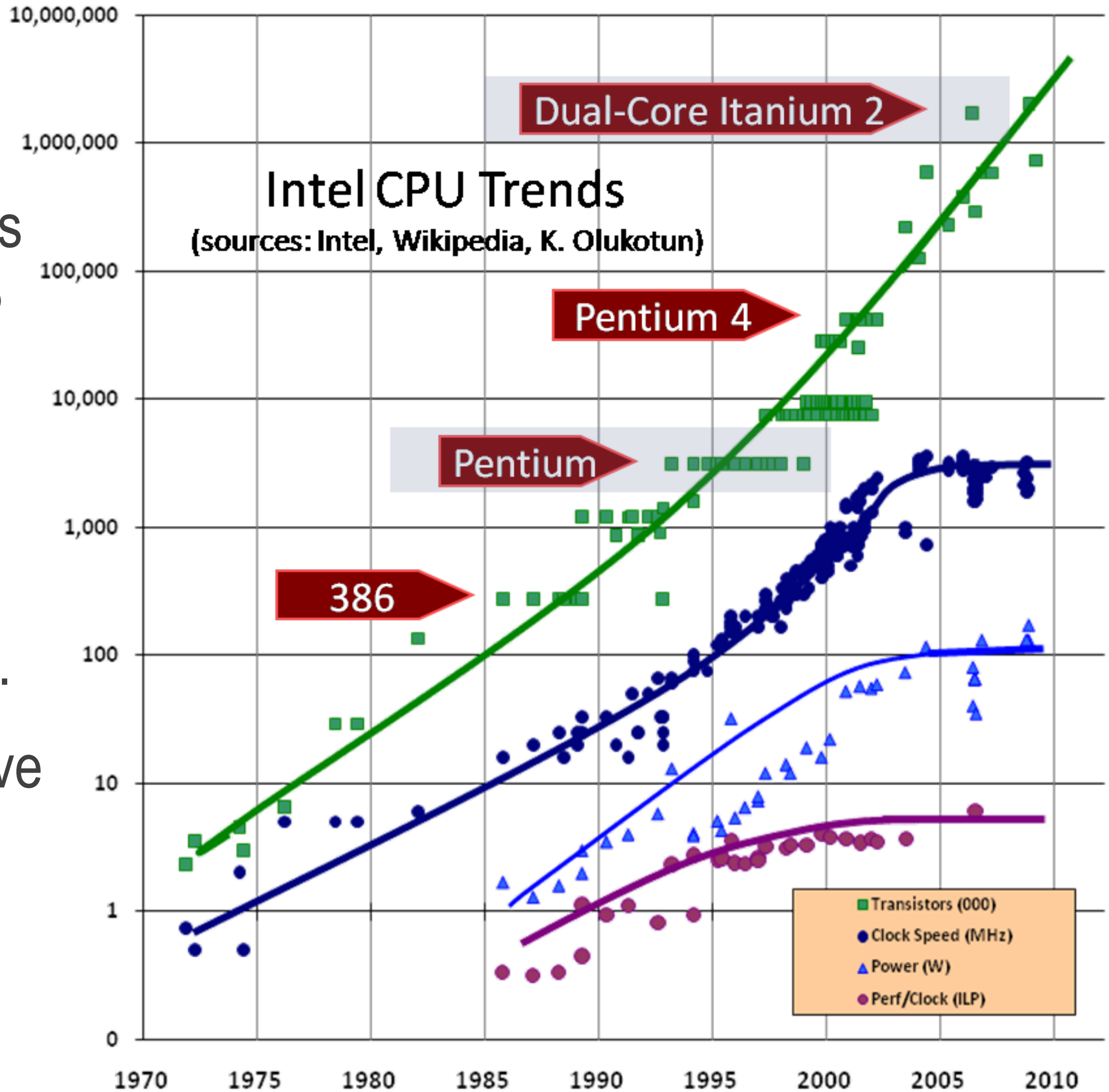
- Deliver transforming discoveries in climate, materials, biology, energy technologies, etc.
- Enabling investigation of otherwise inaccessible systems, from regional climate impacts to energy grid dynamics

- World-class computational resources and specialized services for the most computationally intensive problems
- Stable hardware/software path of increasing scale to maximize productive applications development



Architectural Trends – No more free lunch

- Moore's Law continues (green) but CPU clock rates stopped increasing in 2003 (dark blue) due to power constraints (blue).
- Power is capped by heat dissipation and $$$. Confronting the *power wall*.
- Performance increases have been coming through increased parallelism.



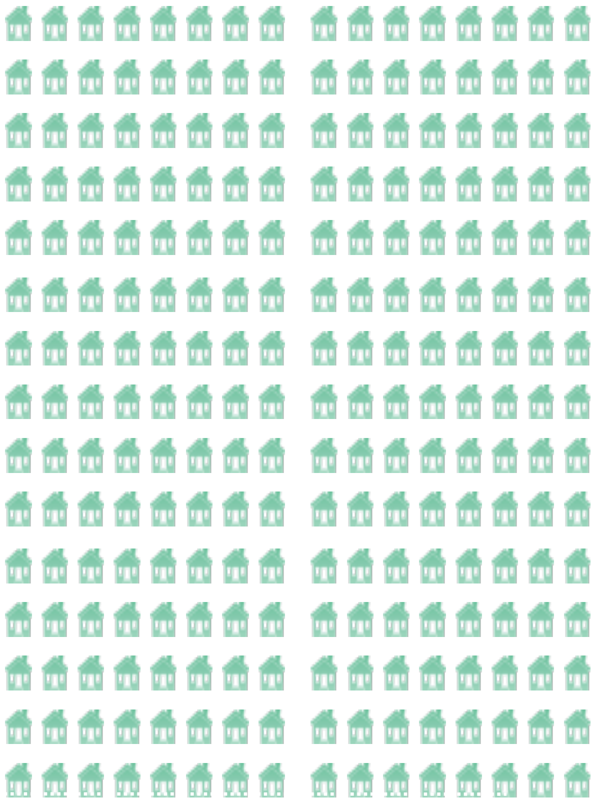
Herb Sutter, Dr. Dobb's Journal:
http://www.gotw.ca/publications/concurrency-ddj.htm

Oak Ridge National Laboratory Argonne National Laboratory

Power is THE problem

Power consumption of 2.3 PF Jaguar:
7 megawatts, equivalent to that of a small city (5,000 homes)

Scaling via traditional CPUs is no longer economically feasible



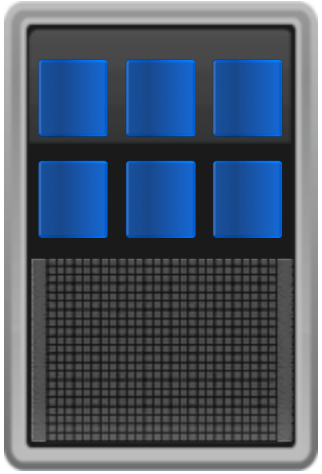
20 PF+ system:
30 megawatts (30,000 homes)

Why GPUs?

High performance & power efficiency towards exascale

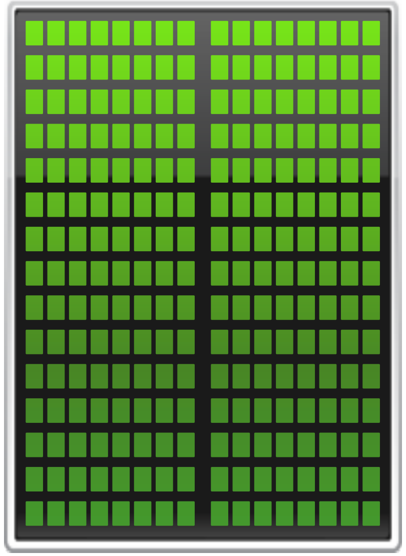
- Hierarchical parallelism improves scalability of applications
- Expose more parallelism through code refactoring and source code directives
  - Doubles performance of many codes
- Heterogeneous multicore processor architecture: Using right type of processor for each task
- Data locality: Keep data near processing
  - GPU has high bandwidth to local memory for rapid access
  - GPU has large internal cache
- Explicit data management: Explicitly manage data movement between CPU and GPU memories

CPU



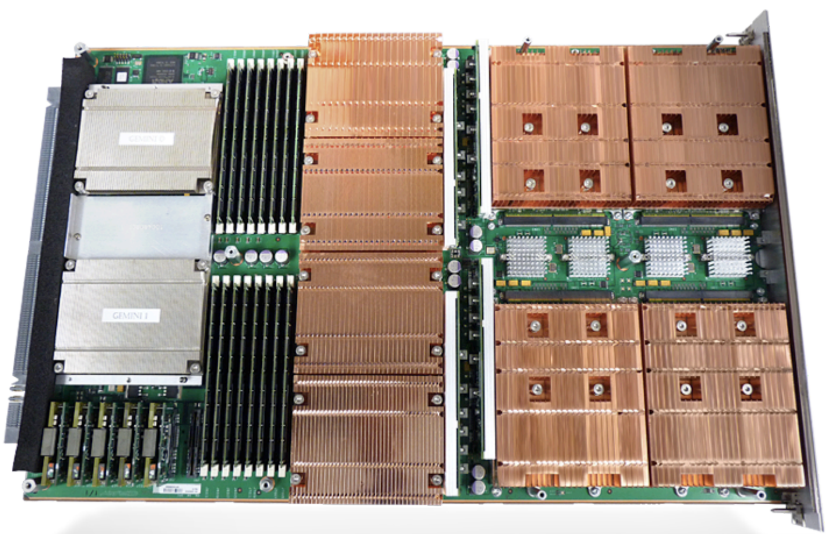
- Optimized for sequential multitasking

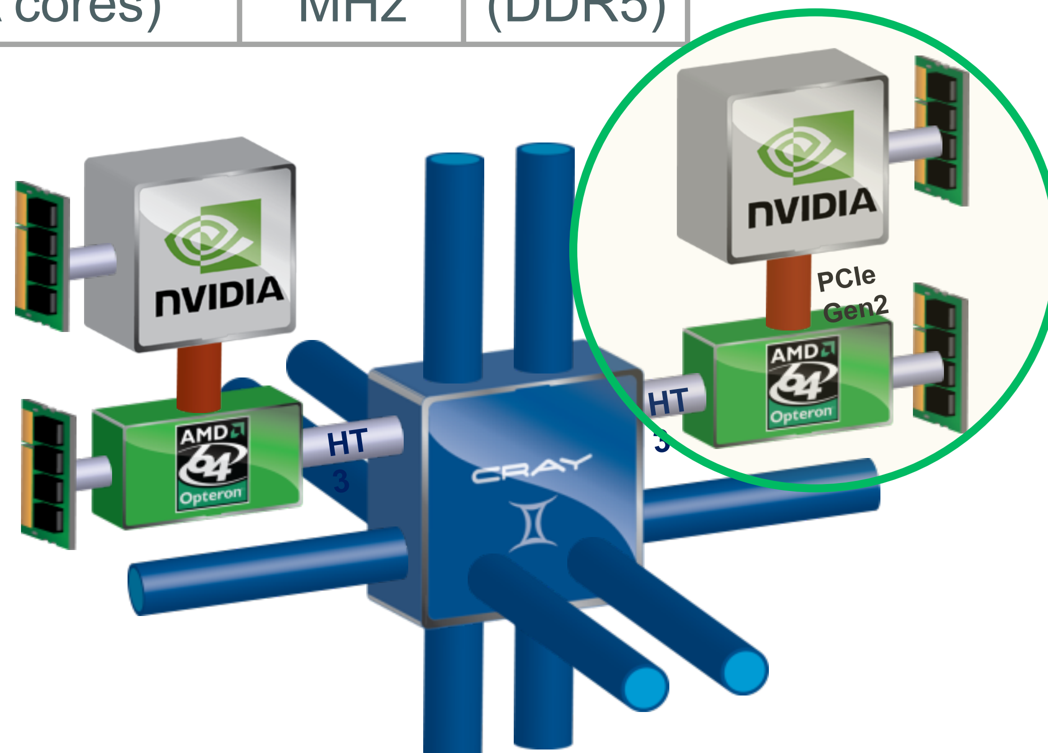
GPU Accelerator

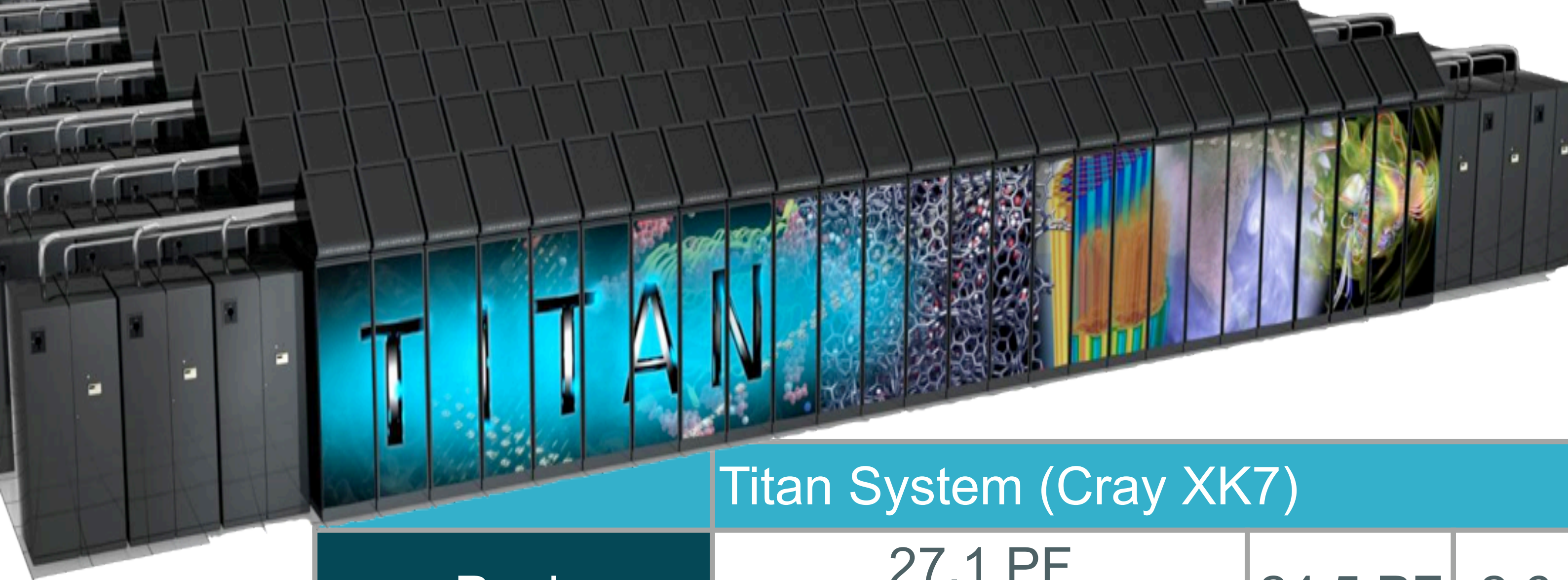


- Optimized for many simultaneous tasks
- 10× performance per socket
- 5× more energy-efficient systems

	Titan Nodes (Cray XK7)		
Node	AMD Opteron 6200 Interlagos (16 cores)	2.2 GHz	32 GB (DDR3)
Accelerator	Tesla K20x (2688 CUDA cores)	732 MHz	6 GB (DDR5)







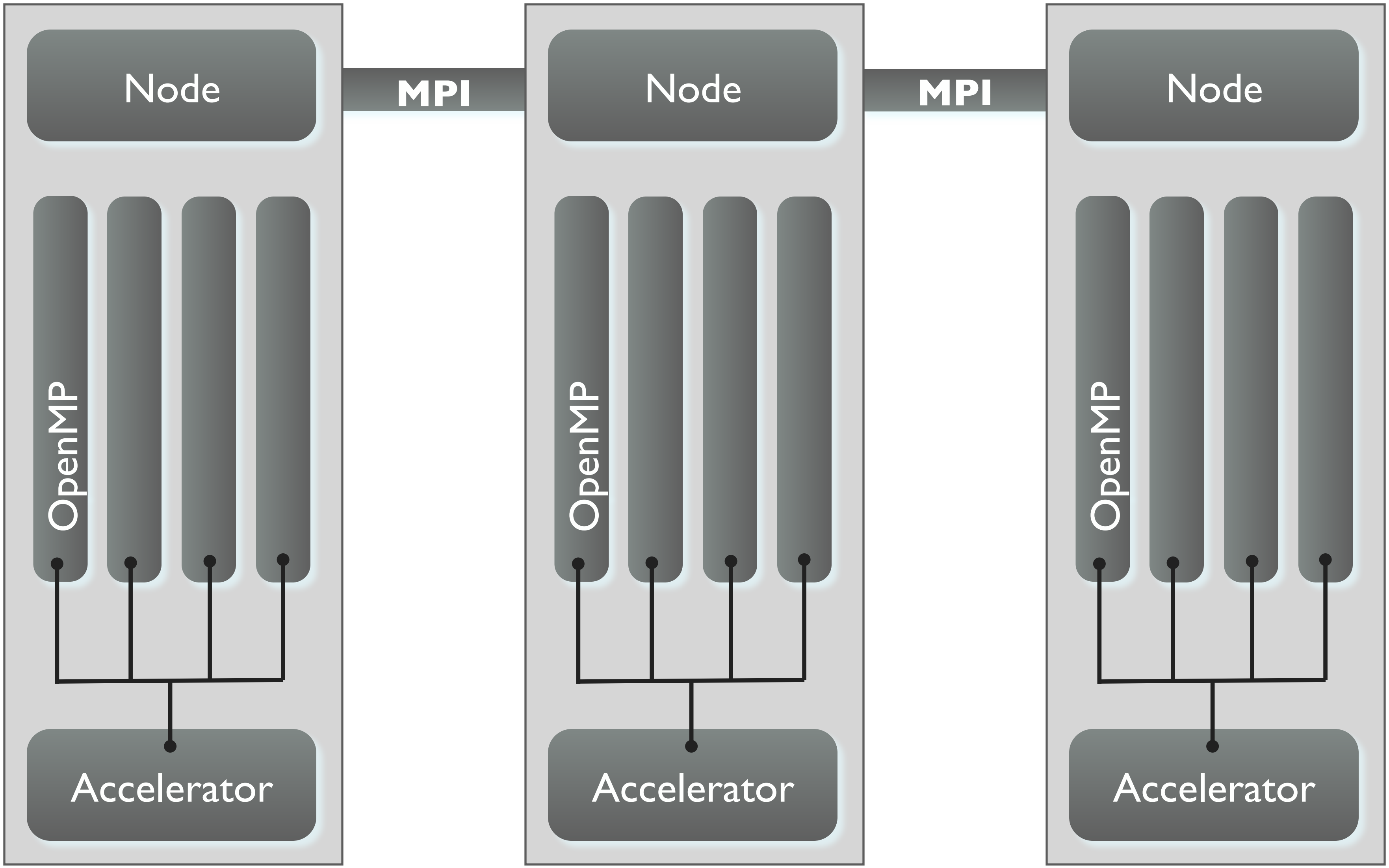
	Titan System (Cray XK7)		
Peak Performance	27.1 PF 18,688 compute nodes	24.5 PF GPU	2.6 PF CPU
System memory	710 TB total memory		
Interconnect	Gemini High Speed Interconnect	3D Torus	
Storage	Lustre Filesystem	32 PB	
Archive	High-Performance Storage System (HPSS)	29 PB	
I/O Nodes	512 Service and I/O nodes		

#1 TOP500 SUPERCOMPUTER SITES

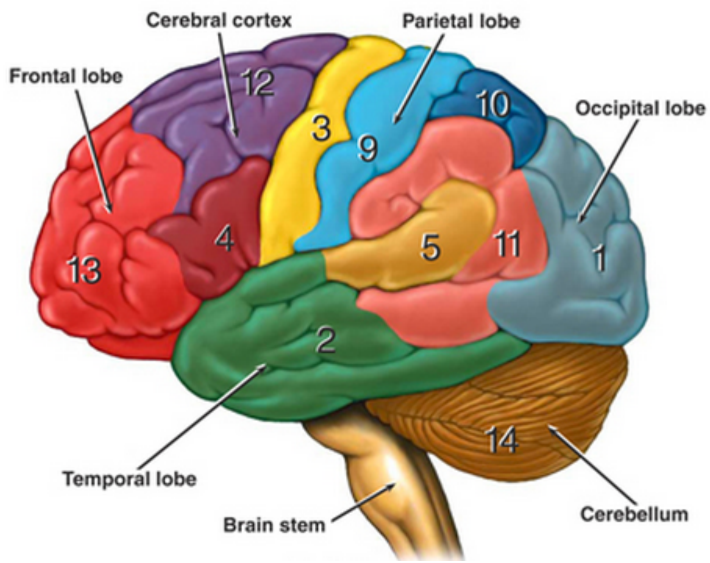
17.59 PF
8.2 Megawatts

THE GREEN 500 #3

Hybrid Programming Model



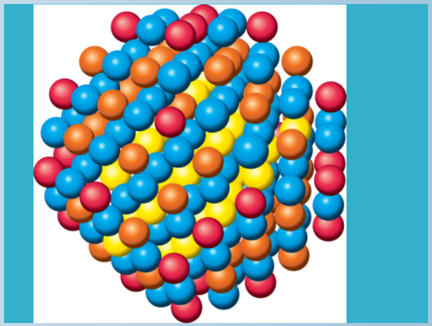
Hybrid Architecture → Scientific Discovery

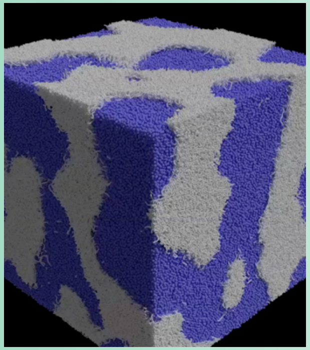


- Evolving hybrid architectures with dedicated centers for specialized tasks surpass solutions from straightforward scaling.
- Significant benefits for those projects that can use it well.
- A growing number of scientific disciplines are benefitting by the resultant speedups.

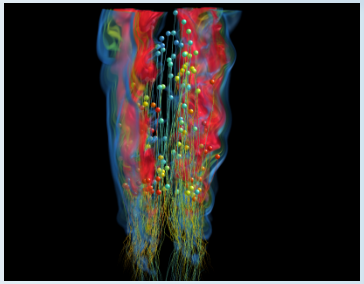
Early Science Challenges for Titan

WL-LSMS
Illuminating the role of material disorder, statistics, and fluctuations in nanoscale materials and systems.



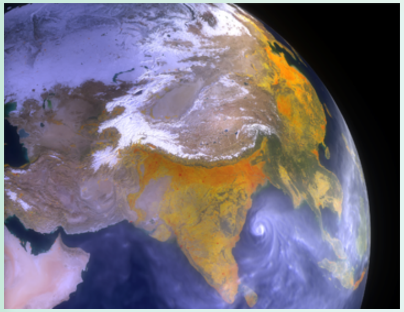


LAMMPS
A molecular dynamics simulation of organic polymers for applications in organic photovoltaic heterojunctions , de-wetting phenomena and biosensor applications

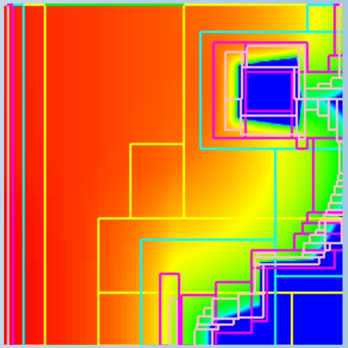


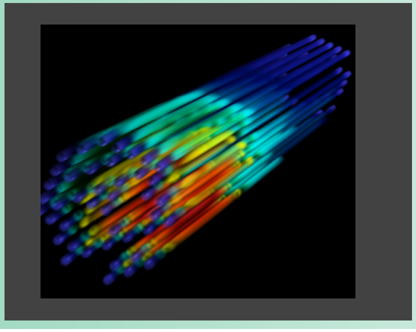
S3D
Understanding turbulent combustion through direct numerical simulation with complex chemistry.
.

CAM-SE
Answering questions about specific climate change adaptation and mitigation scenarios; realistically represent features like precipitation patterns / statistics and tropical storms.



NRDF
Radiation transport – important in astrophysics, laser fusion, combustion, atmospheric dynamics, and medical imaging – computed on AMR grids.





Denovo
Discrete ordinates radiation transport calculations that can be used in a variety of nuclear energy and technology applications.

How Effective are GPUs on Scalable Appications?

OLCF-3 Early Science Codes − *Early* Performance on Titan XK7

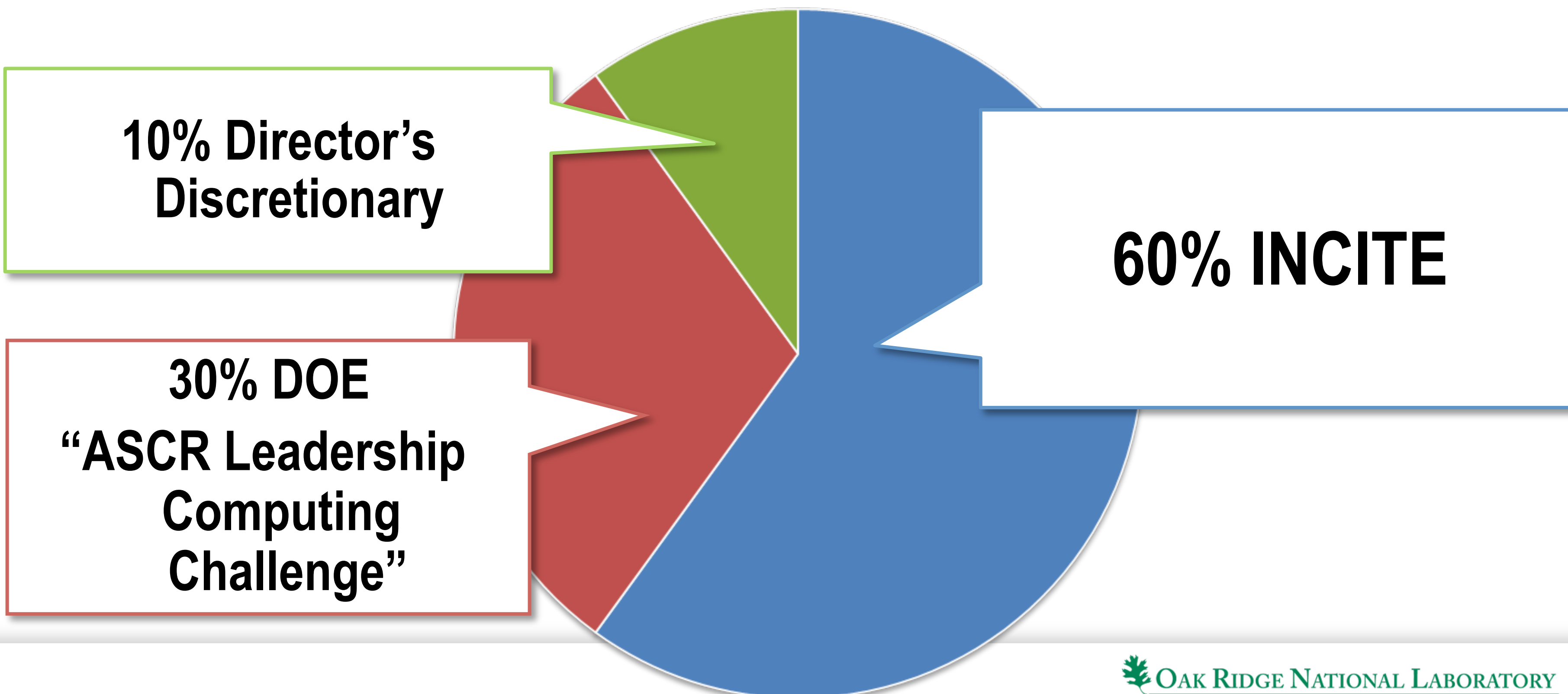
Application	Cray XK7 vs. Cray XE6 Performance Ratio*
LAMMPS* Molecular dynamics	**7.4**
S3D Turbulent combustion	**2**
Denovo 3D neutron transport for nuclear reactors	**3.8**
WL-LSMS Statistical mechanics of magnetic materials	**3.5**

Titan: Cray XK7 (Kepler GPU plus AMD 16-core Opteron CPU)
Cray XE6: (2X AMD 16-core Opteron CPUs)
*Performance depends strongly on specific problem size chosen

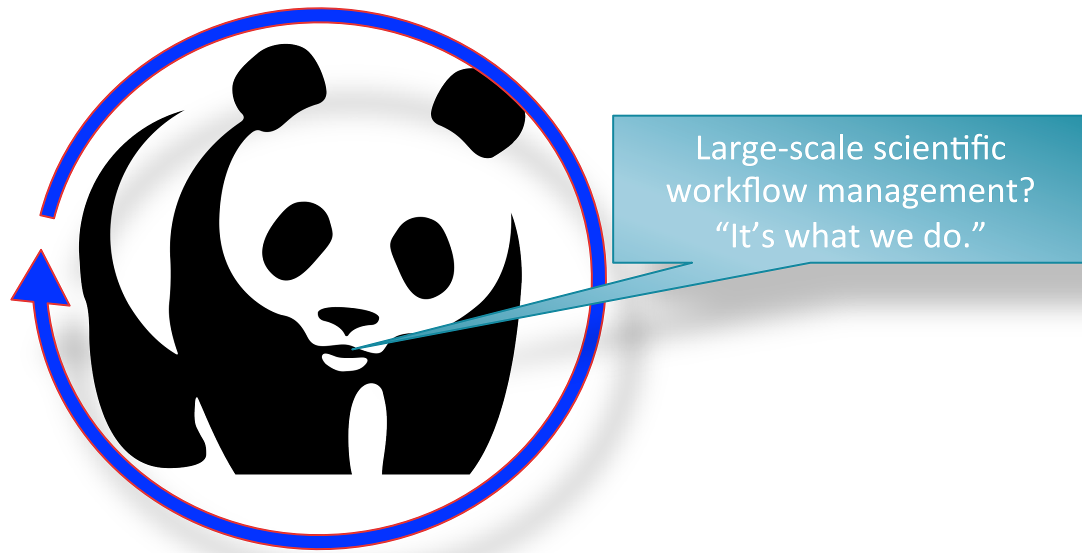
DOE Computational Facilities Allocation Policy for Leadership Facilities

Primary Objective:

- *“Provide substantial allocations to the open science community through a peered process for a small number of high-impact scientific research projects.”*



PanDA Project at the OLCF

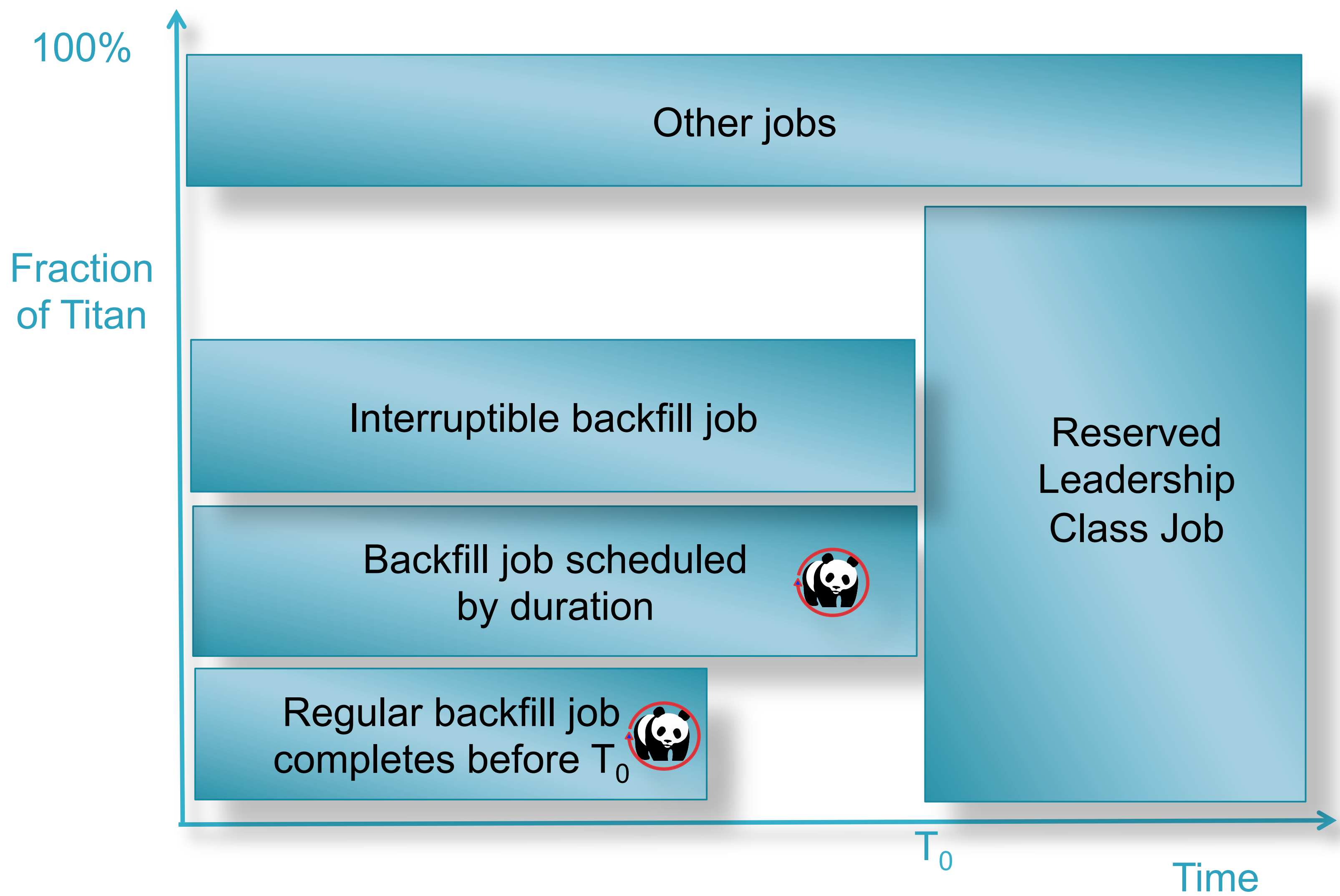


- ATLAS uses sophisticated Production ANd Data Analysis (PanDA) workload management system to optimize data production and availability on the GRID.
- Project underway now testing PanDA agent at OLCF. Pioneers connection of Titan to the LHC/OSG GRID.
- See talks by S. Panitikin, J. Porter, and D. Oleynik.

PanDA Project at the OLCF

- From A. Klimentov's opening presentation at the OLCF "Processing and Analysis of Very Large Data Sets" workshop August 6 – 8, 2013:
  - PanDA processes 1.8 M jobs/day.
  - ATLAS throws away 99.9999% of their data. "We couldn't have found the Higgs without PanDA. It's like searching for one drop of water in the Geneva Jet d'Eau over a 48 hour period."
  - The 150 PB ATLAS managed dataset is larger than the collected written works of all mankind, larger than the Google search space, larger than the YouTube video collection. The only larger re-queried dataset on the planet is . . . Facebook!
- The Titan PanDA project was reviewed August 9 at the Univ. Tennessee by OLCF experts from Cyber Security, operations, and resources.
- Encouraged to move forward, ideally achieving a "Highlight" for reporting to DOE in fall 2013 concerning management of otherwise idle cycles.

Titan Queue Backfill



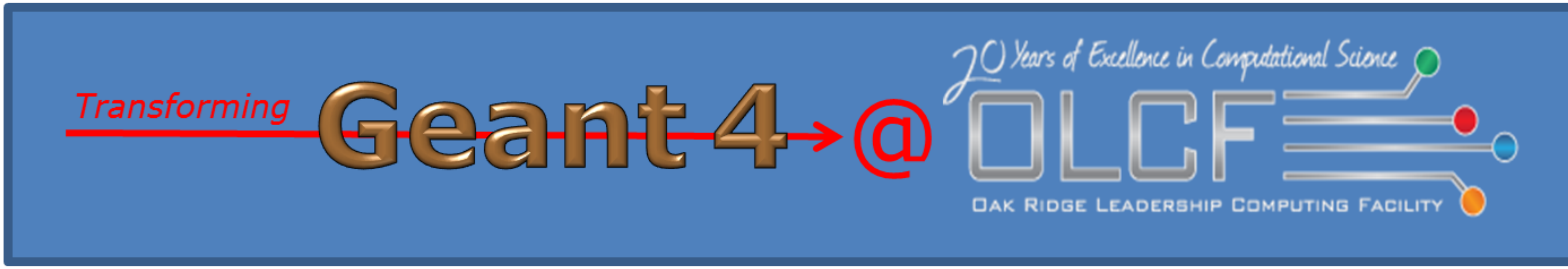
PanDA Backfill

- PanDA has potential to *generate* 300 M Titan hours in 2014 and again in 2015. Estimated to represent between $10 M to $15 M per year worth of computing.
- Jobs are *naturally parallel* (no longer *embarrassed*!) and can be short. Can backfill *arbitrary* amount of Titan queue.
- Exploring whether PanDA pilot can receive more extensive information concerning schedule than available via “qstat”.
- Multiple “payloads” now at varying stages of functionality. “Payloads” need VALIDATION of performance and correctness.
- Testing possibility of remotely-compiled binaries running with associated shared libraries provided by CERN Virtual Machine Filesystem (CVMFS), presumably only appropriate for backfill jobs. This can greatly reduce the effort for validation. Need to check possible performance penalties and scaling.
- Question: ATLAS code validation? Timescale?

Available PanDA Payloads on Titan

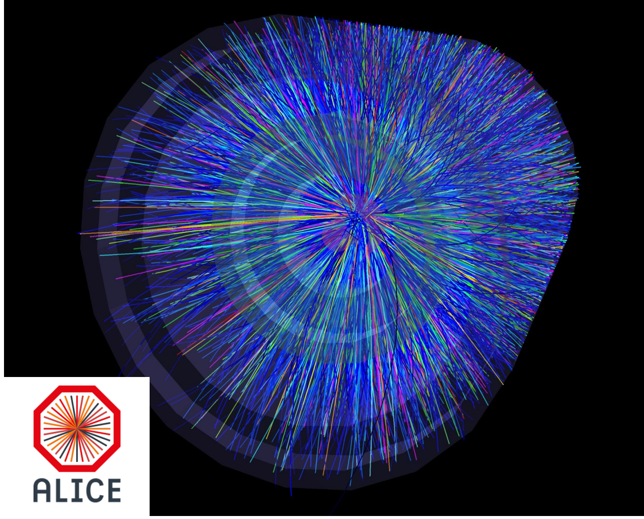
- Multiple shovel-ready HEP and NP codes potentially available for Titan backfill.
- These codes now run on Titan compute nodes:
  - ROOT
  - Geant3, Geant4
  - AliRoot
  - CL-SHASTA (see below)
  - and more…
- CVMFS can deliver code (and binaries) on the external login nodes (using FUSE).
- For the longer term, jobs which use a node fully (multi-threaded with some GPU acceleration) and scale well, may be very competitive at the Leadership Level, especially considering the scientific importance.

Geant4



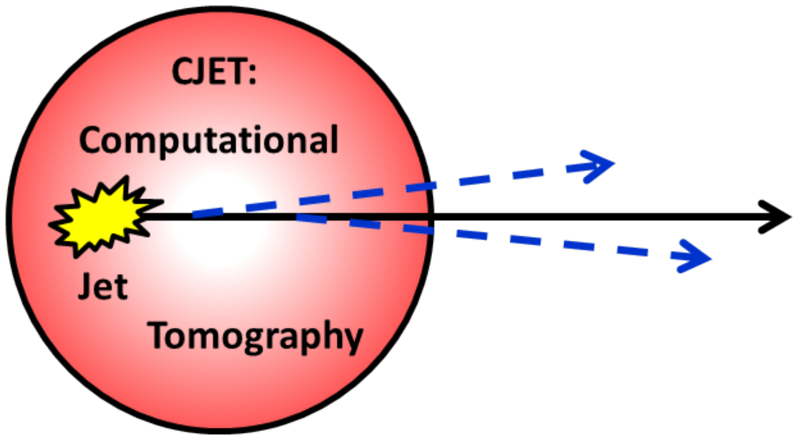
- Simulation of the passage of elementary particles through matter constitutes the greatest computational requirement for
  - ALICE and sPHENIX simulations.
  - ATLAS Higgs simulations.
  - nEDM Experiment (at ORNL SNS) simulations.
  - ORNL SNS 2nd Target Station design.
  - Particle beam radiotherapy simulation validation for cancer treatment.
- Ideal payload since Geant4 now has multi-threaded implementation and a roadmap towards hardware acceleration. Testing underway at several sites with acceleration of Geant4 kernels via GPU and Intel MIC.

AliROOT



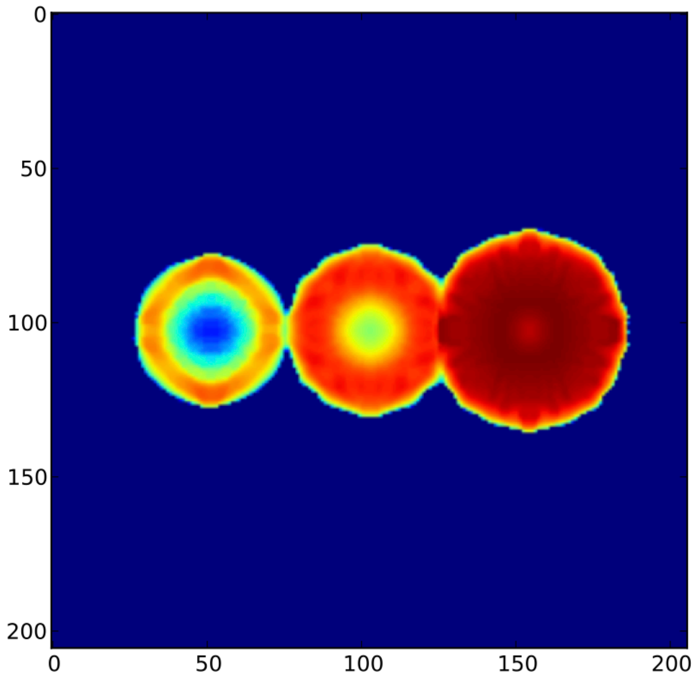
- AliROOT now runs natively on Titan compute nodes. Needed to adjust some scripts and develop a tailored build scheme. Did not need to change source code.
- Available with Geant4 via ROOT Virtual MC.
- Production running requires VALIDATION and DATA PRESERVATION. Evolving code can require on-going validation.
- Validation can be curtailed significantly if (initially?) run externally compiled binaries via CVMFS. CVMFS can be part of a DATA PRESERVATION scheme. Potential performance and scaling penalties to be explored.

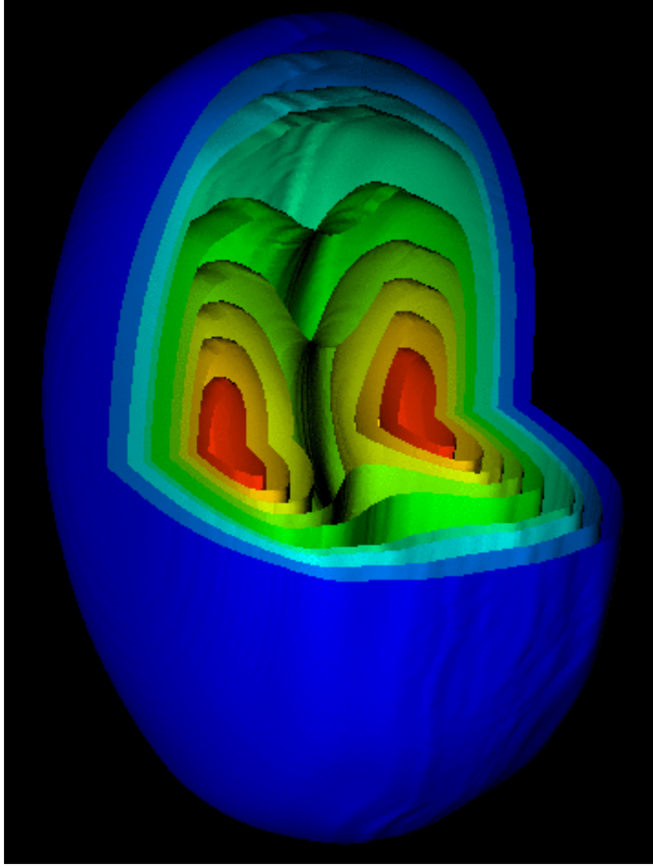
Computational Jet Tomography



- New relativistic hydrodynamics code CL-SHASTA is a complete *re-factorization* of CPU-SHASTA using Open CL and best practices of GPU acceleration.
- 11X speedup in port from FORTRAN77 to C++/OpenCL. Subsequent 14X speedup due to GPU acceleration. An overall improvement of 160X.

Open CL SHASTA





- CL-SHASTA now running on Titan. An exploding fireball (left) from a heavy ion collision with relativistic expansion.
- Relativistic hydrodynamical evolution (right) efficiently computed by CL-SHASTA on a Titan GPU for collision of two (Lorentz-contracted) heavy ions. The passing ions move through each other and separate as lateral expansion and cooling proceeds. Colors indicate contours of constant energy density.

Open CL SHASTA

ELSEVIER

Computer Physics Communications

Volume 184, Issue 2, February 2013, Pages 311–319

Relativistic hydrodynamics on graphic cards

Jochen Gerhard[a, b], Volker Lindenstruth[a, b], Marcus Bleicher[a, c]

[a] Frankfurt Institute for Advanced Studies, Ruth-Moufang-Straße 1, 60438 Frankfurt am Main, Germany

[b] Institut für Informatik, Johann Wolfgang Goethe-Universität, Robert-Mayer-Straße 11–15, 60054 Frankfurt am Main, Germany

[c] Institut für Theoretische Physik, Johann Wolfgang Goethe-Universität, Max-von-Laue-Straße 1, 60438 Frankfurt am Main, Germany

PanDA Outlook

- Today:
  - PanDA and multiple payloads are nearing readiness, some of which are multi-threaded and GPU-accelerated.
  - Testing and optimization of candidate payloads proceeding.
- This year:
  - Prospect of PanDA OLCF "Highlight" later this fall could be of special interest to DOE ASCR.
  - Need to prepare for production testing, code validation, data relocation, and data preservation.
- 2014 and 2015:
  - Potentially, 300 M Titan hours per year could be available.
  - ROOT, Geant4, ATLAS, and ALICE software and computing roadmaps advance further over this period, with increasingly improved node utilization.

Conclusions

- *Leadership computing* is for critically important problems requiring the most powerful compute and data infrastructure. OLCF resources are available to academia and industry through open, peer-reviewed programs.
- Computer system performance increases through parallelism
  - Clock speeds trending flat to slower over coming years
  - Applications *must* utilize all inherent parallelism
- Accelerated, hybrid-multicore computing solutions are performing very well on real, complex scientific applications. Such solutions now appearing on the evolving computing roadmaps for ROOT, GEANT4, ALICE, ATLAS, and CMS.
- For further information
  - https://sites.google.com/site/xgeant4
  - https://sites.google.com/site/cjetsite
  - https://sites.google.com/site/openclshasta
  - http://www.olcf.ornl.gov/

Acknowledgements

- Thanks to J. Wells, OLCF Director of Science, for general OLCF material.
- OLCF-3 Vendor Parters: Cray, AMD, and NVIDIA
- This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.