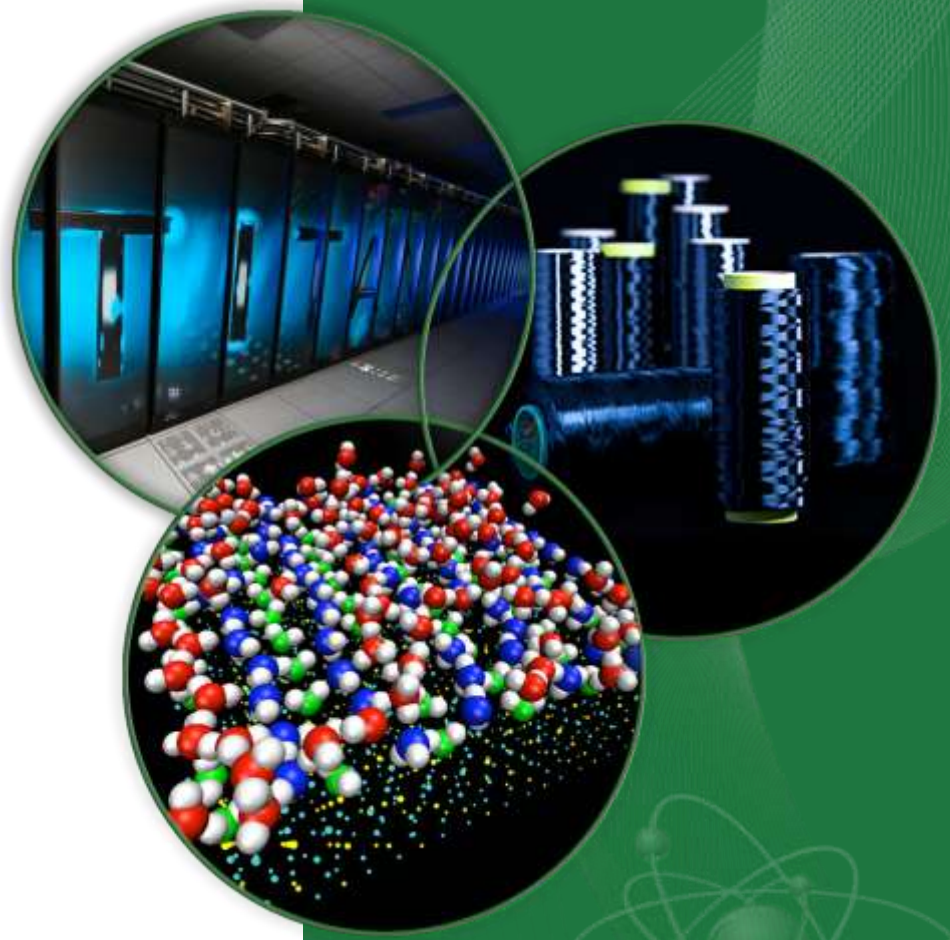
# Oak Ridge Leadership Computing Facility



Jack Wells

Director of Science

Oak Ridge Leadership Computing Facility

Oak Ridge National Laboratory

SciDAC PI Meeting

23 July 2015

Bethesda, MD

ORNL is managed by UT-Battelle
for the US Department of Energy

OAK RIDGE National Laboratory

# (draft) Outline

- Introduction to US DOE Leadership Computing Program
- Allocation programs
- Titan
  - Application Readiness for Titan
  - SciDAC Science on Titan
- CORAL Procurement
- Summit
  - Hardware plans
  - Software plans
  - Application Readiness for Summit

# What is the Leadership Computing Facility (LCF)?

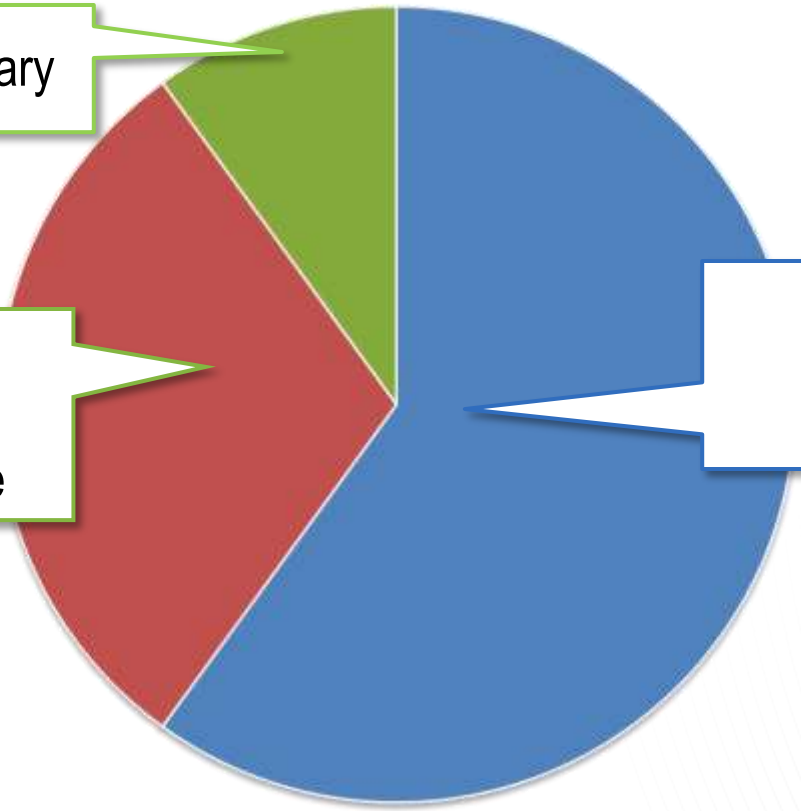
- Collaborative DOE Office of Science user-facility program at ORNL and ANL
- Mission: Provide the computational and data resources required to solve the most challenging problems.
- 2-centers/2-architectures to address diverse and growing computational needs of the scientific community
- Highly competitive user allocation programs (INCITE, ALCC).
- Projects receive 10x to 100x more resource than at other generally available centers.
- LCF centers partner with users to enable science & engineering breakthroughs (Liaisons, Catalysts).



OAK RIDGE National Laboratory
Argonne NATIONAL LABORATORY

# Three primary user programs for access to LCF

## Distribution of allocable hours



10% Director's Discretionary

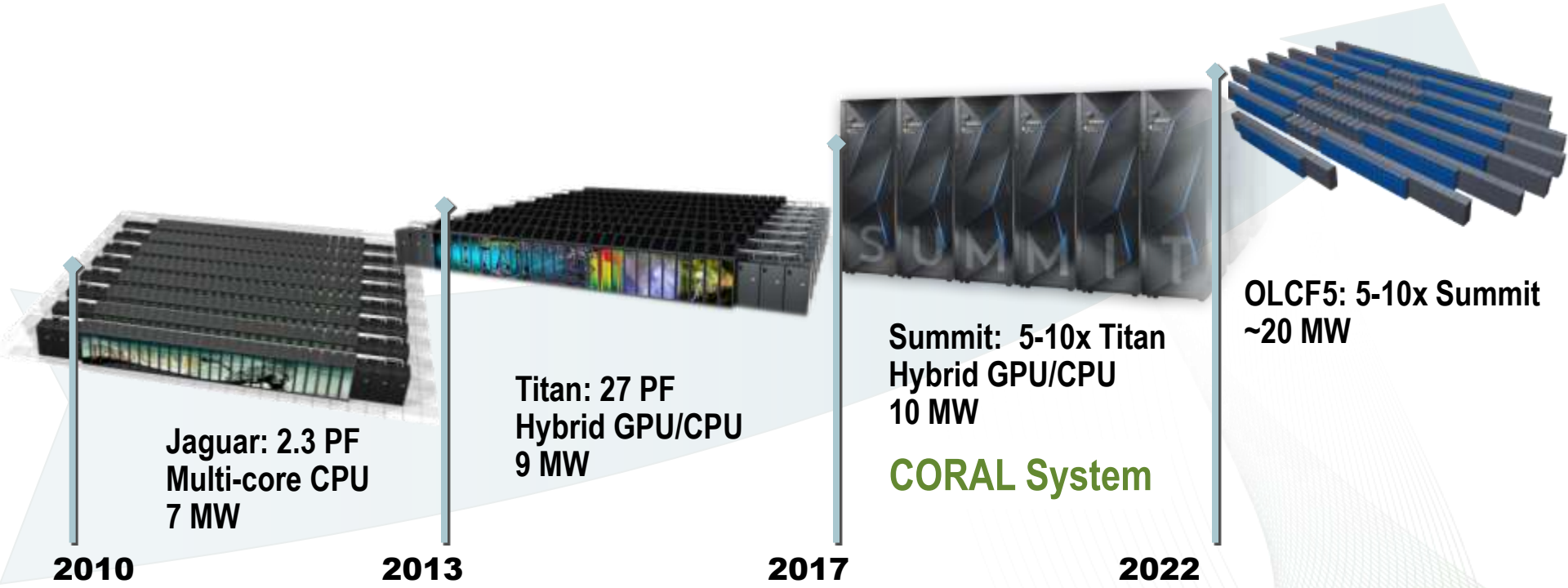
**30% ALCC**

ASCR Leadership Computing Challenge

**60% INCITE**

# Our Science requires that we continue to advance our computational capability over the next decade on the roadmap to Exascale.

| Since clock-rate scaling ended in 2003, HPC performance has been achieved through increased parallelism. Jaguar scaled to 300,000 cores. | Titan and beyond deliver hierarchical parallelism with very powerful nodes. MPI plus thread level parallelism through OpenACC or OpenMP plus vectors |
|---|---|



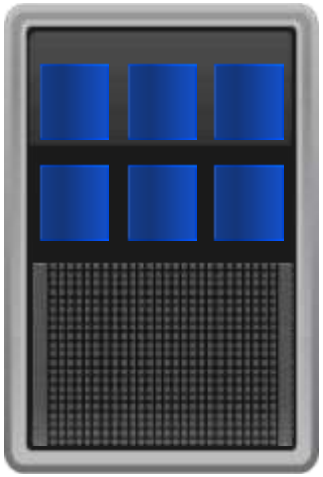



OAK RIDGE National Laboratory

# Why GPUs? Hierarchical Parallelism
## High performance and power efficiency on path to exascale

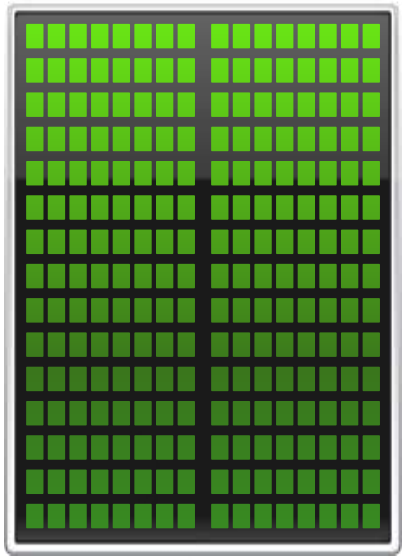
- Expose more parallelism through code refactoring and source code directives
  - Doubles CPU performance of many codes
- Use right type of processor for each task
- Data locality: Keep data near processing
  - GPU has high bandwidth to local memory for rapid access
  - GPU has large internal cache
-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

**CPU**



- Optimized for sequential multitasking

**GPU Accelerator**



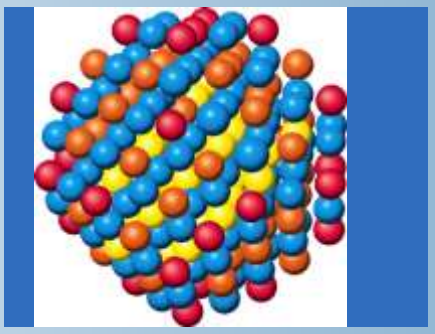
- Optimized for many simultaneous tasks
- 10× performance per socket
- 5× more energy-efficient systems

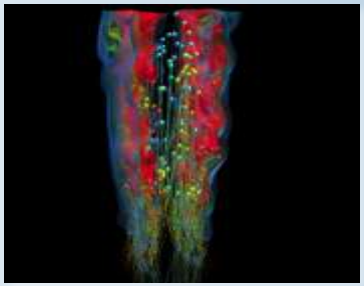
# Center for Accelerated Application Readiness (CAAR)

- We created CAAR as part of the Titan project to help prepare applications for accelerated architectures
- Goals:
  - Work with code teams to develop and implement strategies for exposing hierarchical parallelism for our users applications
  - Maintain code portability across modern architectures
  - Learn from and share our results
- We selected six applications from across different science domains and algorithmic motifs

# Application Readiness for Titan

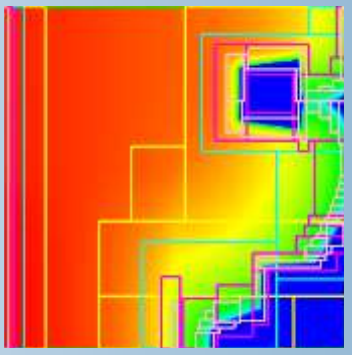
**WL-LSMS**
Illuminating the role of material disorder, statistics, and fluctuations in nanoscale materials and systems.

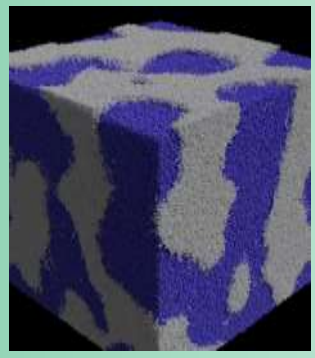




**S3D**
Understanding turbulent combustion through direct numerical simulation with complex chemistry.
.

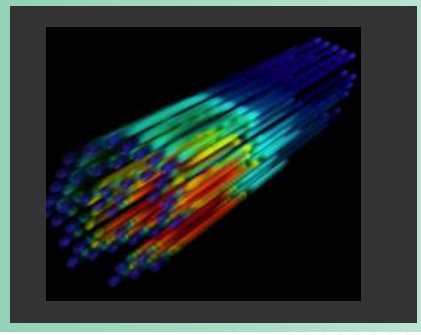
**NRDF**
Radiation transport – important in astrophysics, laser fusion, combustion, atmospheric dynamics, and medical imaging – computed on AMR grids.





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

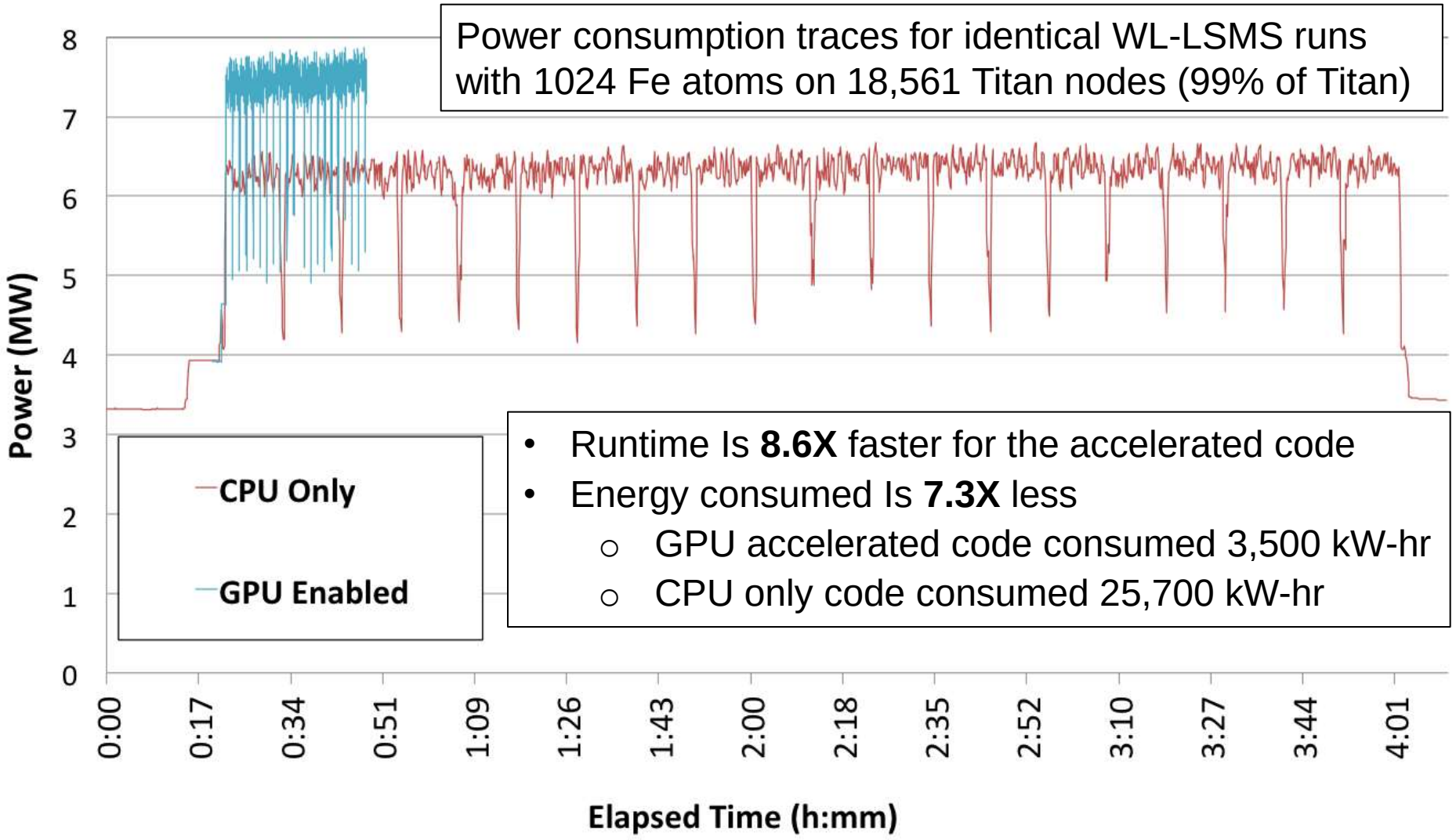
**CAM-SE**
Answering questions about specific climate change adaptation and mitigation scenarios; realistically represent features like precipitation patterns / statistics and tropical storms.



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

# Application Power Efficiency of the Cray XK7

## *WL-LSMS for CPU-only and Accelerated Computing*



# CAAR Lessons Learned

- Up to 1-3 person-years required to port each code
  - Takes work, but an unavoidable step required for exascale
  - Also pays off for other systems—the ported codes often run significantly faster CPU-only (Denovo 2X, CAM-SE >1.7X)
- An estimated 70-80% of developer time is spent in code restructuring, regardless of whether using CUDA, OpenCL, OpenACC, …
- Each code team must make its own choice of using CUDA vs. OpenCL vs. OpenACC, based on the specific case—may be different conclusion for each code
- Science codes are under active development—porting to GPU can be pursuing a "moving target," challenging to manage
- More available flops on the node should lead us to think of new science opportunities enabled—e.g., more DOF per grid cell

# Science Accomplishments Highlights

## *All from 2014 INCITE Program on Titan*

### Cosmology

**Salman Habib**
Argonne National Laboratory

Habib and collaborators used its HACC Code on Titan's CPU–GPU system to conduct today's largest cosmological structure simulation at resolutions needed for modern-day galactic surveys.

K. Heitmann, 2014. *arXiv.org,* 1411.3396

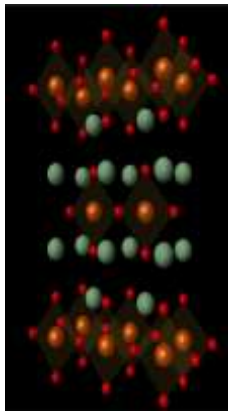
### Combustion

**Jacqueline Chen**
Sandia National Laboratory

Chen and collaborators for the first time performed direct numerical simulation of a jet flame burning dimethyl ether (DME) at new turbulence scales over space and time.

A. Bhagatwala, et al. 2014. *Proc. Combust. Inst.* **35**.

### Superconducting Materials



**Paul Kent**
ORNL

Paul Kent and collaborators performed the first ab initio simulation of a cuprate. They were also the first team to validate quantum Monte Carlo simulations for high-temperature superconductor simulations.

K. Foyevtsova, et al. 2014. *Phys. Rev. X* **4**

### Molecular Science

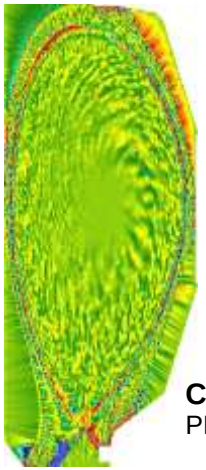
**Michael Klein**
Temple University

Researchers at Procter & Gamble (P&G) and Temple University delivered a comprehensive picture in full atomistic detail of the molecular properties that drive skin barrier disruption.

M. Paloncyova, et al. 2014. *Langmuir* **30**

C. M. MacDermaid, et al. 2014. *J. Chem. Phys.* **141**

### Fusion



**C.S. Chang**
PPPL

Chang and collaborators used the XGC1 code on Titan to obtain fundamental understanding of the divertor heat-load width physics and its dependence on the plasma current in present-day tokamak devices.

C. S. Chang, et al. 2014. *Proceedings of the 25th Fusion Energy Conference, IAEA*, October 13–18, 2014.

# Where do we go from here?


**Summit: 5-10x Titan**
**Hybrid GPU/CPU**
**10 MW**
**CORAL System**

- Provide the Leadership computing capabilities needed for the DOE Office of Science mission from 2018 through 2022
  - Capabilities for INCITE and ALCC science projects

- CORAL (Consortium of Oak Ridge, Argonne, and Lawrence Livermore) was formed by grouping the three Labs who would be acquiring Leadership computers in the same timeframe (2017).
  - Benefits include:
    - Shared technical expertise
    - Decreases risks due to the broader experiences, and broader range of expertise of the collaboration
    - Lower collective cost for developing and responding to RFP

# CORAL Collaboration ORNL, ANL, LLNL)

**Objective -** Procure 3 leadership computers to be sited at Argonne, Oak Ridge and Lawrence Livermore in 2017. Two of the contracts have been awarded with the Argonne contract in process.

***Current DOE Leadership Computers***

**Titan (ORNL) 2012 - 2017**

**Sequoia (LLNL) 2012 - 2017**

**Mira (ANL) 2012 - 2017**

**Leadership Computers** RFP requests >100 PF, 2 GB/core main memory, local NVRAM, and science performance 4x-8x Titan or Sequoia

**Approach**

- Competitive process - one RFP (issued by LLNL) leading to 2 R&D contracts and 3 computer procurement contracts
- For risk reduction and to meet a broad set of requirements, 2 architectural paths will be selected and Oak Ridge and Argonne must choose different architectures
- Once Selected, Multi-year Lab-Awardee relationship to co-design computers
- Both R&D contracts jointly managed by the 3 Labs
- Each lab manages and negotiates its own computer procurement contract, and may exercise options to meet their specific needs
- Understanding that long procurement lead-time may impact architectural characteristics and designs of procured computers

OAK RIDGE National Laboratory | Argonne NATIONAL LABORATORY | Lawrence Livermore National Laboratory

# Two Architecture Paths for Today and Future Leadership Systems

**Power concerns for large supercomputers are driving the largest systems to either Hybrid or Many-core architectures**

**Hybrid Multi-Core (like Titan)**

- CPU / GPU hybrid systems
- Likely to have multiple CPUs and GPUs per node
- Small number of very powerful nodes
- Expect data movement issues to be much easier than previous systems – coherent shared memory within a node
- Multiple levels of memory – on package, DDR, and non-volatile

**Many Core (like Sequoia/Mira)**

- 10's of thousands of nodes with millions of cores
- Homogeneous cores
- Multiple levels of memory – on package, DDR, and non-volatile
- Unlike prior generations, future products are likely to be self hosted

# 2017 OLCF Leadership System

## Hybrid CPU/GPU architecture

SUMMIT
OAK RIDGE NATIONAL LABORATORY

**Vendor: IBM (Prime) / NVIDIA™ / Mellanox Technologies®**

OpenPOWER

**At least 5X Titan's Application Performance**

**Approximately 3,400 nodes, each with:**

- Multiple IBM POWER9 CPUs and multiple NVIDIA Tesla® GPUs using the NVIDIA Volta architecture
- CPUs and GPUs completely connected with high speed NVLink
- Large coherent memory: over 512 GB (HBM + DDR4)
  - all directly addressable from the CPUs and GPUs
- An additional 800 GB of NVRAM, which can be configured as either a burst buffer or as extended memory
- Over 40 TF peak performance

**Dual-rail Mellanox® EDR-IB full, non-blocking fat-tree interconnect**

**IBM Elastic Storage (GPFS™) - 1TB/s I/O and 120 PB disk capacity.**

# Summit Key Software Components

- **System**
  - Linux®
  - IBM Elastic Storage (GPFS™)
  - IBM Platform Computing™ (LSF)
  - IBM Platform Cluster Manager™ (xCAT)
- **Programming Environment**
  - Compilers supporting OpenMP and OpenACC
    - IBM XL, PGI, LLVM, GNU, NVIDIA
  - Libraries
    - IBM Engineering and Scientific Subroutine Library (ESSL)
    - FFTW, ScaLAPACK, PETSc, Trilinos, BLAS-1,-2,-3, NVBLAS
    - cuFFT, cuSPARSE, cuRAND, NPP, Thrust
  - Debugging
    - Allinea DDT, IBM Parallel Environment Runtime Edition (pdb)
    - Cuda-gdb, Cuda-memcheck, valgrind, memcheck, helgrind, stacktrace
  - Profiling
    - IBM Parallel Environment Developer Edition (HPC Toolkit)
    - VAMPIR, Tau, Open|Speedshop, nvprof, gprof, Rice HPCToolkit

# How does Summit compare to Titan?

| Feature | Summit | Titan |
|---|---|---|
| Application Performance | 5-10x Titan | Baseline |
| Number of Nodes | ~3,400 | 18,688 |
| Node performance | > 40 TF | 1.4 TF |
| Memory per Node | >512 GB (HBM + DDR4) | 38GB (GDDR5+DDR3) |
| NVRAM per Node | 800 GB | 0 |
| Node Interconnect | NVLink (5-12x PCIe 3) | PCIe 2 |
| System Interconnect (node injection bandwidth) | Dual Rail EDR-IB (23 GB/s) | Gemini (6.4 GB/s) |
| Interconnect Topology | Non-blocking Fat Tree | 3D Torus |
| Processors | IBM POWER9<br>NVIDIA Volta™ | AMD Opteron™<br>NVIDIA Kepler™ |
| File System | 120 PB, 1 TB/s, GPFS™ | 32 PB, 1 TB/s, Lustre® |
| Peak power consumption | 10 MW | 9 MW |

# ASCR Computing Upgrades At a Glance

| System attributes | NERSC Now | OLCF Now | ALCF Now | NERSC Upgrade | OLCF Upgrade | ALCF Upgrades | |
|---|---|---|---|---|---|---|---|
| Name<br>Planned Installation | **Edison** | **TITAN** | **MIRA** | **Cori<br>2016** | **Summit<br>2017-2018** | **Theta<br>2016** | **Aurora<br>2018-2019** |
| System peak (PF) | 2.6 | 27 | 10 | > 30 | 150 | >8.5 | 180 |
| Peak Power (MW) | 2 | 9 | 4.8 | < 3.7 | 10 | 1.7 | 13 |
| Total system memory | 357 TB | 710TB | 768TB | ~1 PB DDR4 + High Bandwidth Memory (HBM)+1.5PB persistent memory | > 1.74 PB DDR4 + HBM + 2.8 PB persistent memory | >480 TB DDR4 + High Bandwidth Memory (HBM) | > 7 PB High Bandwidth On-Package Memory Local Memory and Persistent Memory |
| Node performance (TF) | 0.460 | 1.452 | 0.204 | > 3 | > 40 | > 3 | > 17 times Mira |
| Node processors | Intel Ivy Bridge | AMD Opteron Nvidia Kepler | 64-bit PowerPC A2 | Intel Knights Landing many core CPUs Intel Haswell CPU in data partition | Multiple IBM Power9 CPUs & multiple Nvidia Voltas GPUS | Intel Knights Landing Xeon Phi many core CPUs | Knights Hill Xeon Phi many core CPUs |
| System size (nodes) | 5,600 nodes | 18,688 nodes | 49,152 | 9,300 nodes 1,900 nodes in data partition | ~3,500 nodes | >2,500 nodes | >50,000 nodes |
| System Interconnect | Aries | Gemini | 5D Torus | Aries | Dual Rail EDR-IB | Aries | 2nd Generation Intel Omni-Path Architecture |
| File System | 7.6 PB 168 GB/s, Lustre® | 32 PB 1 TB/s, Lustre® | 26 PB 300 GB/s GPFS™ | 28 PB 744 GB/s Lustre® | 120 PB 1 TB/s GPFS™ | 10PB, 210 GB/s Lustre initial | 150 PB 1 TB/s Lustre® |

# Center for Accelerated Application Readiness: Summit

**OLCF-4 issued a call for proposals in FY2015 for application development partnerships between community developers, OLCF staff and the OLCF Vendor Center of Excellence.**

Center for Accelerated Application Readiness (CAAR)

- Performance analysis of community applications
- Technical plan for code restructuring and optimization
- Deployment on Summit

# CAAR Objectives

## Porting and optimizing applications for OLCF's next architectures

- Support current and new applications on future systems
- User-ready applications when Summit goes into production
- Develop applications in diverse set of science domains to expand the user programs
- Where feasible develop architecture and performance portable applications

## Development experience to support future users and developers

- Focus on a variety of programming modules, languages, etc.
- Focus on diverse mathematical models

## Software development environment testing

- Development environments for new systems are often not robust

## Hardware testing with production science runs at scale

- Identifying hardware stability issues is best done with runs at scale

# CAAR in Preparation of Summit

**Application Developer Team involvement**

- Knowledge of the application
- Work on application in development “moving target”
- Optimizations included in application release

**Vendor technical support through the IBM/NVIDIA Center of Excellence is crucial**

- Programming environment often not mature
- Best source of information on new hardware features

**Early Science Project**

- Demonstration of application on real problems at scale
- Shake-down on the new system hardware and software
- Large-scale science project is strong incentive to participate

**Access to multiple resources, including early hardware**

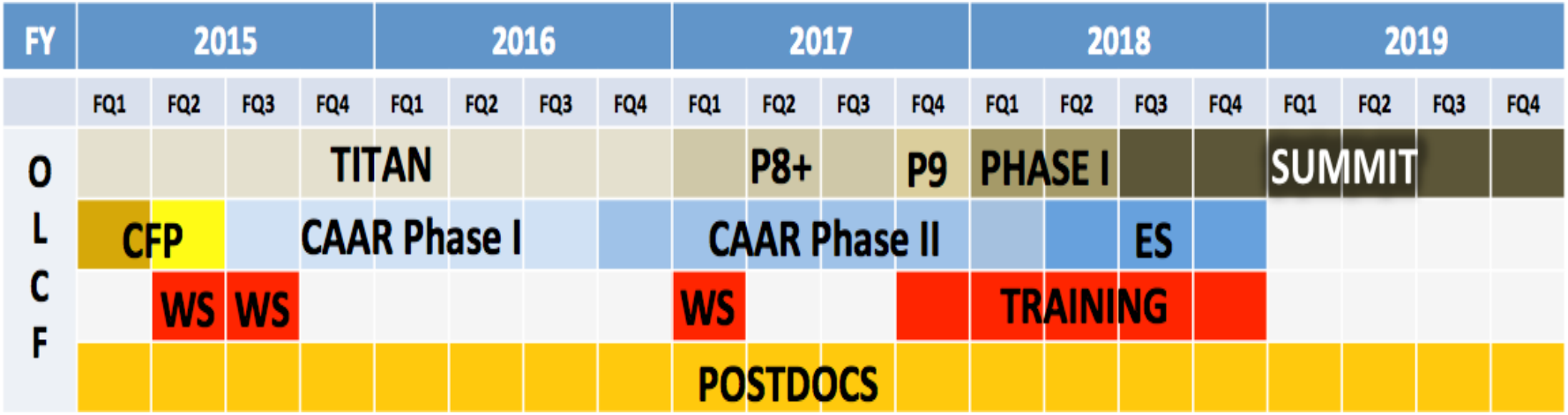
**Joint training activities**

**Portability is a critical concern**

# New CAAR Applications

| Application | Domain | Principal Investigator | Institution |
|---|---|---|---|
| ACME (N) | *Climate Science* | David Bader | Lawrence Livermore National Laboratory |
| DIRAC | *Relativistic Chemistry* | Lucas Visscher | Free University of Amsterdam |
| FLASH | *Astrophysics* | Bronson Messer | Oak Ridge National Laboratory |
| GTC (NE) | *Plasma Physics* | Zhihong Lin | University of California – Irvine |
| HACC(N) | *Cosmology* | Salman Habib | Argonne National Laboratory |
| LSDALTON | *Chemistry* | Poul Jørgensen | Aarhus University |
| NAMD (NE) | *Biophysics* | Klaus Schulten | University of Illinois – Urbana Champaign |
| NUCCOR | *Nuclear Physics* | Gaute Hagen | Oak Ridge National Laboratory |
| NWCHEM (N) | *Chemistry* | Karol Kowalski | Pacific Northwest National Laboratory |
| QMCPACK | *Materials Science* | Paul Kent | Oak Ridge National Laboratory |
| RAPTOR | *Engineering* | Joseph Oefelein | Sandia National Laboratory |
| SPECFEM | *Seismic Science* | Jeroen Tromp | Princeton University |
| XGC (N) | *Plasma Physics* | CS Chang | Princeton Plasma Physics Laboratory |

# CAAR Timeline




1. November 2014: Call for CAAR applications
2. February 20, 2015: CAAR proposal deadline
3. March 2015: Selection of CAAR application teams
4. **April 2015: CAAR application training workshop**
5. April 2015: CAAR application teams start
6. June 2016: CAAR project review
7. October 2017: Call for Early Science projects
8. November 2017: Selection Early Science projects
9. January 2018: Early Science projects start
10. October 2018: Early Science project ends

# Architecture and Performance Portability

**Application portability among NERSC, ALCF and OLCF architectures is critical concern of ASCR**

- Application developers target wide range of architectures
- Maintaining multiple code version is difficult
- Porting to different architectures is time-consuming
- Many Principal Investigators have allocations on multiple resources
- Applications far outlive any computer system

**Improve data locality and thread parallelism**

- GPU or many-core optimizations improve performance on all architectures
- Exposed fine grain parallelism transitions more easily between architectures
- Data locality optimized code design also improves portability

**Use portable libraries**

- Library developers deal with portability challenges
- Many libraries are DOE supported

**MPI+OpenMP 4.0 could emerge as common programming model**

- Significant work is still necessary
- All ASCR centers are on the OpenMP standards committee

**Encourage portable and flexible software development**

- Use open and portable programming models
- Avoid architecture specific models such as Intel TBB, NVIDIA CUDA
- Use good coding practices: parameterized threading, flexible data structure allocation, task load balancing, etc.

# Acknowledgements

OLCF Users

Titan Vendor Partners:  Cray, AMD, NVIDIA

Summit Vendor Partners:  IBM, NVIDIA, Mellanox

*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.*

# Questions?

WellsJC@ornl.gov