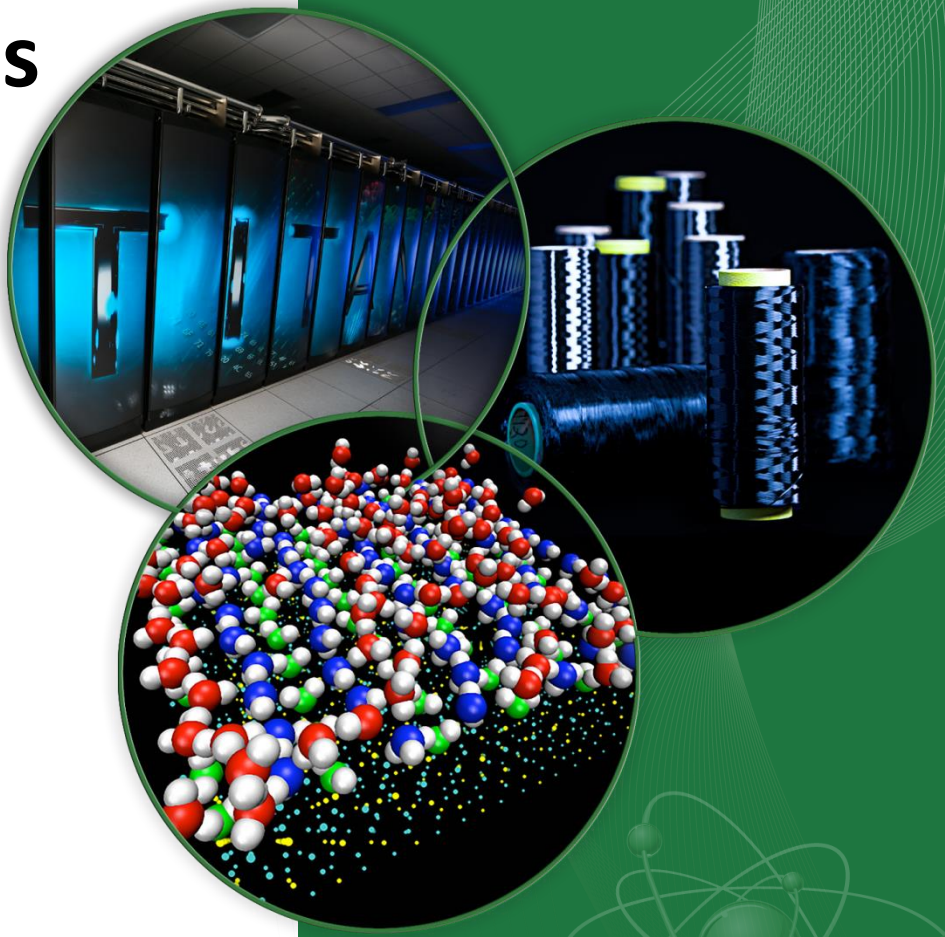
Center for Accelerated Application Readiness

Preparing Applications for Summit



Tjerk Straatsma

OLCF Scientific Computing Group

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

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

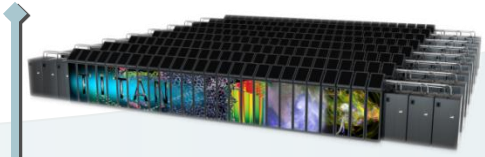
OLCF on the Road to Exascale

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

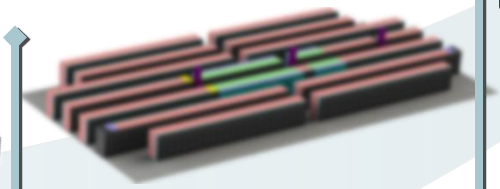
Jaguar: 2.3 PF
Multi-core CPU
7 MW

2010



Titan: 27 PF
Hybrid GPU/CPU
9 MW

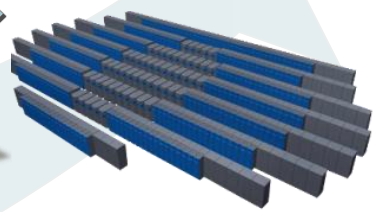
2012



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

CORAL System

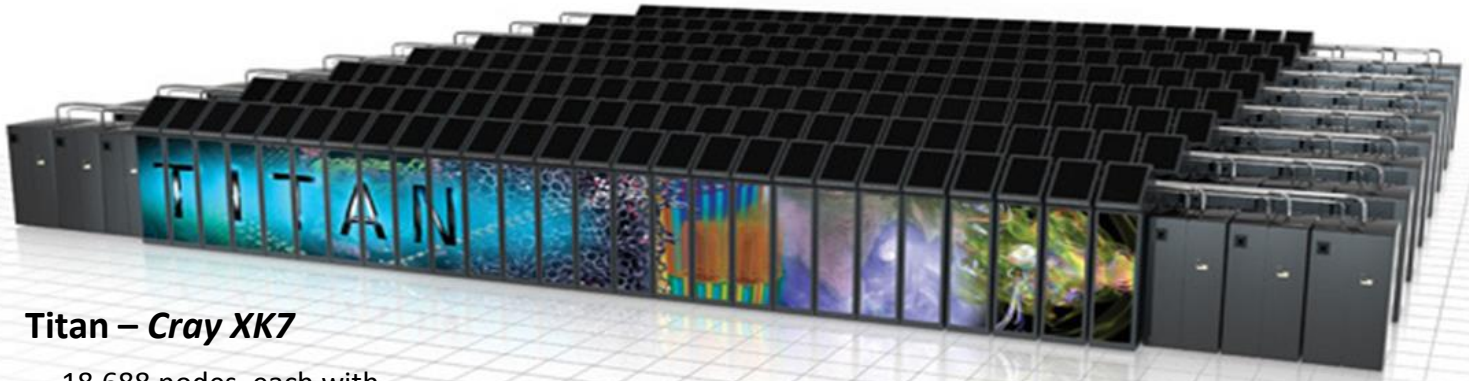
2017



OLCF5: 5-10x Summit
~20 MW

2022

OLCF Current Systems



Titan – *Cray XK7*

- 18,688 nodes, each with
  - AMD™ Opteron™- 141 GF, 32 GB DDR3 memory
  - NVIDIA™ Kepler™ K20X GPU - 1,311 GF, 6 GB GDDR5 memory
  - PCIe2 link between GPU and CPU
- Cray Gemini 3-D Torus Interconnect
- 688 TB of memory
- Peak flop rate: 27 PF

Eos – *Cray XC30*

- 744 nodes, Intel Xeon E5-2670
- 48 TB of memory
- 248 TF

Rhea

- Pre- and post-processing cluster
- 512 nodes, dual 8c Xeon, 64 GB

EVEREST– Visualization Laboratory

- Stereoscopic 6x3 1920x1080 Display Wall, 30.5’ x 8.5’
- Planar 4x4 1920x1080 Display Wall
- Distributed memory Linux cluster

Storage – Spider Lustre® filesystem: 40 PB, >1 TB/s BW; HPSS archival mass storage: 240PB, 6 tape libraries

OLCF Next System: Summit

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

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.

OLCF 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

Summit compared 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 NVIDIA Volta™	AMD Opteron™ NVIDIA Kepler™
File System	120 PB, 1 TB/s, GPFS™	32 PB, 1 TB/s, Lustre®
Peak power consumption	10 MW	9 MW

Center for Accelerated Application Readiness

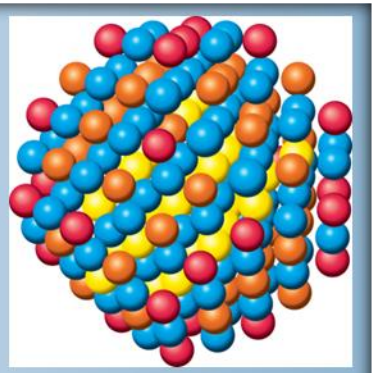
Main Goals:

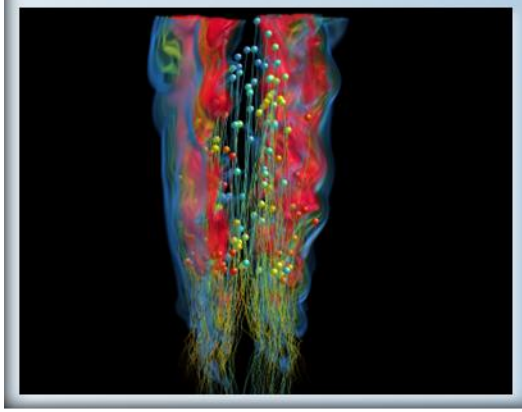
- **Porting and optimizing applications for OLCF's next architectures**
  - Support current applications on future systems
  - Develop applications in diverse set of science domains to expand the user programs
- **Development experience to support future users and developers**
  - Focus on a variety 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 for Titan

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

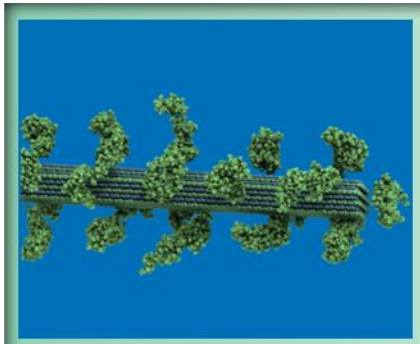
3.8 performance ratio XK7/XE6





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

2.2 performance ratio XK7/XE6

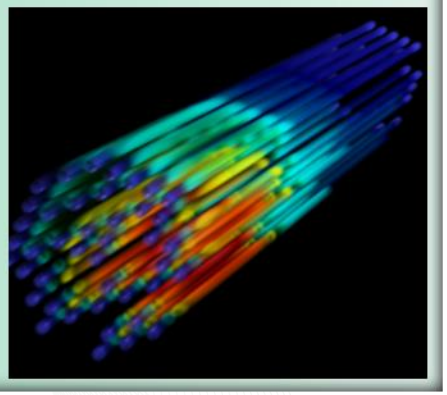


LAMMPS
A molecular description of membrane fusion, one of the most common ways for molecules to enter or exit living cells.

7.4 performance ratio XK7/XE6

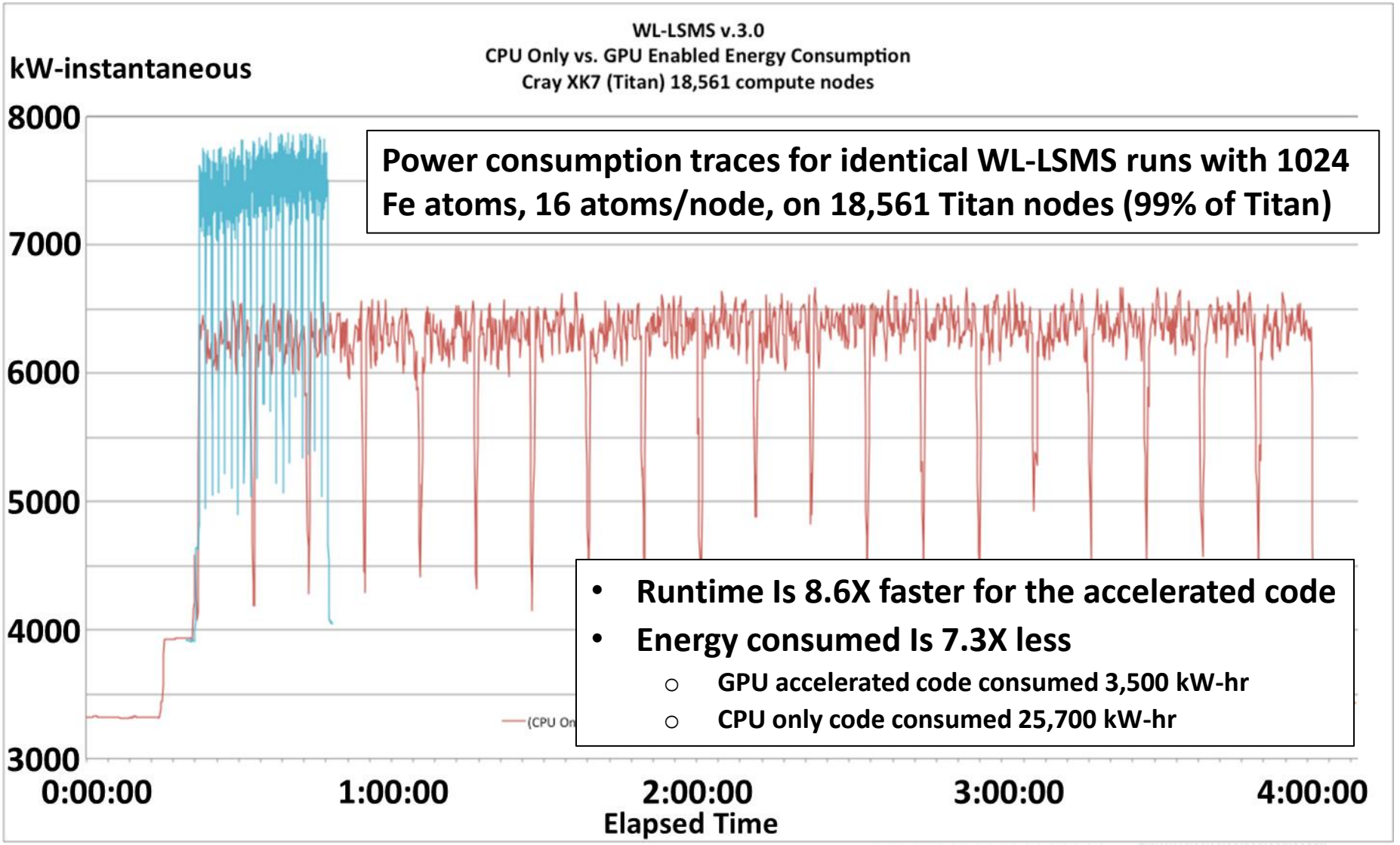
Denovo
Discrete ordinates radiation transport calculations that can be used in a variety of nuclear energy and technology applications.

3.8 performance ratio XK7/XE6



Titan: Cray XK7 (Kepler GPU plus AMD 16-core Opteron CPU)
Cray XE6: (2x AMD 16-core Opteron CPUs)

OLCF Center for Accelerated Application Readiness



Best Practices from CAAR for Titan

- Repeated themes in the code porting work:
  - finding more threadable work for the GPU
  - Improving memory access patterns
  - making GPU work (kernel calls) more coarse-grained if possible
  - making data on the GPU more persistent
  - overlapping data transfers with other work
  - use asynchronicity to extract performance
  - unoptimized MPI communications need to be addressed first
- Code changes that have global impact on the code are difficult to manage, e.g., data structure changes. An abstraction layer may help, e.g., C++ objects/templates
- Two common code modifications are:
  - Permuting loops to improve locality of memory reference
  - Fusing loops for coarser granularity of GPU kernel calls
- The difficulty level of the GPU port was in part determined by:
  - Structure of the algorithms—e.g., available parallelism, high computational intensity
  - Code execution profile—flat or hot spots
  - The code size (LOC)
- Tools (compilers, debuggers, profilers) were lacking early on in the project but are becoming more available and are improving in quality

Best Practices

- 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
- We may need to look in unconventional places to get another ~30X thread parallelism that may be needed for exascale—e.g., parallelism in time

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
- 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
- Vendor technical support is crucial
  - Programming environment often not mature
  - Best source of information on new hardware features
- Access to multiple resources, including early hardware
- Joint training activities

- Portability is a critical concern
- Experience benefits other developers and users
  - Coverage of scientific domains
  - Coverage of algorithmic methods and programming models
- Persistent culture of application readiness
  - More computational ready applications available
  - Experience of science liaisons and catalysts for user programs
  - Synergy with libraries and tools projects
- Success Metric
  - INCITE Computational Readiness

CAAR Projects Overview

Call for Proposals for eight Partnership Projects

- Partnership between
  - Application Development Team
  - OLCF Scientific Computing group
  - IBM/NVIDIA Center of Excellence
- Application Readiness phase for restructuring and optimization
- Early Science phase for grand-challenge scientific campaign
- OLCF Postdoctoral Associate per project
- Extensive training on hardware and software development environment

Portability is critical concern *(vide infra)*

- Coordination with NERSC NESAP and ALCF ESP programs
- Allocations on Titan, early delivery systems and Summit
- Allocations on NERSC and ALCF

CAAR Selection Criteria

- *Composition, experience and commitment of application development team*
- *Assessment of anticipated scientific impact of ported application*
- *Assessment of porting feasibility based on provided porting plan and benchmarks*
- *Application user base*
- *Compelling vision of an Early Science project*

- *Technical domain expertise of OLCF liaison*
- *Technical expertise of IBM/NVIDIA Center of Excellence*
- *Coverage of science domains in CAAR portfolio and support for DOE and US mission*
- *Coverage of algorithms, programming approaches, languages, data models*

- *Consultation with NERSC and ALCF*
- *Consultation with DOE Office of Advanced Scientific Computing Research*

CAAR Partnership Responsibilities

- *Develop and execute a **Technical plan*** for application porting and performance improvement, developed and executed with reviewable milestones
- *Develop and work according to a **Management plan*** with clear description of responsibilities of the CAAR team
- *Develop and execute an **Early Science project*** for compelling scientific grand-challenge campaign
- *Assign an **Application Scientist*** to carry out the Early Science campaign together with the CAAR team
- *Provide **Documentation*** for semi-annual reviews of achieved milestones, and intermediate and final reports

CAAR Partnership Resources

- The core development team of the application, with a stated level of effort
- An ORNL Scientific Computing staff member, who will partner with the core application development team
- A full-time postdoctoral fellow, located and mentored at the OLCF
- Technical support from the IBM/NVidia Center of Excellence
- Allocation of resources on Titan
- Access to early delivery systems and the *Summit*
- Allocation of compute resources on the full *Summit* system for the Early Science campaign

CAAR Partnership Activities

1. **Common training of all Application Readiness teams**
   a. Architecture and performance portability
   b. Avoidance of duplicate efforts
2. **Application Readiness Technical Plan Development and Execution**
   a. ***Code analysis & benchmarking*** to understand application characteristics: code structure, code suitability for architecture port, algorithm structure, data structures and data movement patterns, code execution characteristics ("hot spots" or "flat" execution profile)
   b. ***Develop parallelization and optimization approach*** to determine the algorithms and code components to port, how to map algorithmic parallelism to architectural features, how to manage data locality and motion
   c. ***Decide on programming model*** such as compiler directives, libraries, explicit coding models
   d. ***Execute technical plan***– benchmarking, code rewrite or refactor, porting and testing, managing portability, managing inclusion in main code repository
3. **Development and Execution of and Early Science Project**, i.e., challenging science problem that demonstrates the performance and scientific impact of the developed application port

CAAR Timeline

1. November 2014: Call for CAAR applications
2. February 20, 2015: CAAR proposal deadline (29 proposals submitted)
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. September 2017: Call for Early Science projects
8. November 2017: Selection Early Science projects
9. December 2017: Early Science projects start
10. June 2019: Early Science project ends

Drivers for 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

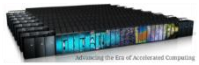
Primary task is exposing parallelism and data locality

Challenge is to find the right abstraction:

- MPI + X (X=OpenMP, OpenACC)
- PGAS + X
- DSL
- …

Two Tracks for Future Large Systems

Tianhe-2 (NUDT): TH-IVB-FEP
Intel Xeon E5-2692 12 C 2.2 GHz
TH Express-2
Intel Xeon Phi 31S1P



Titan (Cray): Cray XK7
AMD Opteron 6274 16C 2.2 GHz
Cray Gemini
NVIDIA K20x

Sequoia (IBM): BlueGene/Q
Power BQC 16C 1.6 GHz

K computer (Fujitsu)
SPARC64 VIIIfx 2.0 GHz
Tofu

Mira (IBM): BlueGene/Q
PowerPC A2 16C 1.6 GHz

Piz Daint (Cray): Cray XC30
Intel Xeon E5-2670 8C 2.6 GHz
Cray Aries
NVIDIA K20x

Edison (Cray): Cray XC30
Intel Xeon E%-2695v2 12C 2.4 GHz
Aries

Many Core

- 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

Hybrid Multi-Core

- CPU / GPU Hybrid systems
- Likely to have multiple CPUs and GPUs per node
- Small number of very fat 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

Cori at NERSC

- Self-hosted many-core system
- Intel/Cray
- 9300 single-socket nodes
- Intel® Xeon Phi™ Knights Landing (KNL)
- 16GB HBM, 64-128 GB DDR4
- Cray Aries Interconnect
- 28 PB Lustre file system @ 430 GB/s
- Target delivery date: June, 2016

Summit at OLCF

- Hybrid CPU/GPU system
- IBM/NVIDIA
- 3400 multi-socket nodes
- POWER9/Volta
- More than 512 GB coherent memory per node
- Mellanox EDR Interconnect
- Target delivery date: 2017

ALCF-3 at ALCF

- TBA
- Target delivery date: 2017-18

Strategies for Portability

- 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.

Synergy Between Application Readiness Programs

NESAP at NERSC

NERSC Exascale Science Application Program

- Call for Proposals – June 2014
- 20 Projects selected
- Partner with Application Readiness Team and Intel IPCC
- 8 Postdoctoral Fellows

Criteria

- An application's computing usage within the DOE Office of Science
- Representation among all 6 Offices of Science
- Ability for application to produce scientific advancements
- Ability for code development and optimizations to be transferred to the broader community through libraries, algorithms, kernels or community codes
- Resources available from the application team to match NERSC/Vendor resources

CAAR at OLCF

Center for Accelerated Application Readiness

- Call for Proposals – November 2014
- 8 Projects to be selected
- Partner with Scientific Computing group and IBM/NVIDIA Center of Excellence
- 8 Postdoctoral Associates

Criteria

- Anticipated impact on the science and engineering fields
- Importance to the user programs of the OLCF
- Feasibility to achieve scalable performance on Summit
- Anticipated opportunity to achieve performance portability for other architectures
- Algorithmic and scientific diversity of the suite of CAAR applications.
- Optimizations incorporated in master repository
- Size of the application's user base

ESP at ALCF

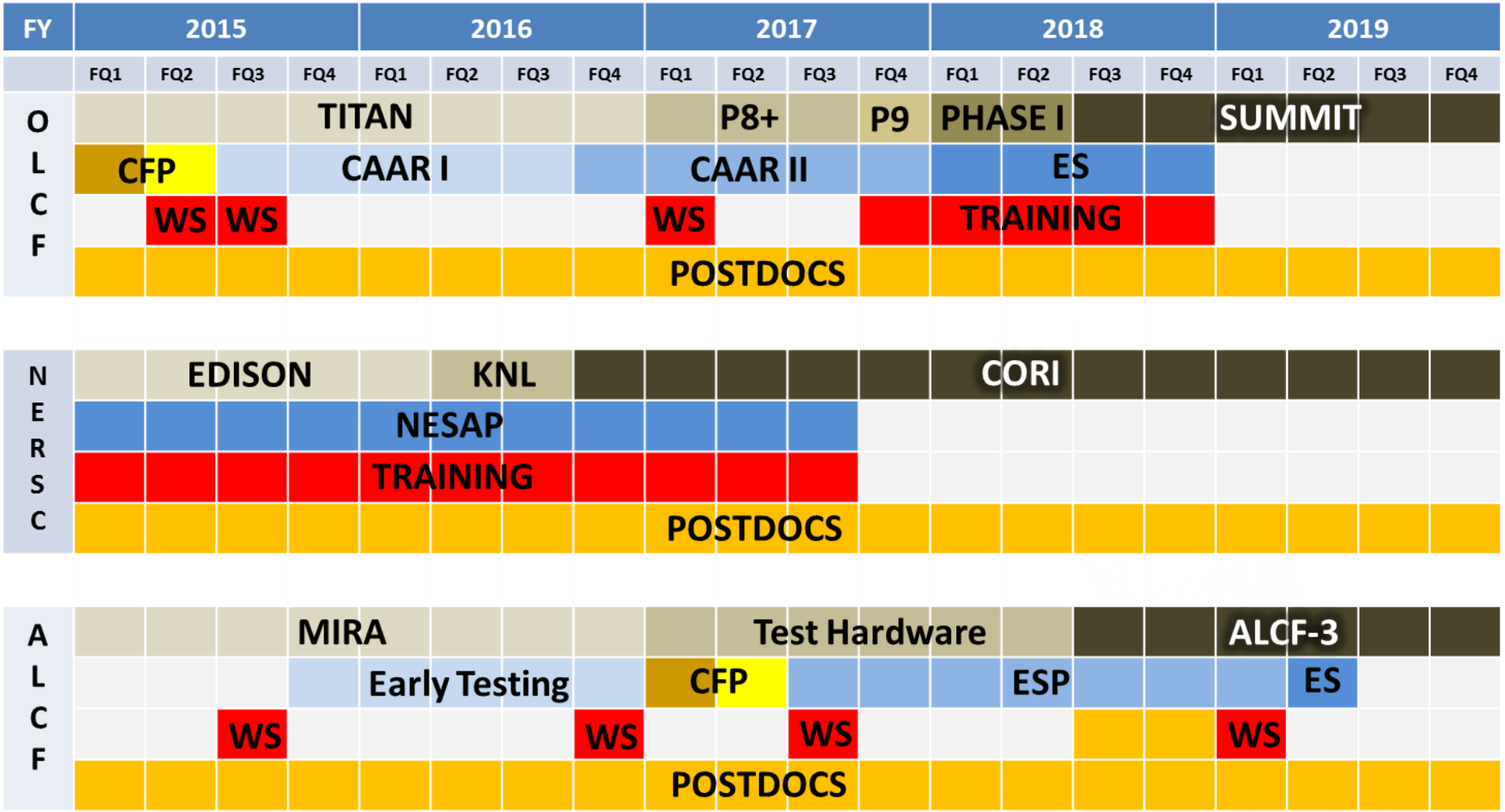
Early Science Program

- Call for Proposals
- 10 Projects to be selected
- Partner with Catalyst group and ALCF Vendor Center of Excellence
- Postdoctoral Appointee per project

Criteria

- Science Impact
- Computational Readiness
- Proposed science problem of appropriate scale to exercise capability of new machine
- Confidence code will be ready in time
- Project code team appropriate
- Willing partner with ALCF & vendor
- Diversity of science and numerical methods
- Samples spectrum of ALCF production apps

Application Readiness Tentative Timelines



Future Computational Scientists for Energy, Environment and National Security – Training Program

ASCR facilities host Distinguished Postdoctoral Associates programs with the objective of training the next generation of computational scientists

These programs provide:

1. Challenging scientific campaigns in critical science mission areas
2. Experience in using ASCR computing capabilities
3. Training in software development and engineering practices for current and future massively parallel computer architectures

Central to achieving these goals is access to leadership computing resources, availability of computational domain scientists to provide adequate mentoring and guidance, and facilities' association with universities with strong computational and computer science programs.

OLCF Distinguished Postdoctoral Associates Program

At the OLCF, **eight** Distinguished Postdoctoral Associate positions are available immediately for candidates interested in and capable of performing leading-edge computational science research and development.

Priorities include the development of methodologies and their efficient massively parallel implementation on current state-of-the-art accelerated computer architectures, as well as their application on large scientific challenge problems.

The center is specifically looking for candidates with strong computational expertise in the following scientific areas:

Astrophysics, Biophysics, Chemistry, Climate Science, Combustion, Fusion Energy Science, Materials Science, and Nuclear Physics.

For information about the positions or to apply, visit:
https://www.olcf.ornl.gov/summit/olcf-distinguished-postdoctoral-associates-program/

OLCF Scientific Computing

Ramanan Sankaran, Mike Matheson, George Ostrouchov, Duane Rosenberg, Valentine Anantharaj, Bronson Messer, Mark Berrill, Matt Norman, Ed D'Azevedo, Norbert Podhorski, Wayne Joubert, JJ Chai (postdoc, now in CSM), Judy Hill, Mark Fahey, Hai Ah Nam (now at LANL), Jamison Daniel, Dmitry Liakh, Supada Loasooksathit (postdoc), Markus Eisenbach, Arnold Tharrington, Ying Wai Li, Mingyang Chen (postdoc), Peyton Ticknor, Tjerk Straatsma, Dave Pugmire and Jan-Michael Carrillo (postdoc, now at SNS).

CAAR Projects

Questions & Discussion