

# SCIDAC FASTMATH INSTITUTE ALL-HANDS MEETING



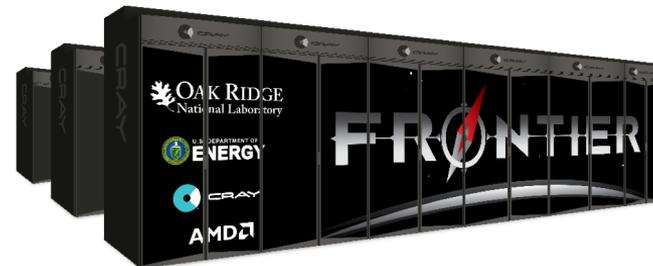
## FACILITIES SESSION - ALCF, NERSC, OLCF



JAEHYUK KWACK (ALCF)

WAYNE JOUBERT (OLCF)

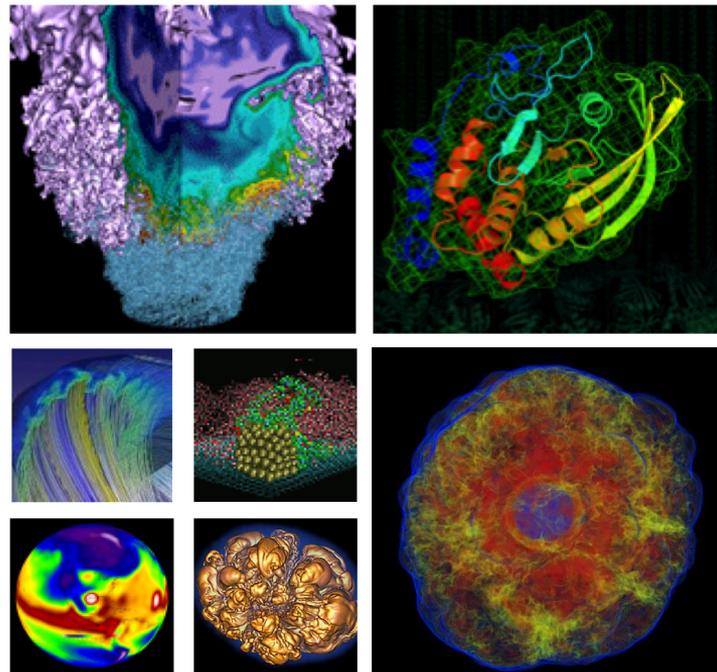
REBECCA HARTMAN-BAKER (NERSC)



# TOPICS

- Facility Roadmaps and Quick Hardware Overview for New Systems
  - NERSC
  - ALCF
  - OLCF
- Software Overview for New Systems
- Expectations for Post-Exascale Systems
- Application Readiness Programs and Lessons Learned

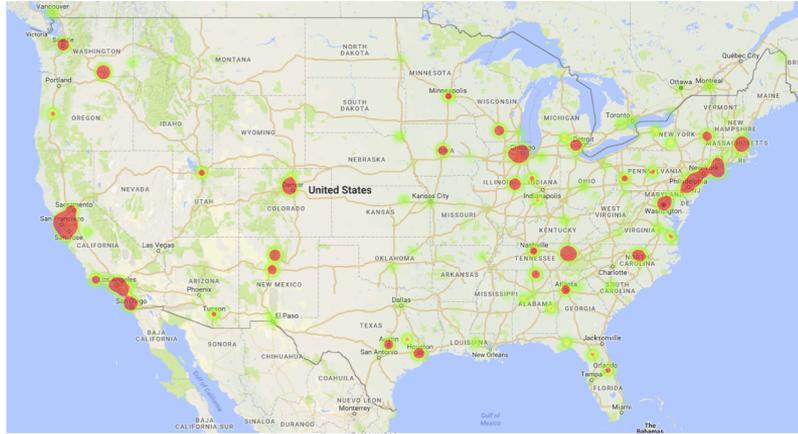
# NERSC Overview



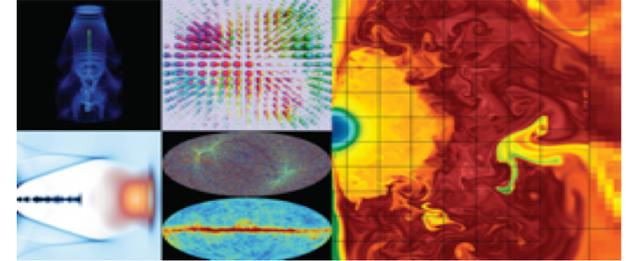
**Rebecca Hartman-Baker**

**NERSC User Engagement Group Lead**

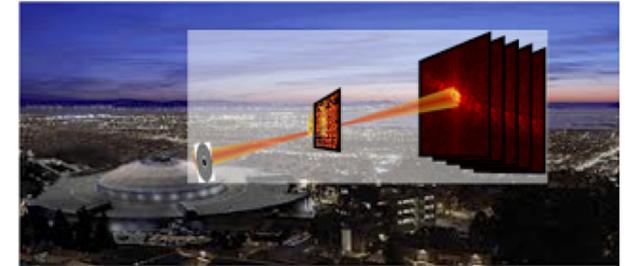
# NERSC: Mission HPC Facility for DOE Office of Science



7,000 Users  
800 Projects  
700 Codes  
~2000 publications per year

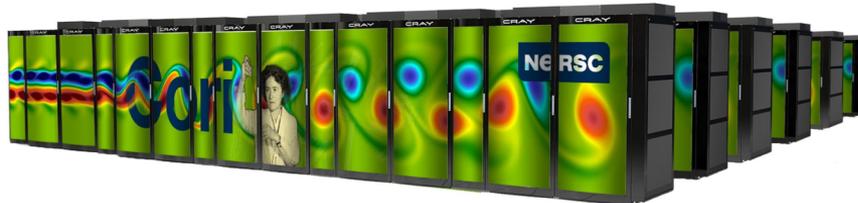


Simulations at scale



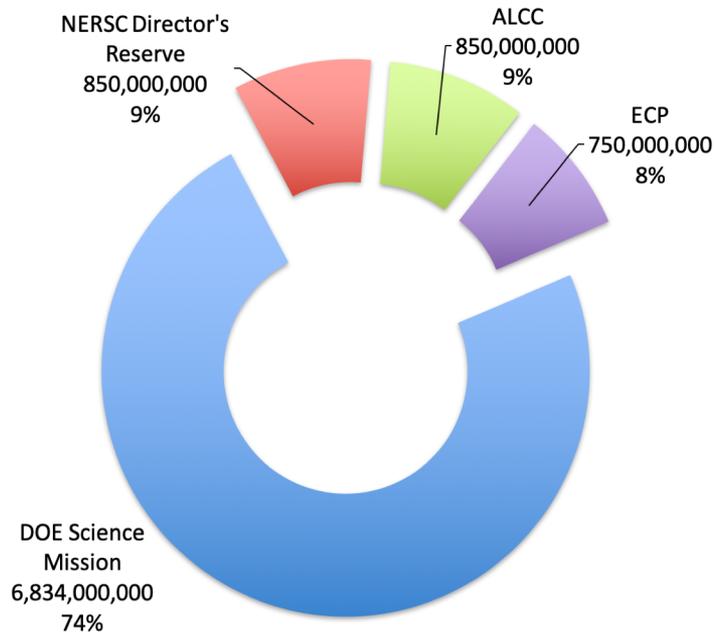
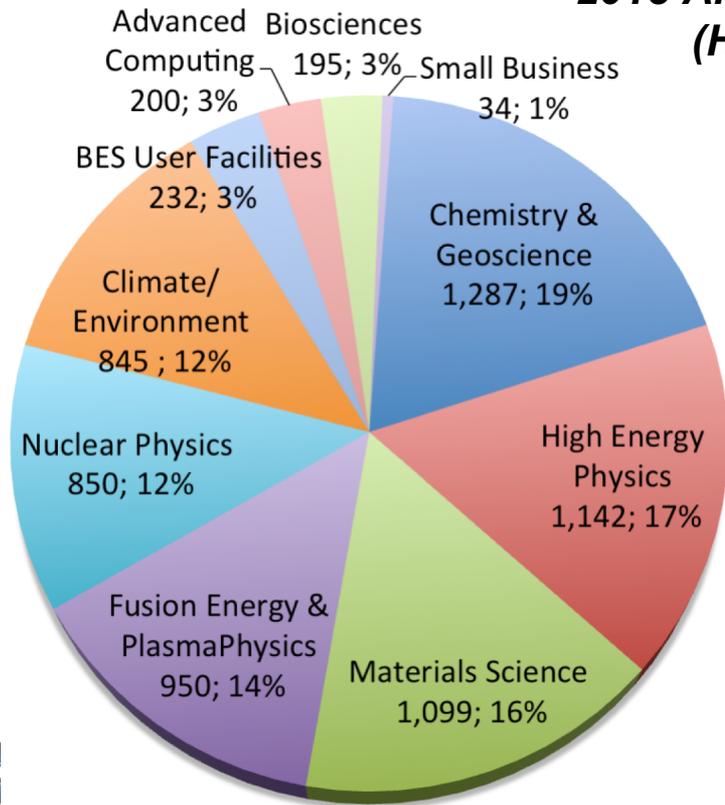
Data analysis support for  
DOE's experimental and  
observational facilities

Photo Credit: CAMERA



# NERSC Directly Supports Office of Science Priorities

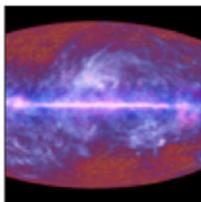
## 2018 Allocation Breakdown (Hours Millions)



# NERSC Supports Many Users & Projects from DOE SC's Experimental & Observational Facilities



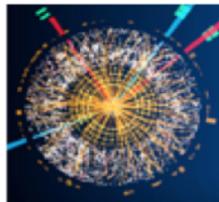
Palomar Transient Factory Supernova



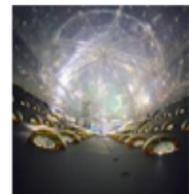
Planck Satellite Cosmic Microwave Background Radiation



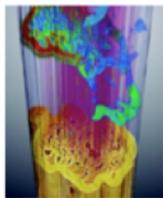
Alice Large Hadron Collider



Atlas Large Hadron Collider



Dayabay Neutrinos



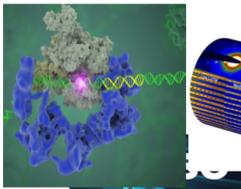
ALS Light Source



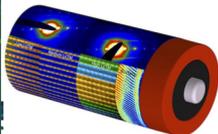
LCLS Light Source



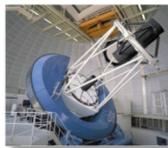
Joint Genome Institute Bioinformatics



Cryo-EM



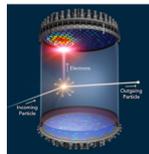
NCEM



DESI

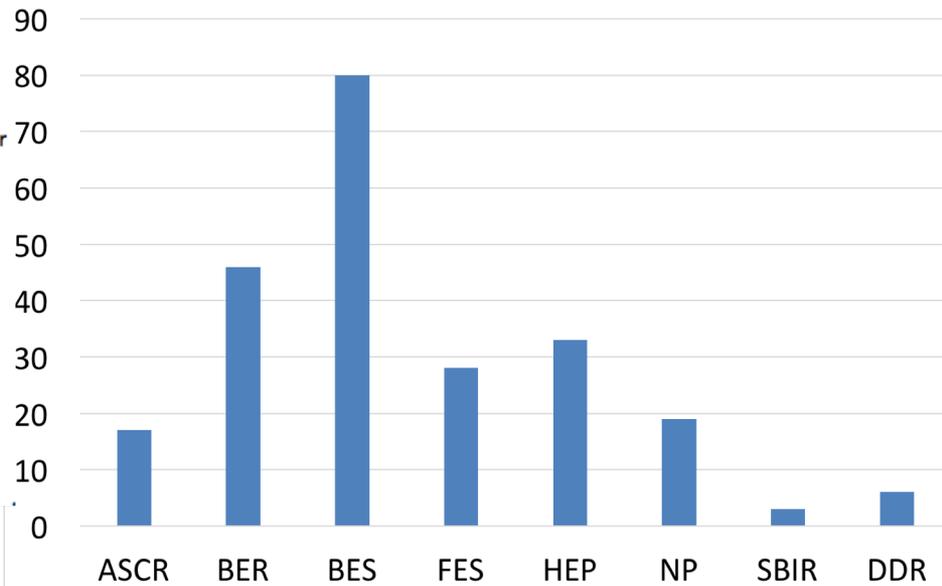


LSST-DESC



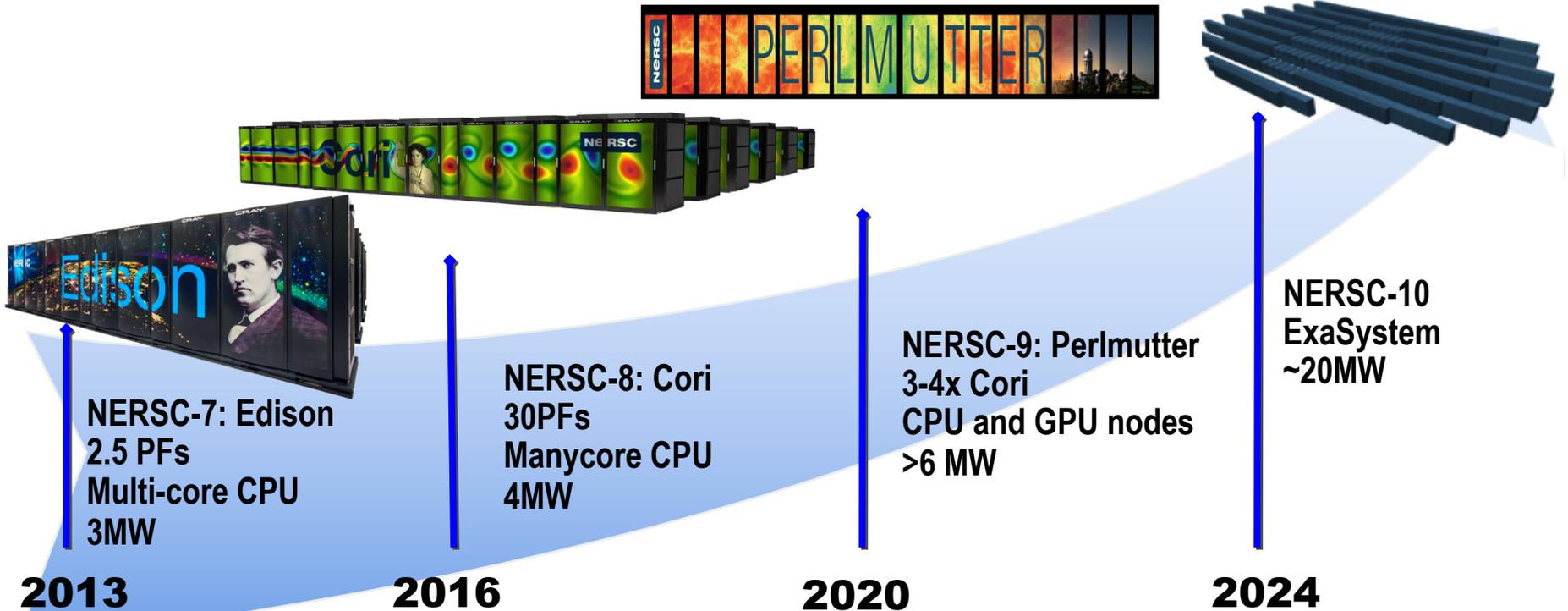
LZ

*# of Projects Analyzing Experimental Data or Combining Modeling and Experimental Data by SC Office*

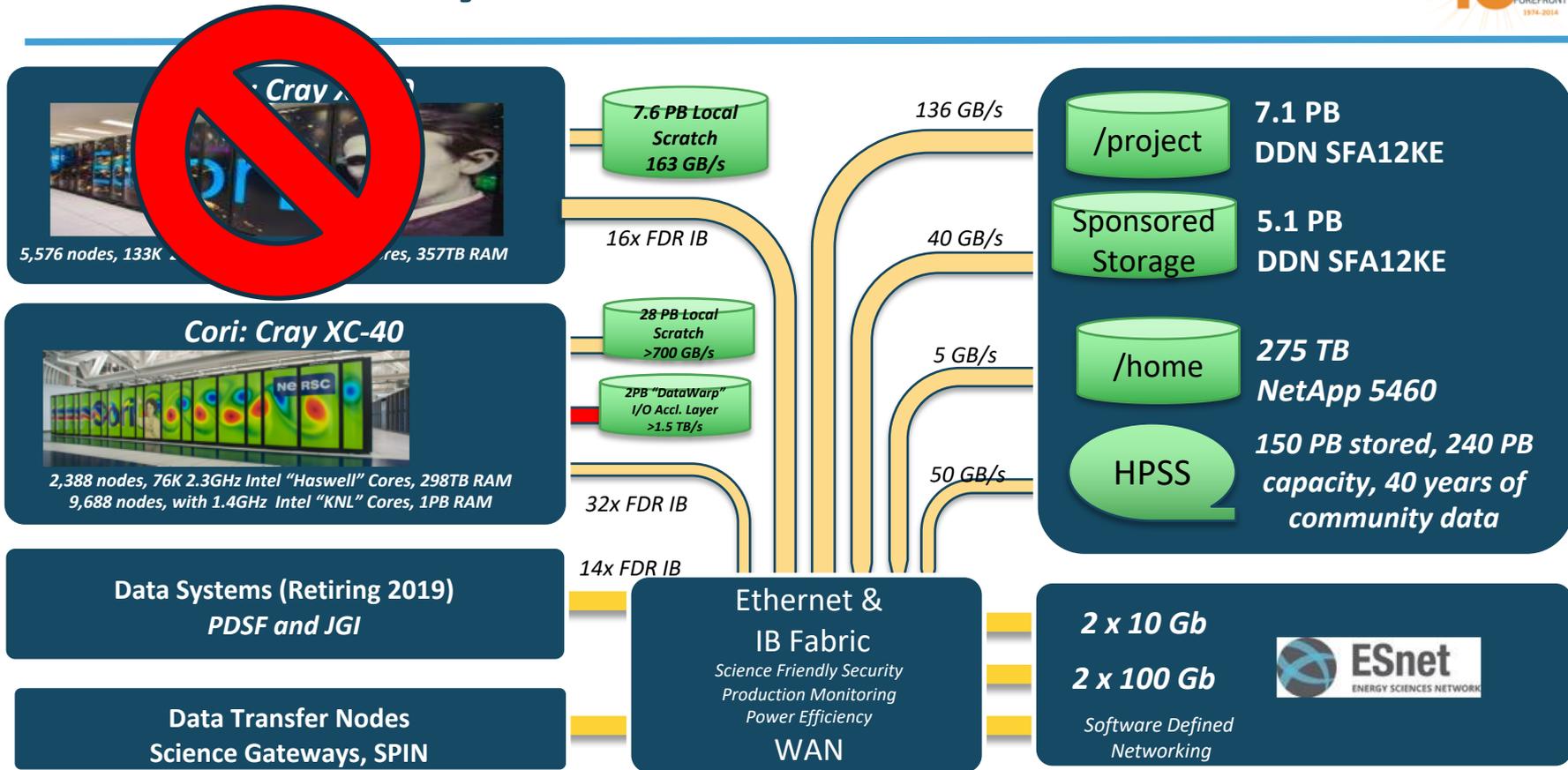


~35% (235) of ERCAP projects self identified as confirming the primary role of the project is to 1) analyze experimental data or; 2) create tools for experimental data analysis or; 3) combine experimental data with 4 simulations and modeling

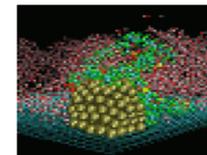
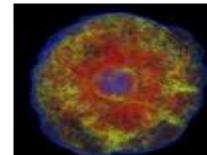
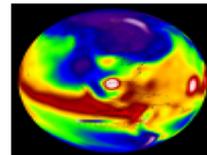
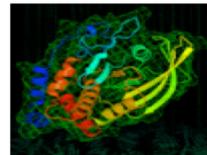
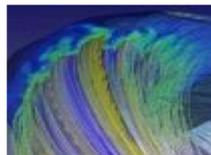
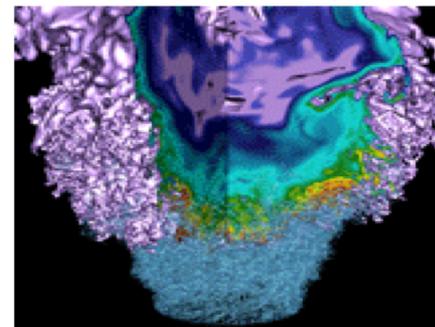
# NERSC Systems Roadmap



# NERSC Facility Architecture 2019



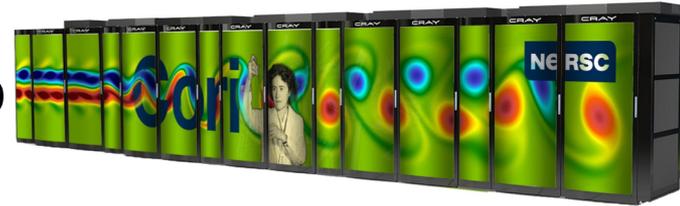
# Cori System



# Cori: Pre-Exascale System for DOE Science

---

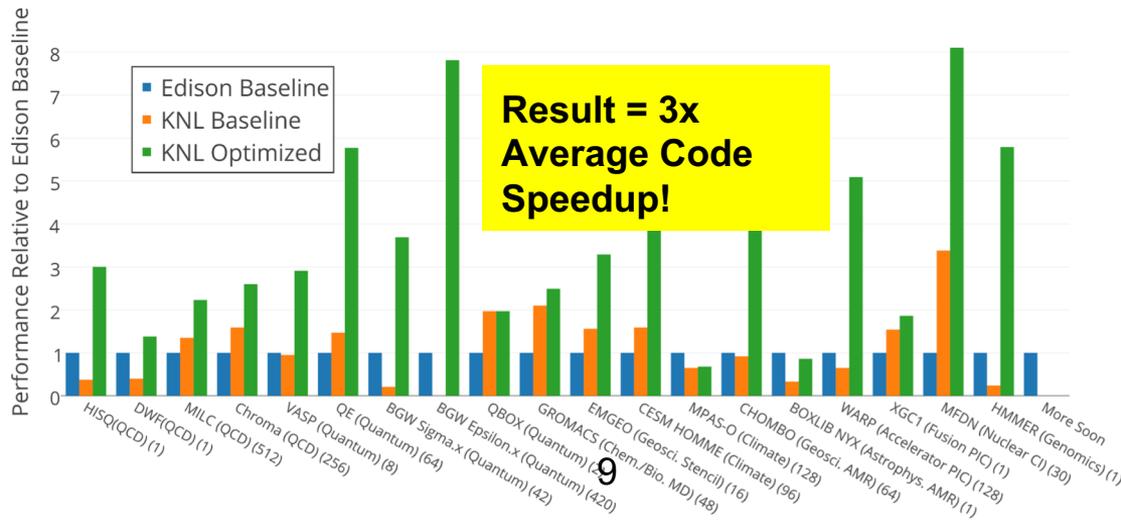
- Cray XC System - heterogeneous compute architecture
  - 9600 Intel KNL compute nodes, >2000 Intel Haswell nodes
- Cray Aries Interconnect
- NVRAM Burst Buffer, 1.6PB and 1.7TB/sec
- Lustre file system 28 PB of disk, >700 GB/sec I/O
- Investments to support large scale data analysis
  - High bandwidth external connectivity to experimental facilities from compute nodes
  - Virtualization capabilities (Shifter/Docker)
  - More login nodes for managing advanced workflows
  - Support for real time and high-throughput queues



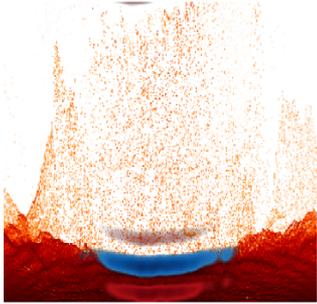
# NERSC Exascale Scientific Application Program (NESAP)



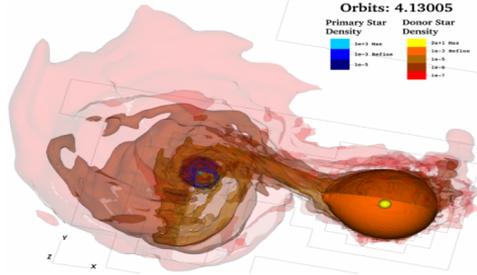
- Prepare DOE SC users for advanced architectures like Cori and Perlmutter
- Partner closely with 20-40 application teams and apply lessons learned to broad NERSC user community.



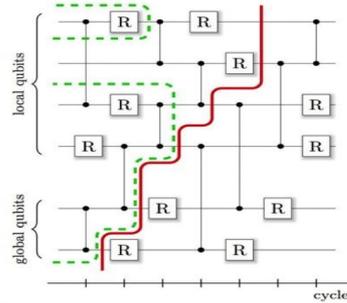
# Users Demonstrate Groundbreaking Science Capability



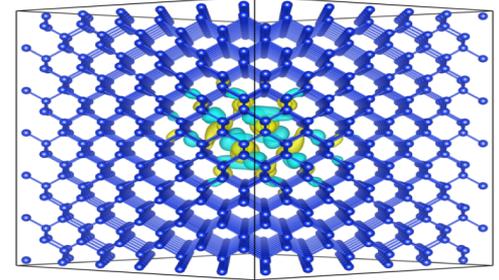
Large Scale Particle in Cell Plasma Simulations



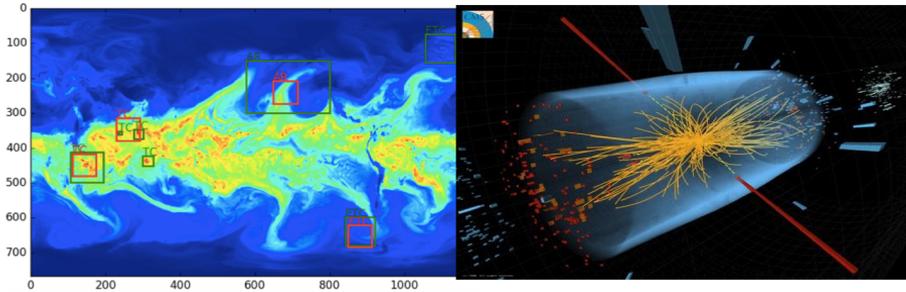
Stellar Merger Simulations with Task Based Programming



Largest Ever Quantum Circuit Simulation



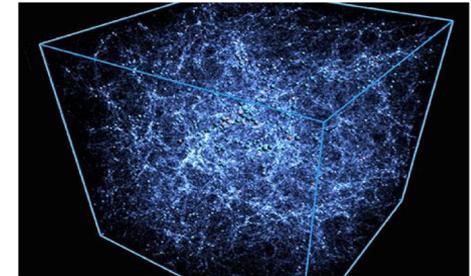
Largest Ever Defect Calculation from Man Body Perturbation Theory > 10PF



Deep Learning at 15PF (SP) for Climate and HEP



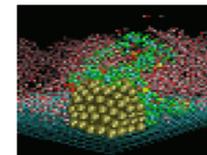
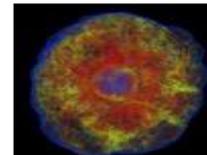
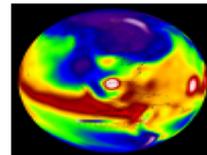
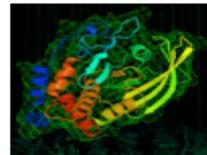
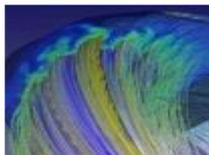
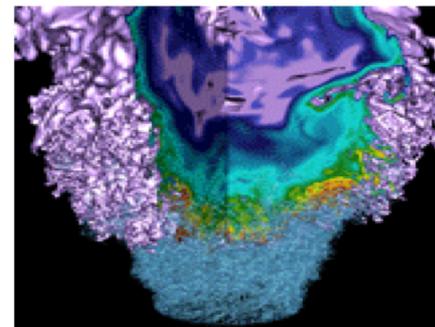
Celeste: 1<sup>st</sup> Julia app to achieve 1 PF



Galactos: Solved 3-pt correlation analysis for Cosmology @9.8PF



# Perlmutter System



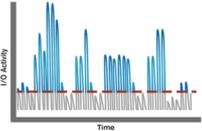
# NERSC-9: A System Optimized for Science

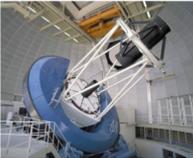
---

- **Cray Shasta System providing 3-4x capability of Cori system**
- **First NERSC system designed to meet needs of both large-scale simulation and data analysis from experimental facilities**
  - Includes both NVIDIA GPU-accelerated and AMD CPU-only nodes
  - Cray Slingshot high-performance network will support Terabit rate connections to system
  - Optimized data software stack enabling analytics and ML at scale
  - All-Flash filesystem for I/O acceleration
- **Robust readiness program for simulation, data and learning applications and complex workflows**
- **Delivery in late 2020**

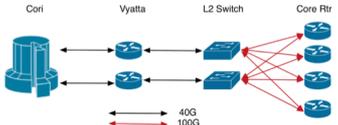


Data Features	Cori experience	N9 enhancements
---------------	-----------------	-----------------

<h2>I/O and Storage</h2>	<p>Burst Buffer</p> 	<p>All-flash file system: performance with ease of data management</p> 
--------------------------	--	--

<h2>Analytics</h2> <ul style="list-style-type: none"> <li>- Production stacks</li> <li>- Analytics libraries</li> <li>- Machine learning</li> </ul>	<p>User defined images with Shifter NESAP for data</p>  <p>New analytics and ML libraries</p> 	 <p>Optimised analytics libraries and deep learning application benchmarks</p>
---	---	---

<h2>Workflow integration</h2>	<p>SchedMD</p> <p>Real-time queues</p>	 <p>SLURM co-scheduling Workflow nodes integrated</p>
-------------------------------	--	--

<h2>Data transfer and streaming</h2> 	<p>SDN</p> 	<p>Slingshot ethernet-based converged fabric</p>  <p>13</p>
--	--	---

# NESAP for Perlmutter

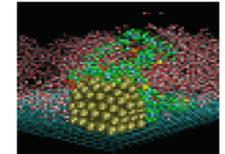
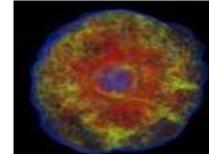
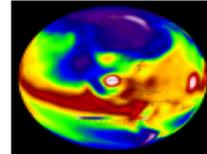
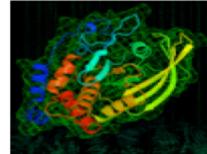
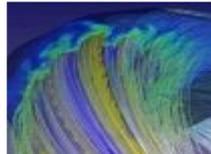
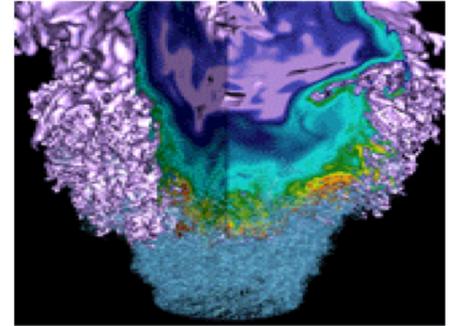
**Simulation  
12 Apps**

**Data Analysis  
8 Apps**

**Learning  
5 Apps**

- **5 ECP Apps Jointly Selected (Participation Funded by ECP)**
- **Open call for proposals, closed in December 2018, evaluated in January 2019.**
  - **App selection contains multiple applications from each SC Office and algorithm area**
  - **Additional applications (beyond 25) selected for second-tier NESAP with access to vendor/training resources and early access**
- **Access to Cori GPU rack for application readiness efforts.**

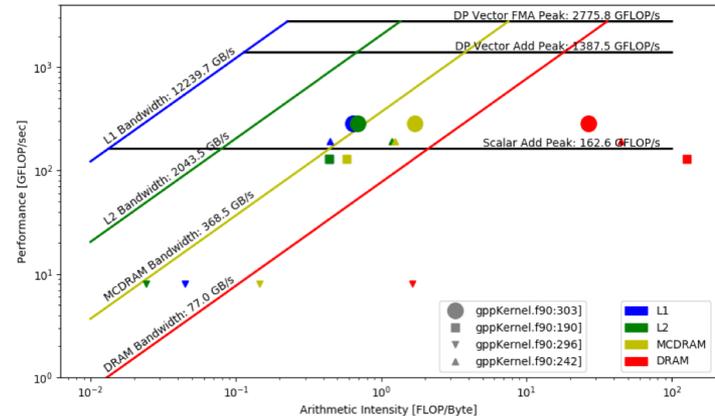
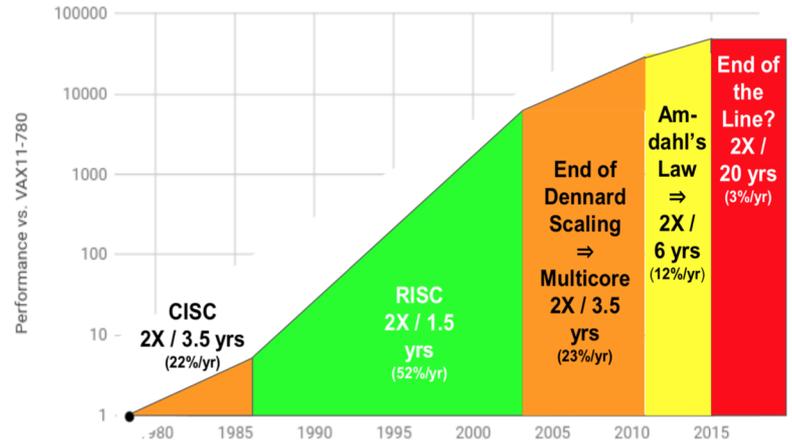
# Future Plans



# Transitioning SC Workload to Advanced Architectures and Exploring Technologies for NERSC-10 and NERSC-11

**Advanced Architectures will be necessary to meet energy-efficiency and performance requirements for NERSC. Our strategy is to:**

1. Engage with the user community to transition applications to advanced architectures
2. Deploy a few new technologies into Perlmutter system and explore its applicability to the NERSC workload
3. Develop techniques for quantitative application analysis to assist design of future accelerators
4. Partner w/ the research community to understand science use cases for specialized accelerators



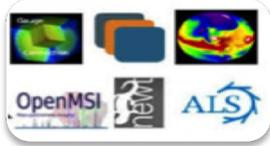
# Supporting the Superfacility model at NERSC

- **NERSC continues to support experimental science**
  - Experience of our users drives the requirements of the Superfacility initiative → focus on making workflows *seamless*



## Applications

Engage with experimental users to design and deploy the infrastructure they need  
e.g. federated ID, Jupyter, real-time workflows



## Data

Provide storage systems and tools for easy analysis, movement and storage of experimental data  
e.g. Storage2020, Perlmutter all-flash storage



## Automation

Manage allocation of NERSC resources and expose system information to experimental users  
e.g. Perlmutter Slingshot network, Superfacility API



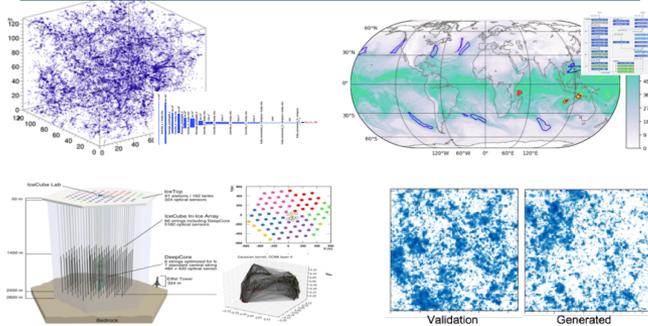
## Edge devices

Support specialised computing devices for experimental workflows  
e.g. NCEM DAQ device, NERSC10 pathfinding



# Machine Learning and Analytics for Science

## Methods and Applications



## Deployment

Automation

Hubs

Notebooks

Software Frameworks and Libraries for Scale

Hardware ML System and Accelerators



## Empowerment

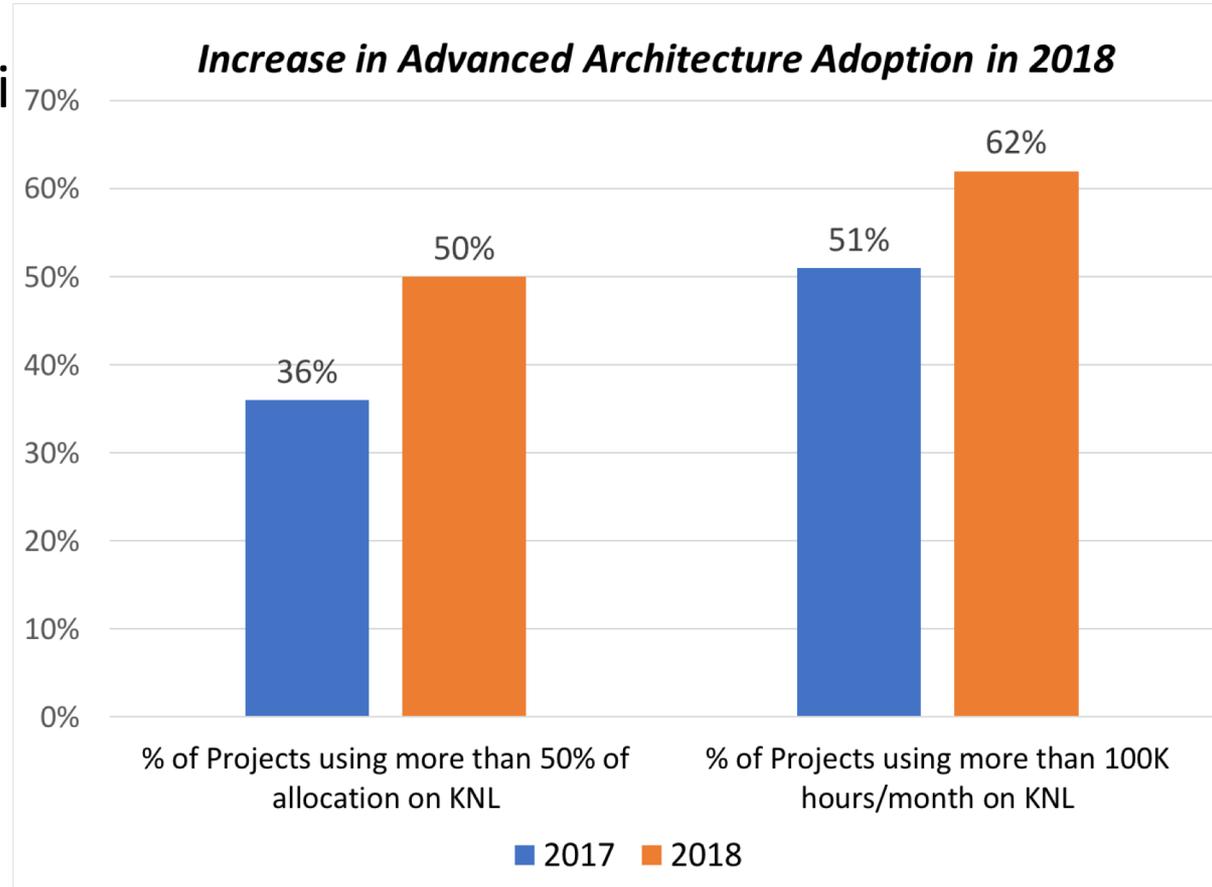


## Our strategy:

- Engage with Office of Science users who want to integrate ML into applications and workloads
- Take a leadership role in training the SC community on ML for Science techniques
- Deploy optimized hardware and software for ML/analytics at scale
- Apply ML for science using cutting-edge methods

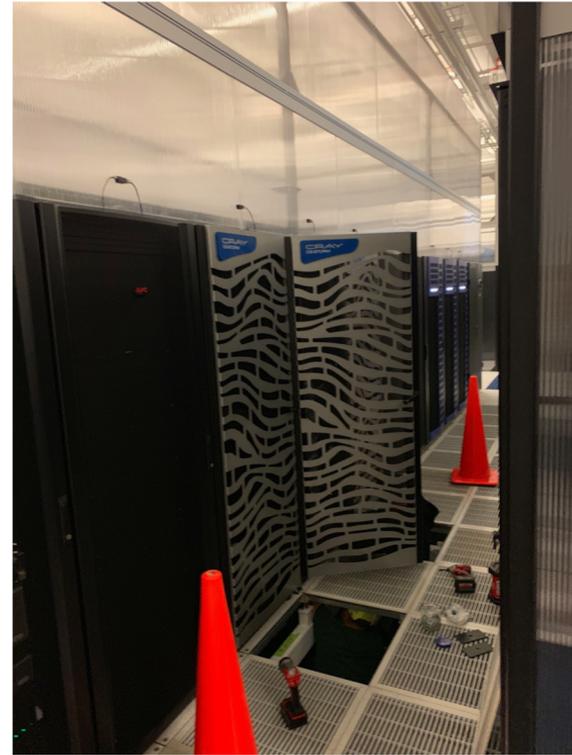
# Transitioning NERSC Workload to Advanced Architectures

To effectively use Cori KNL, users must exploit parallelism, manage data locality and utilize longer vector units. All features that will be present on exascale era systems



# GPU Partition added to Cori for NERSC-9 preparations

- GPU partition added to Cori to enable users to prepare for Perlmutter system
- 18 nodes each with 8 GPUs
- Software support for both HPC simulations and Machine Learning



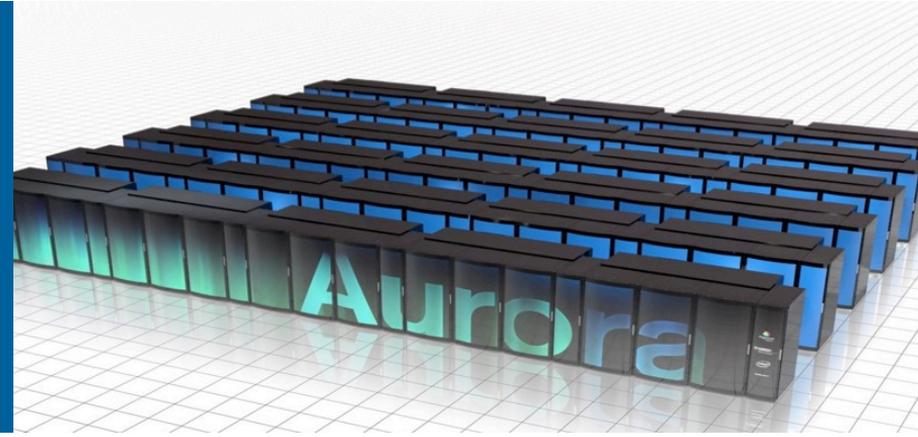
GPU cabinets being integrated into Cori  
Sept. 2018

# User requirements drive our process

- First time users from DOE experimental facilities broadly included
- Focus on broad ecosystem, beyond compute and flops
- Requirements Reviews highlighted common data related themes across SC offices, domains and facilities



## ALCF SYSTEMS AND AURORA HARDWARE



**JAEHYUK KWACK**  
ALCF Perf. Engr. Group

# ALCF SYSTEMS AND ROADMAP



Argonne National Laboratory is a  
U.S. Department of Energy laboratory  
managed by UChicago Argonne, LLC.

# ALCF COMPUTING RESOURCES



**Mira IBM BG/Q**  
49,152 nodes  
786,432 cores  
768 TB RAM  
5D Torus Interconnect  
Peak flop rate: 10 PF



**Theta Cray XC40**  
4,392 compute nodes  
281,088 Intel KNL cores  
70.272 TB MCDRAM  
843.264 TB DDR4  
Aries interconnect - Dragonfly  
Peak flop rate: 11.69 PF



**Cooley Cray/NVIDIA**  
126 compute nodes  
1512 Intel Haswell CPU cores  
126 NVIDIA Tesla K80 GPUs  
48 TB RAM / 3 TB GPU  
FDR Infiniband interconnect

## Storage Capability

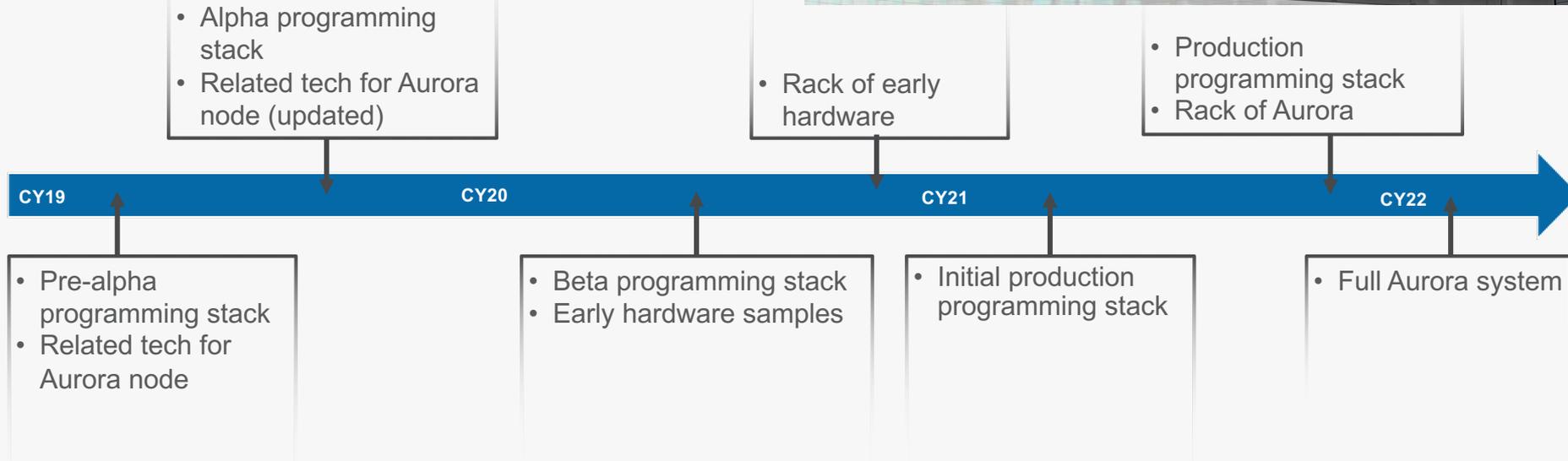
### Disk

- Mira: ~27 PB of GPFS file system capacity with performance of 240 GB/s on the largest file system (19PB).
- Theta: ~18 PB of GPFS/Lustre file system capacity; 9PB is GPFS and 9.2PB is Lustre.

### Tape

- The ALCF has three 10,000-slot libraries using LTO 6 tape technology. The LTO tape drives have built-in hardware compression for an effective capacity of 36-60 PB.

# 2021 AURORA SYSTEM



Application evaluation and projection happens constantly through this  
Constant testing of application, software, and hardware  
Well before hardware, we know how applications will perform

# AURORA HARDWARE OVERVIEW

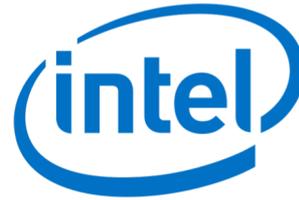


Argonne National Laboratory is a  
U.S. Department of Energy laboratory  
managed by UChicago Argonne, LLC.



# HIGH LEVEL VIEW OF AURORA

- Aurora is an Intel/Cray machine

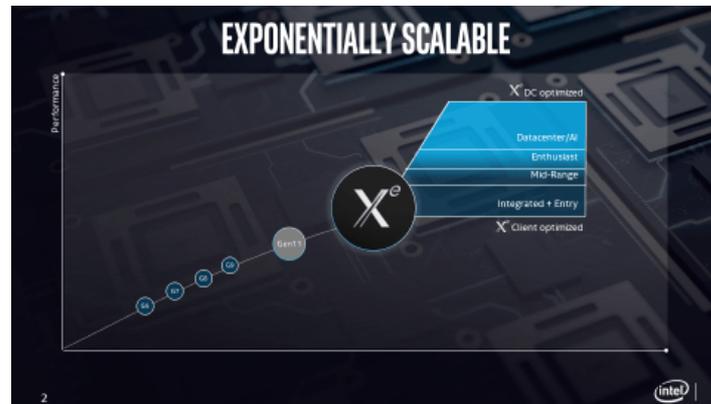


- The system is to be delivered to Argonne in 2021
  - Limited availability to applications in 2021
  - Expect broader availability to applications in 2022
- Aurora will be an Exa-scale system
  - Will have a peak performance of over 1 ExaFlop/s in Double-Precision
  - Much higher than 1 ExaFlop/s in Half-Precision
  - Target for applications performance is an 50x speedup over Titan/Sequoia

# AURORA HARDWARE OVERVIEW

## Compute Node and Memory

- Processor
  - Future Intel® Xeon® Scalable Processor
- Accelerator
  - New Intel® X<sup>e</sup> Compute Architecture



# AURORA HARDWARE OVERVIEW

## Platform, Fabric and I/O

- Compute Platform
  - Cray Shasta next generation supercomputing platform
- System Interconnect
  - Cray Slingshot interconnect
- I/O system will have:
  - Intel's DAOS (Distributed Application Object Storage)
  - Traditional parallel filesystem augments DAOS (bulk storage, legacy support)

The infographic is divided into several sections. The top section, 'Performance', features a speedometer icon and lists: 'Highest power CPUs (500W+) supported via direct liquid cooling', 'Up to 16 Slingshot injection ports per compute blade', and 'Hardware & Software scalable to Exascale class systems'. The middle section, 'TCO', features a bar chart icon and lists: 'Warm water cooling (ASHRAE W3 and W4 temps supported)', 'Efficient power conversion from mains to point-of-load', and 'Upgradable for multiple technology generations'. The bottom-left section, 'Rosetta', features a satellite icon and lists: 'Multiple QoS levels', 'Aggressive adaptive routing', 'Advanced congestion control', and 'Very low average and tail latency'. The bottom-right section, 'NIC', features a network interface card icon and lists: 'Cray MPI stack', 'Ethernet functionality', 'RDMA offload', and '~50M MPI messages/sec'. A small grid icon labeled '64 ports x 200 Gbps' is also present. The bottom-most section features the Intel logo and 'OPTANE DC PERSISTENT MEMORY' with an image of a memory module.

# MORE TO COME...



# OLCF Frontier System Overview

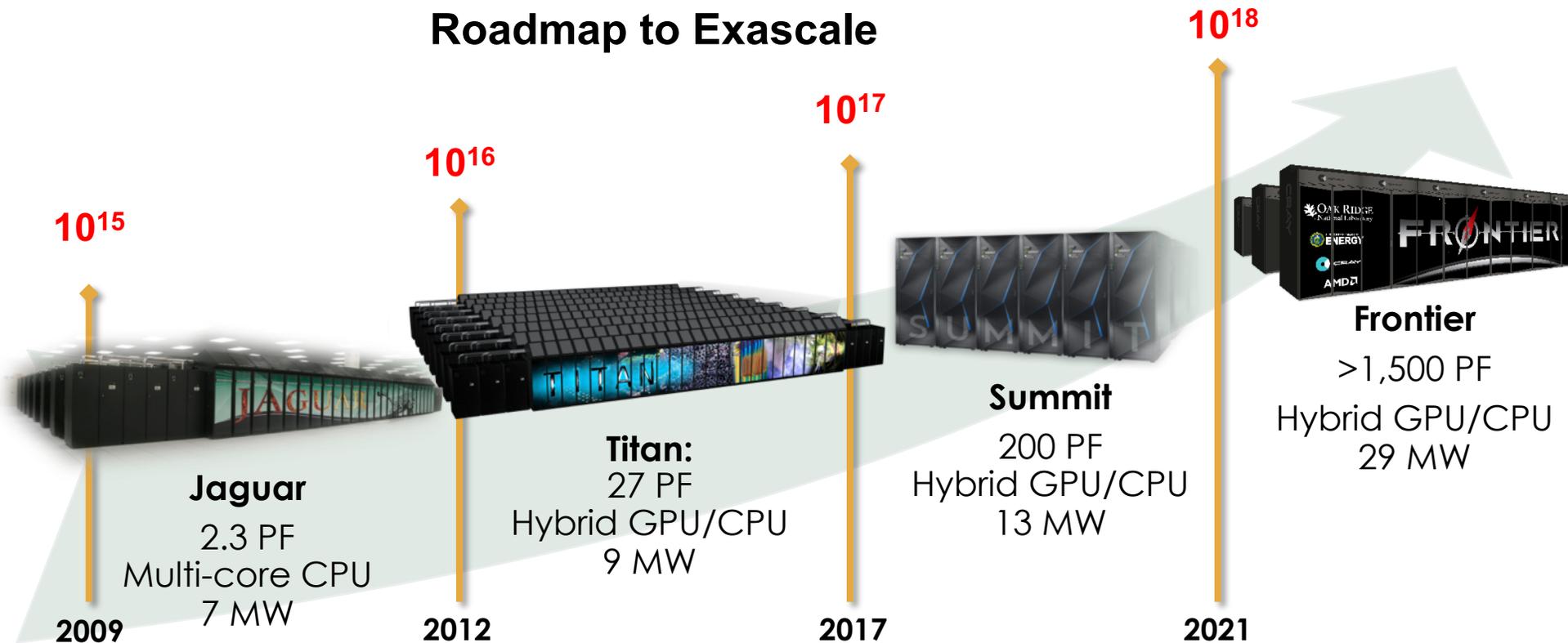
ORNL is managed by UT-Battelle, LLC for the US Department of Energy

# Oak Ridge Leadership Computing Facility – a DOE Office of Science User Facility

**Mission:** Providing world-class computational resources and specialized services for the most computationally intensive global challenges

**Vision:** Deliver transforming discoveries in energy technologies, materials, biology, environment, health, etc.

## Roadmap to Exascale



# Frontier Overview

Partnership between ORNL, Cray, and AMD

The Frontier system will be delivered in 2021

Peak Performance greater than 1.5 EF

Composed of more than 100 Cray Shasta cabinets

- Connected by Slingshot™ interconnect with adaptive routing, congestion control, and quality of service

Node Architecture:

- An AMD EPYC™ processor and four Radeon Instinct™ GPU accelerators purpose-built for exascale computing
- Fully connected with high speed AMD Infinity Fabric links
- Coherent memory across the node
- 100 GB/s injection bandwidth
- Near-node NVM storage

Researchers will harness Frontier to advance science in such applications as systems biology, materials science, energy production, additive manufacturing and health data science.



# Comparison of Titan, Summit, and Frontier Systems

System Specs	Titan	Summit	Frontier
Peak	27 PF	200 PF	~1.5 EF
# cabinets	200	256	> 100
Node	1 AMD Opteron CPU 1 NVIDIA K20X Kepler GPU	2 IBM POWER9™ CPUs 6 NVIDIA Volta GPUs	1 HPC and AI Optimized AMD EPYC CPU 4 Purpose-Built AMD Radeon Instinct GPU
On-node interconnect	PCI Gen2 No coherence across the node	NVIDIA NVLINK Coherent memory across the node	AMD Infinity Fabric Coherent memory across the node
System Interconnect	Cray Gemini network 6.4 GB/s	Mellanox Dual-port EDR IB network 25 GB/s	Cray four-port Slingshot network 100 GB/s
Topology	3D Torus	Non-blocking Fat Tree	Dragonfly
Storage	32 PB, 1 TB/s, Lustre Filesystem	250 PB, 2.5 TB/s, IBM Spectrum Scale™ with GPFS™	2-4x performance and capacity of Summit's I/O subsystem.
Near-node NVM (storage)	No	Yes	Yes

# FASTMATH INSTITUTE ALL-HANDS MEETING

## SOFTWARE STACK FOR NEW SYSTEMS (ALCF, NERSC, OLCF) & EXPECTATIONS FOR POST- EXASCALE SYSTEMS

JAEHYUK KWACK (ALCF)

REBECCA HARTMAN-BAKER (NERSC)

WAYNE JOUBERT (OLCF)



# SOFTWARE OVERVIEW FOR NEW SYSTEMS (ALCF, NERSC AND OLCF)



Argonne National Laboratory is a  
U.S. Department of Energy laboratory  
managed by UChicago Argonne, LLC.



# THREE PILLAR SOFTWARE REQUIREMENTS

Simulation	Data	Learning
HPC Languages	Productivity Languages	Productivity Languages
Directives	Big Data Stack	DL Frameworks
Parallel Runtimes	Statistical Libraries	Statistical Libraries
Solver Libraries	Databases	Linear Algebra Libraries
Compilers, Performance Tools, Debuggers		
Math Libraries, C++ Standard Library, libc		
I/O, Messaging		
Containers, Visualization		
Scheduler		

# SOFTWARE OVERVIEW

- The software environment on new systems at ALCF, NERSC, and OLCF
  - will provide a familiar HPC environment with additions for accelerators
  - will be composed of an integrated set of components from vendors, and open source projects
  - will largely be an evolution of today's HPC software stack
  - will provide accelerator programming models that will allow performance to be achieved through efficiently offloading of data and kernels to the accelerators
- Major components of the environment include:
  - Vendors and open source compilers
  - Multiple accelerator programming models
  - Multiple MPI implementations
  - Parallel I/O libraries
  - Optimized libraries for math and machine learning
  - Frameworks and productivity languages for learning and data applications
  - Containers to facilitate software deployment and productivity

# OPTIMIZED HPC SOFTWARE PACKAGES

## HPC Libraries, Frameworks and Tools

- Optimized Vendors and opensource/3<sup>rd</sup>-party software packages will be provided

### LIBRARIES

- Math Libraries
- I/O
- Data Analytics
- MPI
- Parallel Solvers
- Visualization

### TOOLS

- Performance Analysis
- Debugging

### FRAMEWORKS & CONTAINERS

- DL/ML frameworks
- Big data stack
- Databases
- Graph analytics
- Containers

# USE OPEN PROGRAMMING MODELS, AVOID PROPRIETARY MODELS

- OpenMP
- OpenCL
- Raja
- Kokkos
- HIP
- SYCL



OpenCL



Kokkos

Kokkos C++ Performance Portability Programming EcoSystem

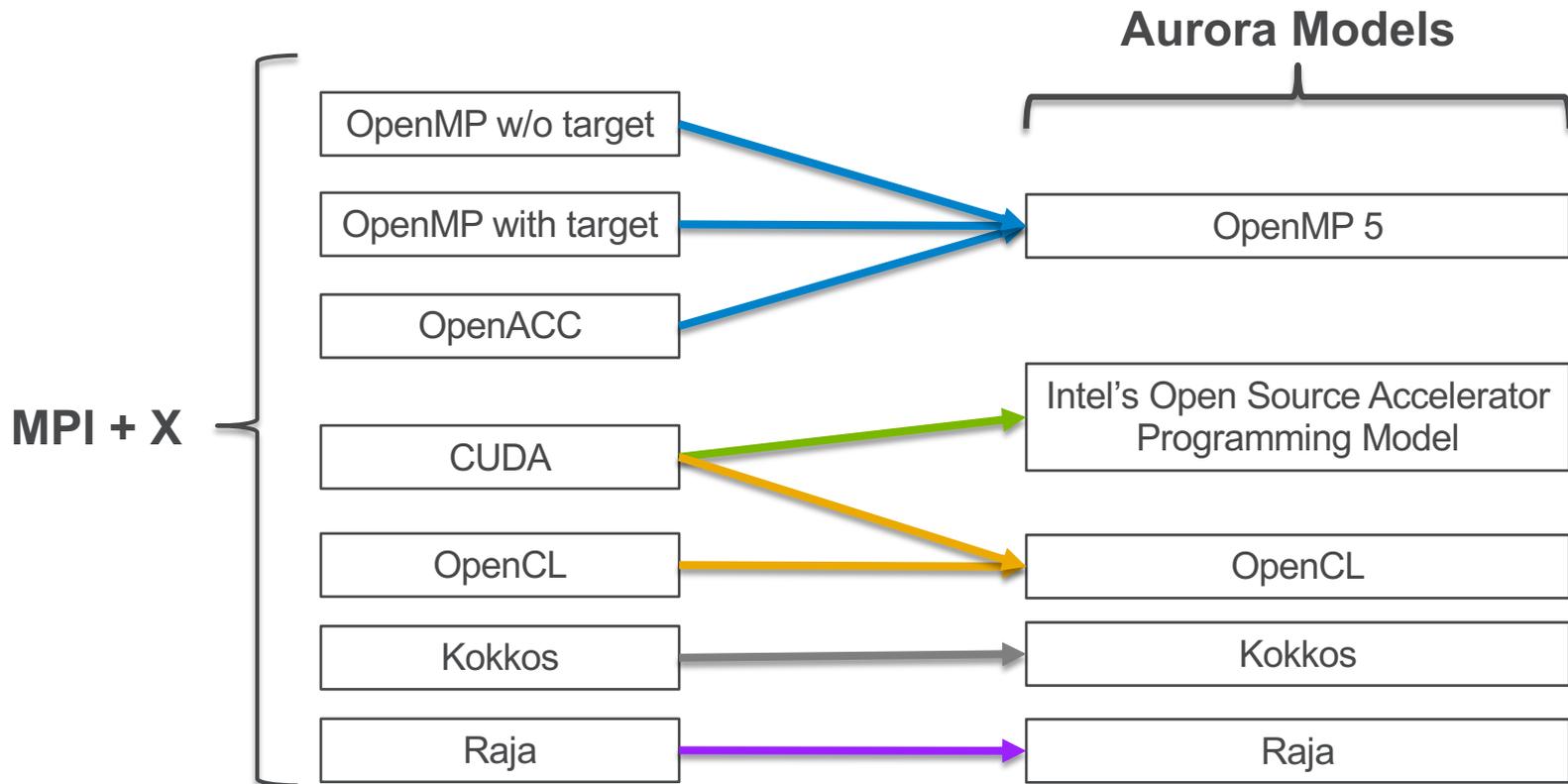
✉ [crtrott@sandia.gov](mailto:crtrott@sandia.gov)



# RAJA

## ROCm-Developer-Tools / HIP

# MAPPING OF EXISTING PROGRAMMING MODELS TO AURORA SYSTEM AT ALCF



# FRONTIER PROGRAMMING ENVIRONMENT

- To aid in moving applications from Titan and Summit to Frontier, ORNL, Cray, and AMD will partner to co-design and develop enhanced GPU programming tools designed for performance, productivity and portability.
- This will include new capabilities in the Cray Programming Environment and AMD's ROCm open compute platform that will be integrated together into the Cray Shasta software stack for Frontier
- In addition, Frontier will support many of the same compilers, programming models, and tools that have been available to OLCF users on both the Titan and Summit supercomputers

Summit is a premier development platform for Frontier

# FRONTIER PORTABLE PROGRAMMING WITH HIP

HIP (Heterogeneous-compute Interface for Portability) is an API developed by AMD that allows developers to write portable code to run on AMD or NVIDIA GPUs. It is a wrapper that uses the underlying CUDA™ or ROCm platform that is installed on a system

The API is very similar to CUDA so transitioning existing codes from CUDA to HIP is fairly straightforward.

AMD has developed a “hipify” tool that automatically converts source from CUDA to HIP.

Developers can specialize for the platform to tune for performance or handle unique cases

OLCF plans to make HIP available on Summit so that users can begin using it prior to its availability on Frontier

# EXPECTATIONS FOR POST-EXASCALE SYSTEMS



Argonne National Laboratory is a  
U.S. Department of Energy laboratory  
managed by UChicago Argonne, LLC.



# EXPECTATIONS FOR POST-EXASCALE SYSTEMS

- ALCF/NERSC/OLCF target bringing new systems every 4-5 years
- A future system may have 5-20x higher performance than exa-scale systems
- Overcoming the system power limitation will be very challenging
- We expect to see further development of accelerator-based systems.
- May see new types of accelerators
- Tighter integration between hosts and accelerators
- May see implementation of advanced photonic networking
- More use of solid-state storage for I/O and closer integration
- We highly recommend developers at the FASTMath Institute use open portable programming models, not proprietary models.
- Your effort for porting your software to exascale systems will be good investment even for post-exascale systems.

# Application Readiness Programs

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

# Application Readiness Programs

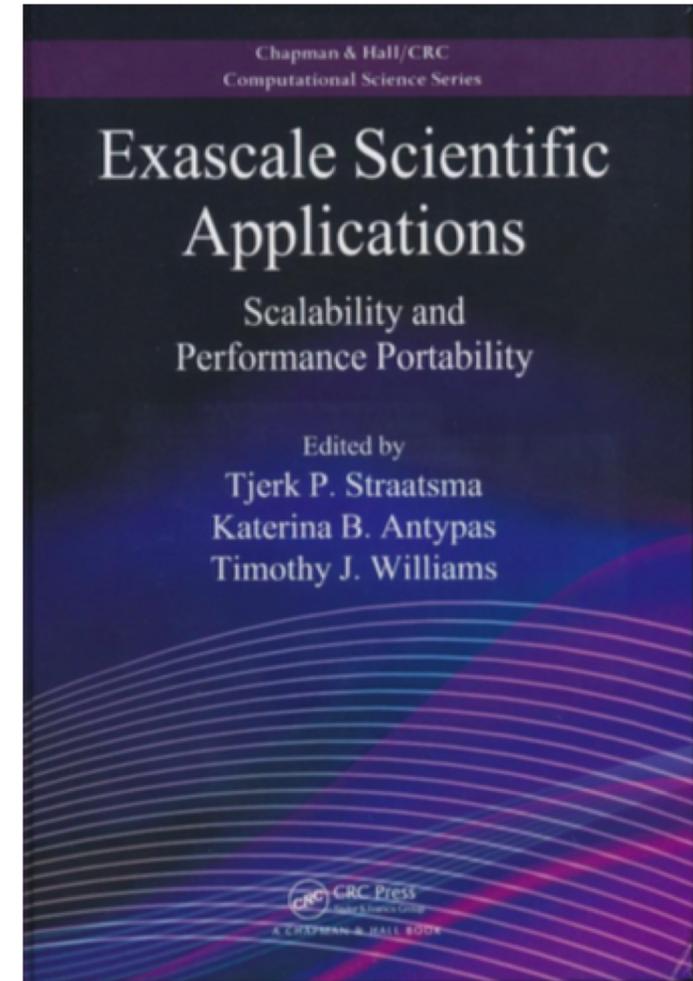
ALCF: *Early Science Program (ESP)*

NERSC: *NERSC Exascale Science Applications Program (NESAP)*

OLCF: *Center for Accelerated Application Readiness (CAAR)*

## Main Goals:

- Porting and Optimization Applications for Next Architectures
  - Support Current Applications on Future Systems
  - Develop Applications in Diverse Set of Science Domains to Expand User Programs
- Development Experience to Support Future Users and Developers
  - Focus on a Variety of Programming Models, Languages, etc.
  - Focus on Diverse Mathematical Models
  - Focus on Performance Portability
- Software Development Environment Testing
  - Development Environment for New Systems are Often Not Robust
- Hardware Hardening with Production Science Runs at Scale
  - Identify Hardware Stability Issues is Best Done with Runs at Scale

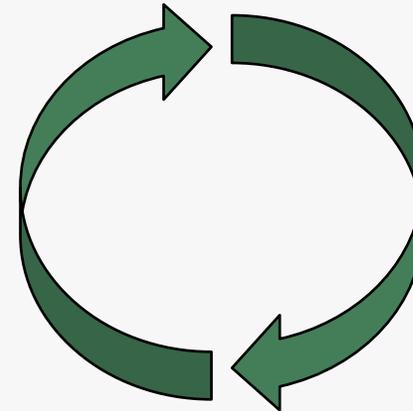


# ALCF Aurora Readiness Efforts

15 ALCF  
Early  
Science

22 ECP AD  
Projects

- Flat Profile at Scale
- Flat Profile for representative problem
- Characterize kernels of interest
- Use hardware specific advising tools
- Run kernels through simulators



- Prepare workflow technologies
- Optimize libraries, frameworks, and tools
- Harden SW stack

Using

Emulation

Early Hardware

Hardware

Feed-  
back

HW

SW

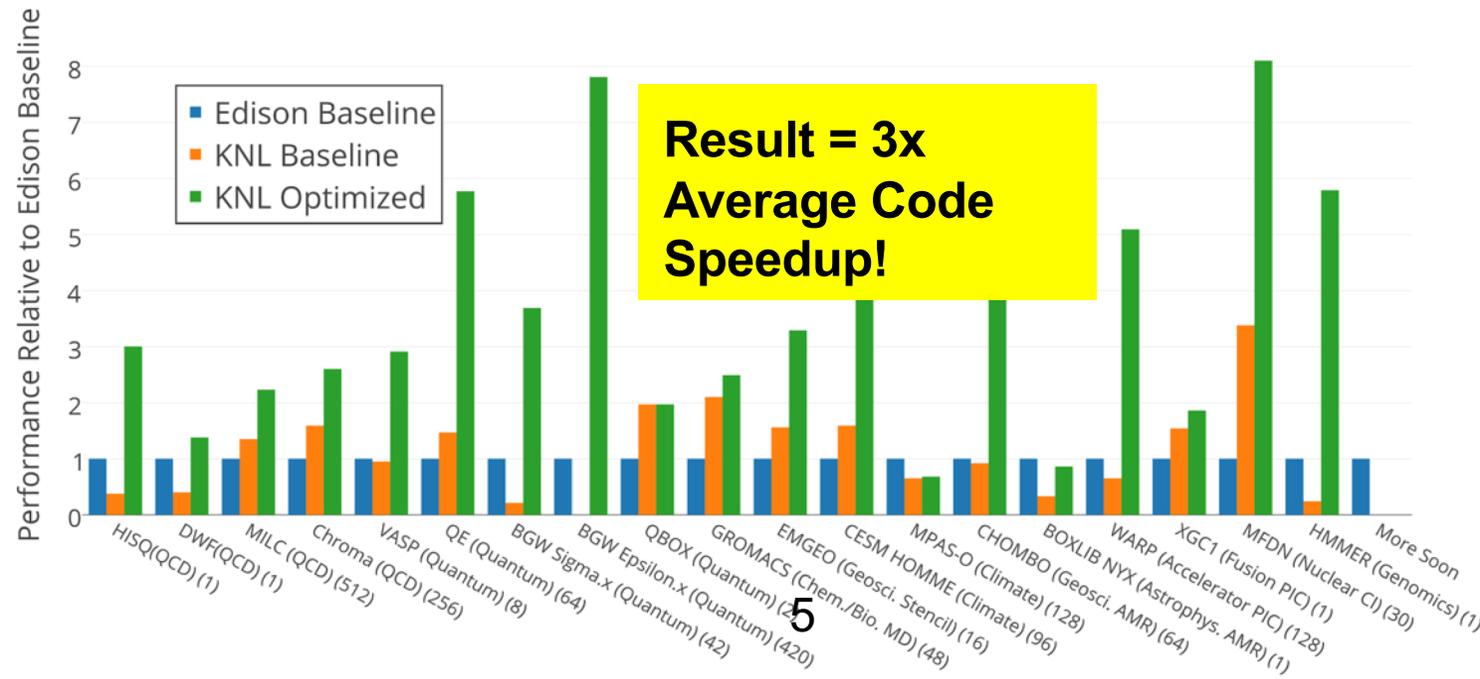
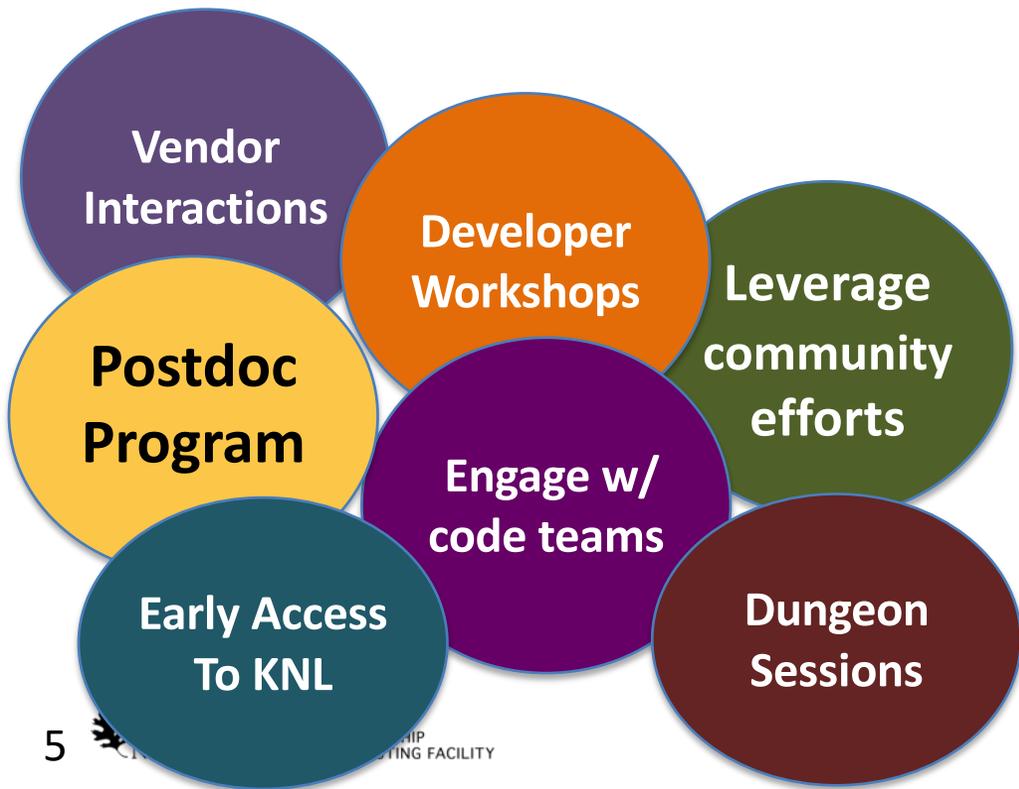
# ALCF Activities Already Underway

- A21 Applications Working Group (NDA)
  - Biweekly meetings
  - Representatives from many apps discuss hardware, software, and porting strategies
  - Source for best practices to be shared as soon as possible
- Community Engagement: Co-Design Workshop
  - First App Co-Design Workshop Sep. 19-21, 2018
  - Intel and Argonne applications staff working closely together on presentations and apps
  - Covered progress to-date and strategies moving forward
  - Model for future workshops and training

# NERSC Exascale Scientific Application Program (NESAP)



- Prepare DOE SC users for advanced architectures like Cori and Perlmutter
- Partner closely with 20-40 application teams and apply lessons learned to broad NERSC user community.



# OLCF Center for Accelerated Application Readiness (CAAR)

Begun in 2009 to prepare selected applications for accelerated systems (Titan, Summit, Frontier)

Frontier CAAR Call for Proposals closed June 8

CAAR Project Kickoff Meeting planned first week of October

<https://www.olcf.ornl.gov/frontier-center-for-accelerated-application-readiness/>

Staff also assisting with readiness for a subset of the 22 ECP applications



## 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

## Technical support from vendor Center of Excellence is crucial

- Programming environment often not mature
- Best source of information on new hardware features
- On-site staff from vendors

## Access to multiple resources, including early hardware

## Joint training activities

## Portability is a critical concern

# Application Readiness for Accelerators – Importance

- High performance on these systems is not possible without using the GPUs
  - Titan, 2012: 90% of flops, 80% of memory bandwidth on the GPUs
  - Summit, 2018: 98% of flops, 96% of memory bandwidth on the GPUs
- Applications that do not use the GPUs are decreasingly used on these systems
  - A computational chemistry code used earlier on Titan has fallen out of use, has now been replaced by QMCPACK by the project PI
- Trend of top systems increasingly using accelerators: “For the first time in history, most of the flops added to the TOP500 list came from GPUs instead of CPUs.” (June 2018)
- Added benefit: optimizing codes for accelerators also helps on conventional processors with similar features, e.g., vector units and hyperthreading
- Several Titan CAAR codes ran 2X faster even on conventional CPUs after restructuring for accelerators

# Porting to Accelerated Systems: Lessons Learned

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

# Lessons Learned: 1. Performance Portability

- Many HPC apps run at multiple centers, so the interest in performance portability is very high
- Several multi-lab workshops have been held to address this issue, <https://performanceportability.org/>
- Approaches:
  - Directives – OpenMP 4.0, 4.5, 5.0 (offload); OpenACC
  - Kokkos, RAJA abstraction layer libraries
  - Lower-level proprietary approaches – CUDA, ROCm, ..., sometimes with custom portability layer – may be required when high performance is critical
  - Drop-in math libraries for GPUs, e.g., BLAS
  - Combinations of the above
- We are pushing on the vendors to support standardized approaches across platforms

## Lessons Learned: 2. Code Porting Themes

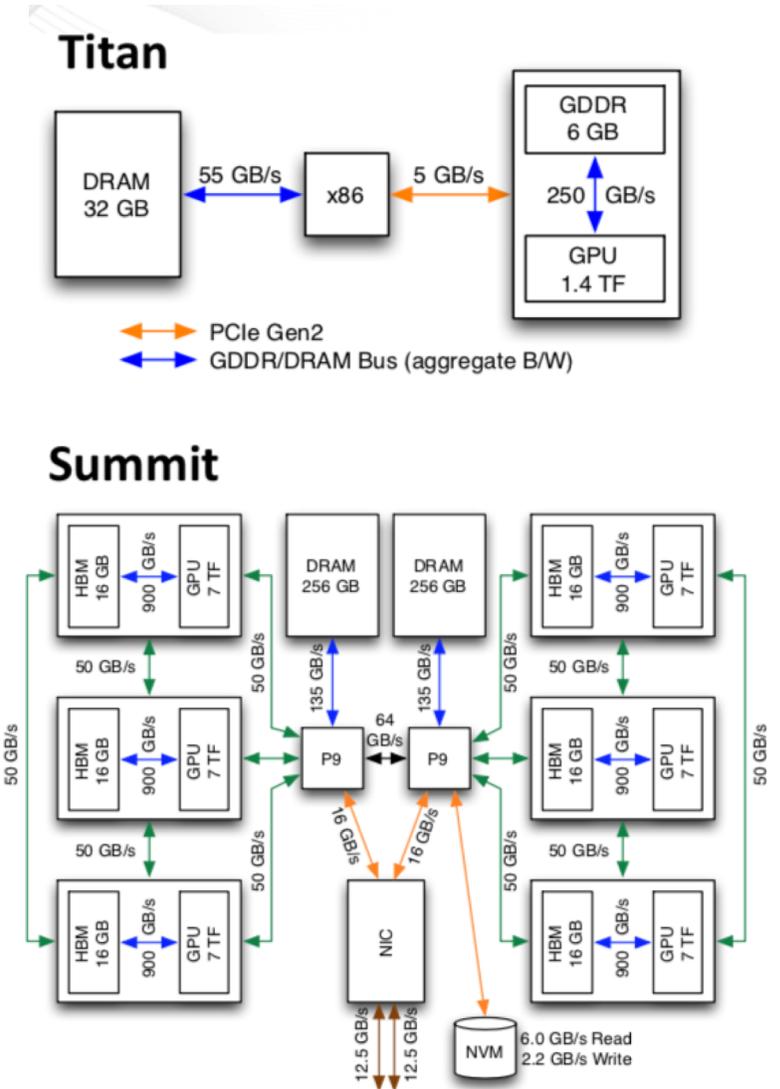
- Repeated themes across multiple apps in code porting work:
  - finding more threadable work for the GPU
  - Improving memory access patterns (stride-1; data reuse)
  - making GPU work (kernel calls) more coarse-grained if possible
  - making data on the GPU more persistent
- The cost of data transfers (MPI comm, CPU-GPU transfers, memory access) will continue to make it challenging to use the flops effectively
- Reducing transfers and asynchronous overlapping of transfers with compute can improve performance
- Some codes are not well-structured for data movement and require substantial refactoring

# Lessons Learned: 3. Level of Effort

- Our experience is 1-3 person-years required to port a full science application to an accelerated system
- Porting to a second accelerated system is generally easier
- Generally, 70%-80% of the work is spent on restructuring the code and algorithms for the accelerated architecture, regardless of target programming model
- Code changes that have global impact on the code are difficult to manage, e.g., data structure changes
- The difficulty level of the GPU port is in part determined by:
  - The code size (LOC)
  - Code execution profile—flat or hot spots
  - Structure of the algorithms—e.g., available parallelism, high computational intensity

# Lessons Learned: 4. System Heterogeneity

- Nodes are becoming more complex, heterogeneous
- Programming is more complex: multiple GPUs per node, CPU SMT threading, NUMA domains, MPS, coordination of host code threading and device selection, NVRAM
- Interoperability of compilers, programming models, libraries an issue
- Most common approach is: 1 MPI rank owns 1 GPU and a fraction of the CPU threads – this considerably simplifies the programming model
- Expect to see more heterogeneity with future systems – already seeing this with new hardware features like the NVIDIA Tensor Cores – will be a challenge for performance portability



# Application Readiness: Lessons Learned

OLCF staff Lessons Learned paper (2015) for our CAAR effort on Titan

Many of these lessons still applicable to newer systems (Summit, Frontier, ...)



Contents lists available at [ScienceDirect](#)

**Computers and Electrical Engineering**

journal homepage: [www.elsevier.com/locate/compeleceng](http://www.elsevier.com/locate/compeleceng)

---

**Accelerated application development: The ORNL Titan experience** ☆☆☆★

Wayne Joubert<sup>a,\*</sup>, Rick Archibald<sup>a</sup>, Mark Berrill<sup>a</sup>, W. Michael Brown<sup>a</sup>, Markus Eisenbach<sup>a</sup>, Ray Grout<sup>c</sup>, Jeff Larkin<sup>d</sup>, John Levesque<sup>b</sup>, Bronson Messer<sup>a</sup>, Matt Norman<sup>a</sup>, Bobby Philip<sup>a</sup>, Ramanan Sankaran<sup>a</sup>, Arnold Tharrington<sup>a</sup>, John Turner<sup>a</sup>

<sup>a</sup> Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831, United States  
<sup>b</sup> Cray, Inc., Knoxville, TN, United States  
<sup>c</sup> National Renewable Energy Laboratory, 15013 Denver West Parkway, Golden, CO 80401, United States  
<sup>d</sup> NVIDIA Corp., P.O. Box 2008, MS-6008, Oak Ridge, TN 37831, United States

# Questions?

