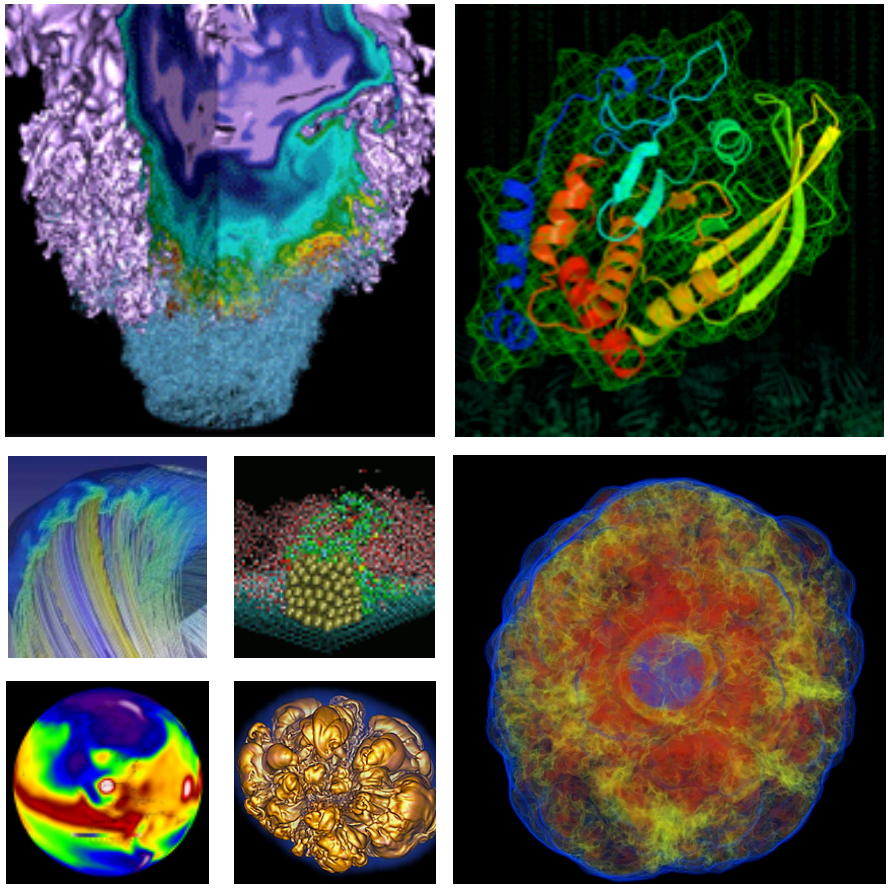
NERSC Now and Into the Future



Richard Gerber

NERSC Senior Science Advisor
HPC Department Head

March 22, 2016

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

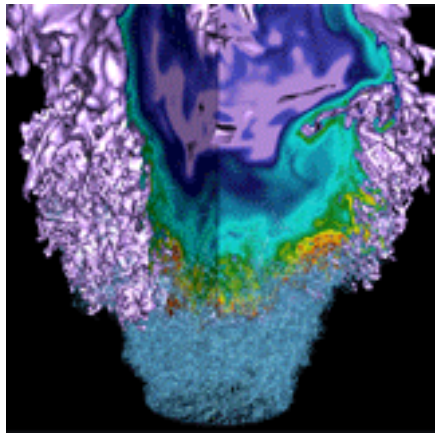
Outline

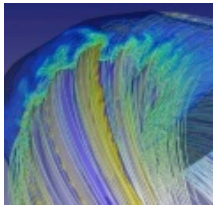
- **NERSC Today**
- **NERSC Tomorrow**
- **Computing Challenges**
- **On the Road to Exascale**

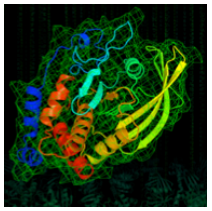
Katie Antypas' Talk on NERSC Data Systems and Strategy will Follow

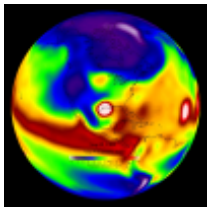
Thanks to Sudip Dosanjh, Katie Antypas and many others for some of these slides.

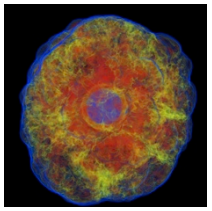
NERSC Today

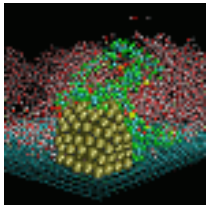












U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Background

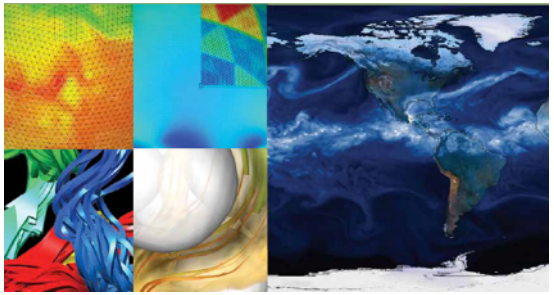
- **NERSC is the National Energy Research Scientific Computing Center**
- **Founded in 1974 at LLNL; moved to LBNL in 1996**
- **Devoted to open science in support of the DOE Office of Science mission**

NERSC's mission is to accelerate scientific discovery at the DOE Office of Science through high performance computing and data analysis.

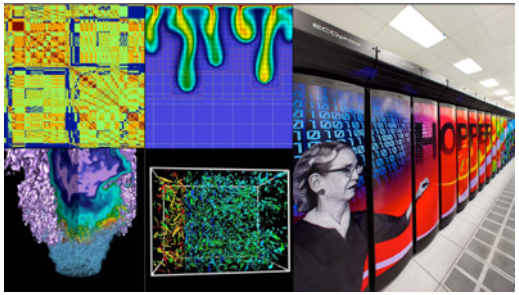
NERSC Provides HPC and Data Resources for DOE Office of Science Research

U.S. DEPARTMENT OF ENERGY | Office of Science

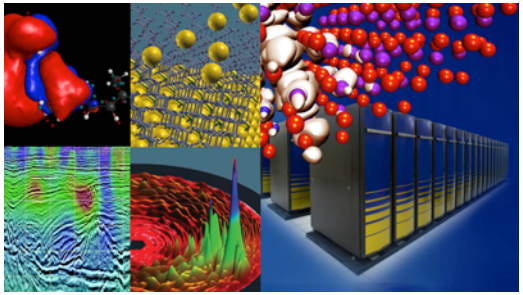
Largest funder of physical science research in U.S.



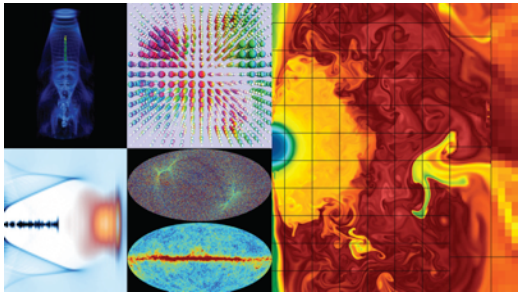
Biology, Environment



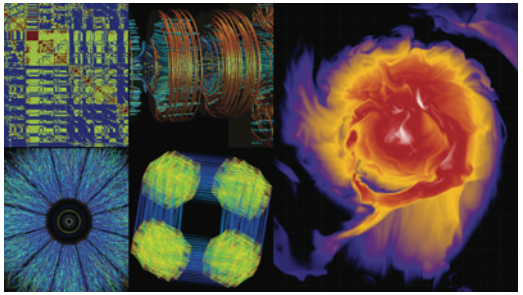
Computing



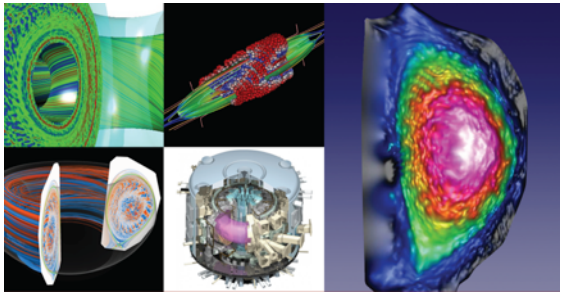
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



Nuclear Physics



Fusion Energy, Plasma Physics

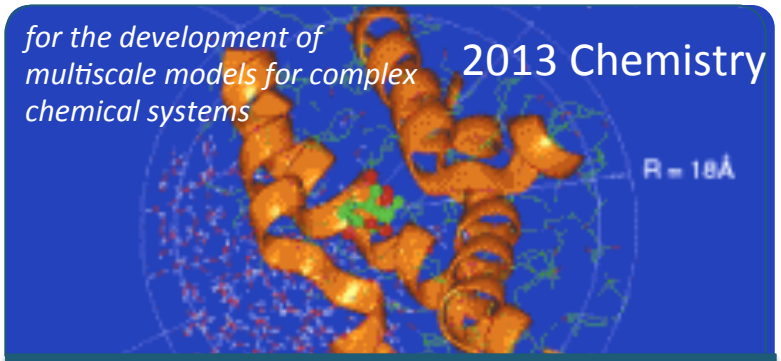
Focus on Science

- **NERSC supports the broad mission needs of the six DOE Office of Science program offices**
- **6,000 users and 750 projects**
- **Supercomputing and data users**
- **NERSC science engagement team provides outreach and POCs**

2,078 refereed publications in 2015

NERSC Nobels

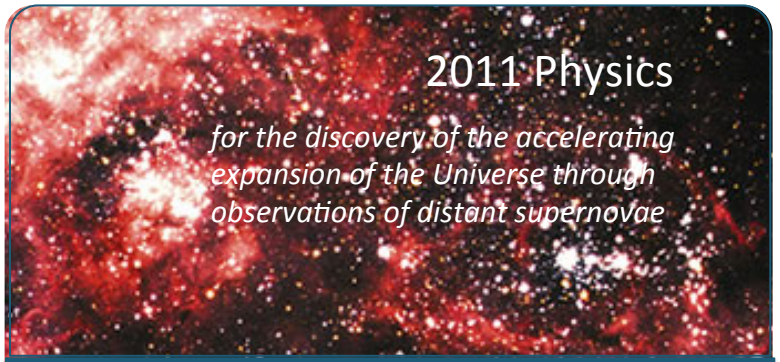
NERSC



for the development of multiscale models for complex chemical systems

2013 Chemistry

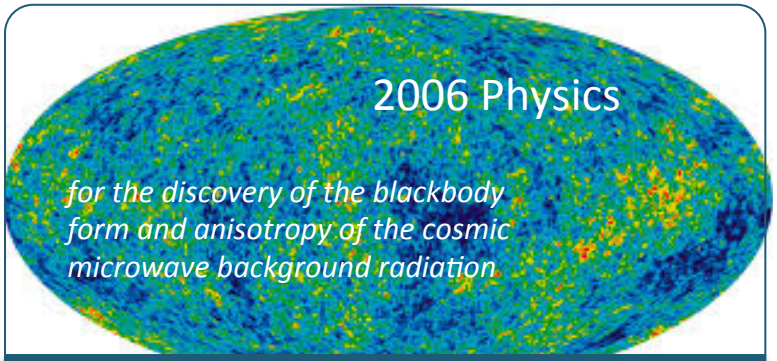
Martin Karplus



2011 Physics

for the discovery of the accelerating expansion of the Universe through observations of distant supernovae

Saul Perlmutter



2006 Physics

for the discovery of the blackbody form and anisotropy of the cosmic microwave background radiation

George Smoot

2007 Peace

for their efforts to build up and disseminate greater knowledge about man-made climate change, and to lay the foundations for the measures that are needed to counteract such change

Warren Washington

Nobel Prize in Physics 2015

Scientific Achievement

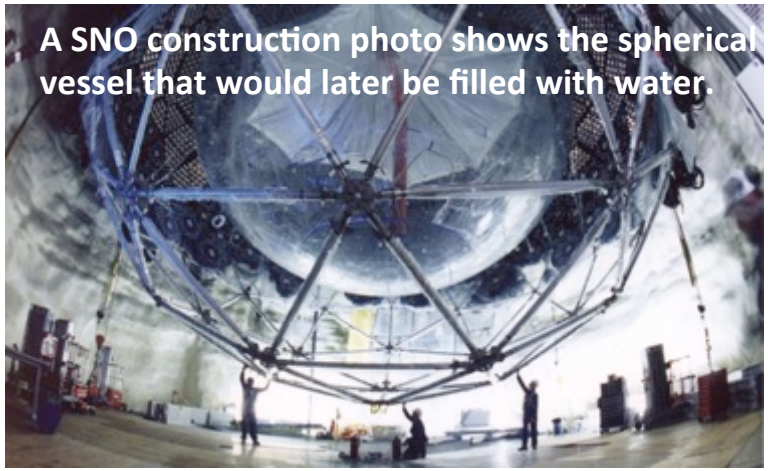
The discovery that neutrinos have mass and oscillate between different types

Significance and Impact

The discrepancy between predicted and observed solar neutrinos was a mystery for decades. This discovery overturned the Standard Model interpretation of neutrinos as massless particles and resolved the "solar neutrino problem"

Research Details

The Sundbury Neutrino Observatory (SNO) detected all three types (flavors) of neutrinos and showed that when all three were considered, the total flux was in line with predictions. This, together with results from the Super Kamiokande experiment, was proof that neutrinos were oscillating between flavors and therefore had mass



A SNO construction photo shows the spherical vessel that would later be filled with water.

NERSC helped the SNO team use PDSF for critical analysis contributing to their seminal PRL paper. HPSS serves as a repository for the entire 26 TB data set.

Q. R. Ahmad et al. (SNO Collaboration). Phys. Rev. Lett. 87, 071301 (2001)

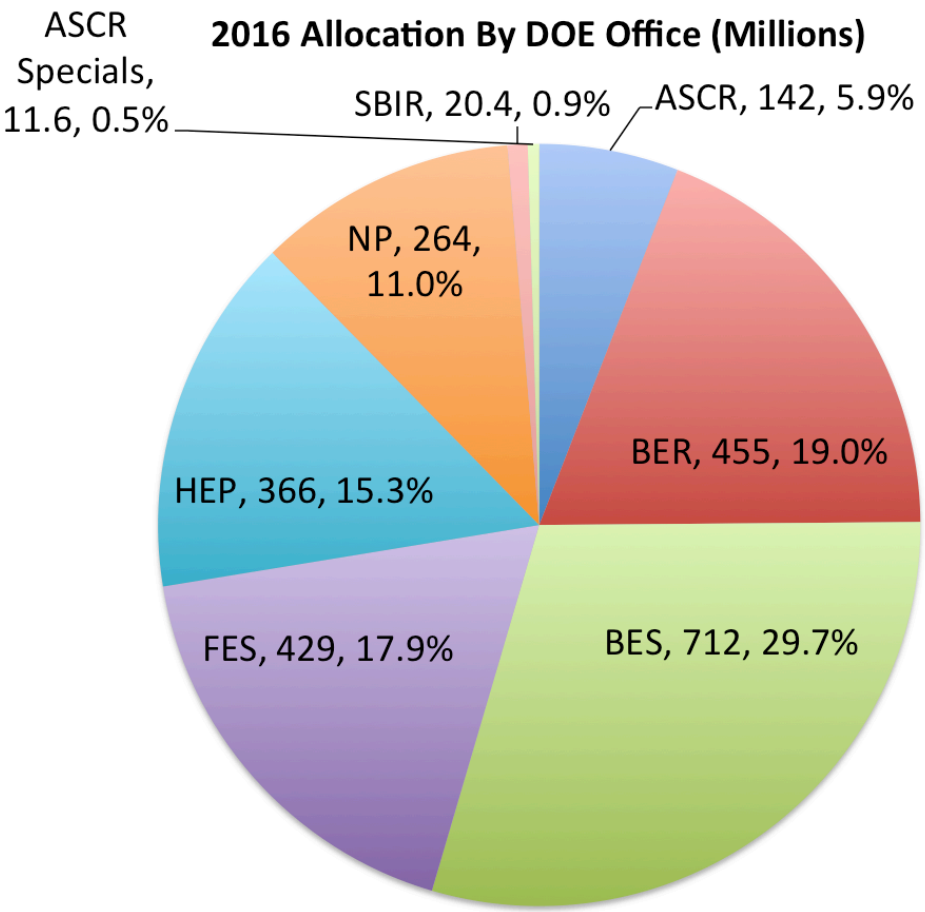
Nobel Recipients: Arthur B. McDonald, Queen's University (SNO)
Takaaki Kajita, Tokyo University (Super Kamiokande)

U.S. DEPARTMENT OF ENERGY | Office of Science

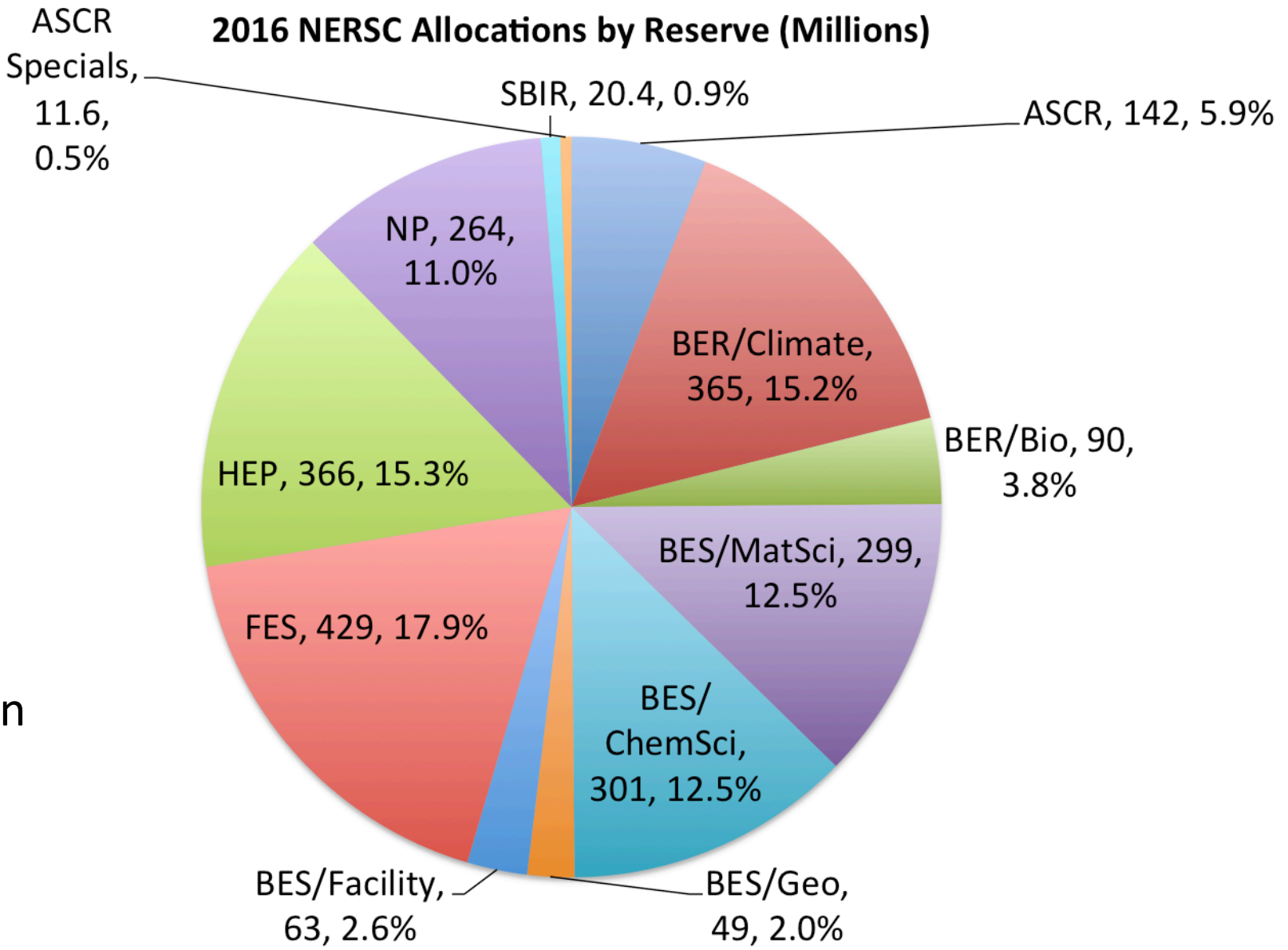
BERKELEY LAB

NERSC directly supports DOE's science mission

- DOE SC offices allocate 80% of NERSC resources
- ASCR Leadership Computing Challenge 10%
- NERSC Director's Reserve 10%

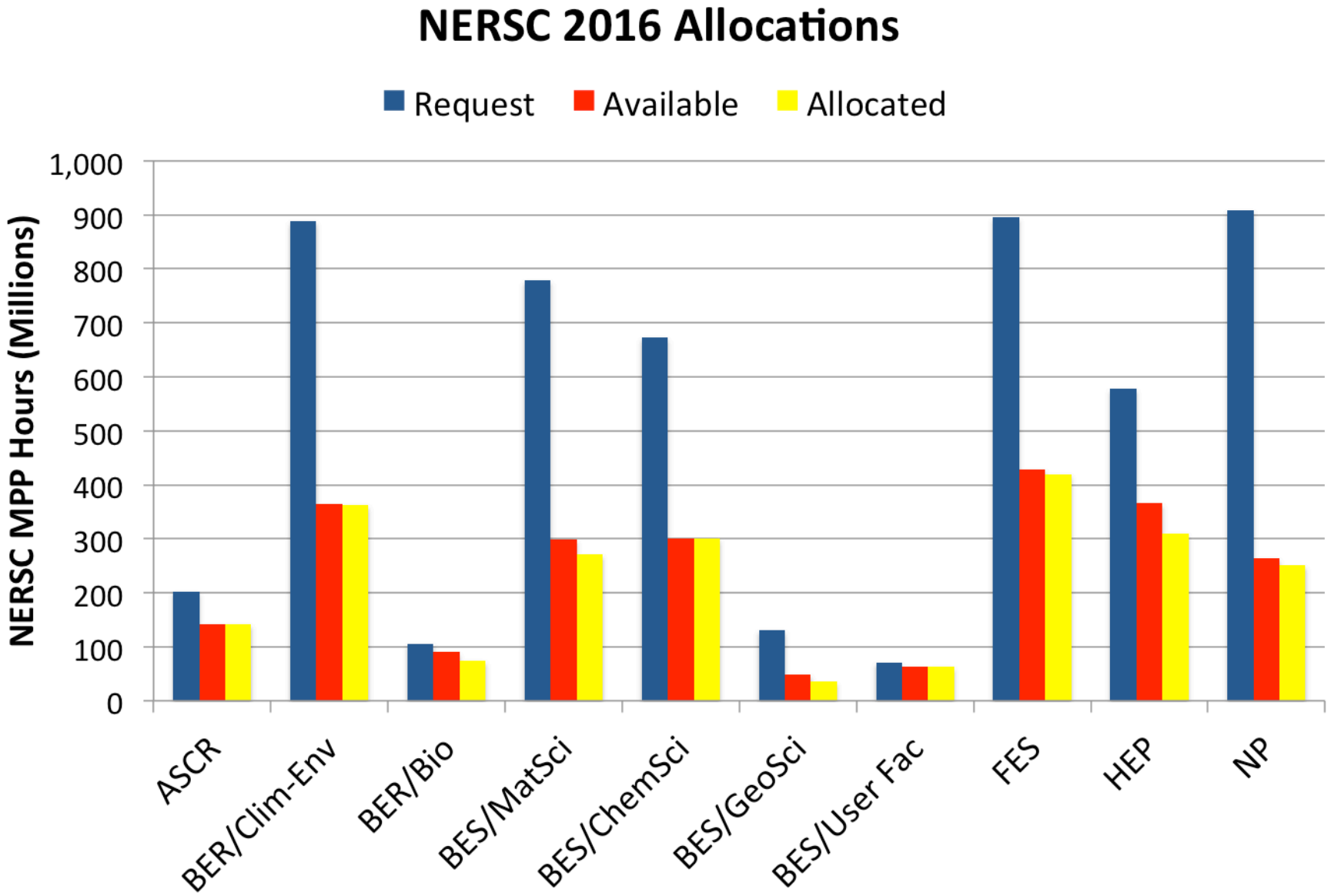


Distribution Among Reserves for 2016



2,400 Million Total

Insatiable Demand for Hours



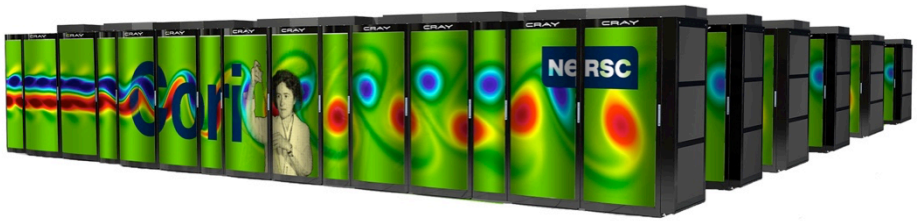
NERSC's Current Big System is Edison

- Edison is the HPCS* demo system (serial #1)
- First Cray Petascale system with Intel processors (Ivy Bridge), Aries interconnect and Dragonfly topology
- Very high memory bandwidth (100 GB/s per node), interconnect bandwidth and bisection bandwidth
- 5,576 nodes, 133K cores, 64 GB/node
- Exceptional application performance

*DARPA High Productivity Computing System program

Cori Phase 1

- **Went into production Jan. 12, 2016**
- **Officially accepted this week**
- **1,630 Compute Nodes (52,160 cores)**
  - Two Haswell processors/node
  - 16 cores/processor at 2.3 GHz
  - 128 GB DDR4 2133 MHz memory/ node
- **Cray Aries high-speed "dragonfly" topology interconnect**
- **22 login nodes for advanced workflows and analytics**
- **SLURM batch system**
- **Lustre File system**
  - 28 PB capacity, >700 GB/sec peak performance
  - 7.6 PB, 84/42/42 on Edison



NERSC moved to Wang Hall (CRT) in late 2015

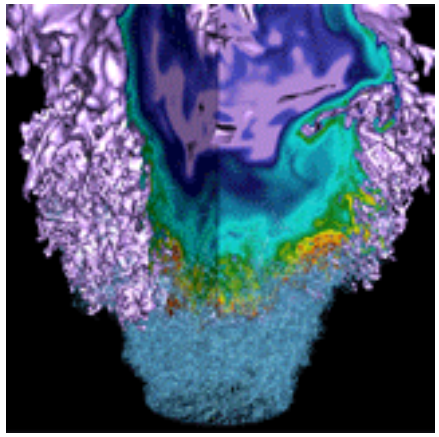
- **Four story, 140,000 GSF, 300 offices, 20Ksf HPC floor, 12.5->40 MW**
- **Located for collaboration**
  - LBNL, CRD, Esnet, UCB
- **Exceptional energy efficiency**
  - Natural air and water cooling
  - Heat recovery
  - PUE < 1.1

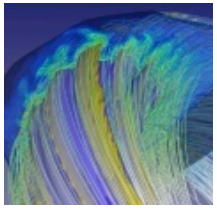
U.S. DEPARTMENT OF ENERGY | Office of Science

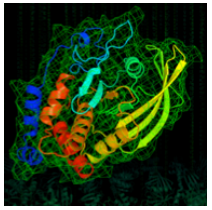
Sign up out front for 4:00 tour

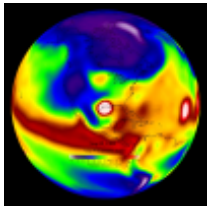
BERKELEY LAB

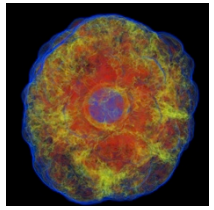
NERSC Tomorrow

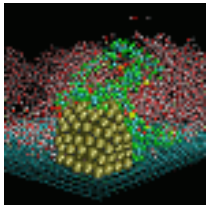












U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

NERSC Users' Needs from Requirements Reviews

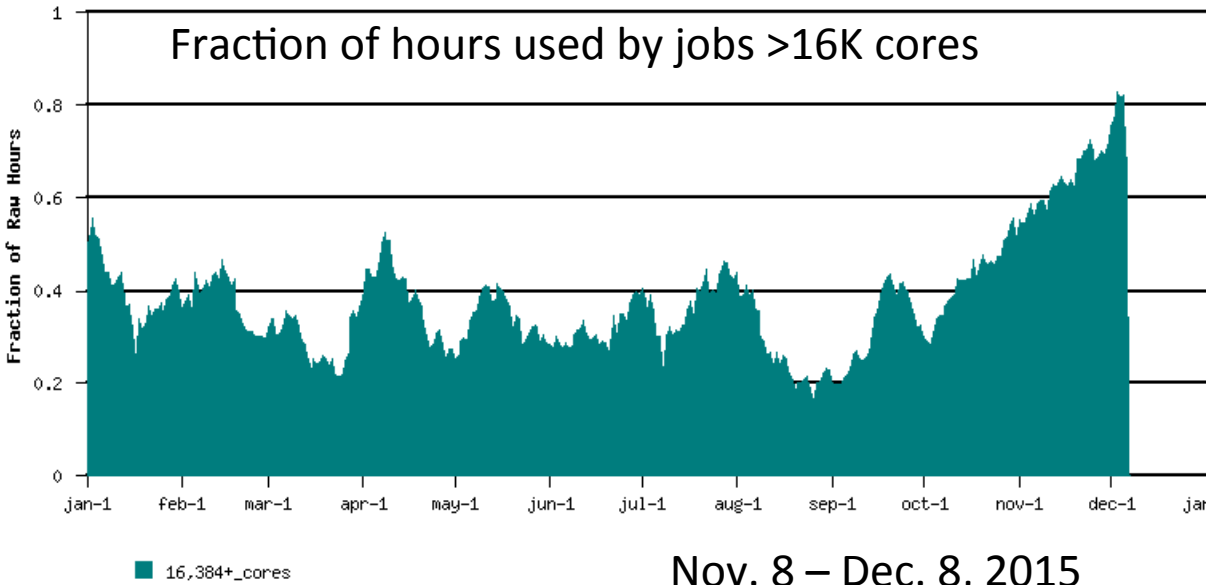
Many more hours to support science

Scalable systems that support jobs at all scales

Support for data-intensive science and complex workflows

Help porting to and using advanced architectures

Large Scale Computing and Data Requirements for Each Office 2014, 2017, 2020/2025



U.S. DEPARTMENT OF ENERGY | Office of Science

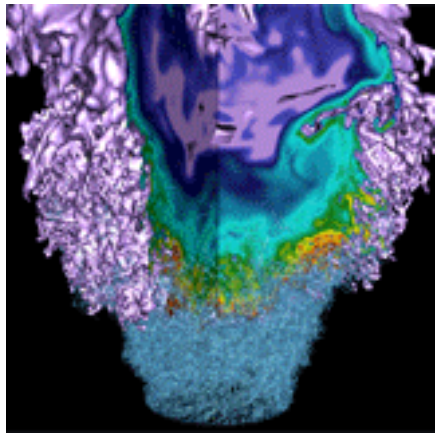
The NERSC-8 System: Cori

- **Cori will support the broad Office of Science research community and begin to transition the workload to more energy efficient architectures**
- **Cray XC system with over 9,300 Intel Knights Landing compute nodes – mid 2016**
  - Self-hosted, (not an accelerator) many-core processor with 68 cores per node
  - On-package high-bandwidth memory, 16 GB, 4-5X node memory BW
  - Low-latency (~1µs), high bandwidth (~15 GB/s) Cray Aries network
- **Data Intensive Science Support – Cori Phase 1**
  - 10 Haswell processor cabinets (Phase 1) to support data intensive applications
- **Robust Application Readiness Plan – NESAP**
  - Outreach and training for user community
  - Application deep dives with Intel and Cray
  - 8 post-docs integrated with key application teams

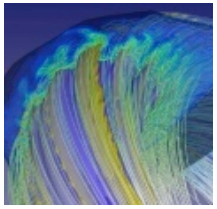
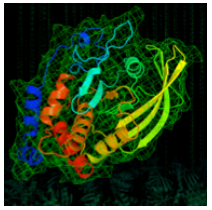
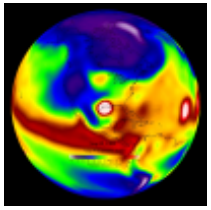
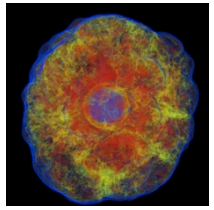
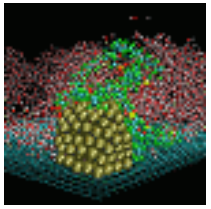
Intel "Knights Landing" Processor

- **Next generation Xeon-Phi, >3TF peak (vs. 460 GF/node Edison)**
  - Single socket processor - Self-hosted, not a co-processor, not an accelerator, one per node
  - 68 cores per processor with support for four hardware threads each; more cores than current generation Intel Xeon Phi™
  - 512b vector units (32 flops/clock – AVX 512)
  - High bandwidth on-package memory (16GB) 5X bandwidth of off-package DDR4 DRAM (96 GB)
- **Presents an application porting challenge to efficiently exploit KNL performance features**

Computing Challenges



NERSC

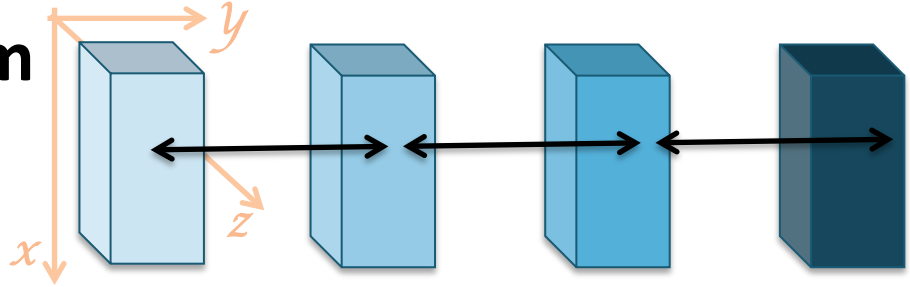






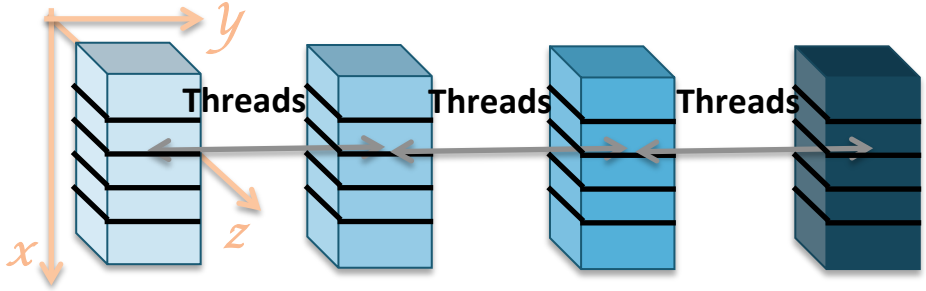
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

To run effectively on Cori users will have to:

- **Manage Domain Parallelism**
  - independent program units; explicit
- **Increase Thread Parallelism**
  - independent execution units within the program; generally explicit
- **Exploit Data Parallelism**
  - Same operation on multiple elements
- **Improve data locality**
  - Cache blocking; Use on-package memory

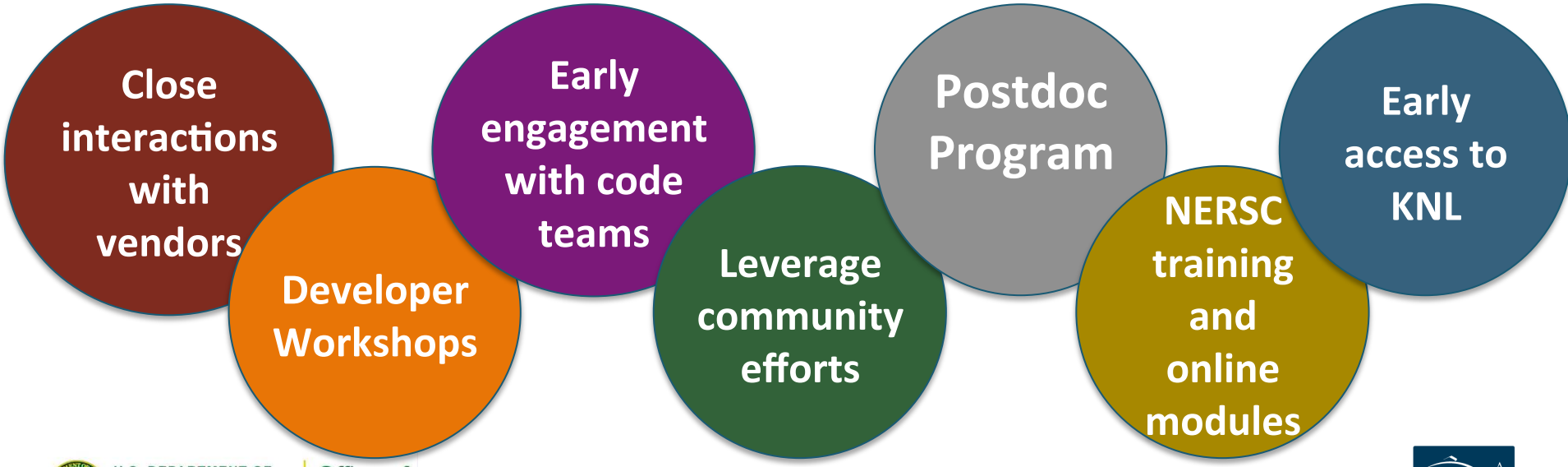





```
|--> DO I = 1, N
|       R(I) = B(I) + A(I)
|--> ENDDO
```

NERSC Exascale Science Application Program

- **Goal: Prepare DOE Office of Science user community for Cori many-core architecture**
- **Partner closely with ~20 application teams and apply lessons learned to broad NERSC user community**
- **NESAP activities include:**



Intel Xeon Phi User Group (IXPUG)

- A forum for the free exchange of information and ideas that enhance the usability and efficiency of scientific applications running on large Xeon Phi-based High Performance Computing (HPC) systems
- NERSC hosted IXPUG 2015 in Sept. at the CRT facility
- Over 100 attendees
- Week long community event with training sessions, hack-a-thons and technical briefings and community meetings

Application Portability

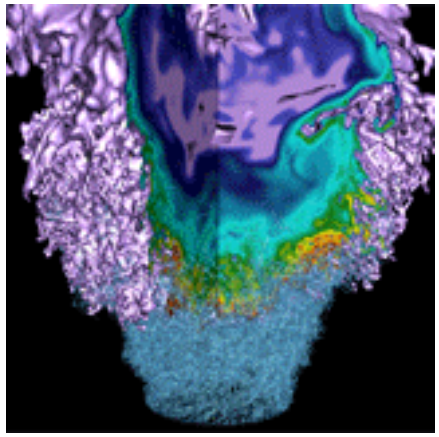
NERSC, OLCF and ALCF have been partnering on application portability and have held a number of workshops in the past 18 months.

Workshop/Meeting	Topic	Date	Location
Application Portability Kick-off meeting	Briefing NERSC, OLCF and ALCF on the other's architectures	Mar. 2014	Oakland, CA NERSC Facility
Application Portability	Programming models for each next generation system. Coordinating NESAP, CAAR and ESP projects	Sept. 2014	Oakland, CA NERSC Facility
Application Portability II	Vendors briefing on tools and programming models for portability (NNSA participants included)	Jan. 2015	Oak Ridge, TN
HPCOR (HPC Operational Review) on Application Performance Portability	Workshop with ~100 participants discussing best practices, emerging practices and opportunities for application performance portability	Sept. 2015	Bethesda, MD
HPC Portability Workshop at SC15	Papers presented on application portability, followed by discussion.	Nov. 2015	SC 15 in Austin, TX

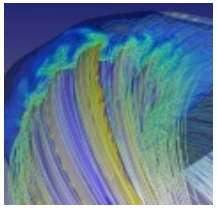
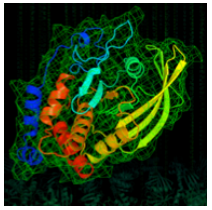
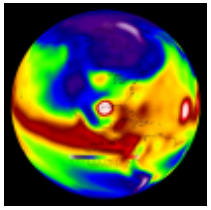
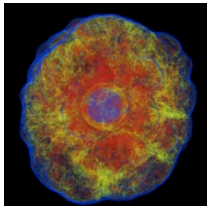
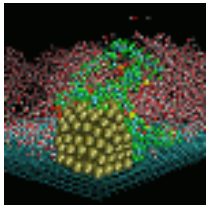
Next steps – New collaboration running from March 2016 – March 2018

- NERSC has new hire identified to work on Application Portability
- Each facility will choose a 'mini-app' and optimize it for their home system
- Then will work with cross facility team, to test portable options on different architecture
- Project timed to correspond with availability of prototype hardware and new testbeds

On the Road to Exascale



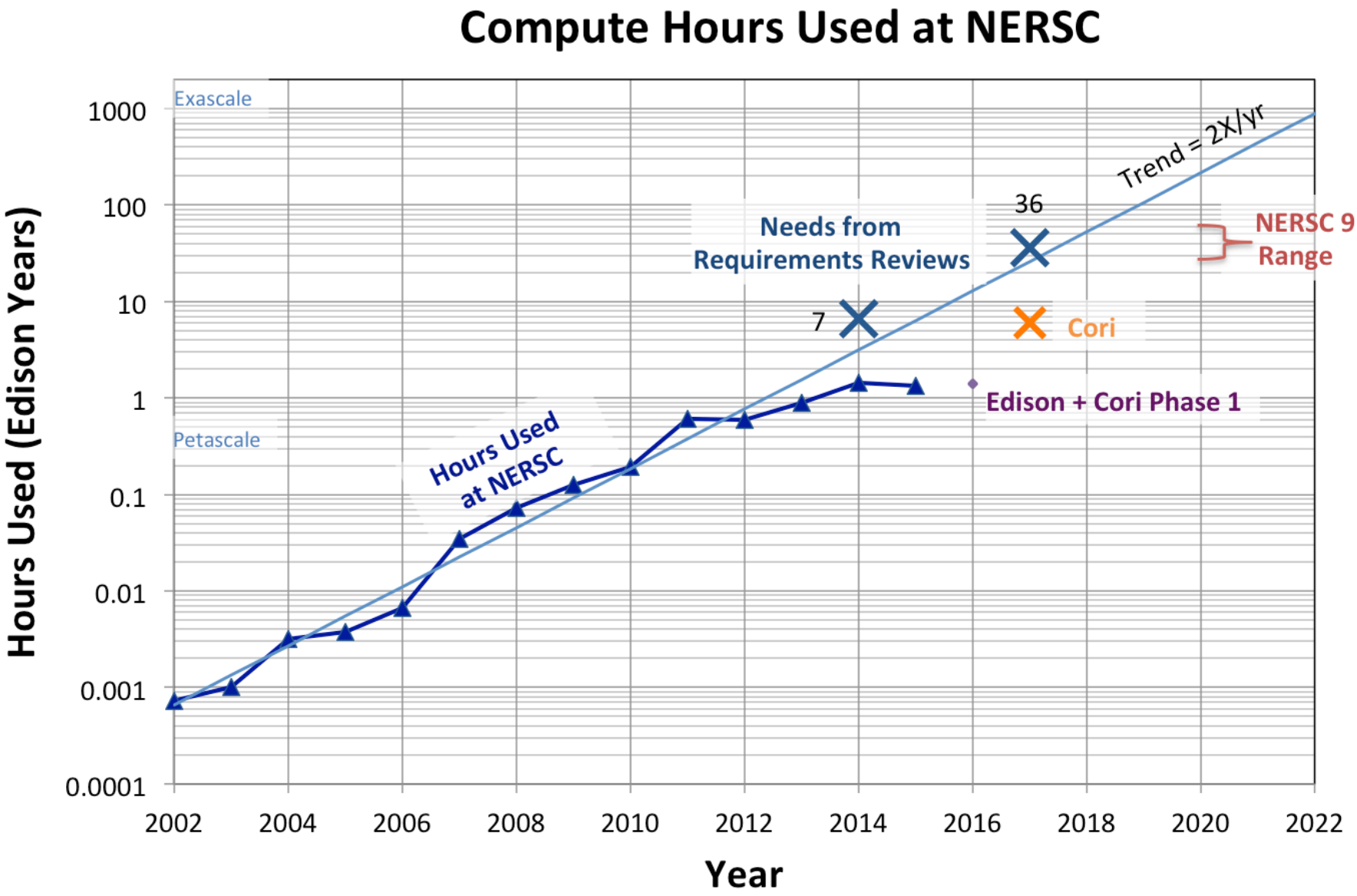
NERSC

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Keeping Up with User Needs Will Be a Challenge



Goals and Objectives for the NERSC-9 Project

1. Provide a significant increase in computational capabilities over the Edison system, 16-25x on a set of representative DOE benchmarks.
2. Meet the needs of extreme computing and data users by accelerating workflow performance.
3. Provide a vehicle for the demonstration and development of exascale-era technologies
4. Delivery in the 2020 time frame

NERSC 9 Activities

- CD0 signed August 24, 2015
- RFP draft technical specs released Nov. 10, 2015
  - Updated V2.0 released March 11, 2016
- Design Review Jan. 19-20, 2016
- Independent Project Review (IPR) Q2CY16
- RFP released late Spring/early Summer 2016

APEX

Alliance for Performance at Extreme Scale – NERSC, LANL, Sandia

ACES (LANL+Sandia) and NERSC have been collaborating since 2010

- **Informal collaboration on Hopper and Cielo**
- **The first formal DOE laboratory collaboration on system acquisition resulted in the procurement of Trinity and Cori (joint RFP, separate contracts)**
- **We have formed the Alliance for application Performance at the Extreme Scale (APEX)**
  - Crossroads and N9 will be deployed in 2020

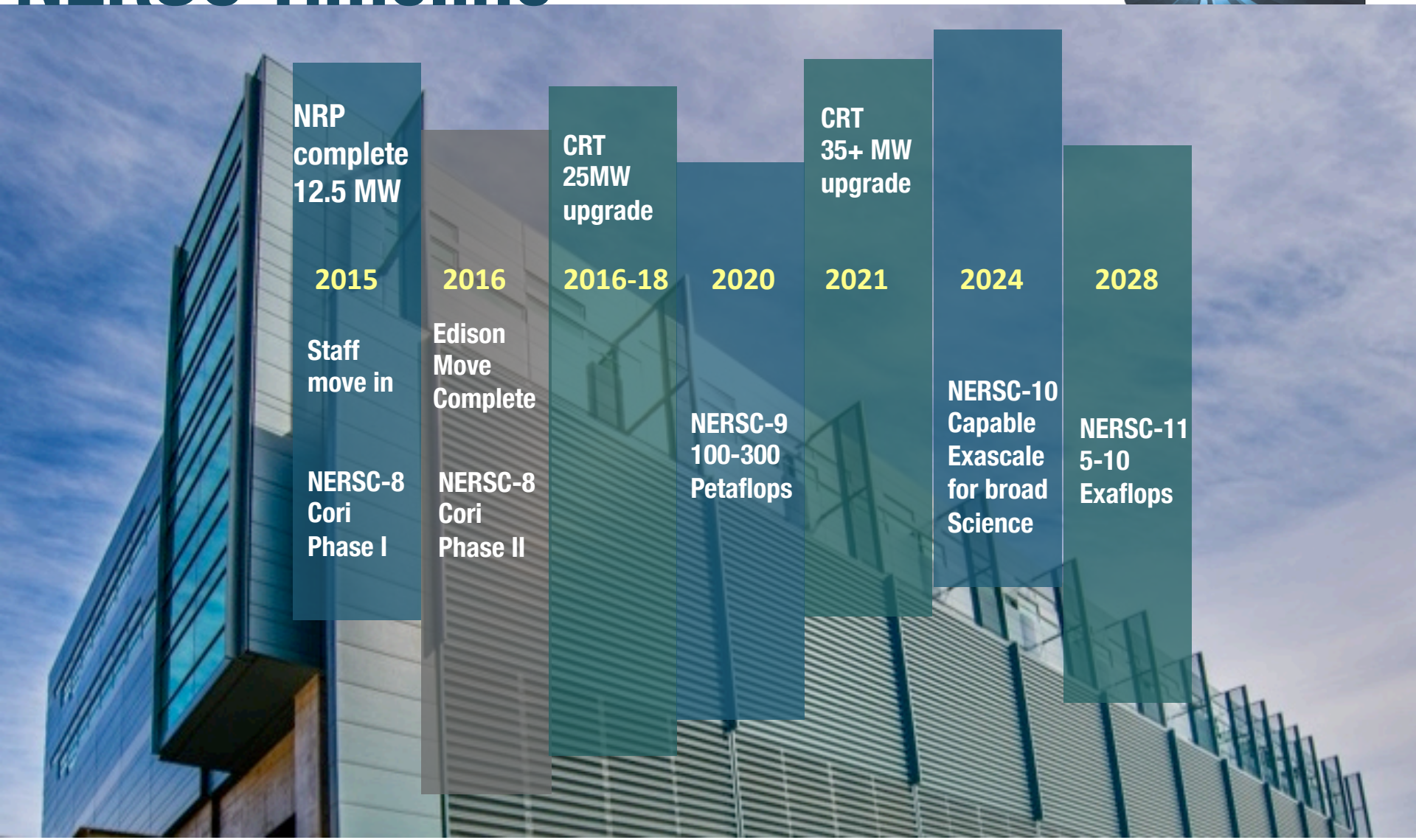
APEX

ALLIANCE FOR APPLICATION PERFORMANCE AT EXTREME SCALE

Benefits of the APEX Collaboration

- A deeper and more productive partnership with vendors
- Risk mitigation to vendors, which allows for a deeper and more fruitful consideration of all technology alternatives
- Broader and different perspectives from partnering labs brought to the technology evaluation process.
- Deeper pool of expertise in system procurement, deployment, and integration activities.

NERSC Timeline



U.S. DEPARTMENT OF ENERGY | Office of Science

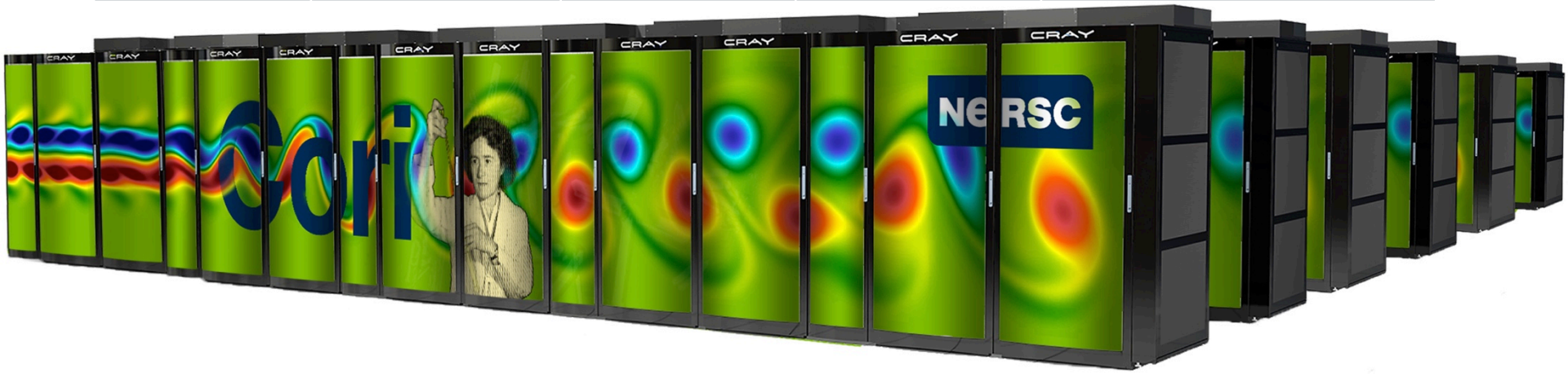
BERKELEY LAB

Summary

- NERSC is moving toward exascale technologies to support the needs of the DOE Office of Science research needs
- Cori (Phase 2), coming this summer, introduces energy-efficient processors to the large-scale NERSC community
- NERSC 9 will provide 16-25X Edison in ~2020
- Exascale for the broad community in ~2024
- NERSC is helping its users take on exascale programming challenges through NESAP: Talk later today.
- Data science and management challenges and opportunities abound: See next talk!

NERSC Systems Timeline

2007/2009	NERSC-5	Franklin	Cray XT4	102/352 TF
2010	NERSC-6	Hopper	Cray XE6	1.28 PF
2014	NERSC-7	Edison	Cray XC30	2.57 PF
2016	NERSC-8	Cori	Cray XC	30 PF
2020	NERSC-9			100PF-300PF
2024	NERSC-10			1EF



NERSC is teaming with the ALCF and OLCF to lead Exascale Requirements Reviews

Office	Date
HEP	**September 2015**
BES	**November 2015**
FES	**January 2016**
BER	March 2016
NP	June 2016
ASCR	September 2016

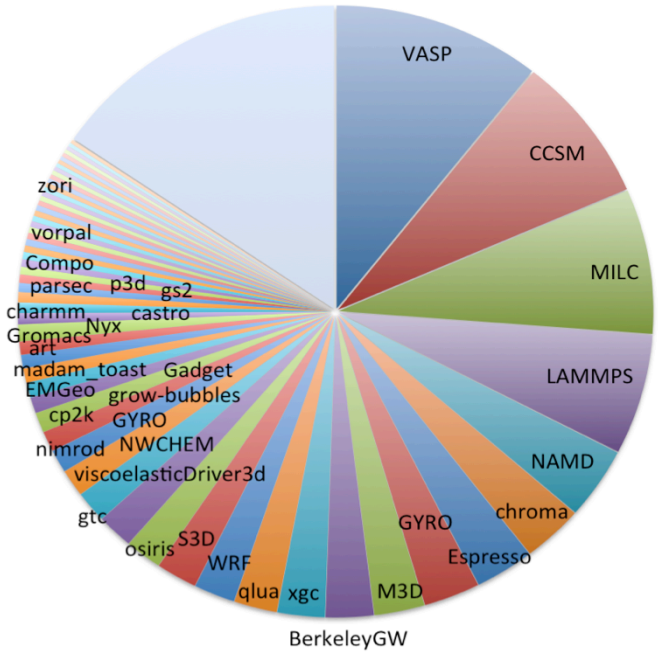
NERSC's Previous requirements reviews

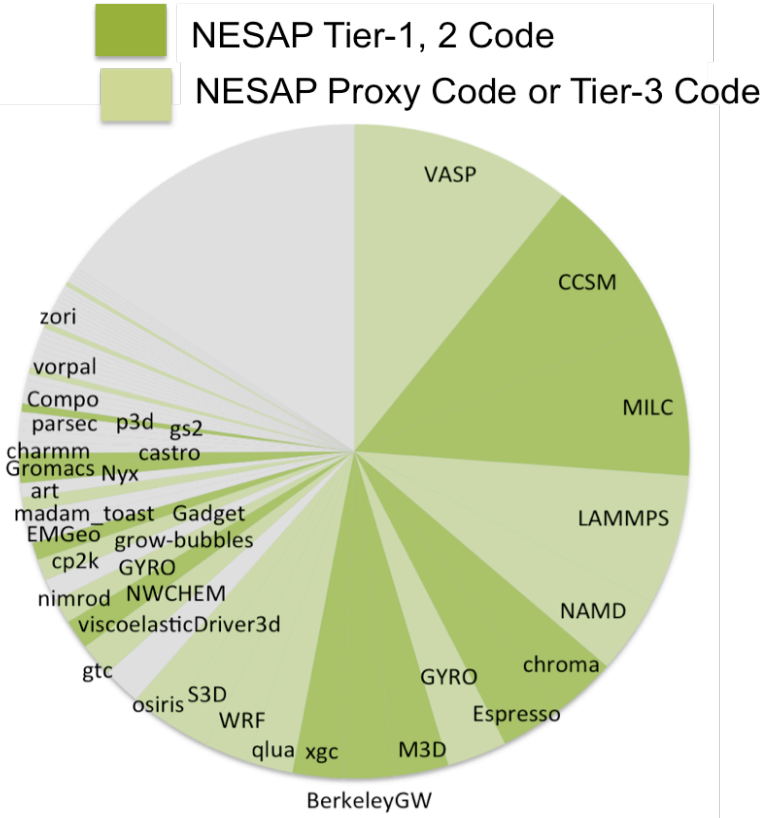
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

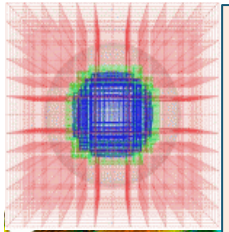
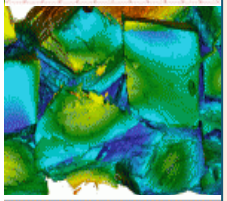
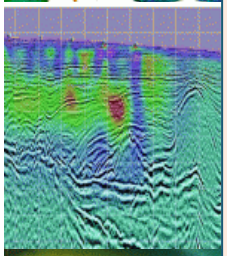
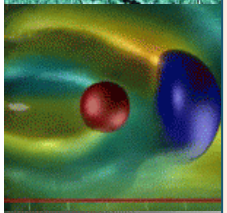
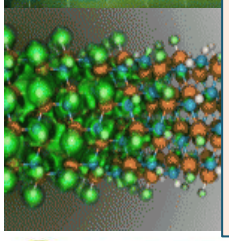
Code Coverage

Breakdown of Application Hours on Hopper and Edison 2013





NESAP Codes

Advanced Scientific Computing Research

Almgren (LBNL) **BoxLib AMR Framework**

Trebotich (LBNL) **Chombo-crunch**

High Energy Physics

Vay (LBNL) **WARP & IMPACT**

Toussaint(Arizona) **MILC**

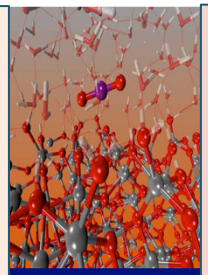
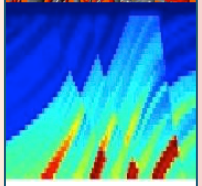
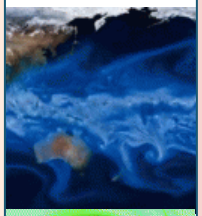
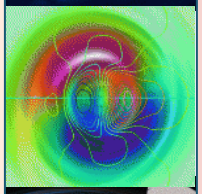
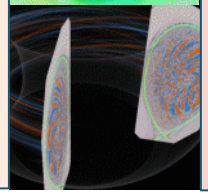
Habib (ANL) **HACC**

Nuclear Physics

Maris (Iowa St.) **MFDn**

Joo (JLAB) **Chroma**

Christ/Karsch (Columbia/BNL) **DWF/HISQ**

Basic Energy Sciences

Kent (ORNL) **Quantum Espresso**

Deslippe (NERSC) **BerkeleyGW**

Chelikowsky (UT) **PARSEC**

Bylaska (PNNL) **NWChem**

Newman (LBNL) **EMGeo**

Biological and Environmental Research

Smith (ORNL) **Gromacs**

Yelick (LBNL) **Meraculous**

Ringler (LANL) **MPAS-O**

Johansen (LBNL) **ACME**

Dennis (NCAR) **CESM**

Fusion Energy Sciences

Jardin (PPPL) **M3D**

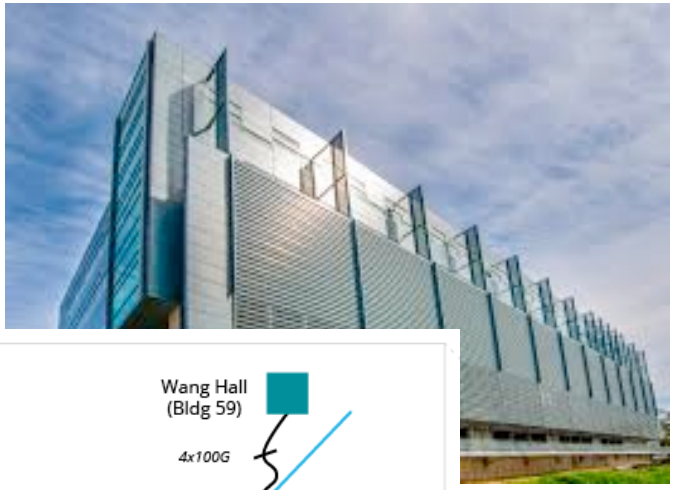
Chang (PPPL) **XGC1**

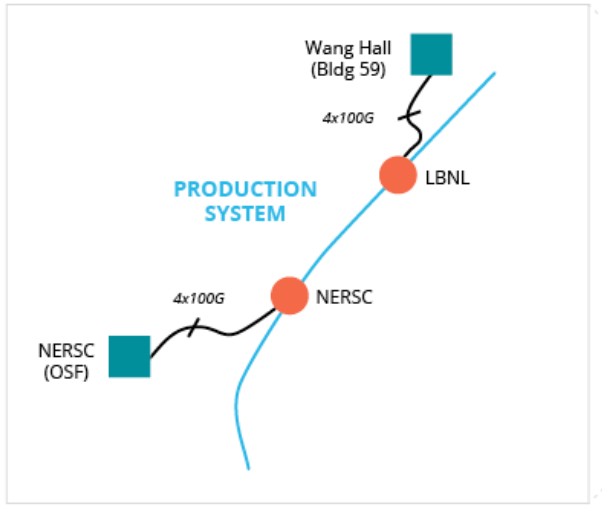
Resources for Code Teams

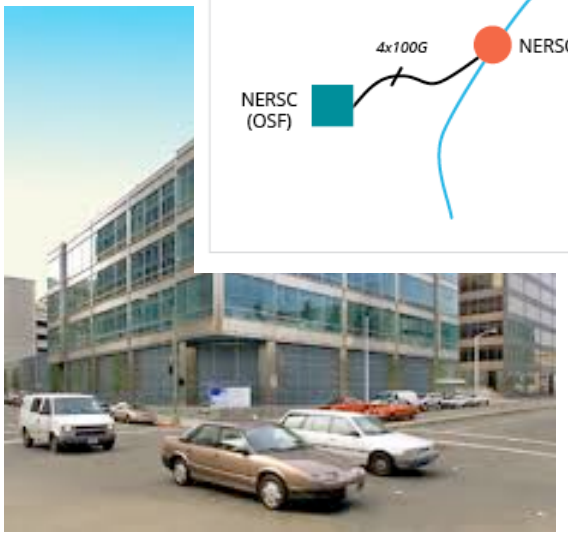
- **Early access to hardware**
  - Access to Babbage (KNC cluster) and early "white box" test systems expected in early 2016
  - Early access and significant time on the full Cori system
- **Technical deep dives**
  - Access to Cray and Intel staff on-site staff for application optimization and performance analysis
  - Multi-day deep dive ('dungeon' session) with Intel staff at Oregon Campus to examine specific optimization issues
- **User Training Sessions**
  - From NERSC, Cray and Intel staff on OpenMP, vectorization, application profiling
  - Knights Landing architectural briefings from Intel
- **NERSC Staff as Code Team Laisons (Hands on assistance)**
  - New Application Performance Group
- **Postdocs (6 of 8 hired) embedded with Tier 1 projects**

We were able to minimize downtime for users during the move

- 400 Gb/sec link between Oakland and Berkeley allowed file systems to be migrated live, without downtime
- Network was down for a total of 50 minutes during the move
- Edison took 5 weeks to uninstall, move, reinstall, retest

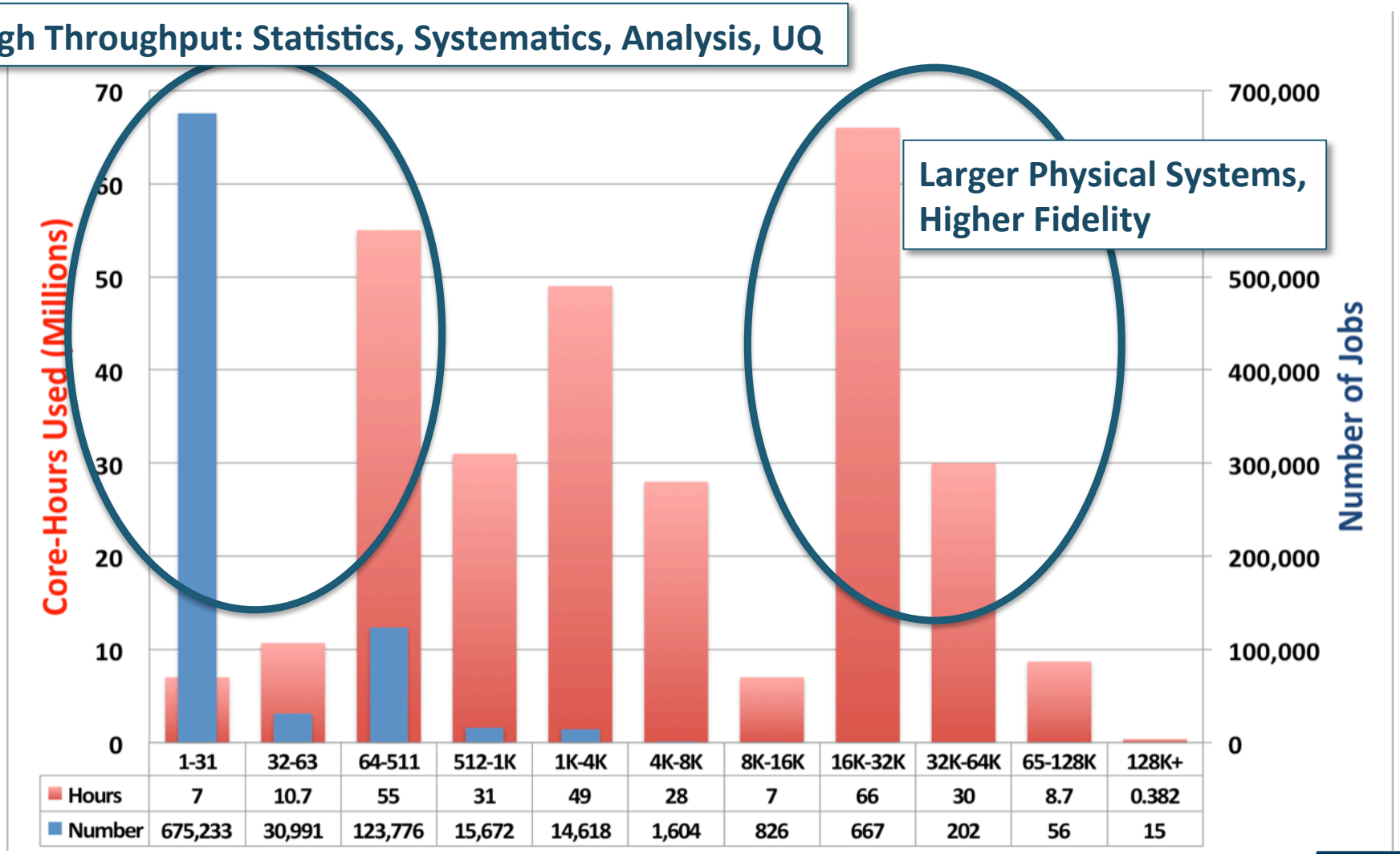






NERSC Supports Jobs of all Kinds and Sizes

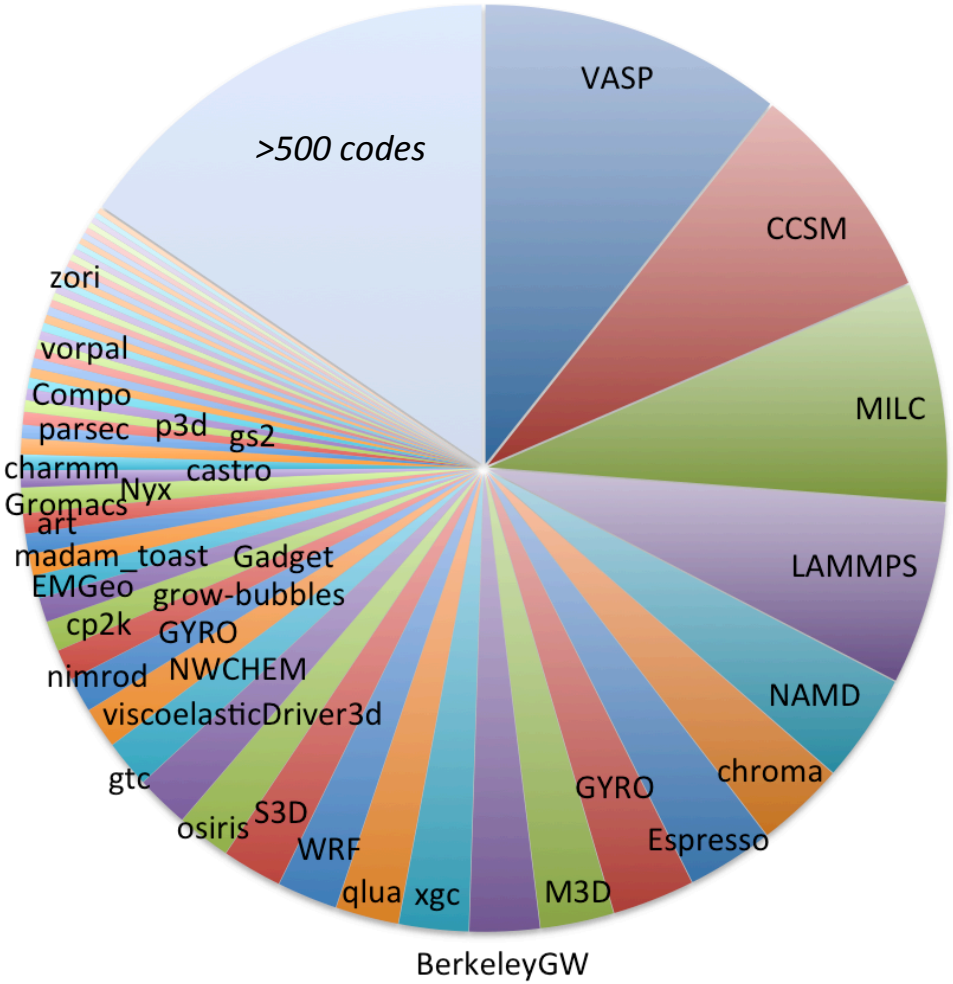
High Throughput: Statistics, Systematics, Analysis, UQ



	1-31	32-63	64-511	512-1K	1K-4K	4K-8K	8K-16K	16K-32K	32K-64K	65-128K	128K+
Hours	7	10.7	55	31	49	28	7	66	30	8.7	0.382
Number	675,233	30,991	123,776	15,672	14,618	1,604	826	667	202	56	15

We are initially focusing on 20 codes

Breakdown of Application Hours on Hopper and Edison 2013



- 10 codes make up 50% of the workload
- 25 codes make up 66% of the workload
- Edison will be available until 2019
- Training and lessons learned will be made available to all application teams

We have a dual mission to advance the state-of-the-art in supercomputing

- We collaborate with computer companies years before a system's delivery to deploy advanced systems with new capabilities at large scale
- We provide a highly customized software and programming environment for science applications
- We are tightly coupled with the workflows of DOE's experimental and observational facilities – ingesting tens of terabytes of data each day
- Our staff provide advanced application and system performance expertise to users