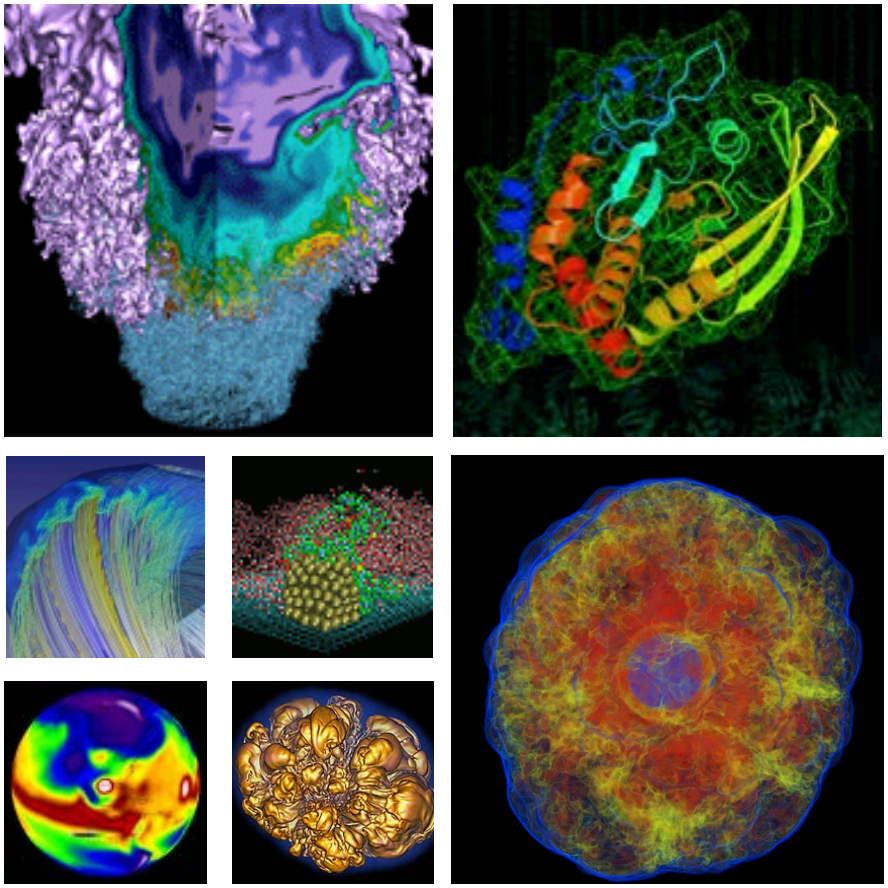
Systems Roadmap and Plans for Supporting Extreme Data Science



Richard Gerber

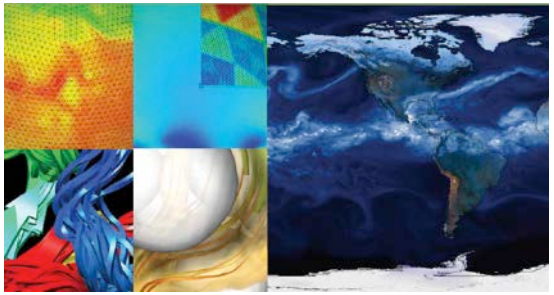
NERSC Senior Science Advisor

December 10, 2015

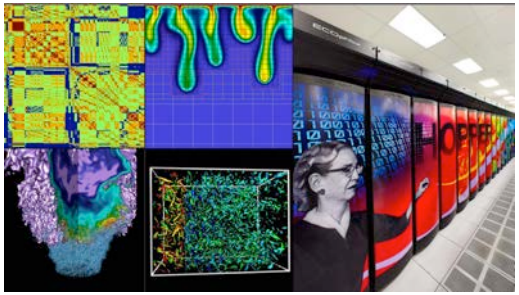
Resources for DOE Office of Science Research

U.S. DEPARTMENT OF ENERGY | Office of Science

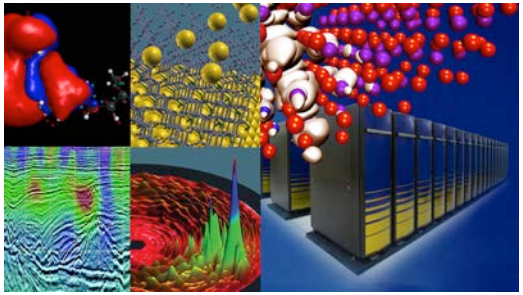
Largest funder of physical science research in U.S.



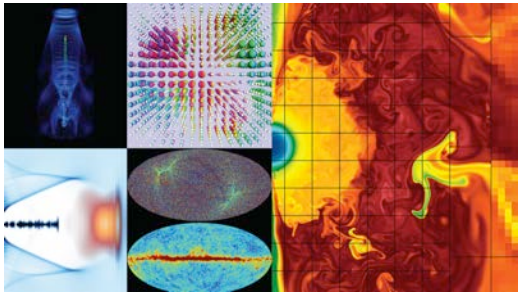
Bio Energy, Environment



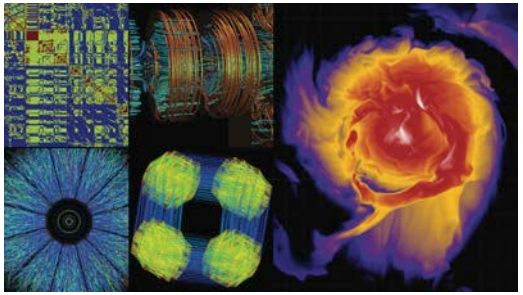
Computing



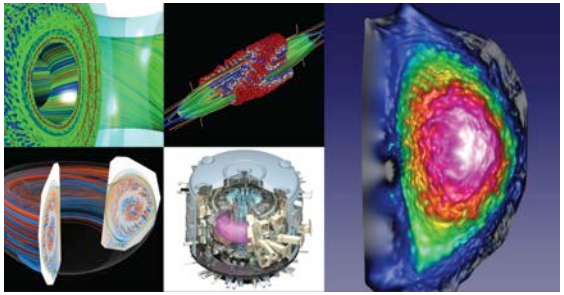
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



Nuclear Physics

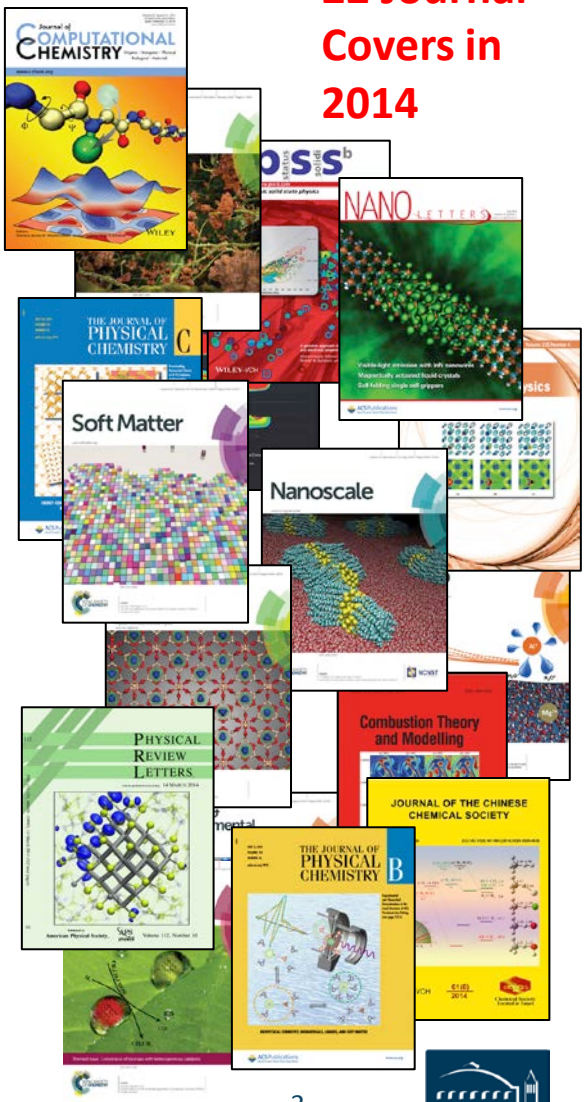


Fusion Energy, Plasma Physics

NERSC overview

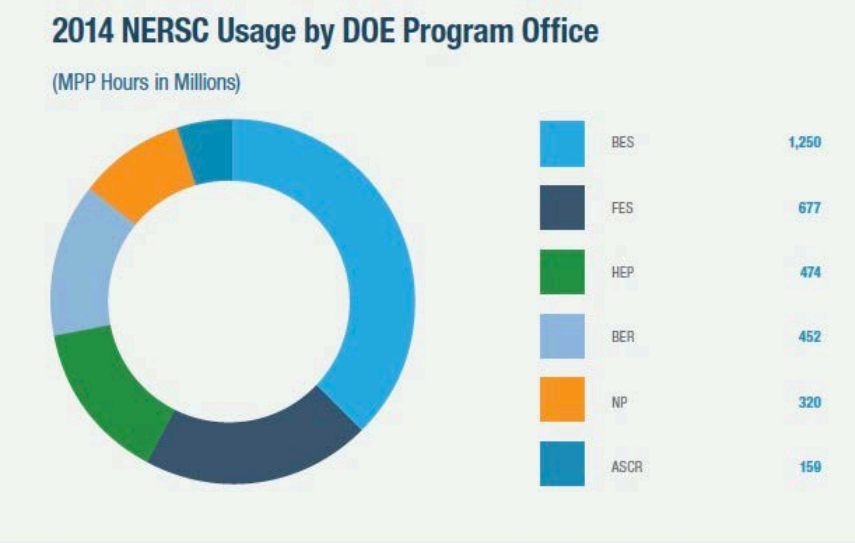
- Strong focus on science
- 1,808 referred publications in 2014
- Deploy first of a kind systems
- Many users (~6,000)
- Users compute at scale and at high volume
- Diversity of algorithms (~600 codes)
- Extreme scale computing and data analysis

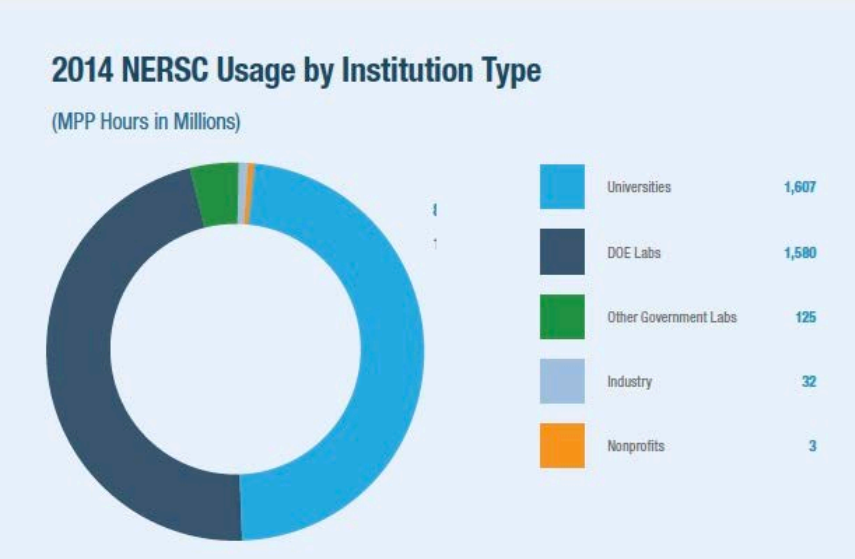
22 Journal Covers in 2014



NERSC directly supports DOE's science mission

- **DOE SC offices allocate 80% of the computing and storage resources at NERSC**
- **ALCC 10%**
- **NERSC Director's Reserve 10%**





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) manycore processor with up to 72 cores per node
  - On-package high-bandwidth memory
- **Data Intensive Science Support**
  - 10 Haswell processor cabinets (Phase 1) to support data intensive applications – Summer 2015
  - NVRAM Burst Buffer to accelerate data intensive applications
  - 28 PB of disk, >700 GB/sec I/O bandwidth
- **Robust Application Readiness Plan**
  - 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

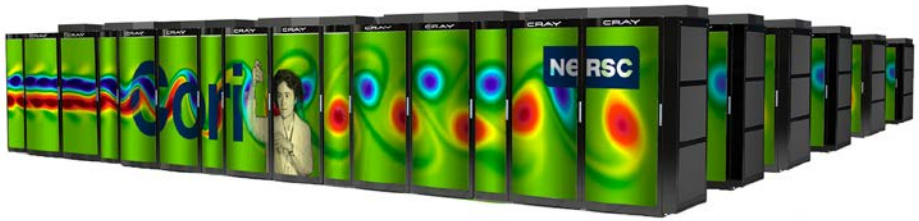
- **Next generation Xeon-Phi, >3TF peak**
- **Single socket processor - Self-hosted, not a co-processor, not an accelerator**
- **Up to 72 cores per processor with support for four hardware threads each; more cores than current generation Intel Xeon Phi™**
- **512b vector units (32 flops/clock – AVX 512)**
- **3X single-thread performance over current generation Xeon Phi co-processor (KNC)**
- **High bandwidth on-package memory, up to 16GB capacity with bandwidth projected to be 5X that of DDR4 DRAM memory**
- **Higher performance per watt**
- **Presents an application porting challenge to efficiently exploit KNL performance features**

Cori will be installed in the Computational Research and Theory (CRT) Facility

- **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

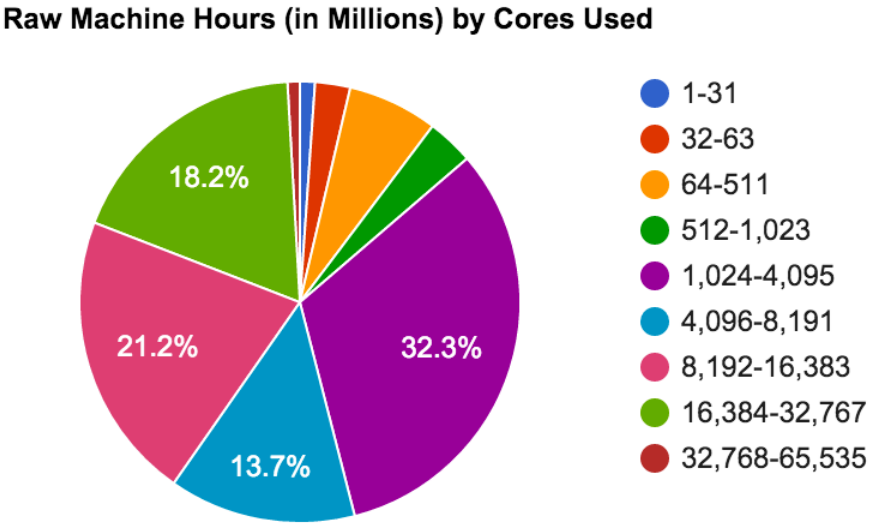
Cori Phase 1

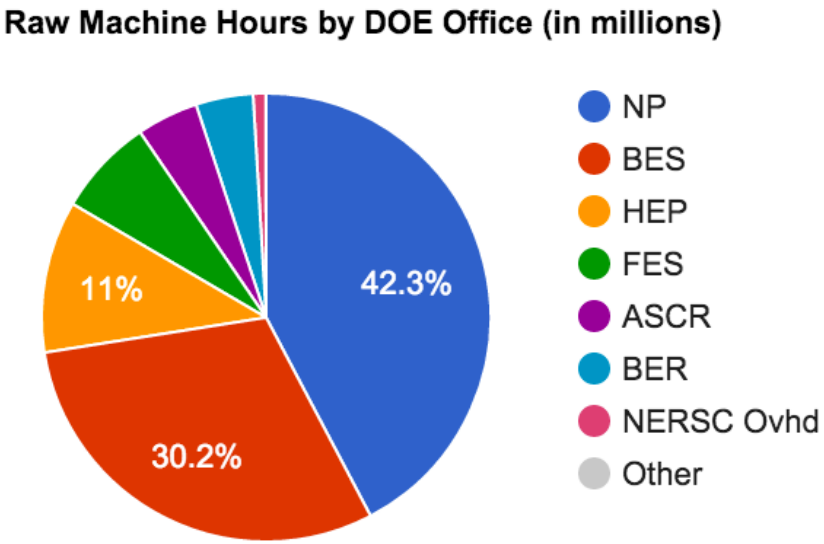
- **Installed in CRT now; running with all NERSC users in pre-production mode**
- **1,630 Compute Nodes (52,160 cores)**
  - Two Haswell processors/node
  - 16 cores/processor at 2.3 GHz
  - 128 GB DDR4 2133 Mhz memory/ node
- **Cray Aries high-speed "dragonfly" topology interconnect**
- **22 login nodes for advanced workflows and analytics**
- **SLURM batch system**
- **Lustre File system**
  - 28 PB capacity, >700 GB/sec peak performance

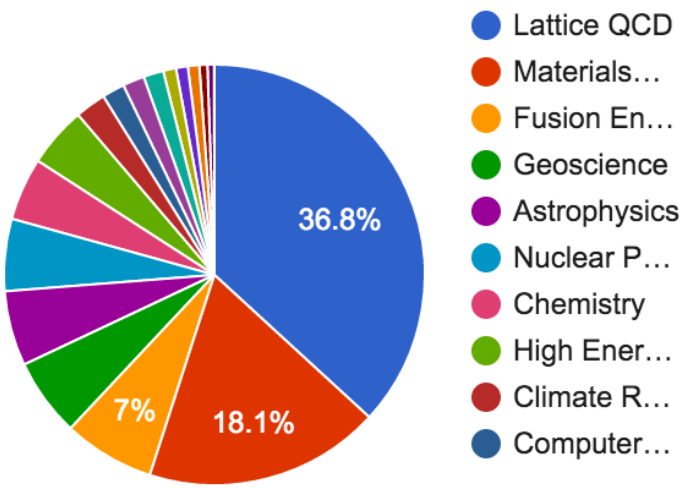


Cori Phase 1: Nov. 8 – Dec. 8, 2015

83 million NERSC MPP hours delivered to science







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

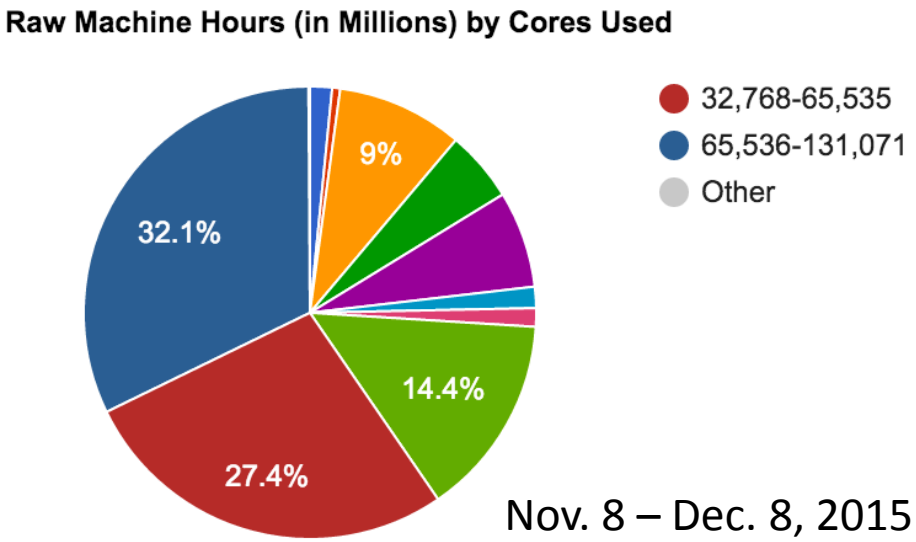
Edison workload

Since Cori Phase 1 came online Edison has been better serving demand to run large jobs

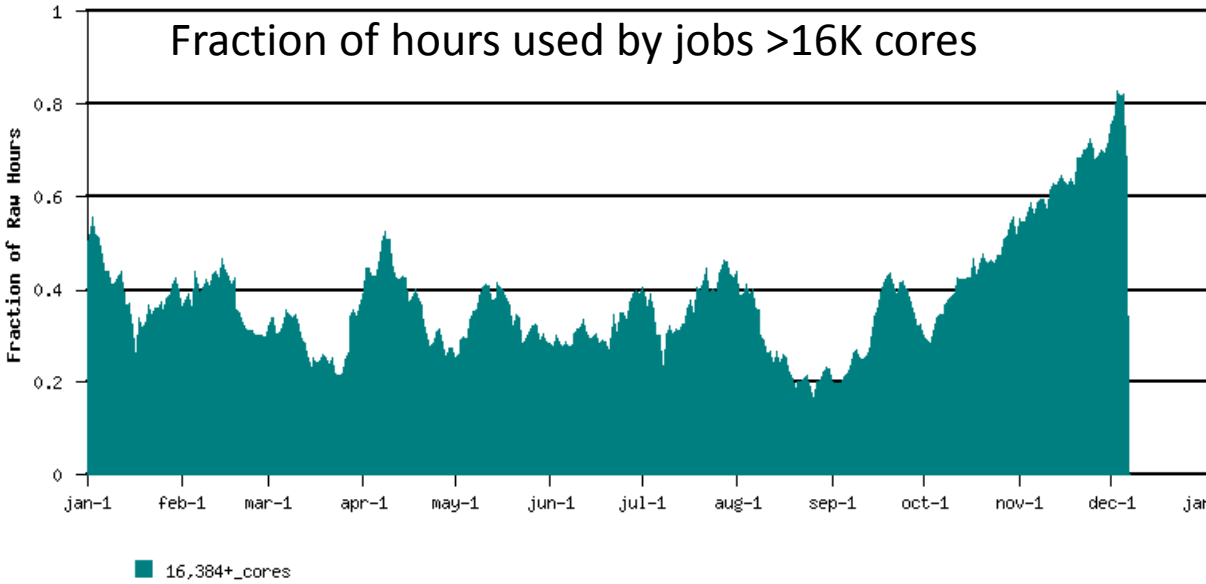
>16K core jobs using 80% of time now

>32K core jobs using 59% of time

>64K core jobs using 32% of time



Nov. 8 – Dec. 8, 2015



U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC Systems Timeline

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
2028	NERSC-11			5-10EF

NERSC 9 Activities

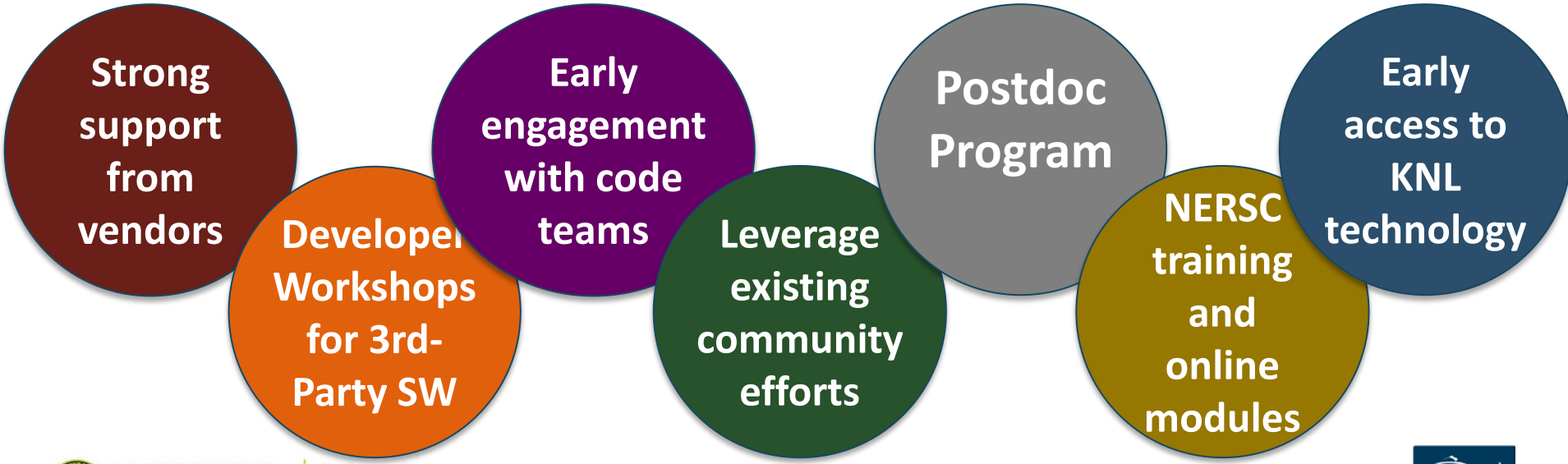
- CD0 signed August 24, 2015
- RFP draft technical specs released Nov. 10, 2015
  - Vendor feedback due Dec. 11, 2015
- Design Review Jan. 19-20, 2016
- Independent Project Review (IPR) Q2CY16
- RFP released late Spring/early Summer 2016

APEX

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

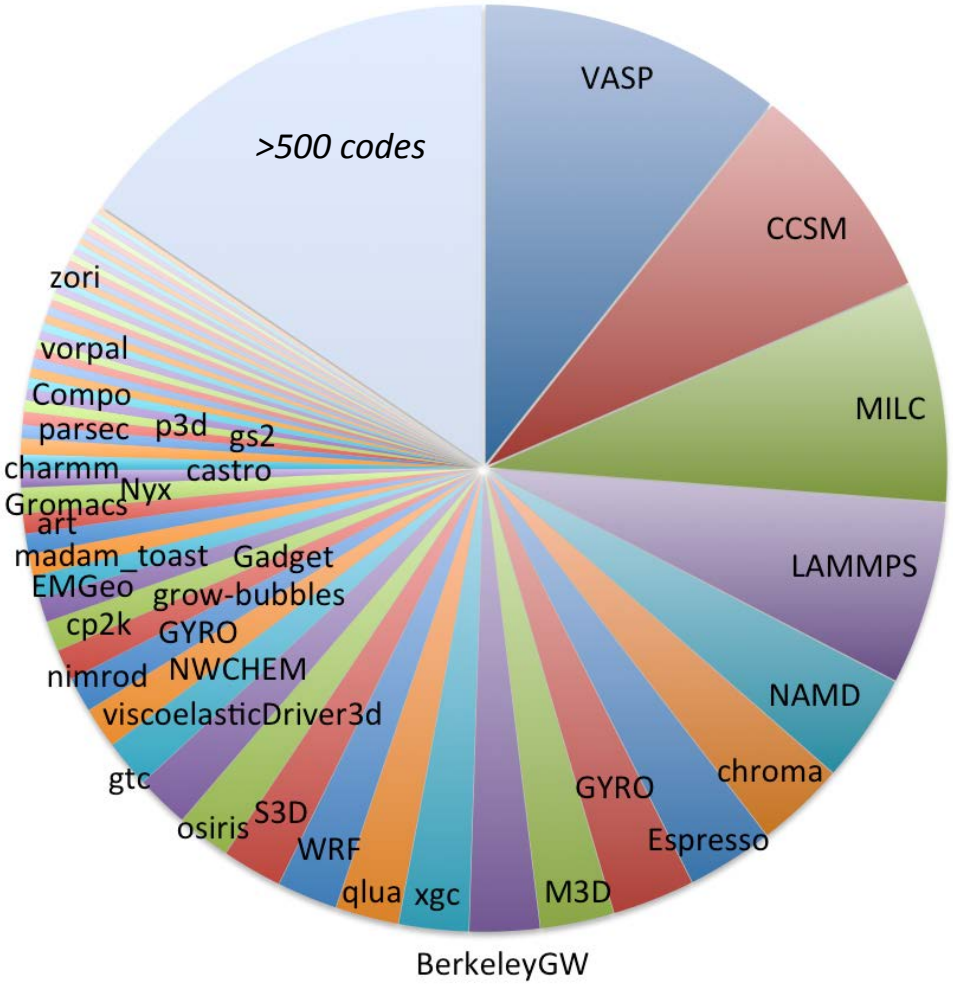
NERSC Exascale Science Application Program

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



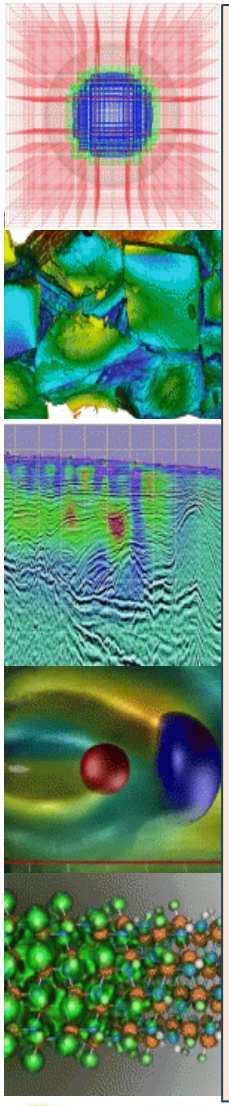
We are initially focussing on 20 codes

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 2019/2020
- Training and lessons learned will be made available to all application teams

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**

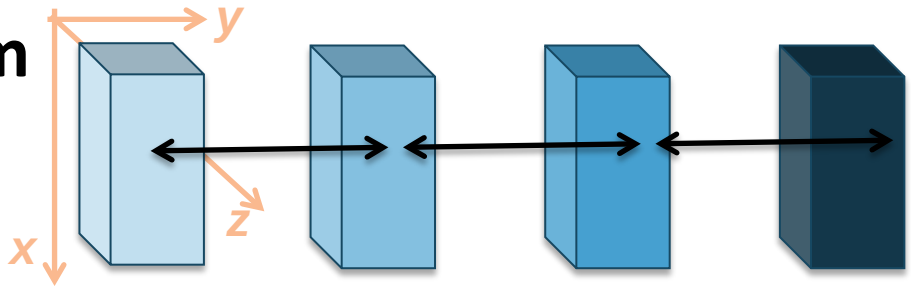
Fusion Energy Sciences

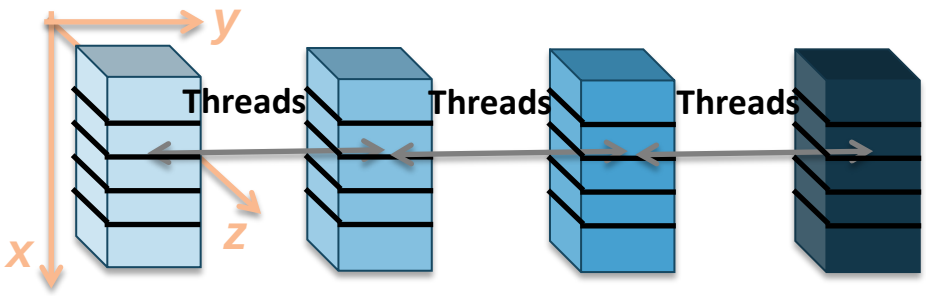
Jardin (PPPL) **M3D**

Chang (PPPL) **XGC1**

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



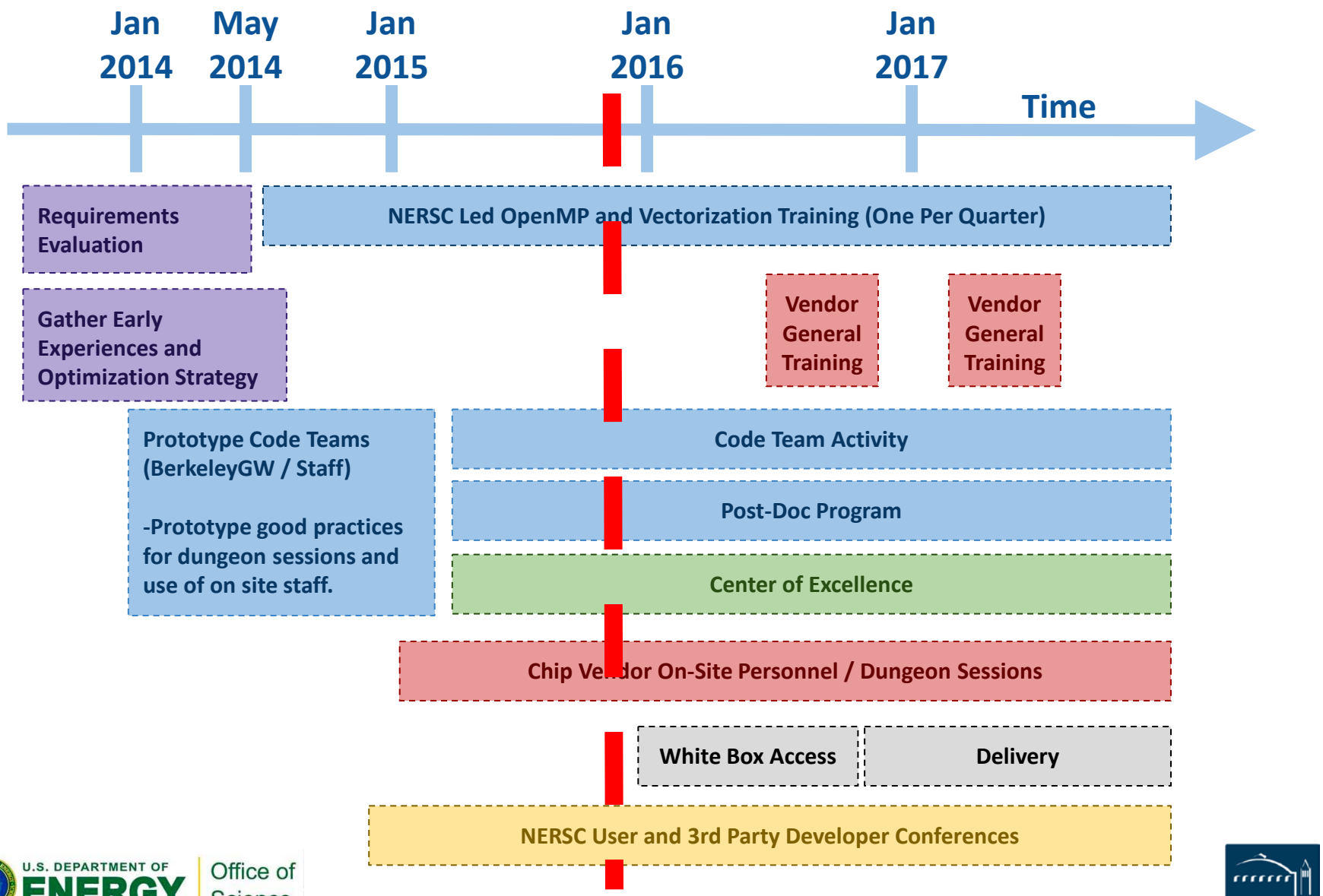



```
|--> DO I = 1, N
|       R(I) = B(I) + A(I)
|--> ENDDO
```

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)**
- **Hiring for application performance expertise**
- **8 Postdocs**

NESAP Timeline



NESAP Code Status

Advanced (waiting for hardware)

- Chroma
- DWF
- Gromacs
- BerkeleyGW
- MILC
- HACC

Lots of Progress

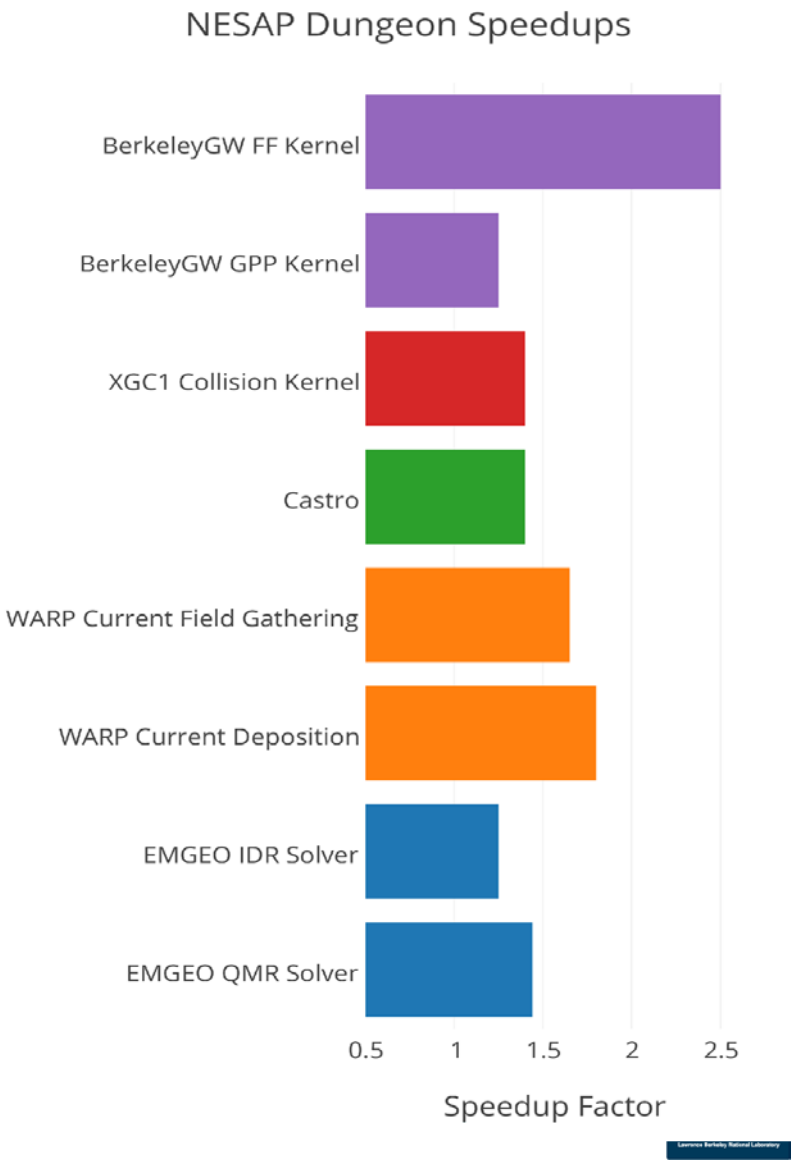
- EMGEO
- Boxlib
- WARP
- XGC1
- VASP
- ESPRESSO

Moving

- PARSEC
- Chombo
- MFDN
- Meraculous
- NWChem

Need Lots of Work

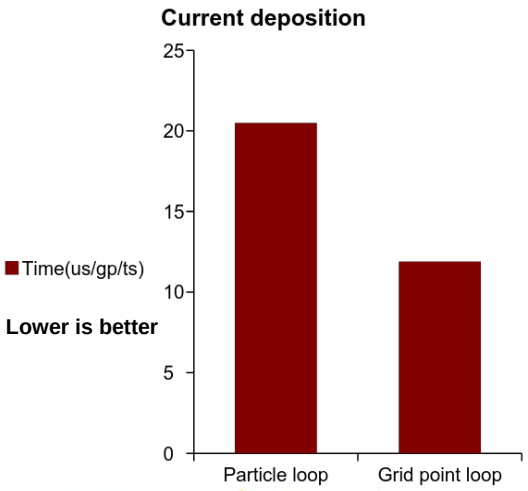
- CESM
- ACME
- MPAS

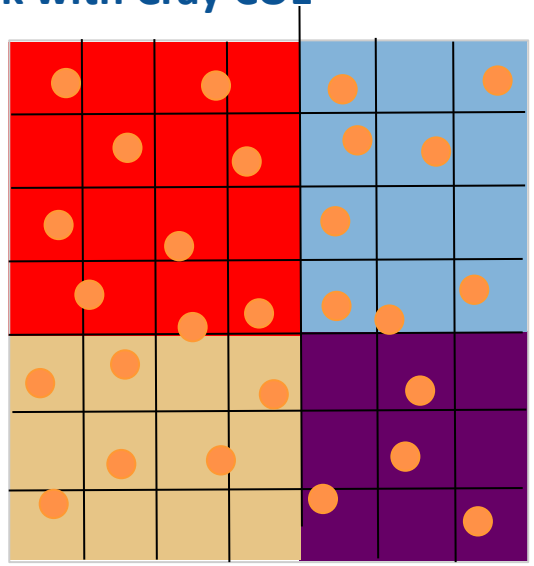


What has gone well

1. **Setting requirements for Dungeon Session (Dungeon Session Worksheet).**
2. **Engagement with IXPUG and user communities (DFT, Accelerator Design for Exascale Workshop at CRT)**
3. **Large number of NERSC and Vendor Training (Vectorization, OpenMP, Tools/Compilers) Well Received**
4. **Learned a Massive Amount about Tools and Architecture**
5. **Cray COE VERY helpful to work with. Very pro-active.**
6. **Pipelining Code Work Via Cray and Intel resources**

Warp Vectorization Improvements at The Dungeon - Directly enabled by tiling work with Cray COE





Techniques and Tools to Target Arrays for Fastmem:

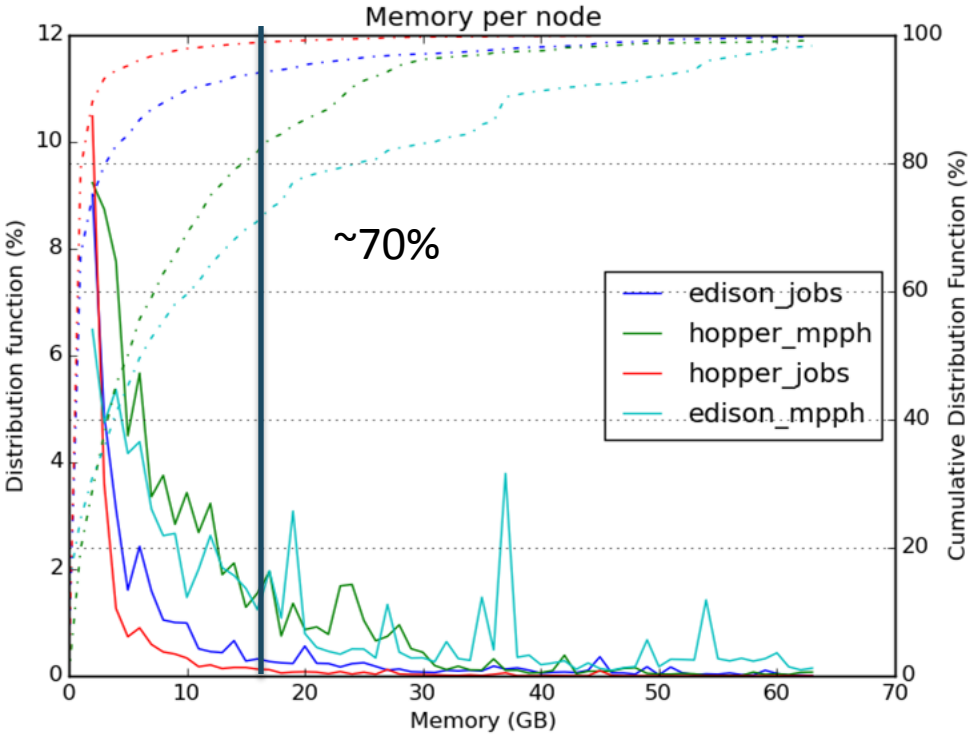
Application	All memory on far memory	All memory on near memory	Key arrays on near memory
BerkeleyGW	baseline	52% faster	52.4% faster
EmGeo	baseline	40% faster	32% faster
XGC1	baseline		24% faster

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

What Has Gone Well (Continued)

7. **Bandwidth sensitive applications that live in HBM expected to perform very well.**
8. **A lot of Lessons Learned: techniques to place key-arrays in fast-memory, improve prefetching effectiveness, coping without L3 cache etc...**
9. **CPU Intensive tasks (BGW GPP Kernel) shown to perform well (> Haswell) on early KNL projections if effectively use L2, vectorize well and can make use of 2 VPUs.**
10. **Postdocs deeply engaged.**

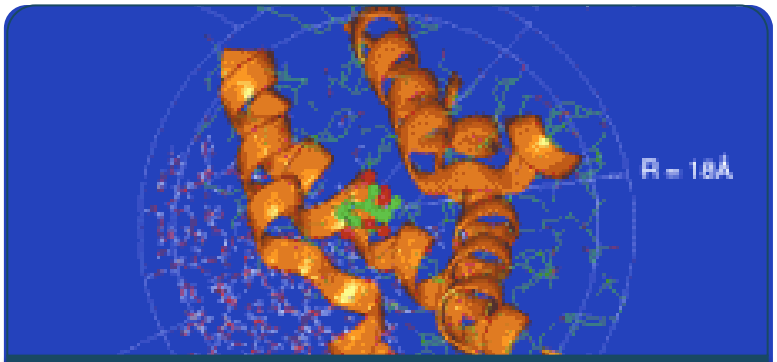


The N9 workload analysis shows a large fraction of jobs use < 16GB of memory per node

NESAP Plans

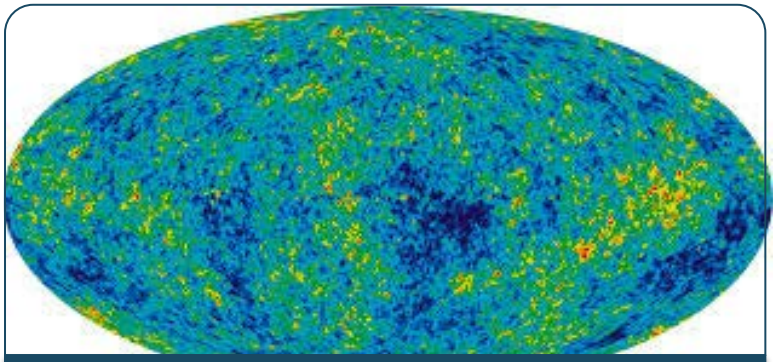
- **Increase excitement and effort in 2016 with extra training events, on-site hackathons and more dungeon sessions with KNL hardware in first 9 months of year (5-6).**
- **Continue successful Cray+Intel pipelining approach.**
- **Continue App-Readiness (and post-doc program) as an ongoing center effort through 2025 (exascale).**
- **Maintain a community database of lessons learned and programming "pearls" for many-core that is searchable by keywords like "vectorization", "latency", "stencil" as a standalone portal**

Extreme Data Science is Playing a Key Role in Scientific Discovery



Martin Karplus

Saul Perlmutter



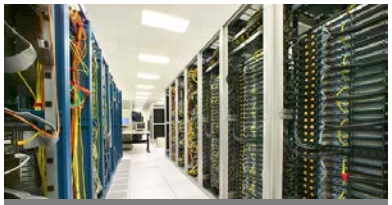
George Smoot

Warren Washington

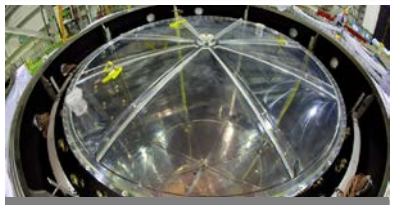
Solving the Puzzle of the Neutrino

- **HPC and ESnet vital in the measurement of the important "θ_{13}" neutrino parameter.**
  - Last and most elusive piece of a longstanding puzzle: why neutrinos appear to vanish as they travel
  - The result affords new understanding of fundamental physics; may eventually help solve the riddle of matter-antimatter asymmetry in the universe.
- **HPC for simulation / analysis; HPSS and data transfer capabilities; NGF and Science Gateways for distributing results**
  - All the raw, simulated, and derived data are analyzed and archived at a single site
  - => Investment in experimental physics requires investment in HPC.
- **One of Science Magazine's Top-Ten Breakthroughs of 2012**

The Daya Bay experiment counts antineutrinos at three detectors (shown in yellow) near the nuclear reactors and calculates how many would reach the detectors if there were no oscillation. transformation.



NERSC's PDSF cluster



Daya Bay detectors

U.S. DEPARTMENT OF ENERGY | Office of Science

HEP

Science AAAS

PI: Kam-Biu Luk (LBNL)

BERKELEY LAB

Nobel Prize in Physics 2015

Scientific Achievement

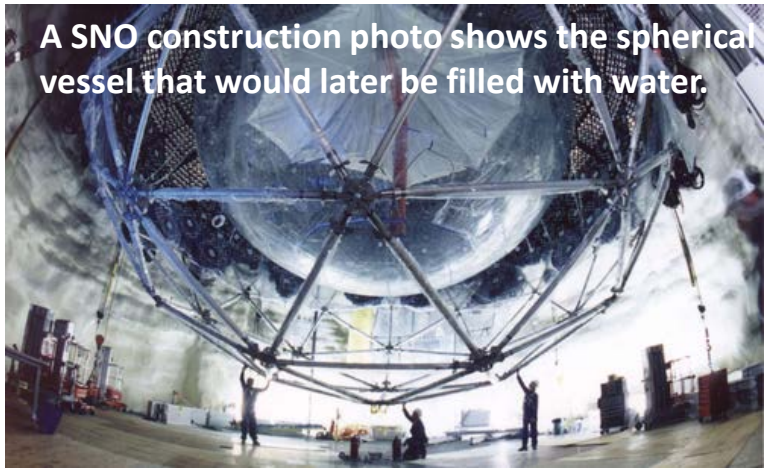
The discovery that neutrinos have mass and oscillate between different types

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.

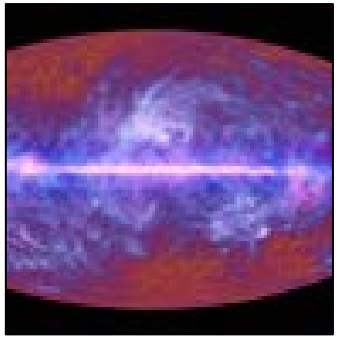
NERSC helped the SNO team use PDSF for critical analysis contributing to their seminal PRL paper. HPSS serves as a repository for the entire 26 TB data set.

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

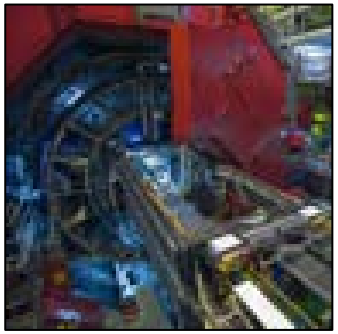
Nobel Recipients: Arthur B. McDonald, Queen's University (SNO)
Takaaki Kajita, Tokyo University (Super Kamiokande)

NERSC has been supporting data intensive science for a long time

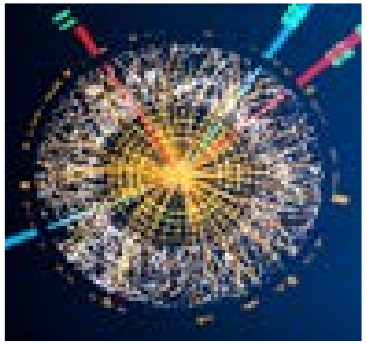
Palomar Transient Factory
Supernova



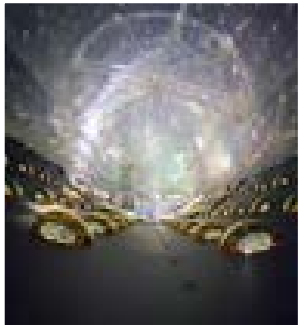
Planck Satellite
Cosmic Microwave Background
Radiation



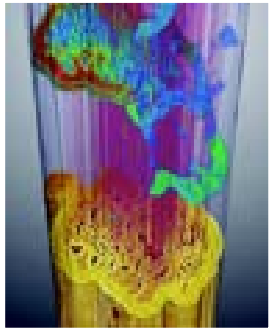
Alice
Large Hadron Collider



Atlas
Large Hadron Collider



Dayabay
Neutrinos



ALS
Light Source

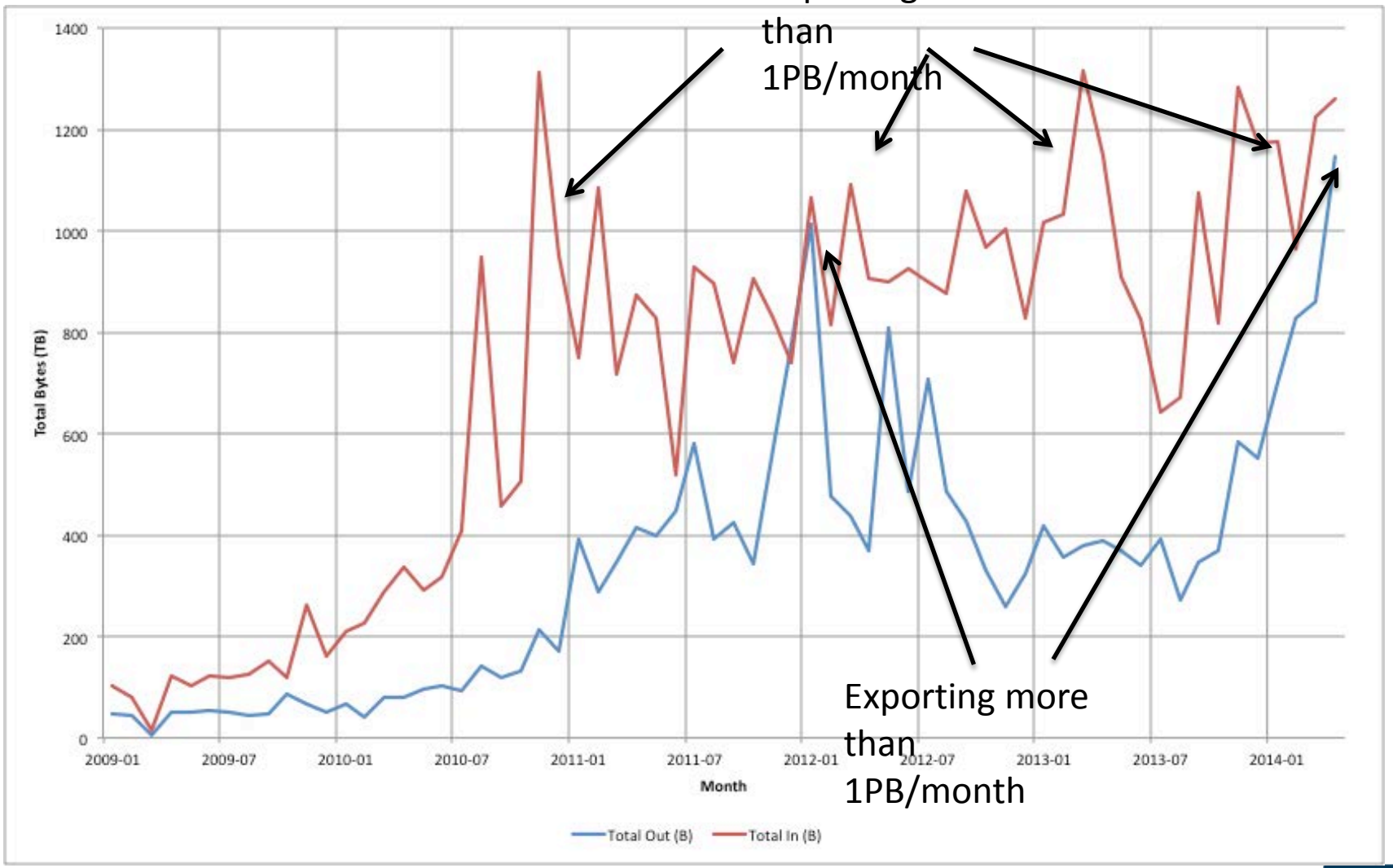
LCLS
Light Source

Joint Genome Institute
Bioinformatics

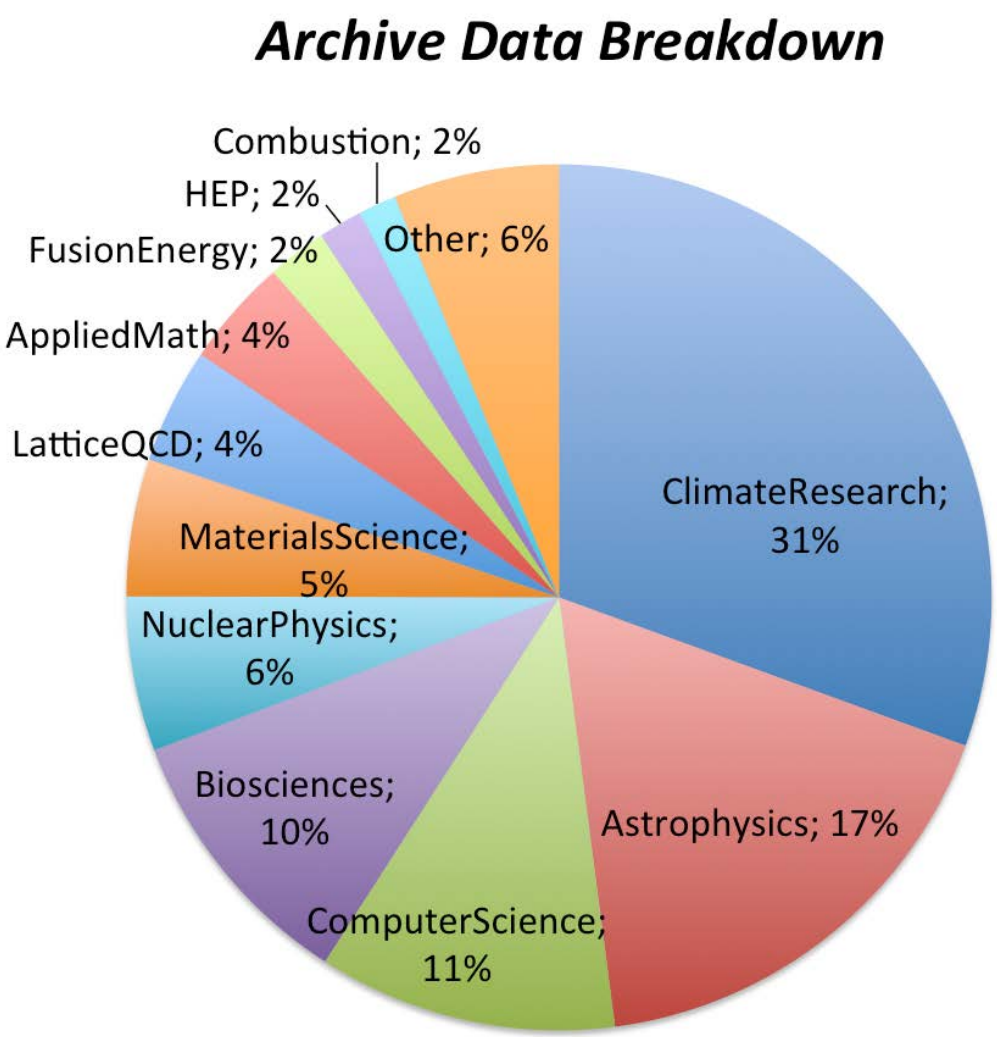
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

NERSC users import more data than they export!



NERSC archives An Enormous Amount of Data for the Scientific Community



60 PB of data are stored in NERSC's HPSS Archive

NERSC's Goal for Data Initiative

Increase the productivity, usability, and impact of DOE's experimental user facilities and other data-intensive science by providing comprehensive data systems and services to store, analyze, manage, and share data.

Through 2013 NERSC deployed separate Compute Intensive and Data Intensive Systems

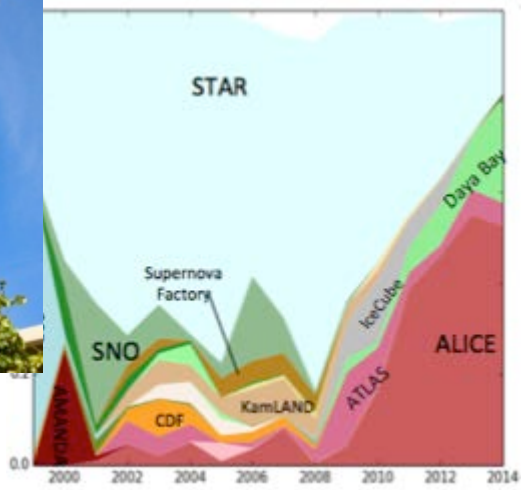
NERSC 40 YEARS at the FOREFRONT

Compute Intensive

Data Intensive

Carver

Genepool



PDSF

But how different really are the compute and data intensive platforms?

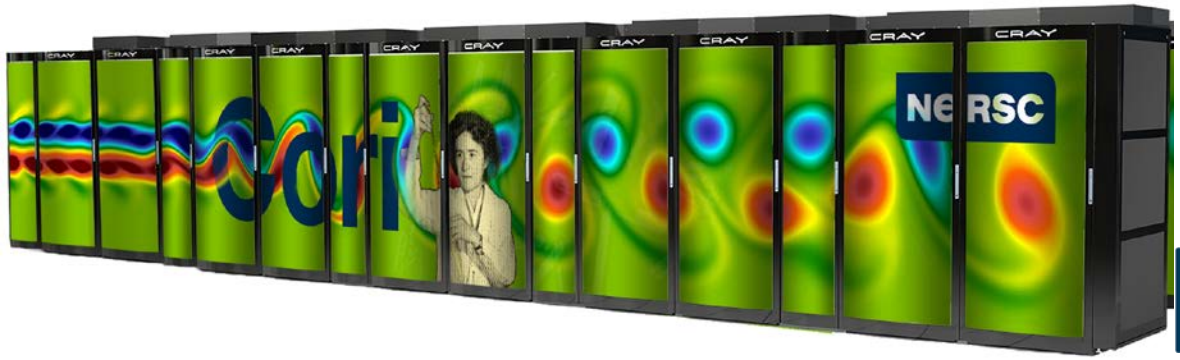
NERSC 40 YEARS at the FOREFRONT

Policies	***Software/Configuration***	***Hardware***
• Fast-turn around time. Jobs start shortly after submitted • Can run large numbers of throughput jobs	• Support for complex workflows • Communication and streaming data from external databases and data sources • Easy to customize user environment	• Local disk for fast I/O • Some systems (not all) have larger memory nodes • Support for advanced workflows (DB, web, etc)

Differences are primarily software and policy issues with some hardware differences in the ratio of I/O, memory and compute

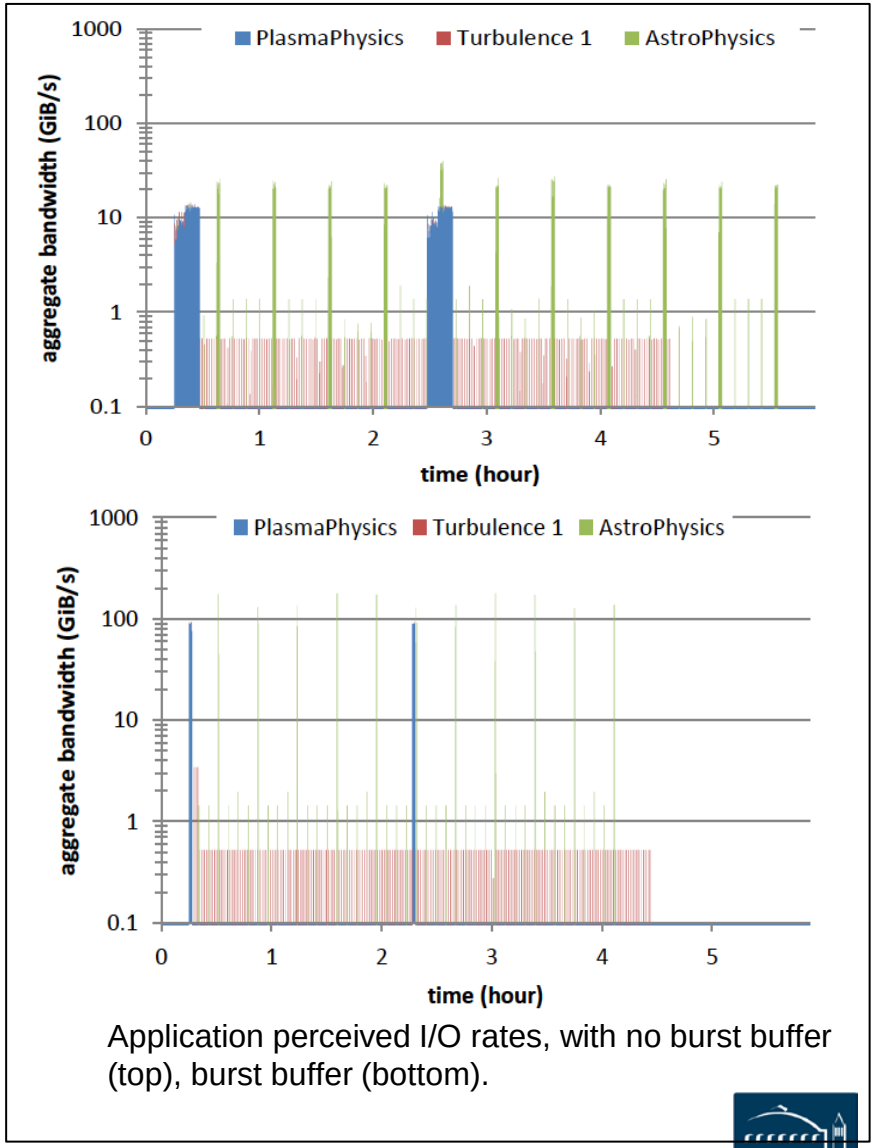
NERSC is making significant investments on Cori to support data intensive science

- **New queue policies: real time, and high throughput queues**
- **High bandwidth external connectivity to databases from compute nodes**
- **More (23) login nodes for managing advanced workflows**
- **Virtualization capabilities (Docker)**
- **NVRAM Flash Burst Buffer as I/O accelerator**
  - 1.5PB, 1.5 TB/sec
  - User can request I/O bandwidth and capacity at job launch time
  - Use cases include, out-of-core simulations, image processing, shared library applications, heavy read/write I/O applications



Burst Buffer Motivation

- **Flash storage is significantly more cost effective at providing bandwidth than disk (up to 6x)**
- **Flash storage has better random access characteristics than disk, which help many SC workloads**
- **Users' biggest request (complaint) after wanting more cycles, is for better I/O performance**



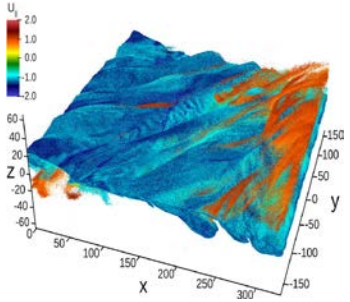
Application perceived I/O rates, with no burst buffer (top), burst buffer (bottom).

Analysis from Chris Carothers (RPI) and Rob Ross (ANL)

Burst Buffer Use Cases

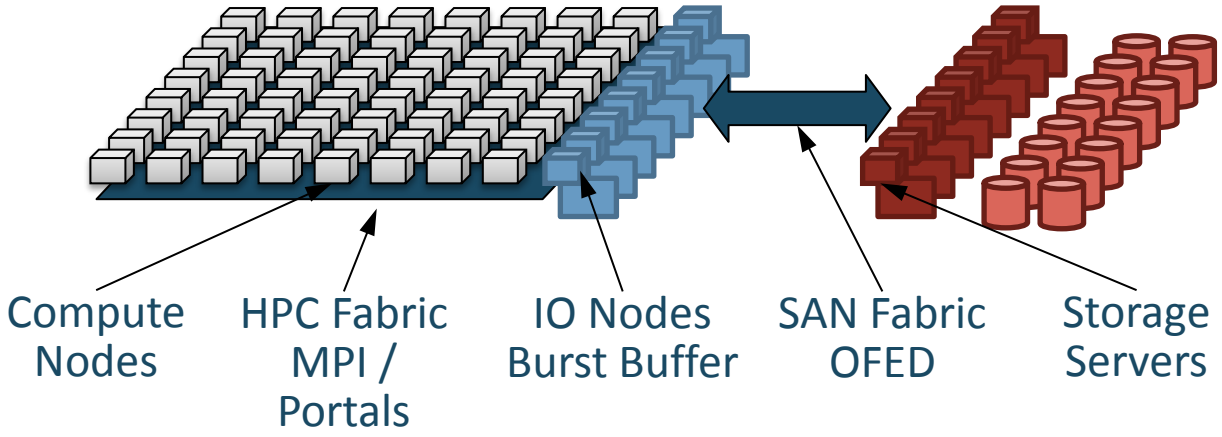
- **Accelerate I/O**
  - Checkpoint/restart or other high bandwidth reads/writes
  - Apps with high IOP/s e.g. non-sequential table lookup
  - Out-of-core applications
  - Fast reads for image analysis
- **Advanced Workflows**
  - Coupling applications, using the Burst Buffer as interim storage
  - Streaming data from experimental facilities
- **Analysis and Visualization**
  - In-situ/ in-transit
  - Interactive visualization

Palomar Transient Factory Pipeline:
Use Burst Buffer as cache for fast reads



VPIC – in situ visualization of a trillion particles

Burst Buffer Software Development Efforts



Create Software to enhance usability and to meet the needs of all NERSC users

- Scheduler enhancements
  - Automatic migration of data to/from flash
  - Dedicated provisioning of flash resources
  - Persistent reservations of flash storage
- Caching mode – data transparently captured by the BB nodes
  - Transparent to user -> no code modifications required
- Enable In-transit analysis
  - Data processing or filtering on the BB nodes – model for exascale

Burst Buffer Early User Program call for proposals

- **Aug 10th: solicited proposals for BB Early Users program.**
  - Award of exclusive early use of BB on Cori P1, plus help of NERSC experts to optimise application for BB.
- **Selection criteria include:**
  - Scientific merit.
  - Computational challenges.
  - Cover range of BB data features.
  - Cover range of DoE Science Offices.
- **Great interest from the community, 29 proposals received. Good distribution across offices...**

Many great applications...

- We're very happy with the response to the program call.
- Decided to support more applications than we'd originally anticipated
- Other applications will not be supported by NERSC staff, but will have early access to Cori P1 and the BB.
- Breakdown by DoE Office:

	ASCR	BER	BES	Fusion	HEP	Nuclear	Total
NERSC Supported	1.5	2.5	2.5	1	4.5	1	13
Early Access	3	7	2.5	0	2.5	0	15

NERSC supported projects

Project	DoE office	BB data features
Nyx/Boxlib cosmology simulations (*Ann Almgren, LBNL*)	HEP	I/O bandwidth with BB; checkpointing; workflow application coupling; in-situ analysis.
Phoenix: 3D atmosphere simulator for supernovae (*Eddie Baron, U. Oklahoma*)	HEP	I/O bandwidth with BB; staging intermediate files; workflow application coupling; checkpointing.
Chombo-Crunch + Visit for carbon sequestration (*David Trebotich, LBNL*)	BES	I/O bandwidth with BB; in-situ analysis/visualization using BB; workflow application coupling.
Sigma/UniFam/Sipros Bioinformatics codes (*Chongle Pan, ORNL*)	BER	Staging intermediate files; high IOPs; checkpointing; fast reads.
XGC1 for plasma simulation (*Scott Klasky, ORNL*)	Fusion	I/O bandwidth with BB; intermediate file I/O; checkpointing.
PSANA for LCLS (*Amadeo Perazzo, SLAC*)	BES/BER	Staging data with BB; workflow management; in-transit analysis.

NERSC supported projects

Project	DoE office	BB data features
ALICE data analysis (*Jeff Porter, LBNL*)	NP	I/O bandwidth with BB; read-intensive I/O.
Tractor: cosmological data analysis (DESI) (*Peter Nugent, LBNL*)	HEP	Intermediate file I/O using BB; high IOPs.
VPIC-IO performance (*Suren Byna, LBNL*)	HEP/ACSR	I/O bandwidth with BB; in-situ data analysis; BB to stage data.
YODA: Geant4 sims for ATLAS detector (*Vakhtang Tsulaia, LBNL*)	HEP	BB for high IOPs; stage small intermediate files.
ALS SPOT Suite (*Craig Tull, LBNL*)	BES/BER	BB as fast cache; workflow management; visualization.
TomoPy for ALS image reconstruction (*Craig Tull, LBNL*)	BES/BER	I/O throughput with BB; workflow management; read-intensive I/O.
kitware: VPIC/Catalyst/ParaView (*Berk Geveci, kitware*)	ASCR	in-situ analysis/visualization with BB; multi-stage workflow.

A variety of use cases are represented by the BB Early Users

Application	I/O bandwidth : reads	I/O bandwidth: writes (checkpointing)	High IOPs	Workflow coupling	In-situ / in-transit analysis and visualization	Staging inter-mediate files/ pre-loading data
Nyx/Boxlib		X		X	X	
Phoenix 3D		X		X		X
Chomo/Crunch + Visit		X		X	X	
Sigma/UniFam/Sipros	X	X	X			X
XGC1	X	X				X
PSANA				X	X	X
ALICE	X					
Tractor			X	X		X
VPIC/IO					X	X
YODA			X			X
ALS SPOT/TomoPy	X			X	X	X
kitware						

Realtime access to HPC systems

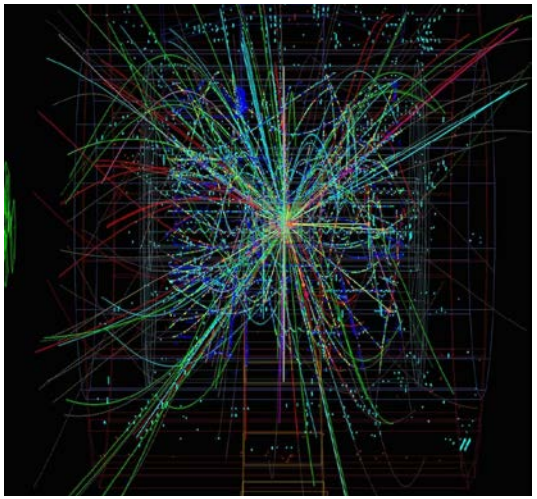
- **We've heard from a number of users that the lack of 'realtime' access to the system is a barrier to scientific productivity**
- **With NERSC's new batch scheduler, SLURM, we have the capability to offer 'immediate' or 'real-time' access on Cori Phase 1, for projects and users with requirements for fast turn around**
- **But how to select and judge projects?**
- **We added a question to ERCAP about realtime needs to assess demand and size realtime resources.**
- **Review with program managers and get feedback**

Immediate Queue – ERCAP Requests

- **19 responses (out of > 700) a small fraction of our workload**
- **Responses from 5 of 6 Offices, SBIR and EERE, demonstrating need is not confined to one scientific domain**
- **Expected responses:**
  - ALS, Palomar Transient Factory, CRD workflow research, MyGreenCar, OpenMSI, KBASE, Materials analysis, 2 PDSF projects
- **A few surprises**
  - 3 similar Fusion responses (MIT, GA, LLNL) noting DIII-D (tokomak fusion reactor run by General Atomics) can be adjusted by real-time codes
  - Industry response, Vertum partners, Predictive Power Grid performance, run simulations daily 12 hours apart.

Shifter brings user defined images to supercomputers

- **Shifter, a container for HPC, allows users to bring a customized OS environment and software stack to an HPC system.**
- **Use cases**
  - High energy physics collaborations that require validated software stacks
  - Cosmology and bioinformatics applications with many 3rd party dependencies
  - Light source applications that with complicated software stacks that need to run at multiple sites



SHIFTER

Upgrading Cori's External Connectivity

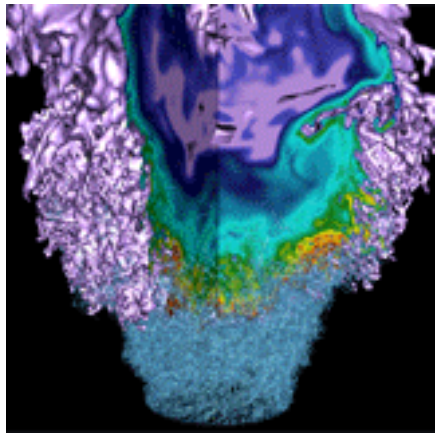
Enable 100Gb+ Instrument to Cori

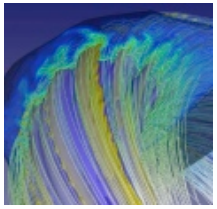
- Streaming data to the supercomputer allows for analytics on data in motion
- Cori network upgrade provides SDN (software defined networking) interface to ESnet. 8 x 40Gb/s bandwidth.
- Integration of data transfer and compute enables workflow automation

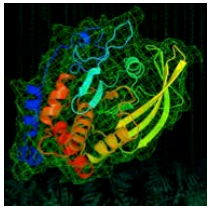
Cori Network Upgrade Use Case:

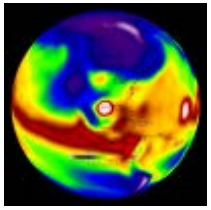
- X-ray data sets stream from detector directly to Cori compute nodes, removing need to stage data for analysis.
- Software Defined Networking allows planning bandwidth around experiment run-time schedules
- 150TB bursts now, LCLS-II has 100x data rates

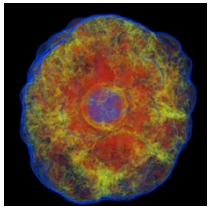
Superfacility Concept

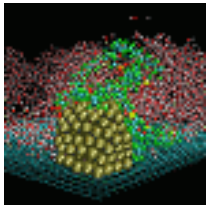












NERSC 40 YEARS at the FOREFRONT 1974-2014

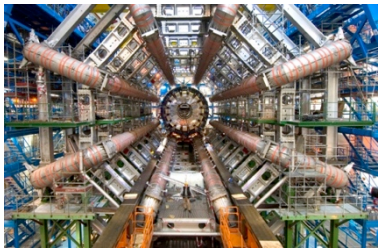
U.S. DEPARTMENT OF ENERGY | Office of Science

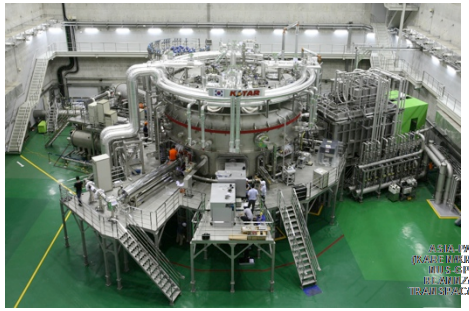
BERKELEY LAB

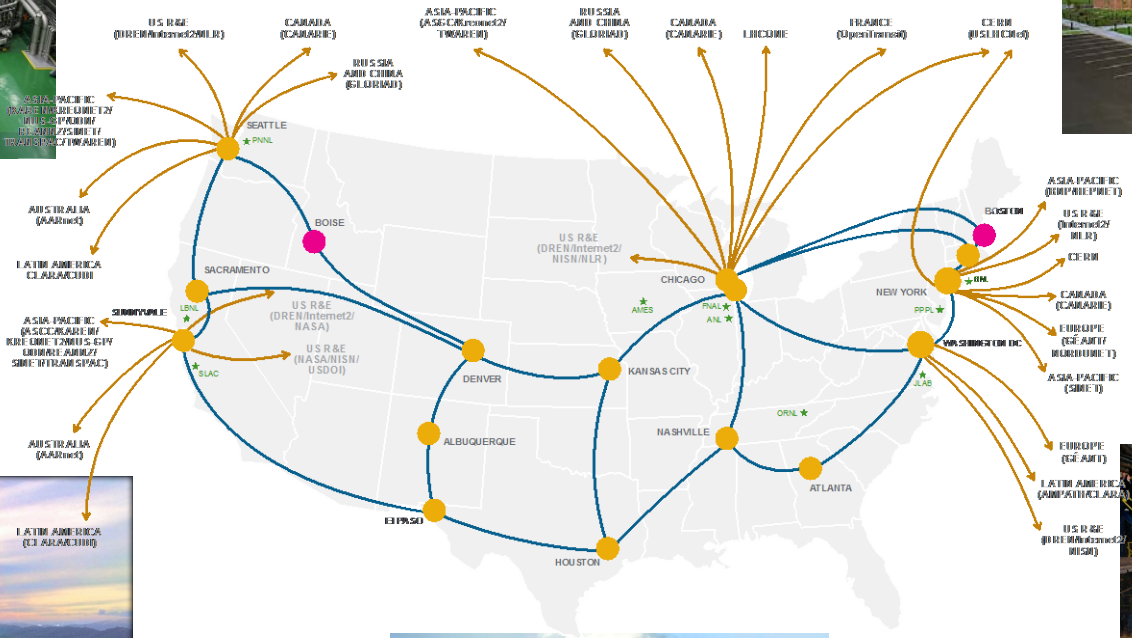
New technology innovations are challenging existing ways of scientific discovery

- Data volumes are increasing faster than Moore's Law
- Facility data exceeds local computing and networking capabilities
- Unfeasible to put a supercomputing center at every experimental facility

Fortunately we have ESnet to connect all our facilities







U.S. DEPARTMENT OF ENERGY | Office of Science

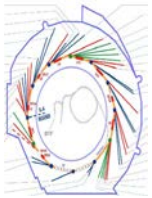
BERKELEY LAB

Based on our experiences supporting science from experimental facilities, we see a Common Design Pattern

predictable data movement

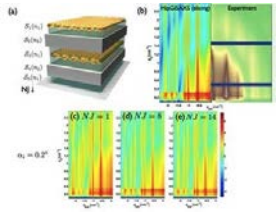
Local Data processing/filtering

Data generation from Sensors

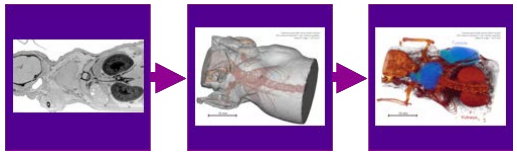


on-the-fly calliberation

Analysis and modeling



Real-time access and Visualization

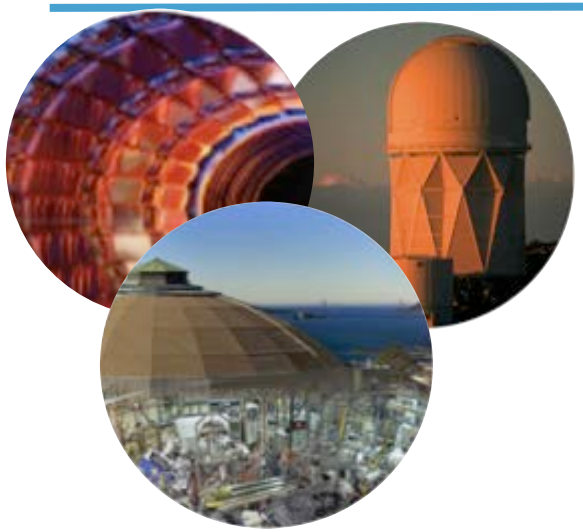


Storage, Archive and Share

On-site Scientist

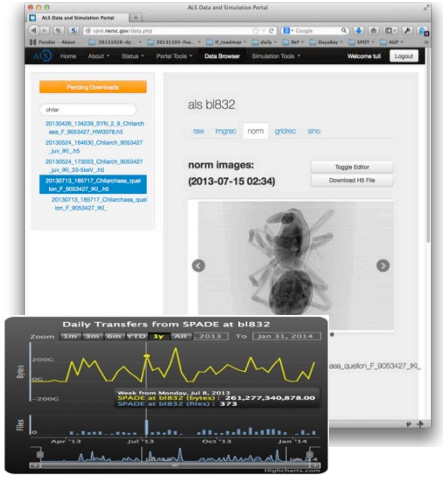
Remote users

Superfacility Vision: A network of connected facilities, software and expertise to enable new modes of discovery

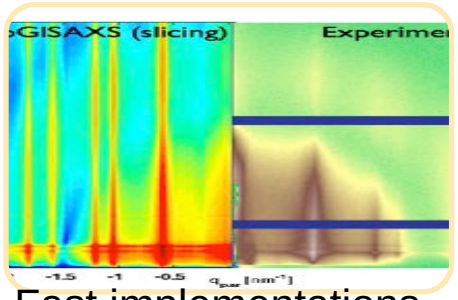


Experimental Facilities

New mathematical analyses



Real-time analysis and data management



Fast implementations on latest computers

Integrated with ESnet: Designed for Big Science Data

Computing Facilities

NERSC has engagements with a number of experimental facilities

- **ALS – Advance Light Source**
- **LCLS – Linac Coherent Light Source**
- **LHC through PDSF collaboration**
- **JGI – Joint Genome Institute**

Thank you!