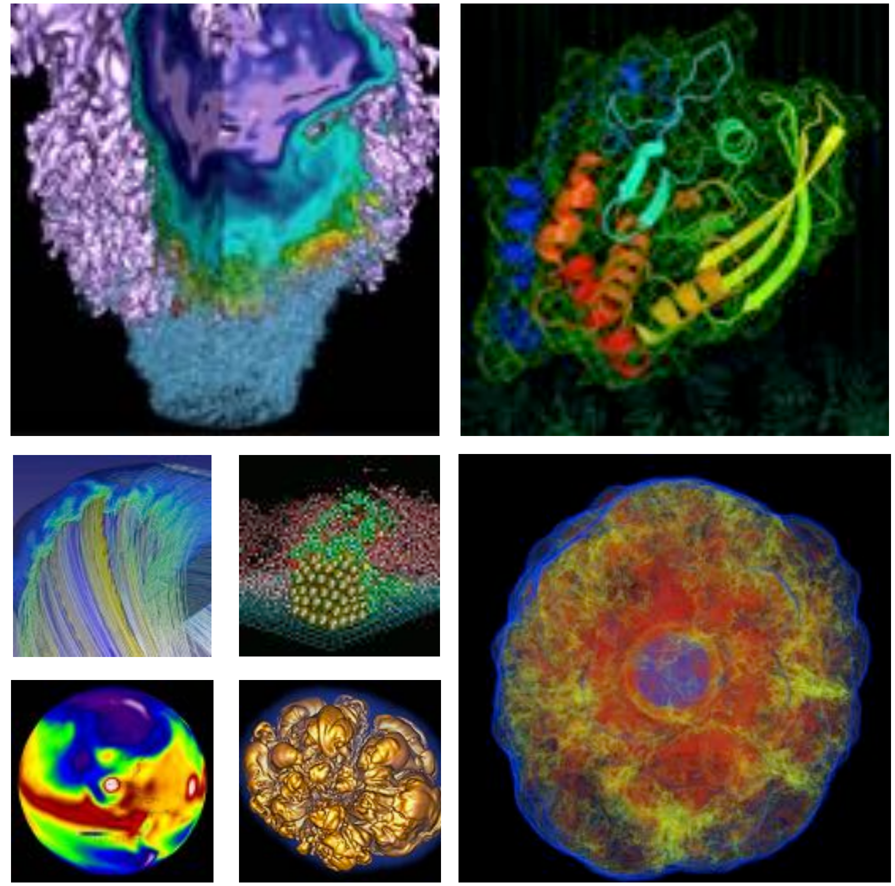
Exascale and Extreme Data Science at NERSC



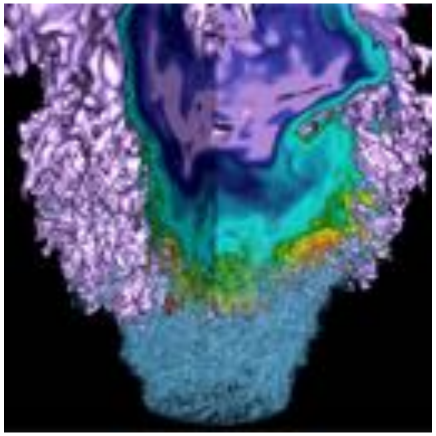
Sudip Dosanjh
NERSC Director

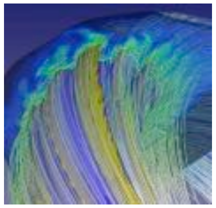
October 17, 2016

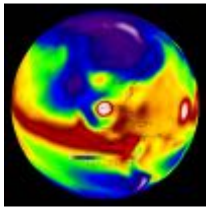
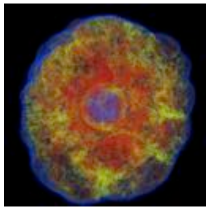
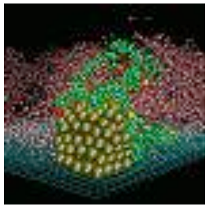
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

NERSC Overview





NERSC has moved to 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

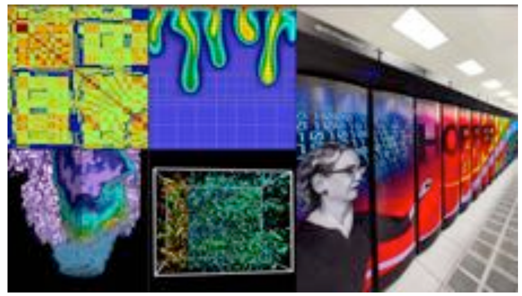
NERSC Provides HPC and Data Resources for DOE Office of Science Research

NERSC

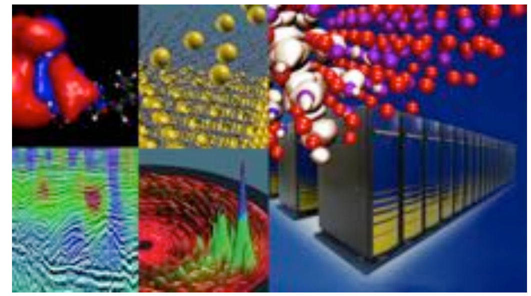
U.S. DEPARTMENT OF ENERGY | Office of Science

Largest funder of physical science research in U.S.

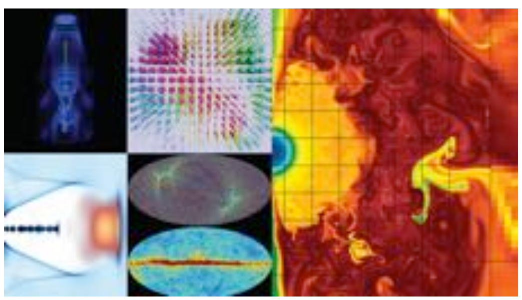
Biology, Environment



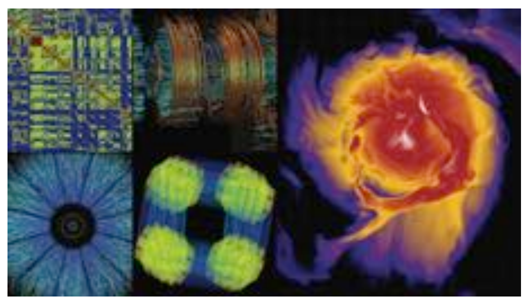
Computing



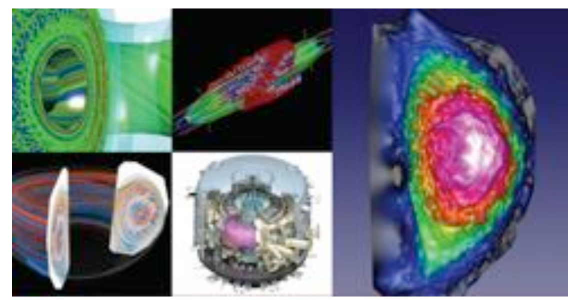
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



Nuclear Physics

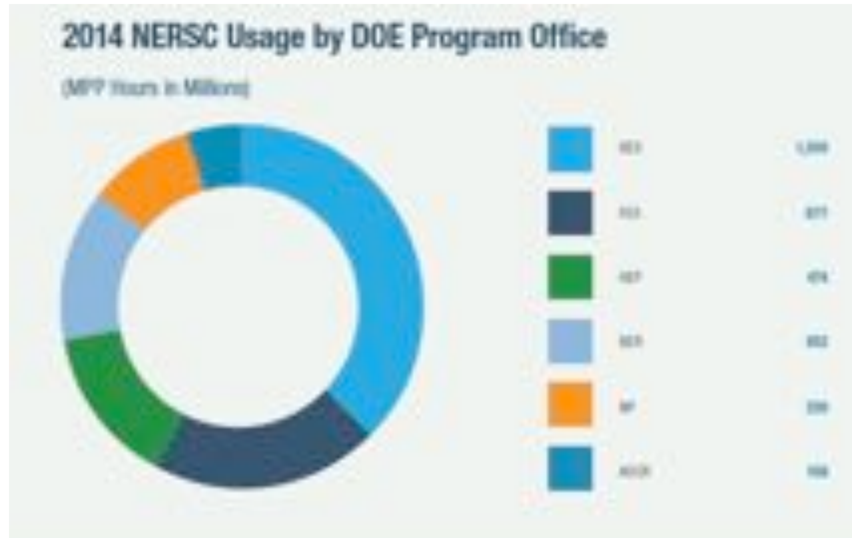


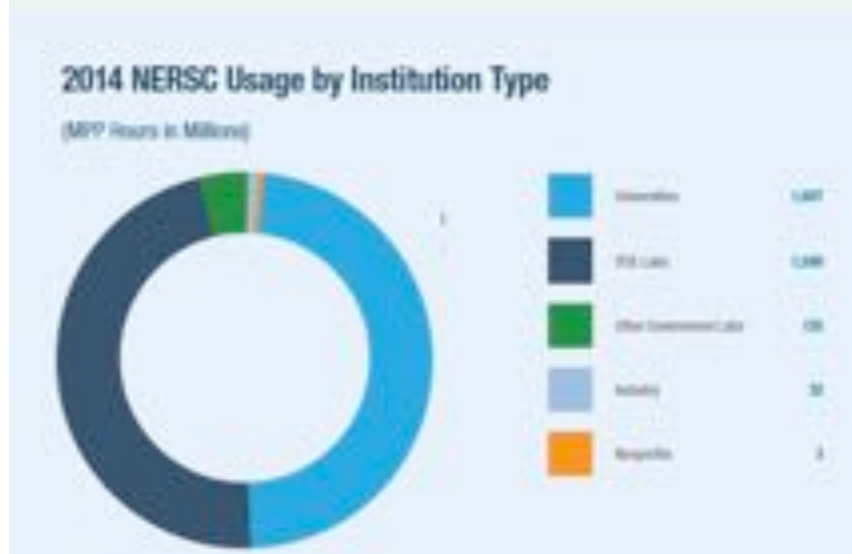
Fusion Energy, Plasma Physics

U.S. DEPARTMENT OF ENERGY | Office of Science

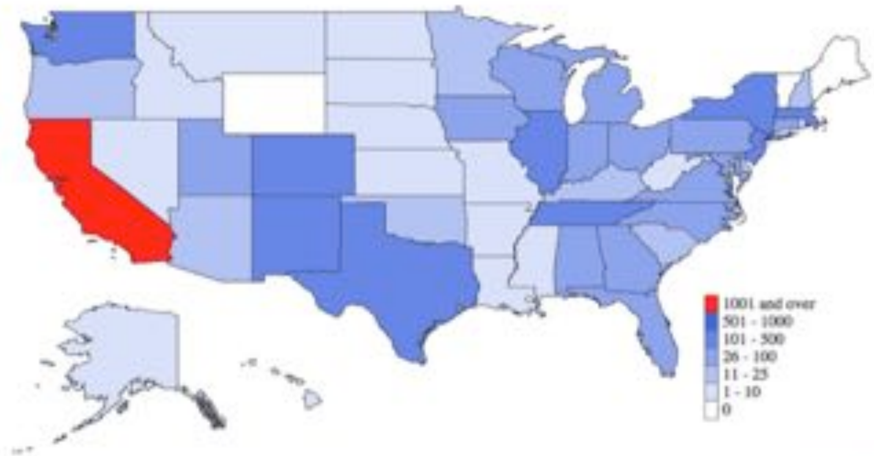
NERSC directly supports DOE's science mission

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



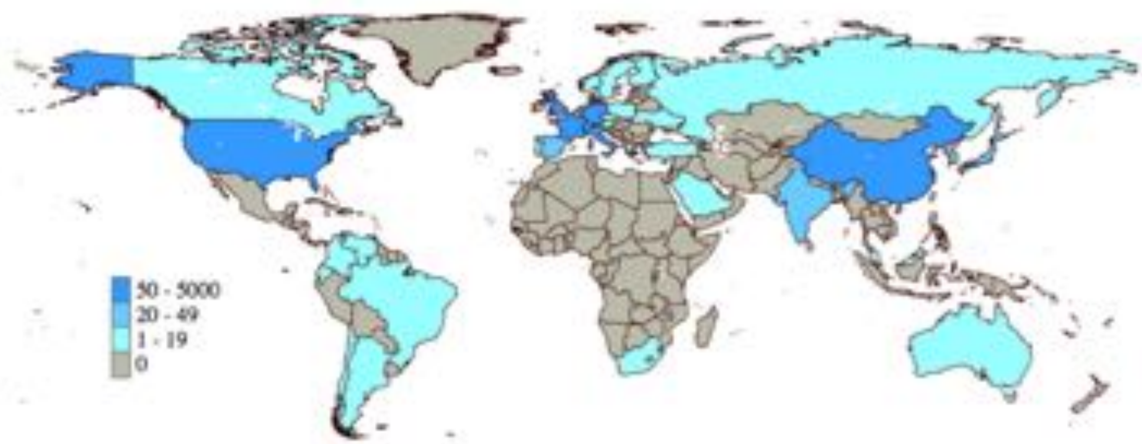


NERSC has a broad user base

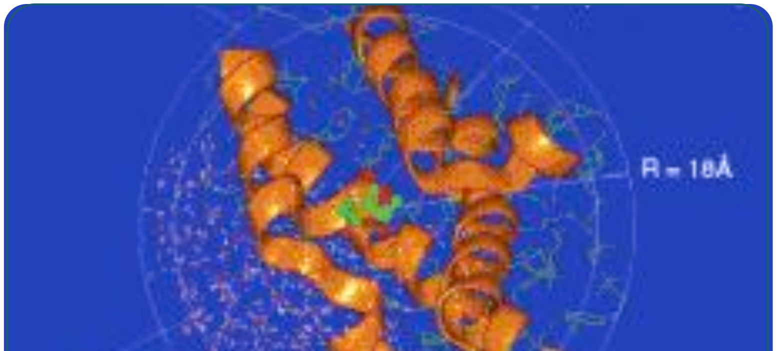


642 international users

5,000 users from 47 U.S. states

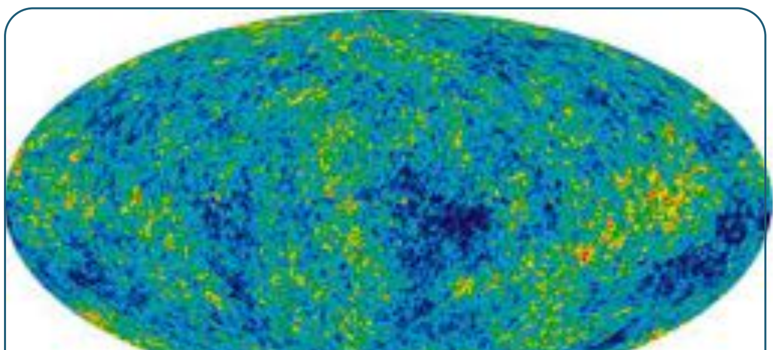


Strong focus on science



Martin Karplus

Saul Perlmutter



George Smoot

Warren Washington

Nobel Prize in Physics 2015

Scientific Achievement

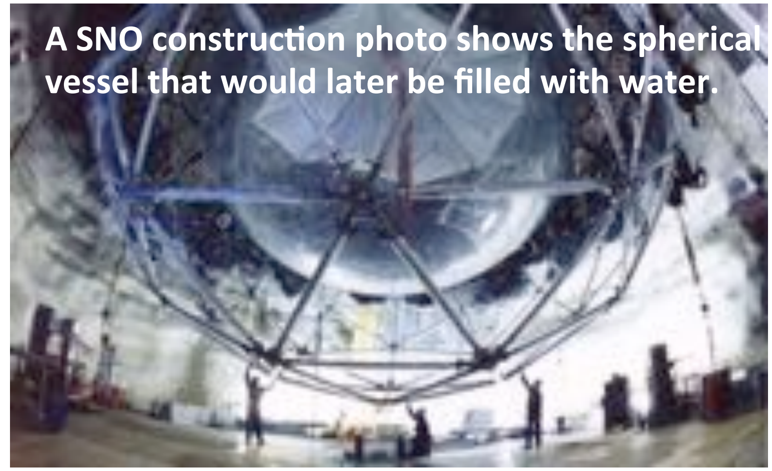
The discovery that neutrinos have mass and oscillate between different types

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.

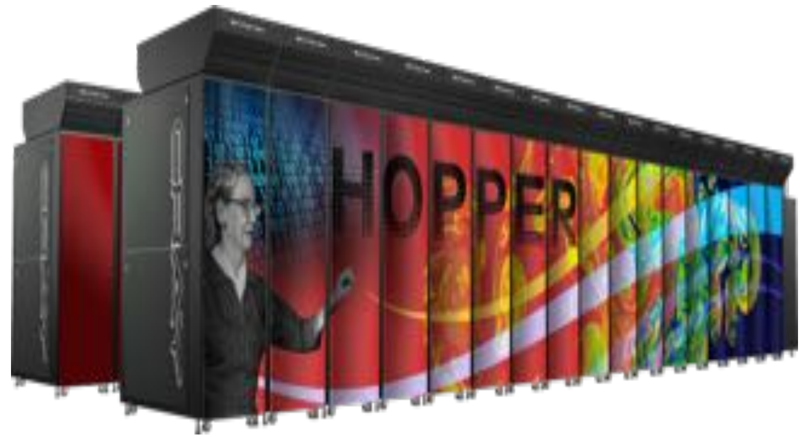
Q. R. Ahmad et al. (SNO Collaboration). Phys. Rev. Lett. 87, 071301 (2001)

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

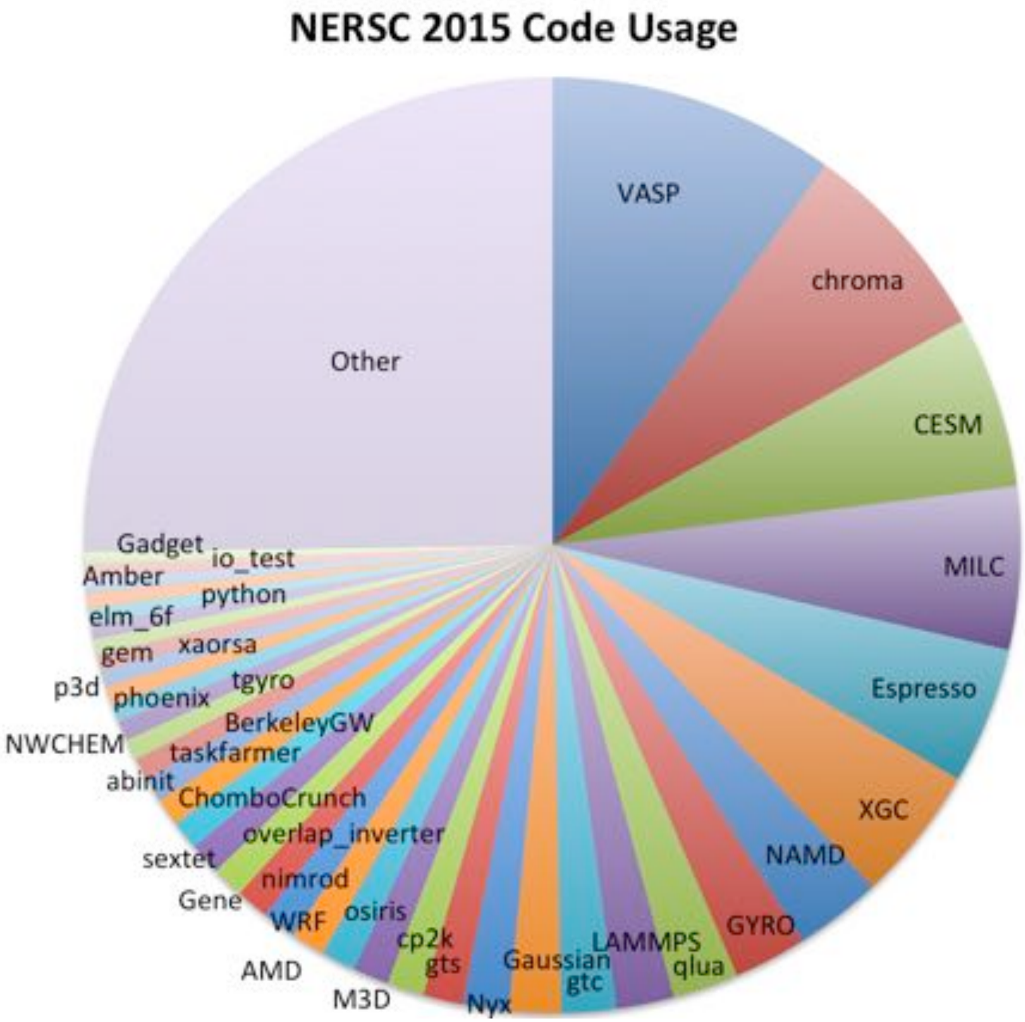
We have a dual mission to advance the state-of-the-art in supercomputing

NERSC

- 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



Over 600 codes run at NERSC



- 10 codes make up 50% of the workload
- 25 codes make up 66% of the workload

NERSC Deployed Edison in 2013

- 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

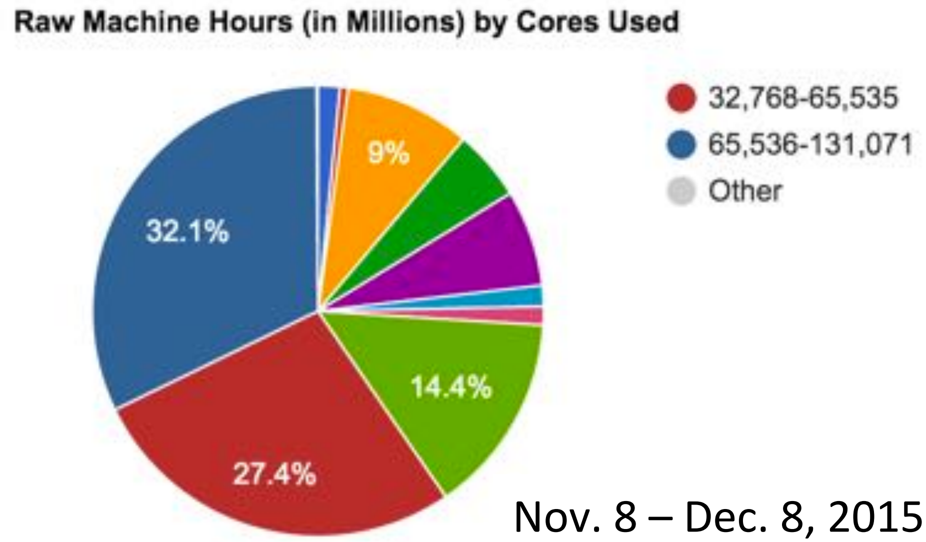
NERSC's Users Require a Low Latency, Highly Scalable Interconnect

The introduction of the new Cori Phase 1 machine has enabled users to run at higher concurrencies on Edison

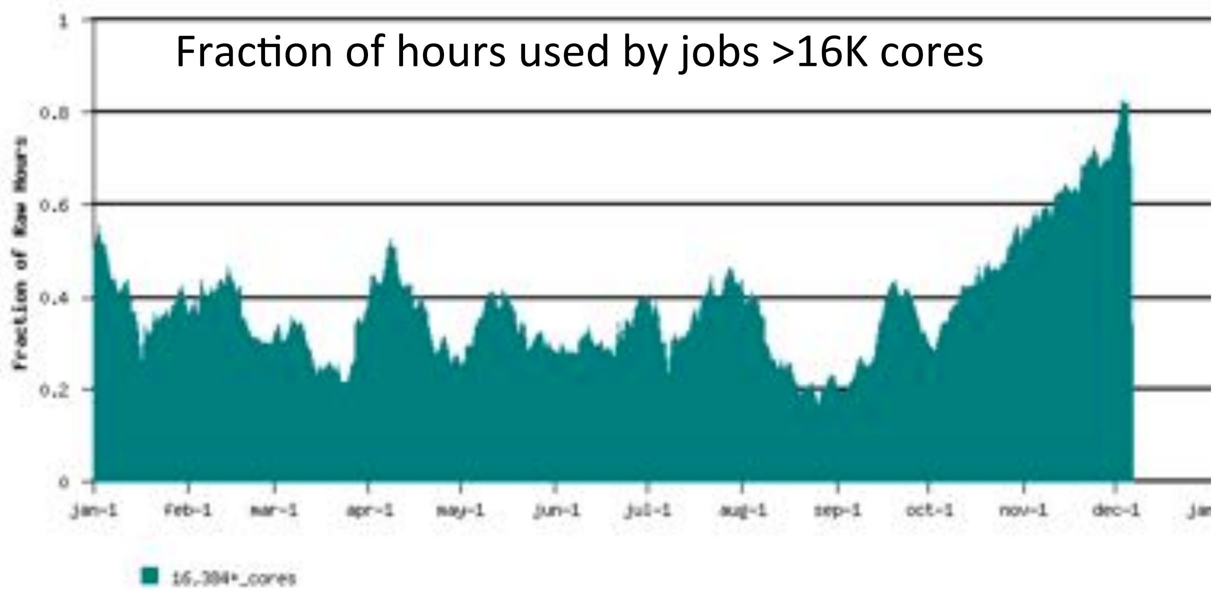
>16K core jobs using 80% of time now

>32K core jobs using 59% of time

>64K core jobs using 32% of time

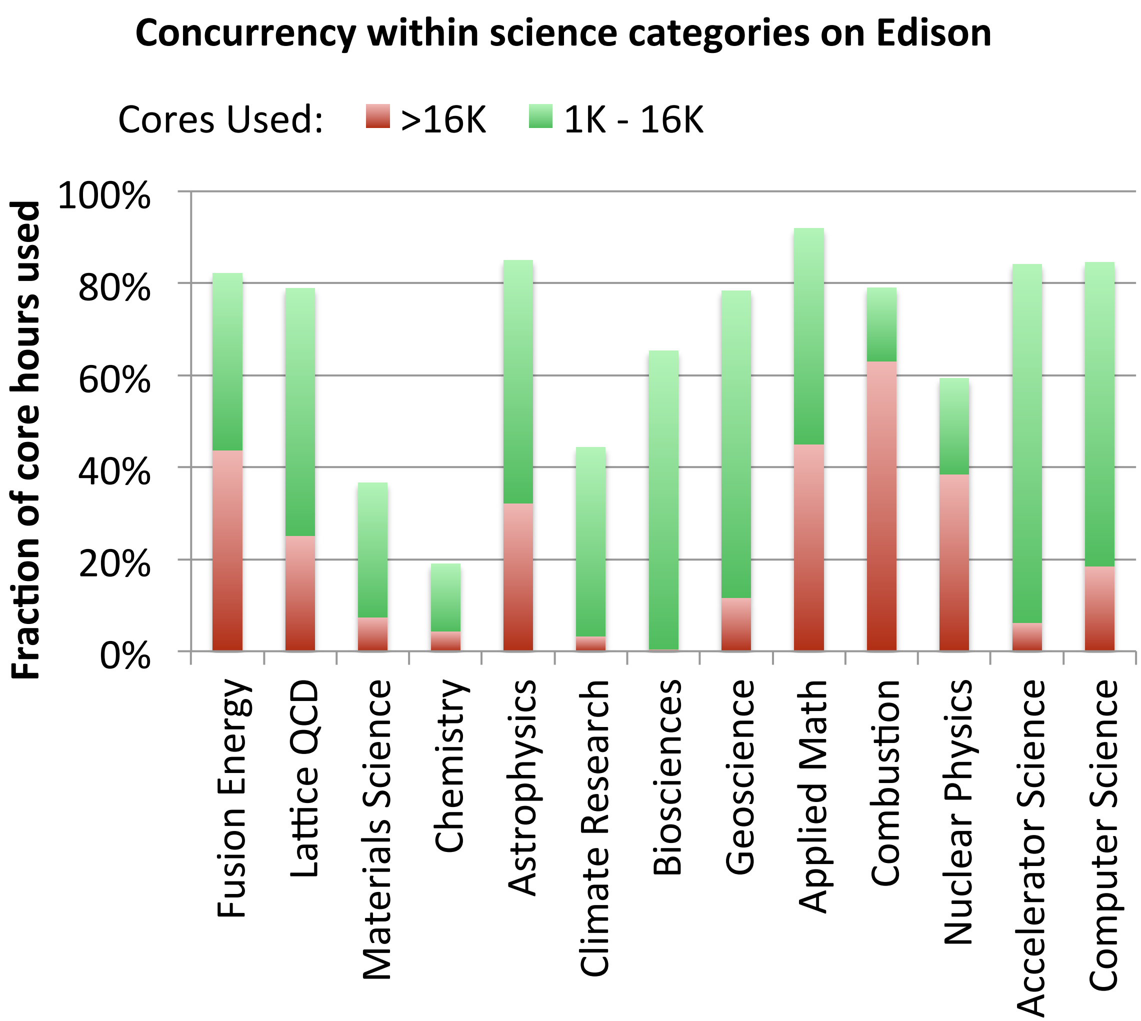


Nov. 8 – Dec. 8, 2015



U.S. DEPARTMENT OF ENERGY | Office of Science

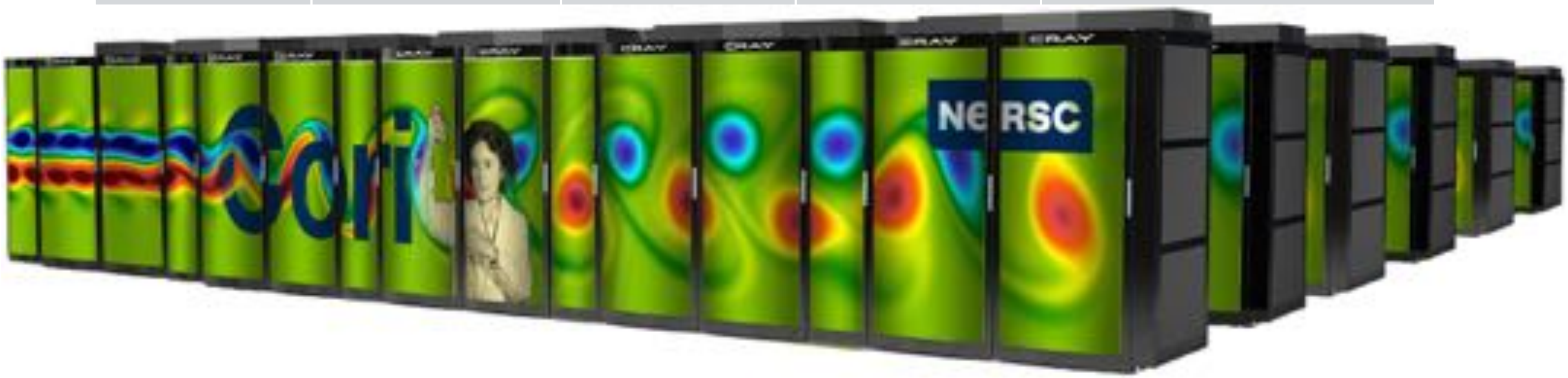
High concurrency jobs are important across all science domains



- **Some fraction of every domain's workload runs with more than 16K cores.**
- **In almost all domains, more than half the workload uses more than 1K cores.**

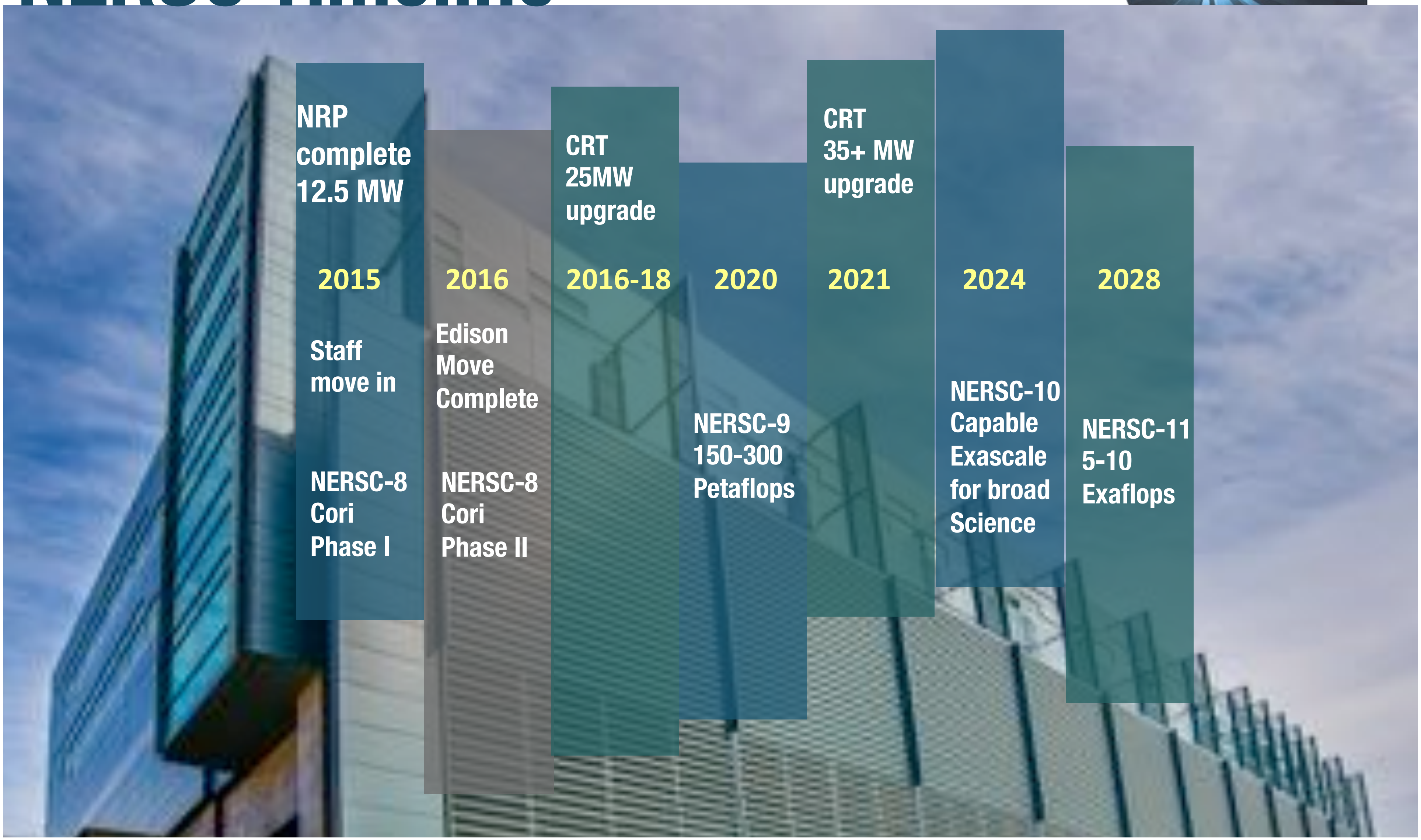
NERSC Systems Timeline

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 Timeline

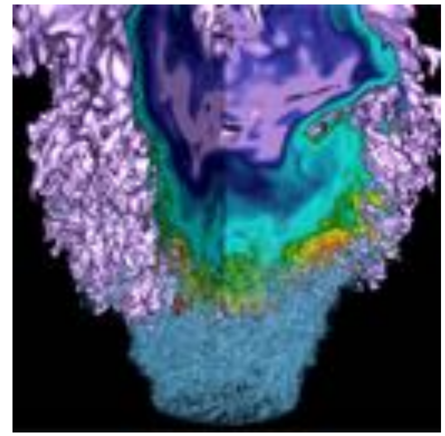
NERSC

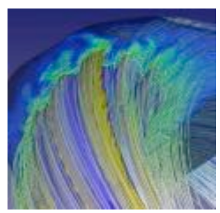


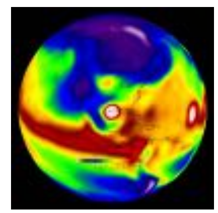
U.S. DEPARTMENT OF ENERGY | Office of Science

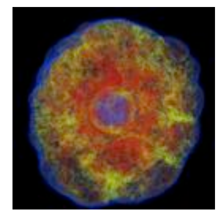
BERKELEY LAB

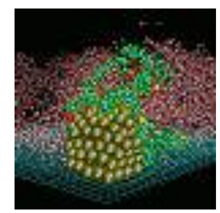
Exascale at NERSC



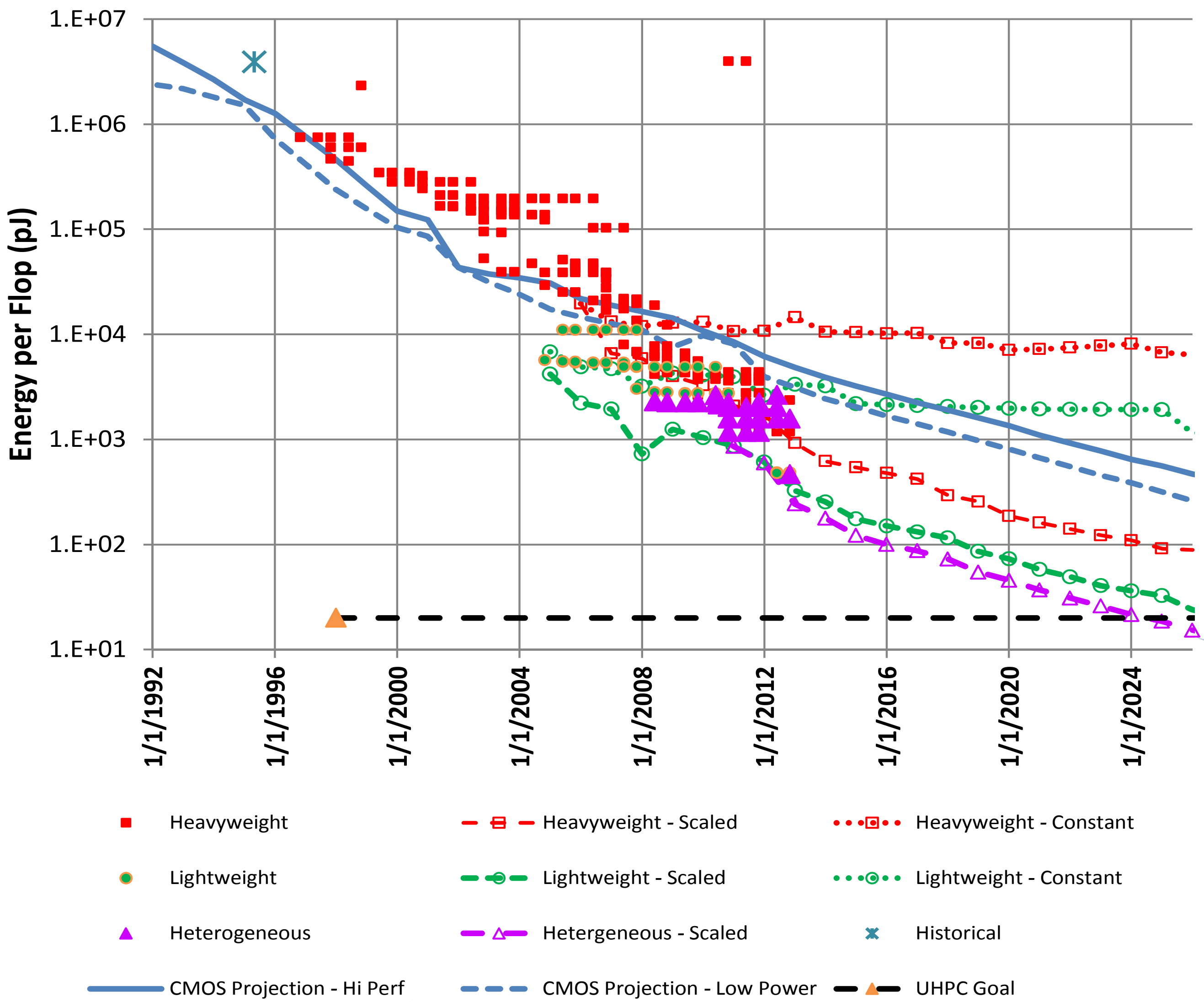








We need to transition to energy efficient architectures to reach Exascale



Manycore or Hybrid is the only approach that crosses the exascale finish line

Cori has an energy-efficient processor

- We didn't deploy accelerators or GPUs in Edison
- Disruptions in programming models are a challenge for NERSC
- What's changed for Cori?
  - Energy-efficient architectures are needed to meet the science needs of our users
  - Heightened awareness about Exascale among application teams
  - Many codes are being adapted for next generation systems
  - Technology changes (e.g., self-hosted many core chips, tighter CPU/GPU integration) are making the transition easier

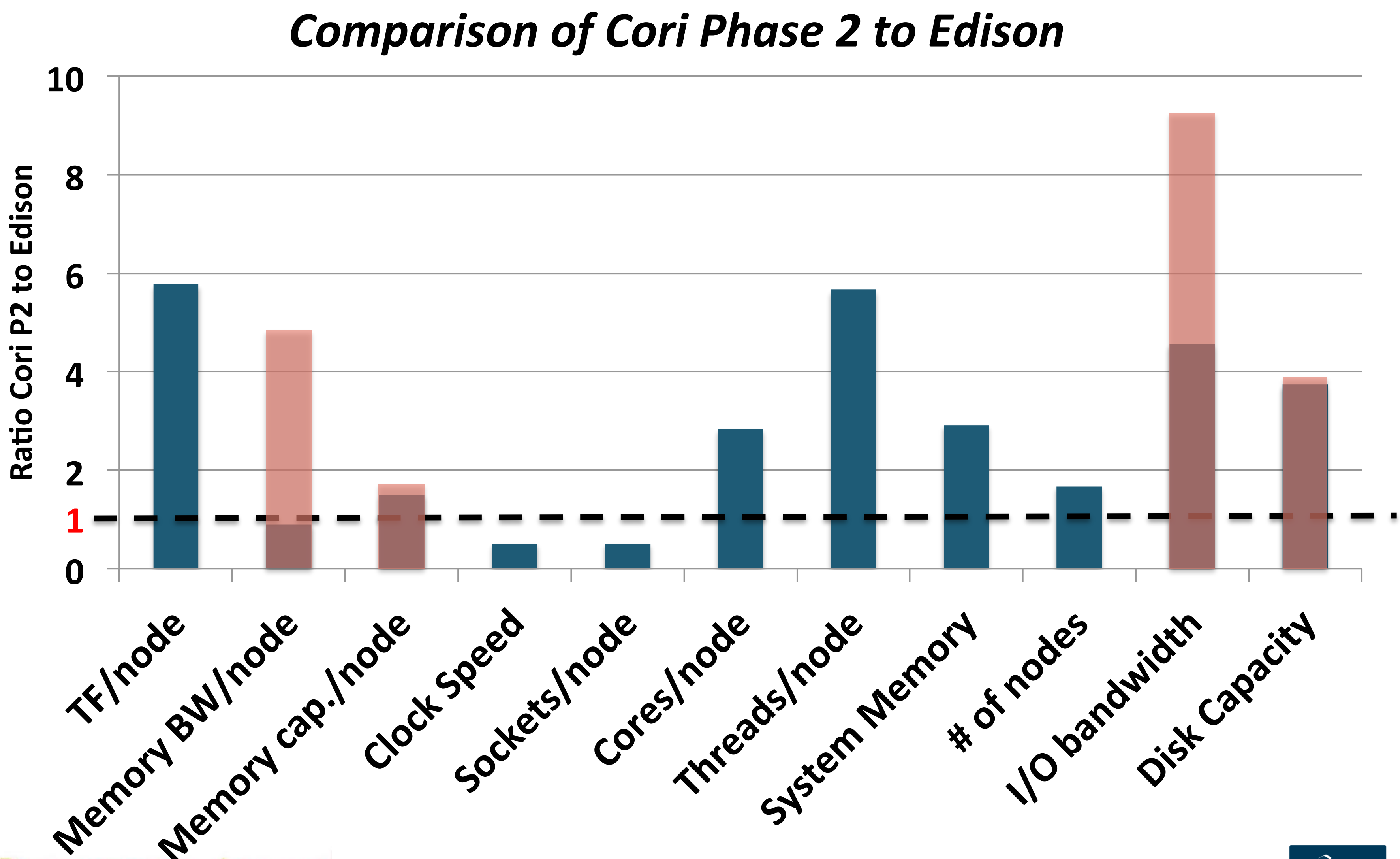
Cori Supports the Office of Science HPC Workload and Data-Intensive Science

- **Cray system with 9,300 Intel Knights Landing compute nodes**
  - Self-hosted, (not an accelerator) manycore processor > 64 cores per node
  - On-package high-bandwidth memory at >400GB/sec
- **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
- **Data Intensive Science Support**
  - 10 Haswell processor cabinets (Phase 1) to support data intensive applications
  - NVRAM Burst Buffer with 1.5PB of disk and 1.5TB/sec
  - 28 PB of disk, >700 GB/sec I/O bandwidth in Lustre bandwith

Intel “Knights Landing” Processor

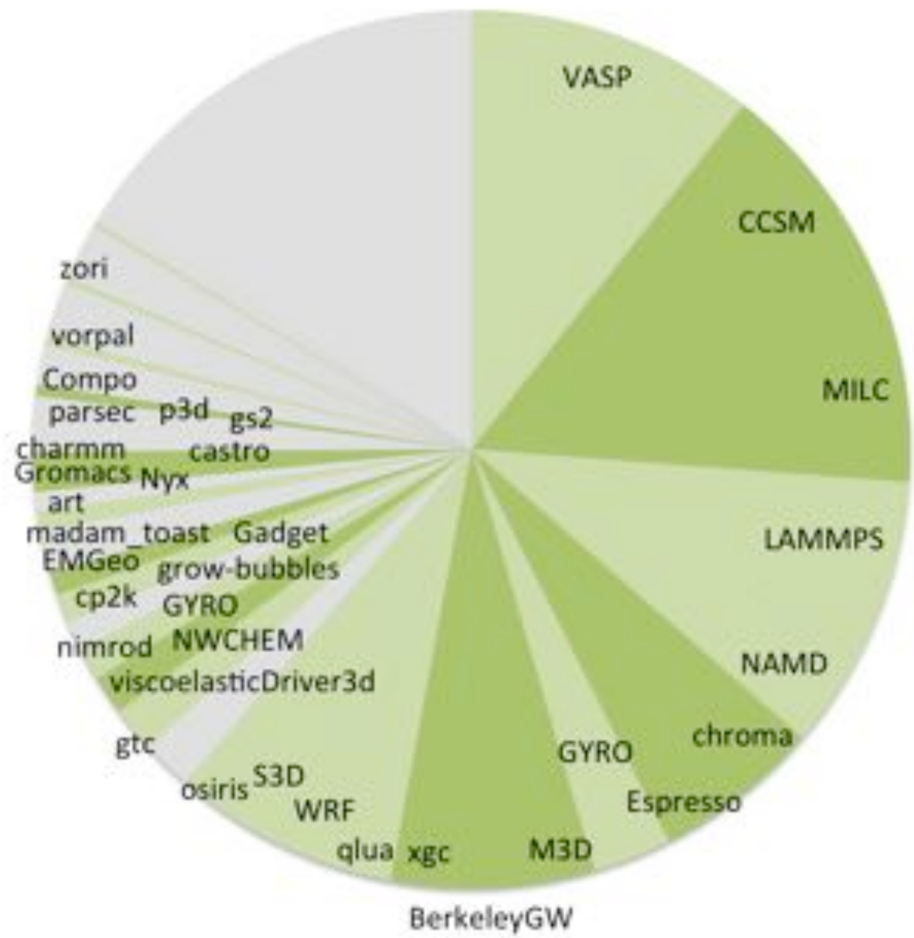
- Next generation Xeon-Phi, 3TF peak
- Single socket processor - Self-hosted, not a co-processor, not an accelerator
- 68 cores per processor with support for four hardware threads each; more cores than current generation Intel Xeon Phi™
- 512b vector units (32 flops/clock – AVX 512)
- 3X single-thread 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

New technologies in Cori Phase 2 increase memory and I/O bandwidth



Transitioning the broad SC workload to energy-efficient architectures

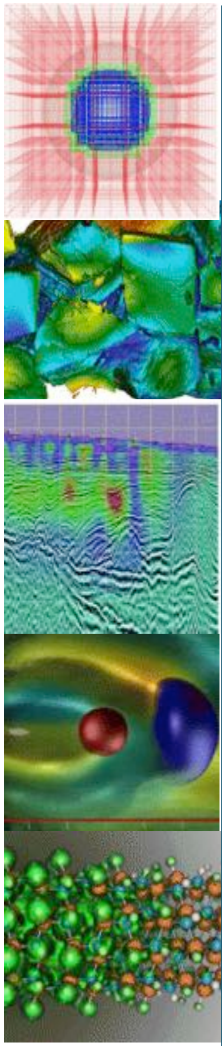
- Collaborate on codes that are proxy's for much of the workload (NESAP)
- Focus on concurrency and locality
- Advocate for adoption of useful features into existing and emerging programming models
- Provide libraries and training to the broader community
- Establish broader forums for sharing information (e.g., the Intel Xeon Phi User Group)



U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

NESAP Codes



Advanced Scientific Computing Research

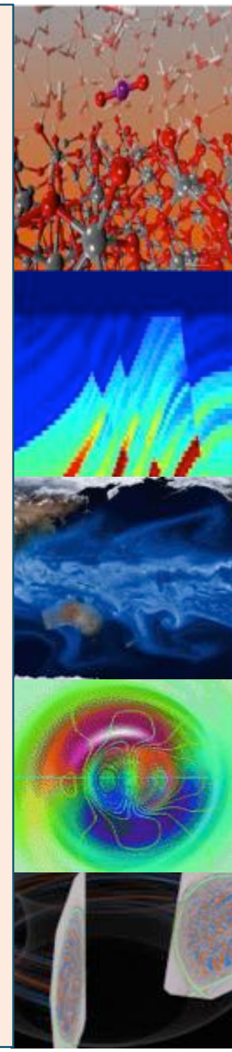
Almgren (LBNL) **BoxLib**
Trebotich (LBNL) **Chombo-crunch**

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

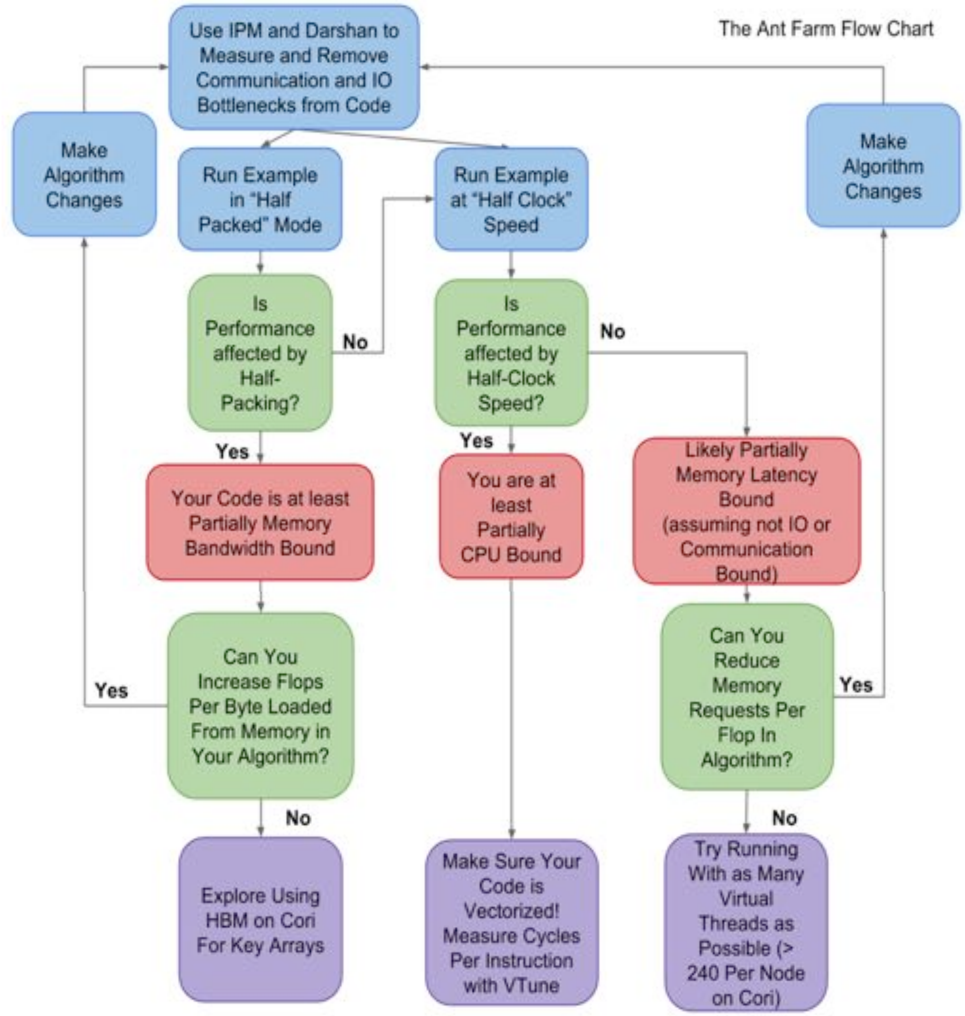
Smith (ORNL) **Gromacs**
Yelick (LBNL) **Meraculous**
Ringler (LANL) **MPAS-O**
Johansen (LBNL) **ACME**
Dennis (NCAR) **CESM**

Fusion Energy Sciences

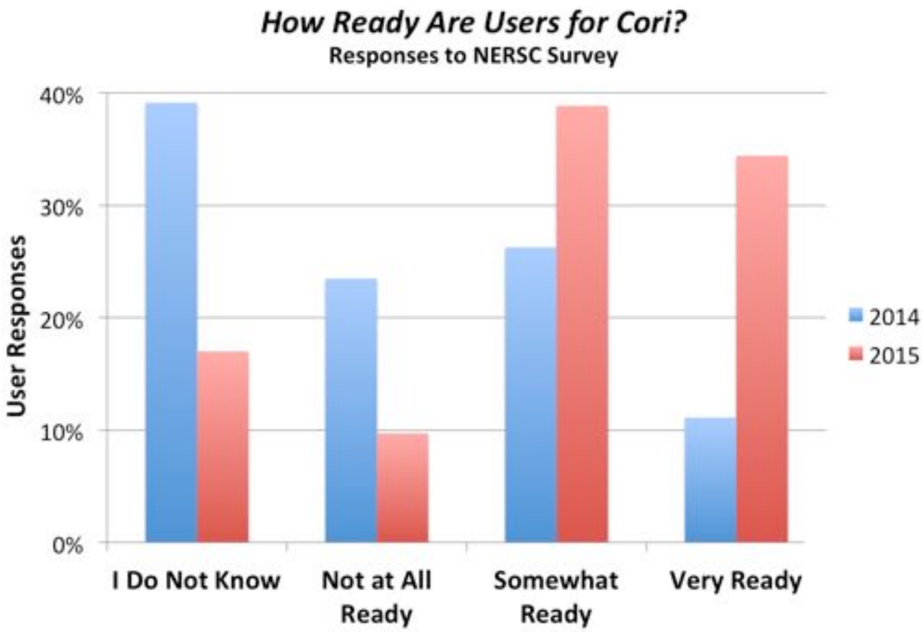
Jardin (PPPL) **M3D**
Chang (PPPL) **XGC1**

Lessons learned

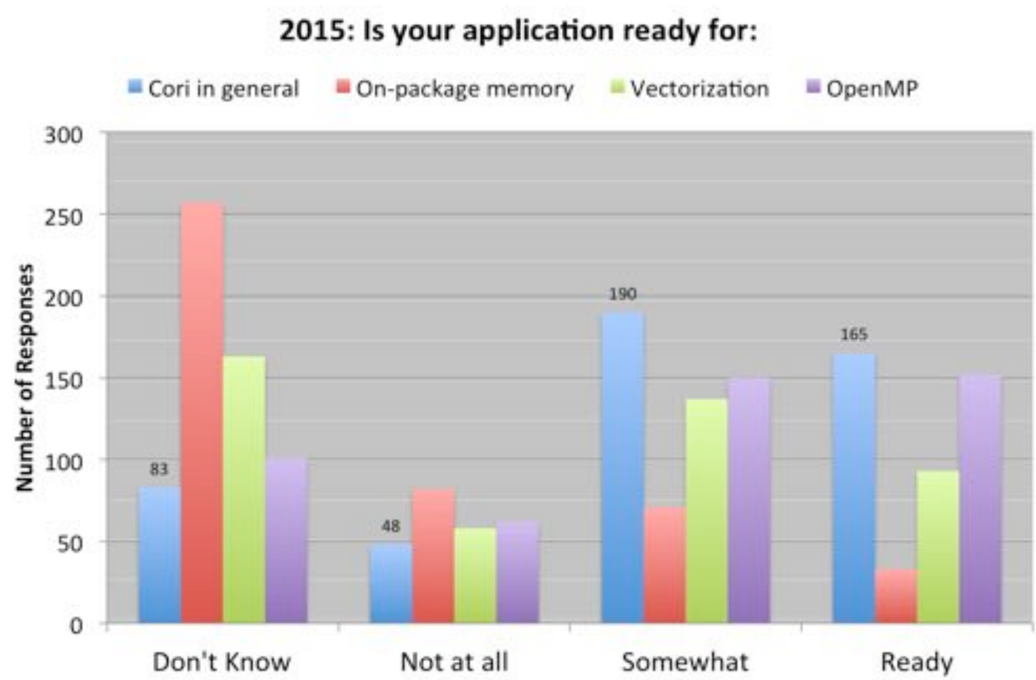
- **Identify/Exploit On-node shared-memory parallelism**
- **Identify/Exploit On-core vector parallelism**
- **Understand and optimize memory bandwidth requirements with MCDRAM**



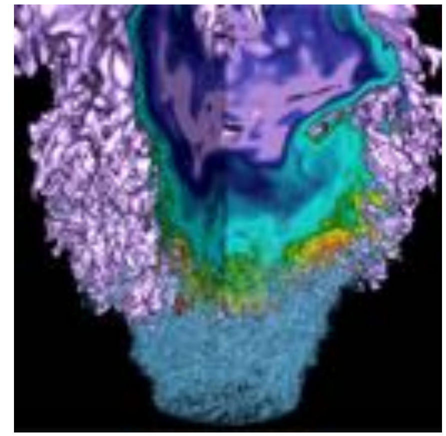
Have our training sessions, outreach and case studies made a difference?

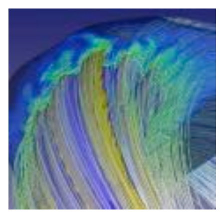


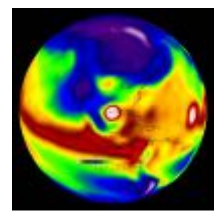
Users report significant increase in readiness and awareness of Cori architecture

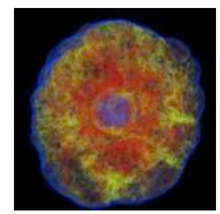


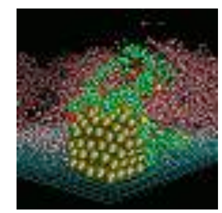
Data Intensive Science









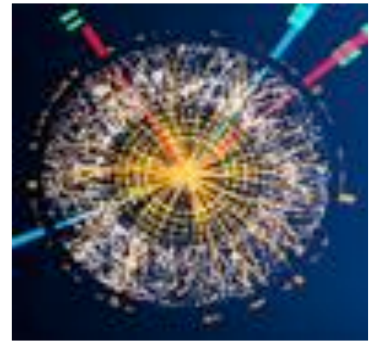


NERSC has been supporting data intensive science for a long time

Palomar Transient Factory
Supernova

Planck Satellite
Cosmic Microwave Background Radiation

Alice
Large Hadron Collider

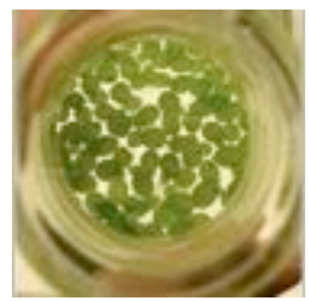


Atlas
Large Hadron Collider

Dayabay
Neutrinos

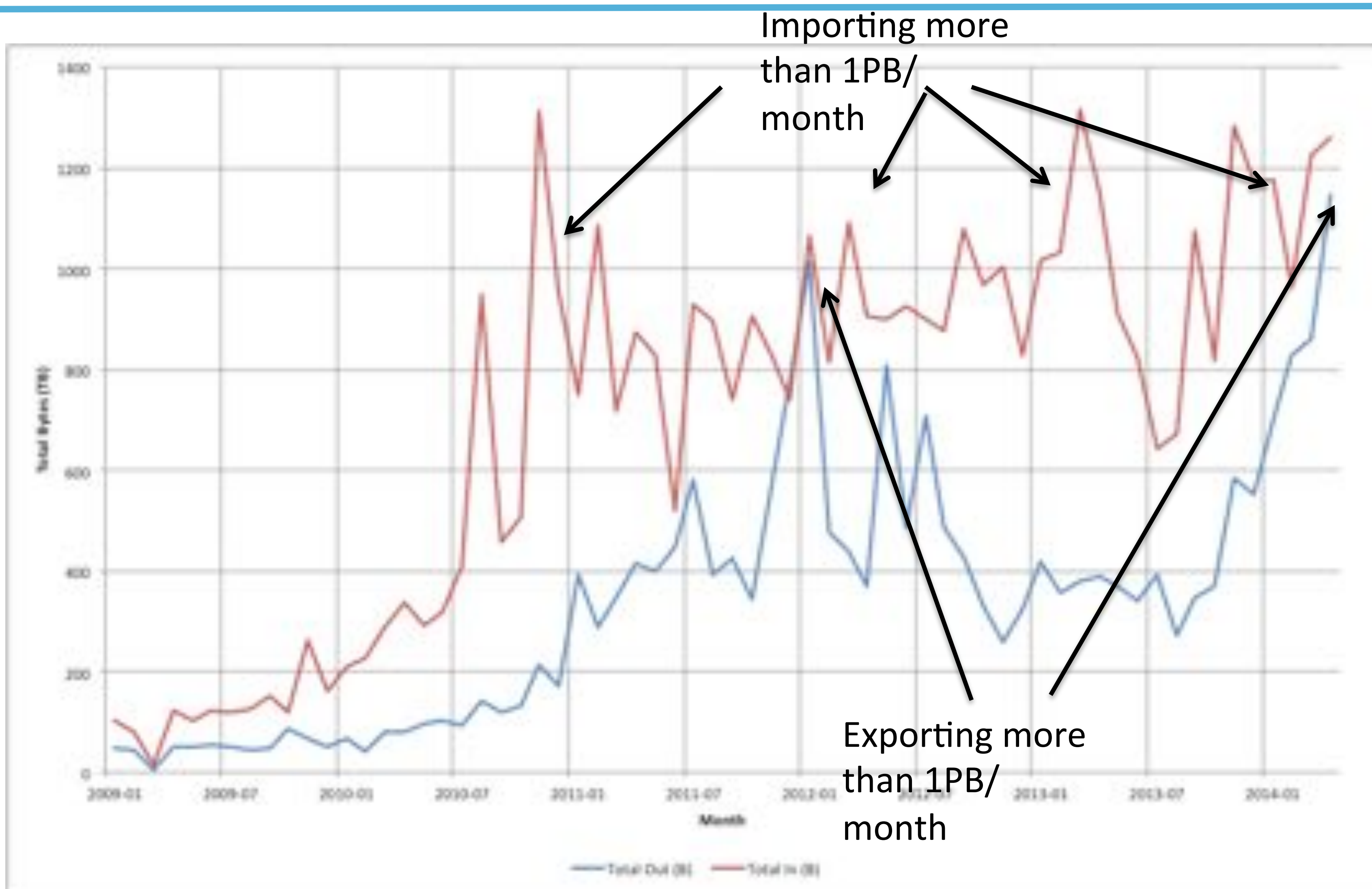
ALS
Light Source

LCLS
Light Source



Joint Genome Institute
Bioinformatics

NERSC users import more data than they export!



U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

We currently deploy separate Compute Intensive and Data Intensive Systems

NERSC

Compute Intensive

Data Intensive

Carver

Genepool

PDSF

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Need for Change

- **Dramatically growing data sets require Petascale+ computing for analysis**
- **We increasingly need to couple large-scale simulations and data analysis**

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

Policies

- Fast-turn around time. Jobs start shortly after submitted
- Can run large numbers of throughput jobs

Software/Configuration

- Support for complex workflows
- Communication and streaming data from external databases and data sources
- Easy to customize user environment

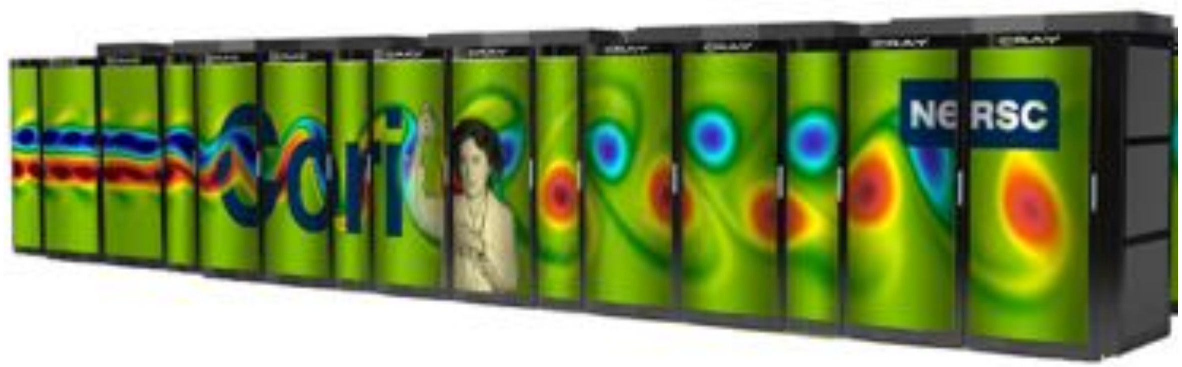
Hardware

- 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

- **High bandwidth external connectivity to experimental facilities from compute nodes (Software Defined Networking)**
- **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
- **Virtualization capabilities (Docker)**
- **More login nodes for managing advanced workflows**
- **Support for real time and high-throughput queues**
- **Big Data Software**



U.S. DEPARTMENT OF ENERGY | Office of Science

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

<u>Cori Network Upgrade Use Case:</u>

- 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

Shifter: Containers for HPC

Challenge and Opportunity

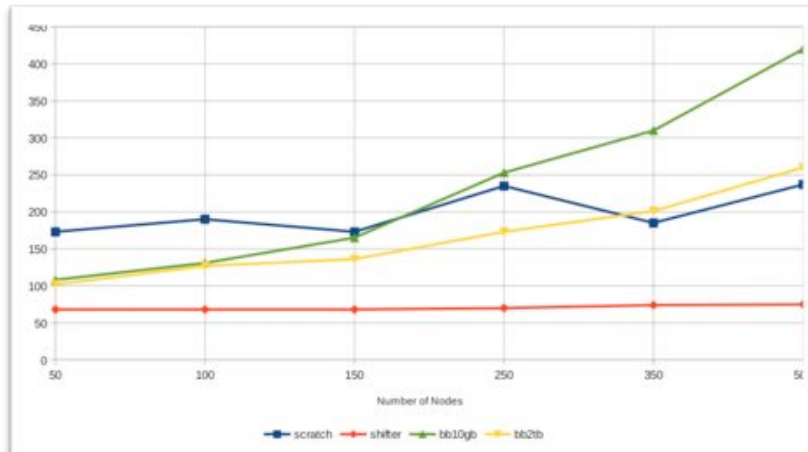
- Data Intensive computing often require large, complex software stacks that are difficult to support in HPC
- Docker is rapidly becoming a new way to package and run applications

Innovation

- Shifter is a NERSC R&D effort, in collaboration with Cray, to support User-created Application images.
- Shifter provides “Docker-like” functionality for HPC

Impact and Early Successes

- Shifter has already enabled multiple projects to quickly make use of NERSC (e.g. LCLS, LHC)
- Shifter can improve job-startup times and application performance (e.g. Python)
- Shifter will be supported by Cray and is already being evaluated by other HPC centers




Biting the hand that feeds IT

DATA CENTER SOFTWARE NETWORKS SECURITY INFRASTRUCTURE DEVOPS BUSINESS HARDWA

Data Center ▸ HPC

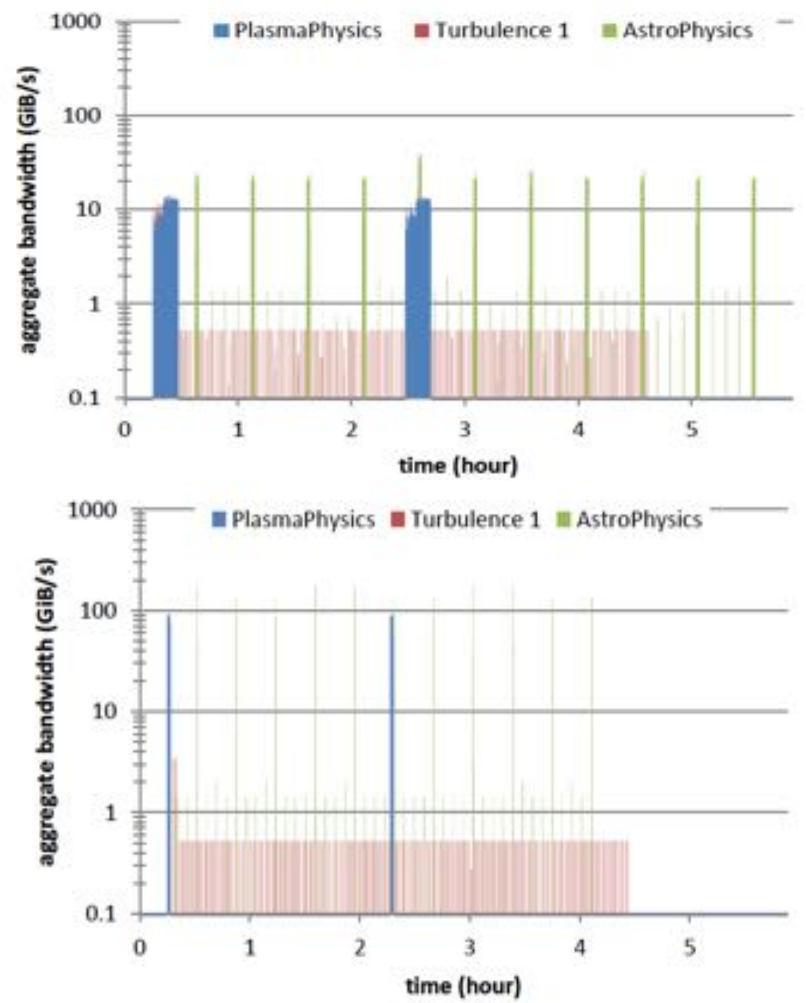
Cray hoists Docker containers into supercomputers

Productivity gains without performance hits

18 Nov 2015 at 00:01, Drew Cullen

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**

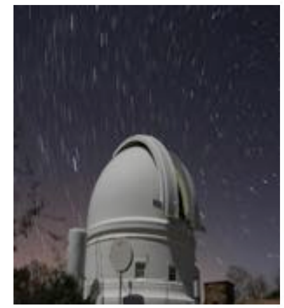


Application perceived I/O rates, with no burst buffer (top), burst buffer (bottom).

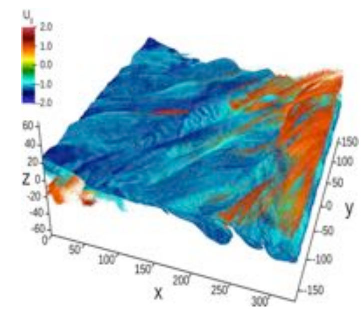
Analysis from Chris Carothers (RPI) and Rob Ross (ANL)

NERSC is exploring Burst Buffer Use Cases beyond checkpoint-restart

- **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

Big Data Software Portfolio

Capabilities	Technology Areas	Tools, Libraries
Data Transfer + Access	Globus, Grid Stack, Authentication	Globus Online, Grid FTP
	Portals, Gateways, RESTful APIs	Drupal/PHP, Django/Python, NEWT
Data Processing	Workflows	Swift, Fireworks, …
Data Management	Formats, Models Databases	HDF5, NetCDF, ROOT MongoDB, SciDB, PostgreSQL, MySQL
	Storage, Archiving	Lustre/GPFS, HPSS, SRM
Data Analytics	Statistics, Machine Learning	python, R, ROOT BDAS/Spark
	Imaging	OMERO, Fiji, MATLAB
Data Visualization	SciVis InfoVis	VisIt, Paraview

Conclusions

- **Progress is being made on a wide range of codes**
  - Focus is on concurrency and locality
- **Increasing engagement from scientific facilities**
  - Burst buffer
  - Real time queues
  - Software defined networking
- **Focus on improved scalability for deep learning**
- **Including data benchmarks and workflows in planning for NERSC-9 (2020)**