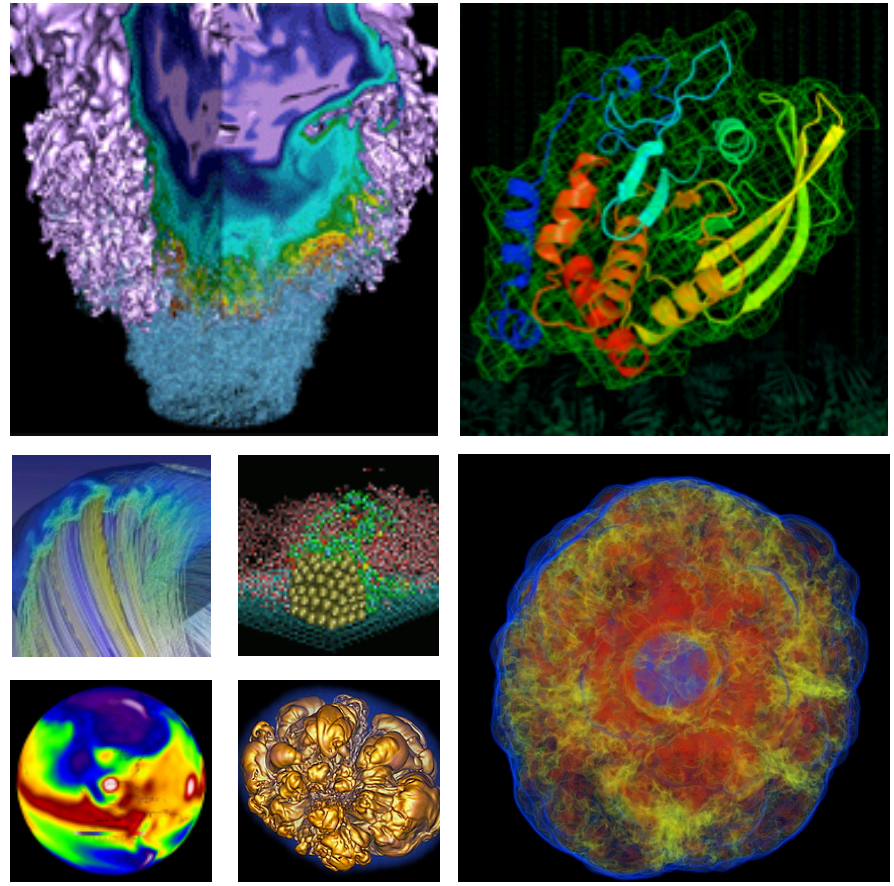
Overview of Cori



Sudip Dosanjh
Director

November 21, 2014

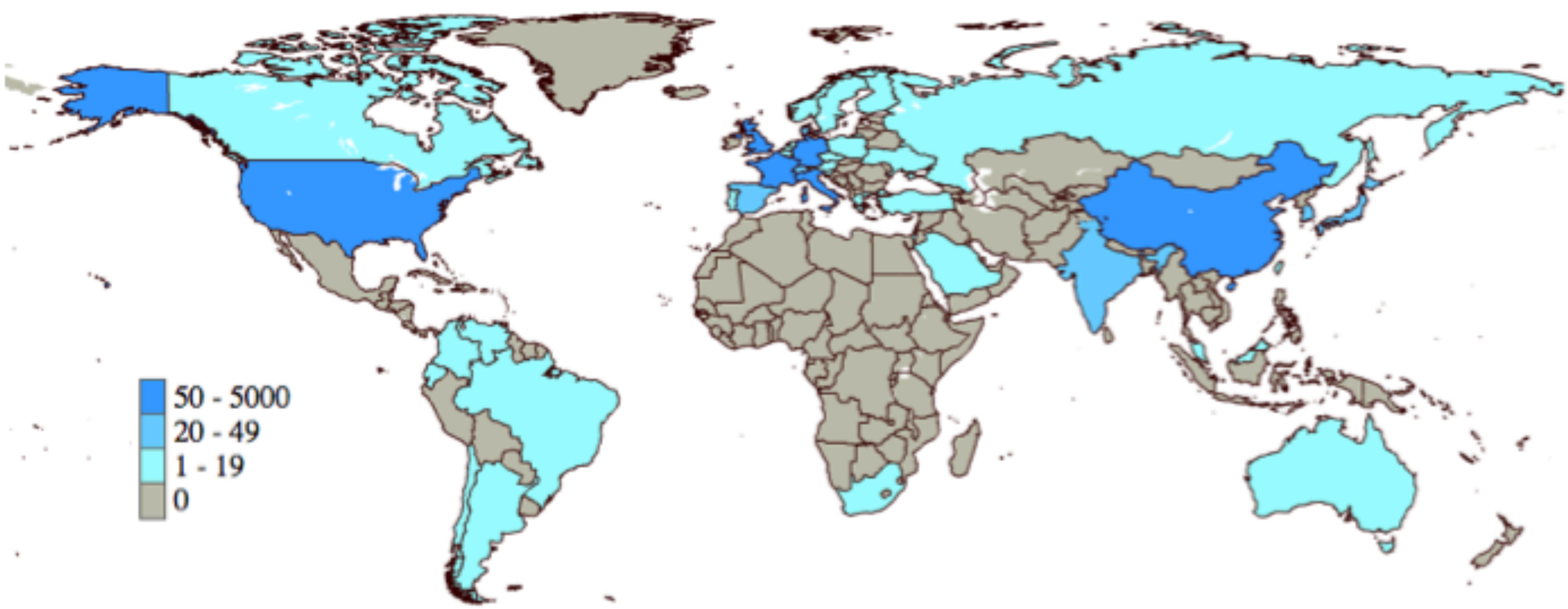
NERSC's latest system is Edison

- Edison is a HPCS demo system (serial #1)
- First Cray Petascale system with Intel processors, Aries interconnect and Dragonfly topology
- Very high memory bandwidth (100 GB/s per node), interconnect bandwidth and bisection bandwidth
- 64 GB/node
- Exceptional application performance

- **Edison doesn't deploy accelerators or GPUs**
- **Disruptions in programming models are a challenge for NERSC**
  - Many users
  - Many codes
  - We don't select our users

We support a broad user base

- **5000 users, and we typically add 350 per year**
- **Geographically distributed: 47 states as well as multinational projects**



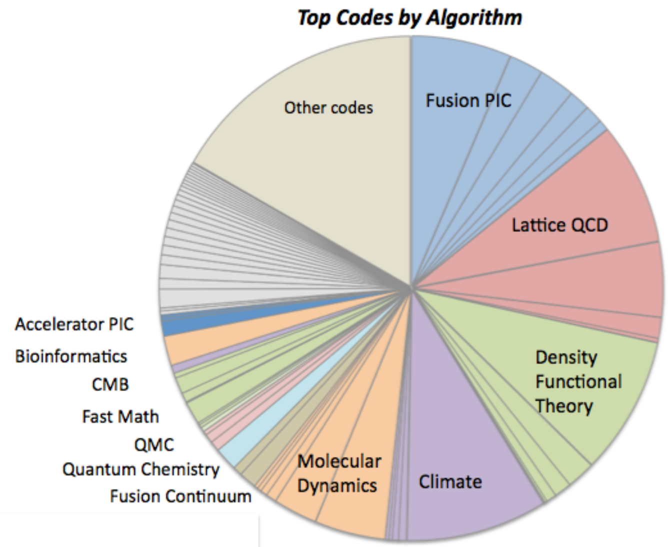
NERSC Users at ASCAC Institutions

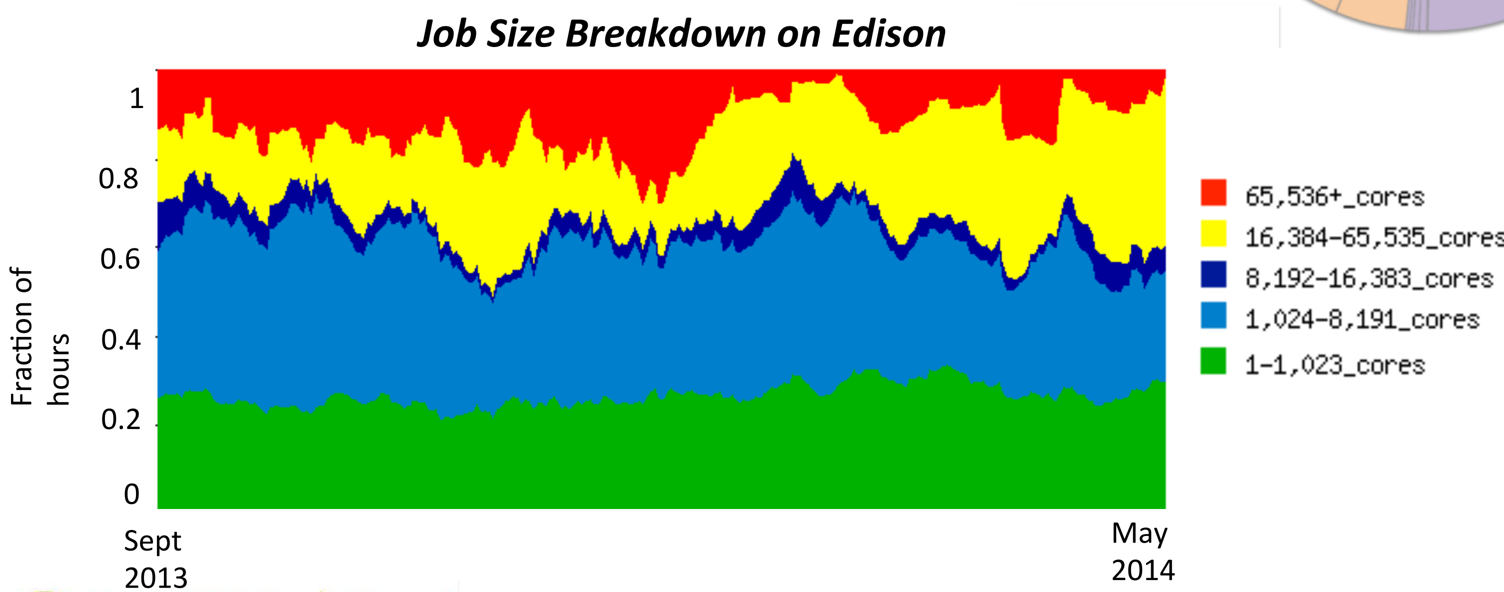
- Boston University – 2
- New York University - 4
- University of Houston – 26
- Google Inc. – 0
- Sandia National Lab – 69
- University of Tennessee – 50
- University of Michigan -60
- Microsoft Research – 0
- Pittsburgh Supercomputer Center – 0 (CMU: 15 - U Pittsburgh: 3)
- MIT– 120
- UC Merced – 4
- UC Santa Barbara – 39
- Rice University – 33
- LLNL– 77

- Total: 467

We support a diverse workload

- **Many codes (600+) and algorithms**
- **Computing at scale and at high volume**

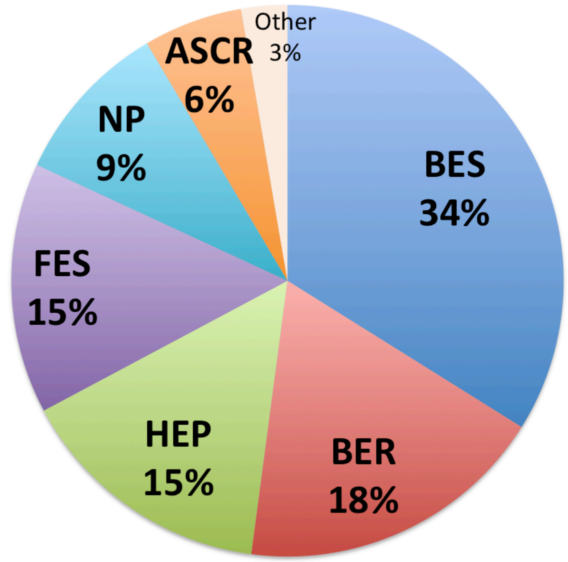


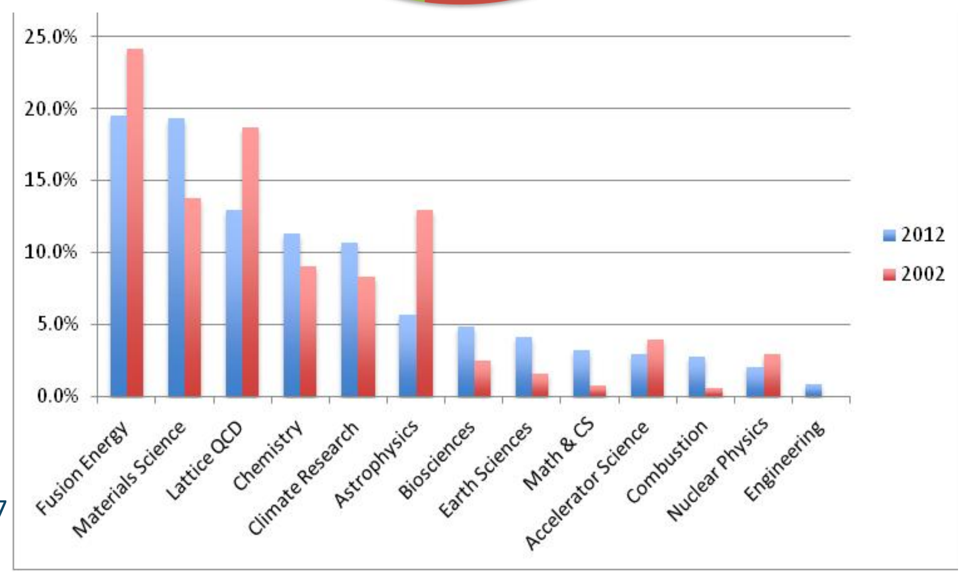


We directly support DOE's science mission

- **We are the primary computing facility for DOE Office of Science**
- **DOE SC allocates the vast majority of the computing and storage resources at NERSC**
  - Six program offices allocate their base allocations and they submit proposals for overtargets
  - Deputy Director of Science prioritizes overtarget requests
- **Usage shifts as DOE priorities change**

2014 Allocation Breakdown

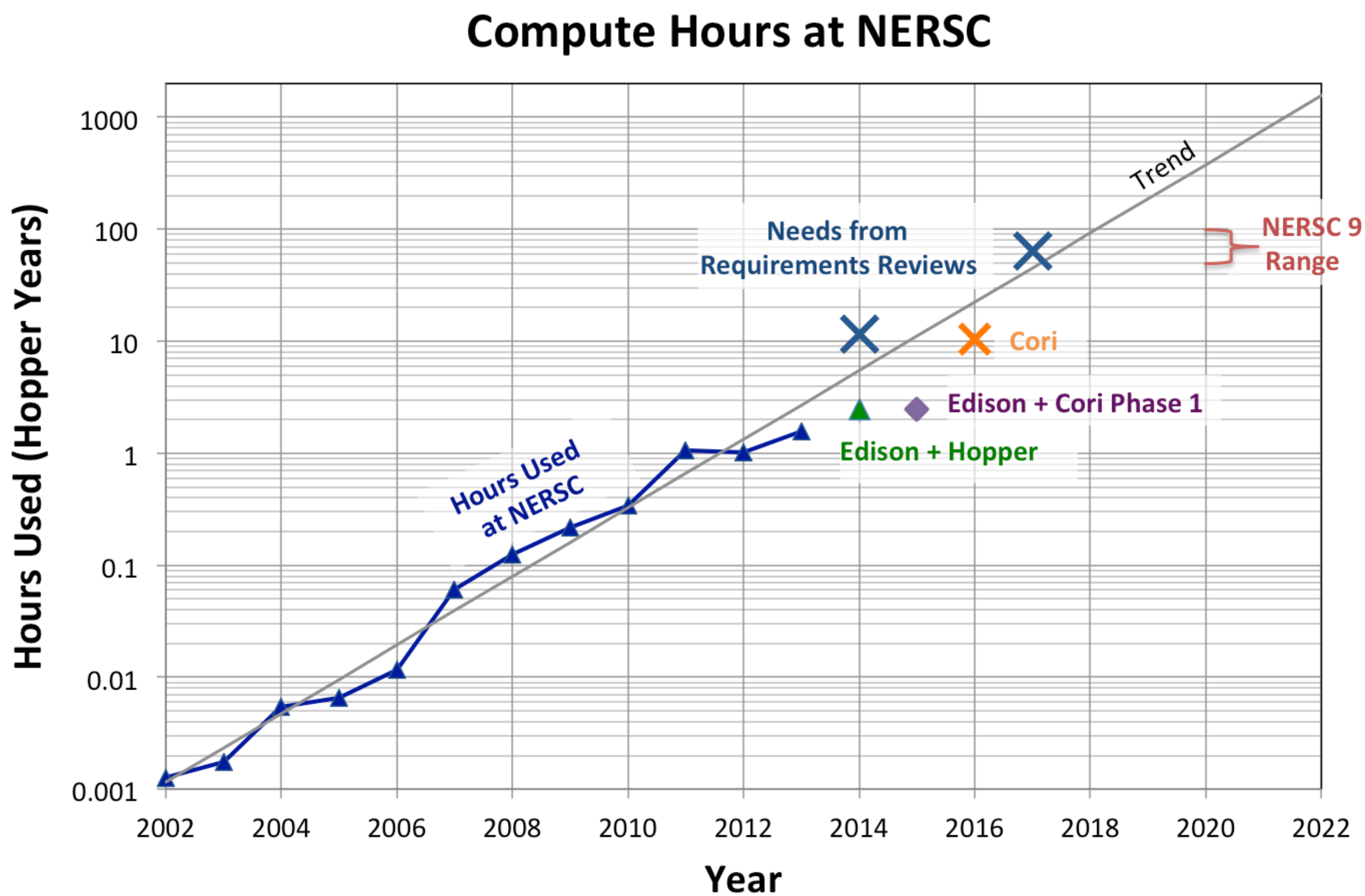




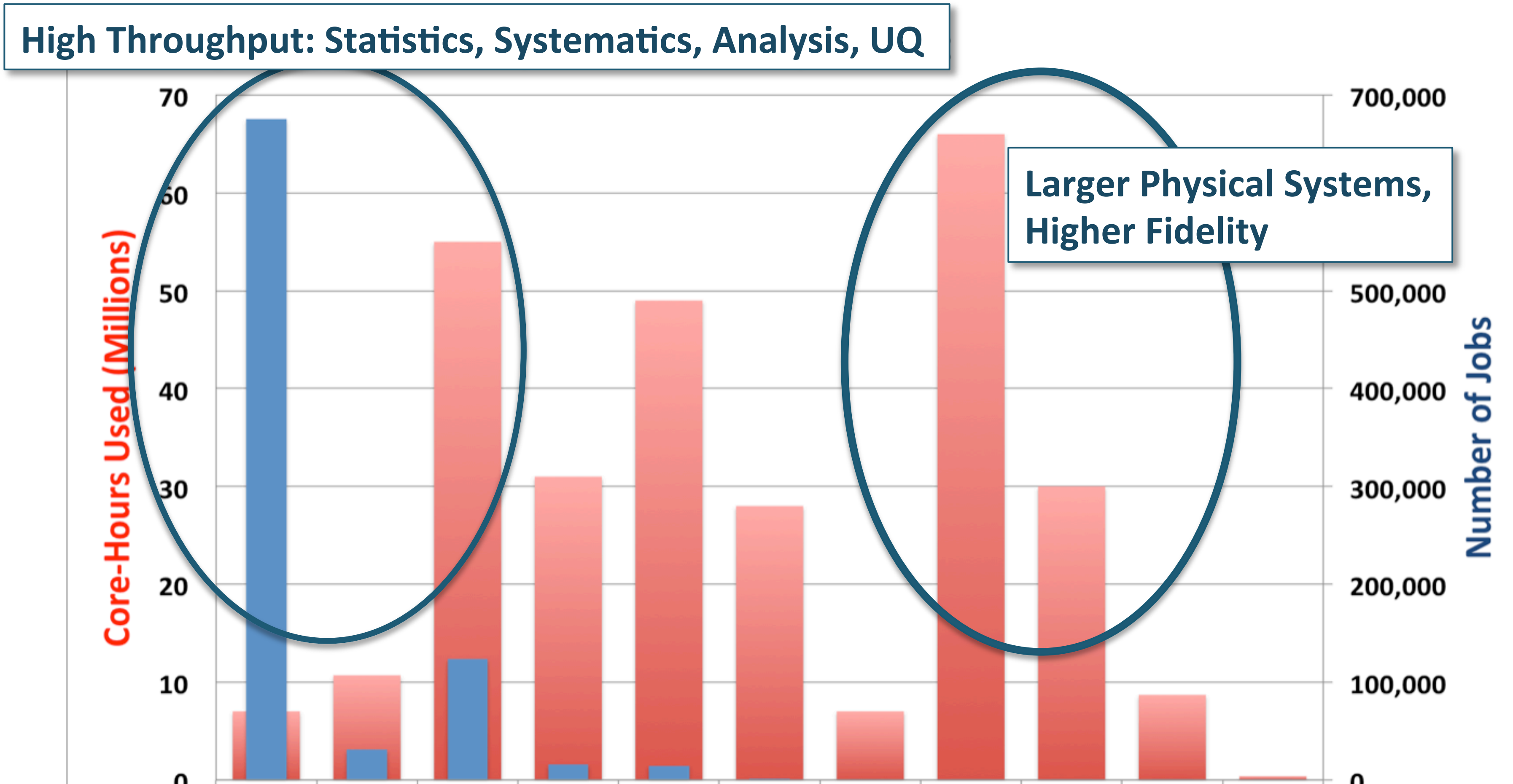
What's changed for Cori?

- **Heightened awareness 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) will make the transition easier**
- **We must transition to energy efficient architectures to meet the science needs of our users**

Keeping up with user needs will be a challenge

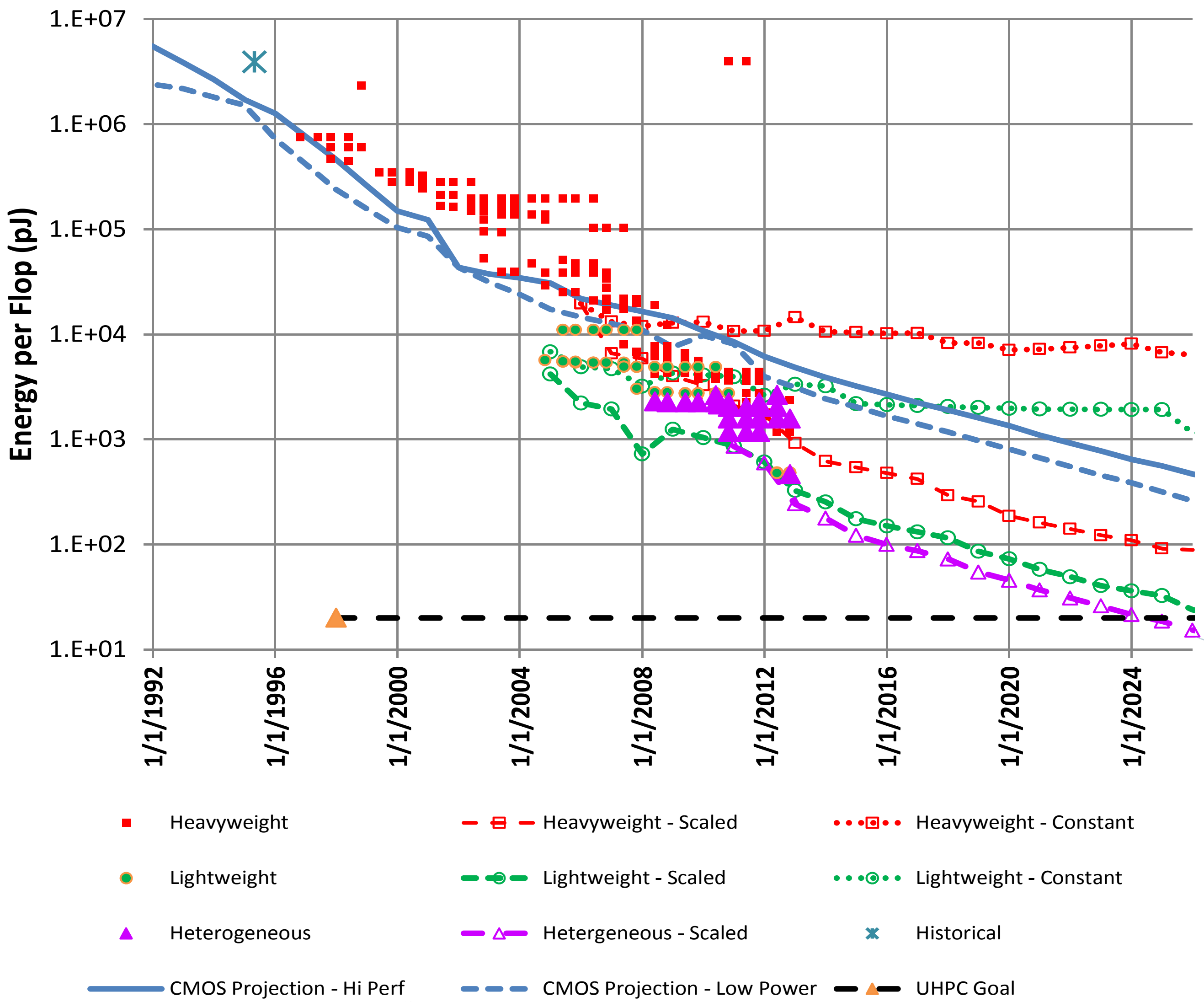


NERSC Supports Science Needs at Many Difference Scales and Sizes



	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

NERSC needs to transition to energy efficient architectures



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

NERSC-8 (Cori) Mission Need

The Department of Energy Office of Science requires an HPC system to support the rapidly increasing computational demands of the entire spectrum of DOE SC computational research.

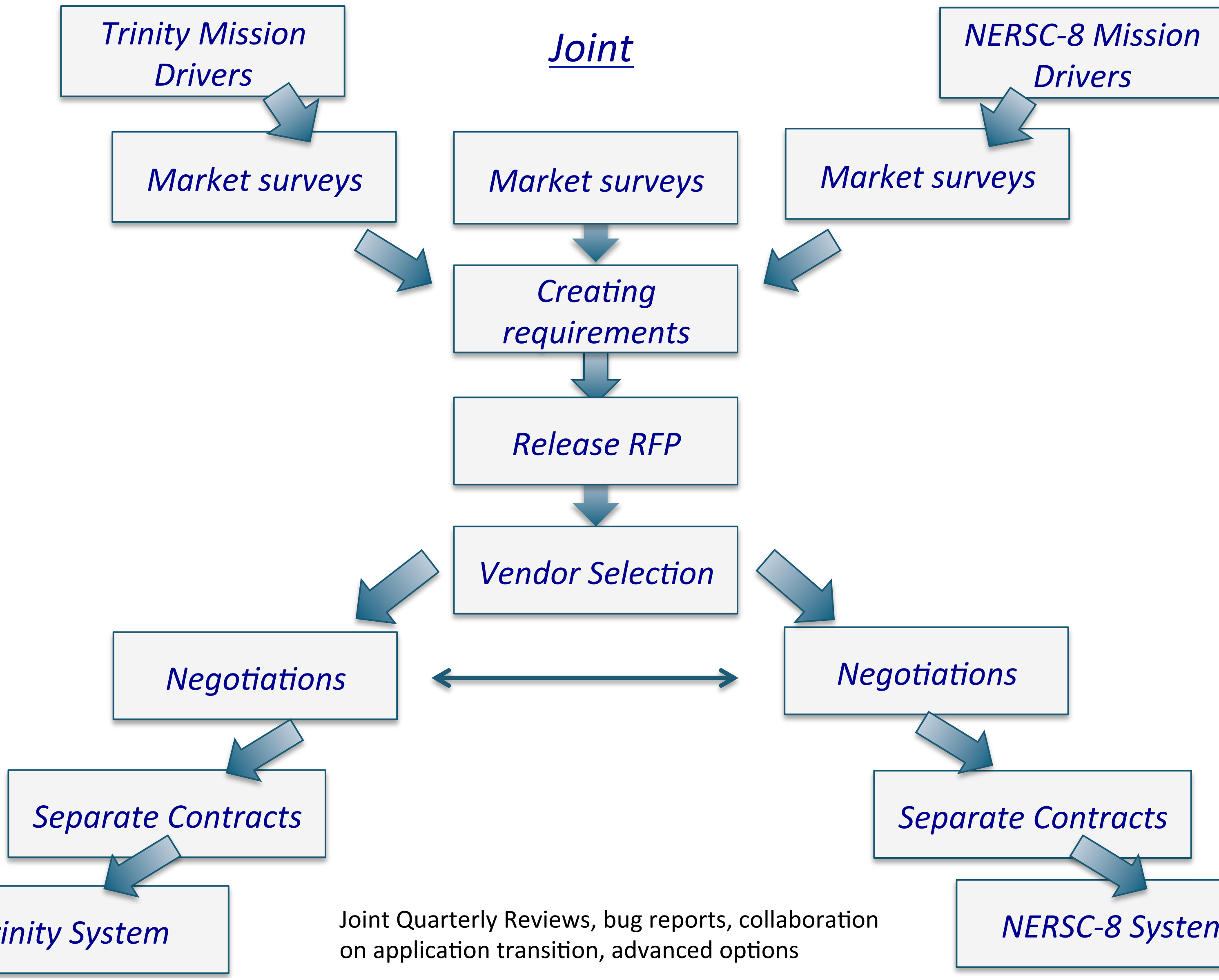
- Provide a significant increase in computational capabilities, at least 10 times the sustained performance of the Hopper system on a set of representative DOE benchmarks
- Delivery in the 2015/2016 time frame
- Provide high bandwidth access to existing data stored by continuing research projects.
- Platform needs to begin to transition users to more energy-efficient many-core architectures.

ACES and NERSC formed a partnership for next-generation supercomputers

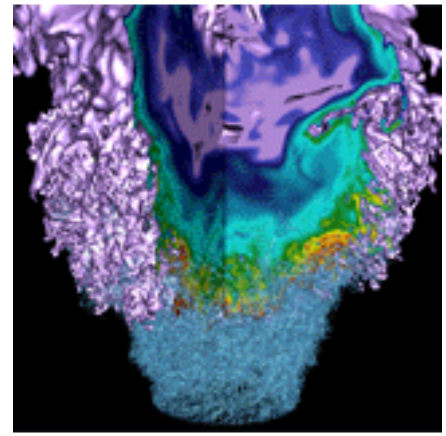
- **Visible collaboration between ASCR and ASC**
- **Strengthen impact on industry**
- **Address challenges transitioning applications to advanced manycore architectures with a broader coalition**
- **Act as a risk mitigation strategy for NERSC-8 and Trinity systems by having a partner to work with on technical challenges deploying and testing NERSC-8 and Trinity**

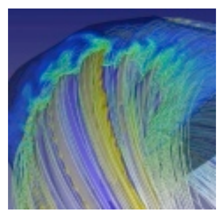
Alliance for application Performance at the EXtreme scale (APEX)

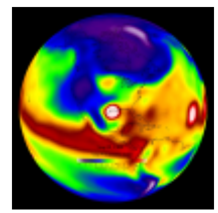
This was a collaboration of two separate projects

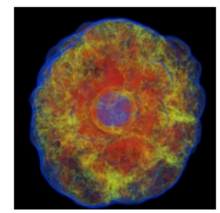


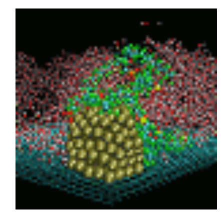
The Cori system











Cori Configuration

- **64 Cabinets of Cray XC System**
  - Over 9,300 'Knights Landing' compute nodes
    - 64-128 GB memory per node
  - Over 1900 'Haswell' compute nodes
    - Data partition
  - 14 external login nodes
  - Aries Interconnect (same as on Edison)
  - > 10x Hopper sustained performance using NERSC SSP metric
- **Lustre File system**
  - 28 PB capacity, 432 GB/sec peak performance
- **NVRAM "Burst Buffer" for I/O acceleration**
- **Significant Intel and Cray application transition support**
- **Delivery in mid-2016; installation in new LBNL CRT**

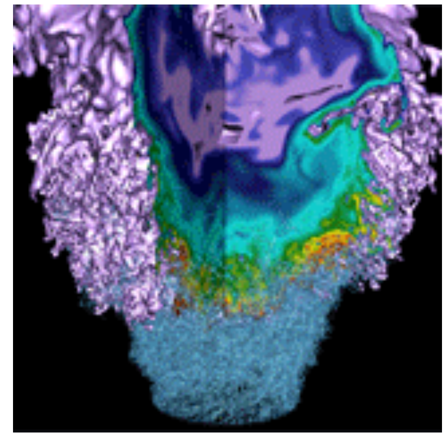
Intel “Knights Landing” Processor

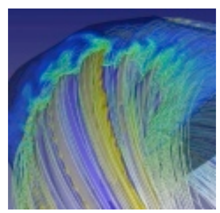
- Next generation Xeon-Phi, >3TF peak
- Single socket processor - Self-hosted, not a co-processor, not an accelerator
- Greater than 60 cores per processor with support for four hardware threads each; more cores than current generation Intel Xeon Phi™
- Intel® "Silvermont" architecture enhanced for high performance computing
- 512b vector units (32 flops/clock – AVX 512)
- 3X single-thread performance over current generation Xeon-Phi co-processor
- High bandwidth on-package memory, up to 16GB capacity with bandwidth projected to be 5X that of DDR4 DRAM memory
- Higher performance per watt

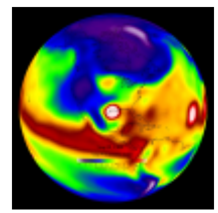
Cori will be installed in the Computational Research and Theory (CRT) Facility

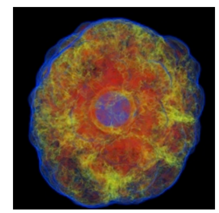
- **Four story, 140,000 GSF**
  - 300 offices on two floors
  - 20K -> 29Ksf HPC floor
  - 12.5MW -> 40 MW to building
- **Located for collaboration**
  - CRD and ESnet
  - UC Berkeley
- **Exceptional energy efficiency**
  - Natural air and water cooling
  - Heat recovery
  - PUE < 1.1
  - LEED gold design
- **Initial occupancy early 2015**

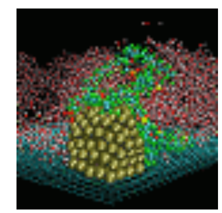
Application Readiness -- Challenges and Strategy





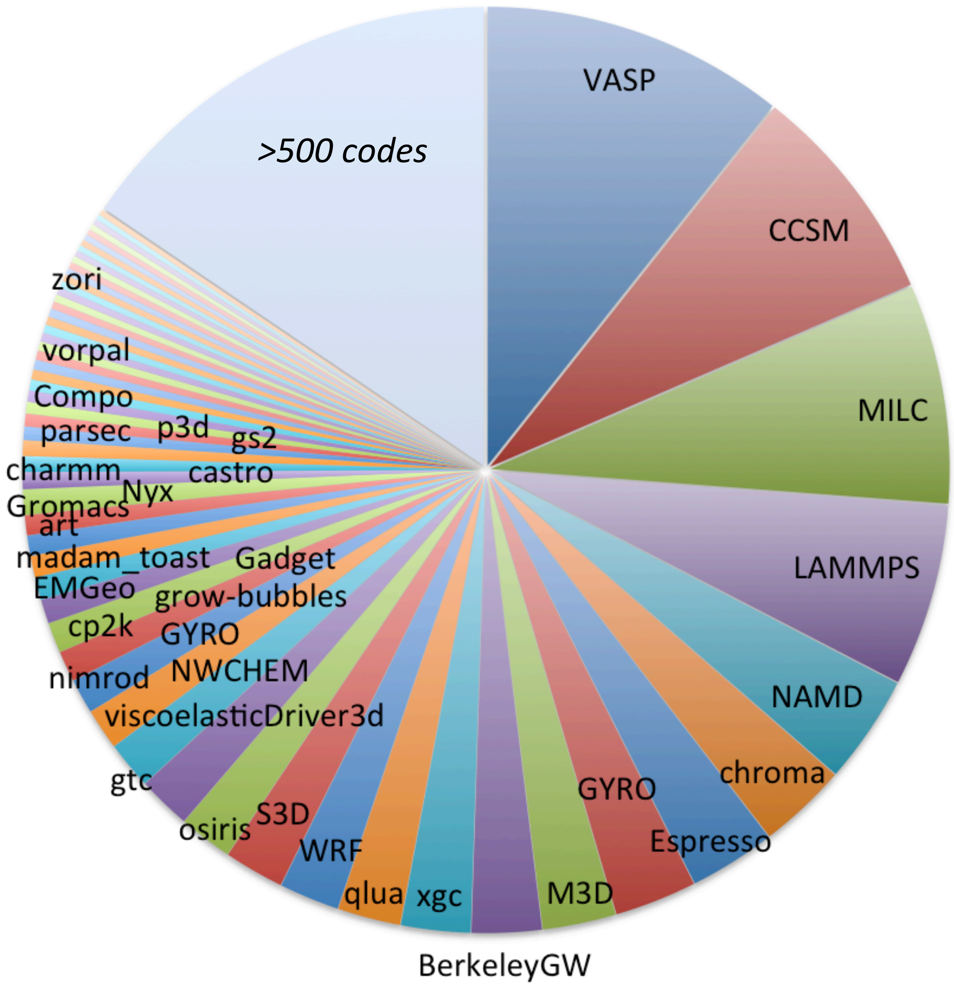






We will initially focus on 20 codes

Breakdown of Application Hours on Hopper and Edison 2013



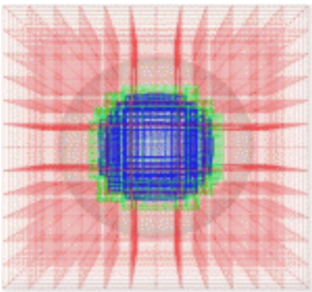
- 10 codes make up 50% of the workload
- 25 codes make up 66% of the workload
- Edison will be available until 2019/2020
- Training and lessons learned will be made available to all application teams

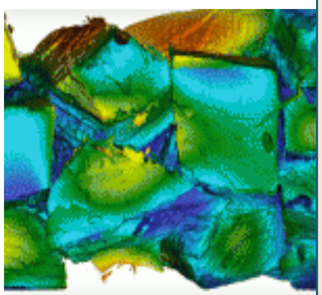
We are launching the NERSC Exascale Science Applications Program (NESAP)

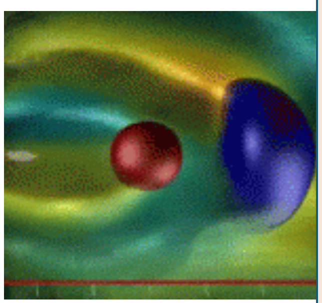
- **NESAP components:**

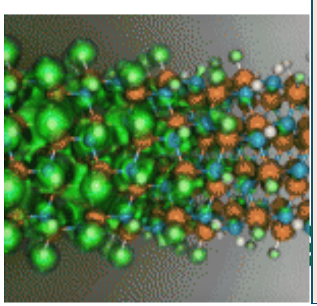
- Strong support from vendors
- Developer Workshops for 3rd-Party SW
- Early engagement with code teams
- Leverage existing community efforts
- Postdoc Program
- NERSC training and online modules
- Early access to N8 technology

20 NESAP Tier-1 and Tier-2 codes









Advanced Scientific Computing Research

Almgren (LBNL) – **BoxLib AMR Framework**

Trebotich (LBNL) – **Chombo-crunch**

High Energy Physics

Vay (LBNL) – **WARP & IMPACT**

Toussaint(Arizona) – **MILC**

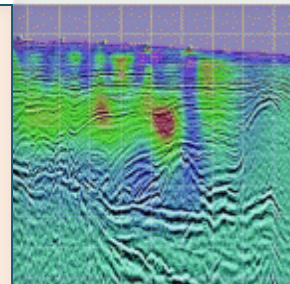
Habib (ANL) – **HACC**

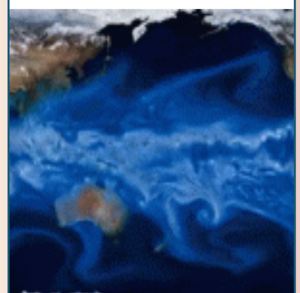
Nuclear Physics

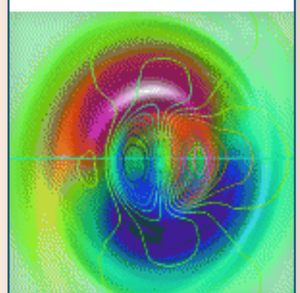
Maris (Iowa St.) – **MFDn**

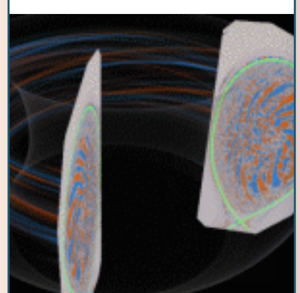
Joo (JLAB) – **Chroma**

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**

Comparison of Selected Apps with 2013 Usage

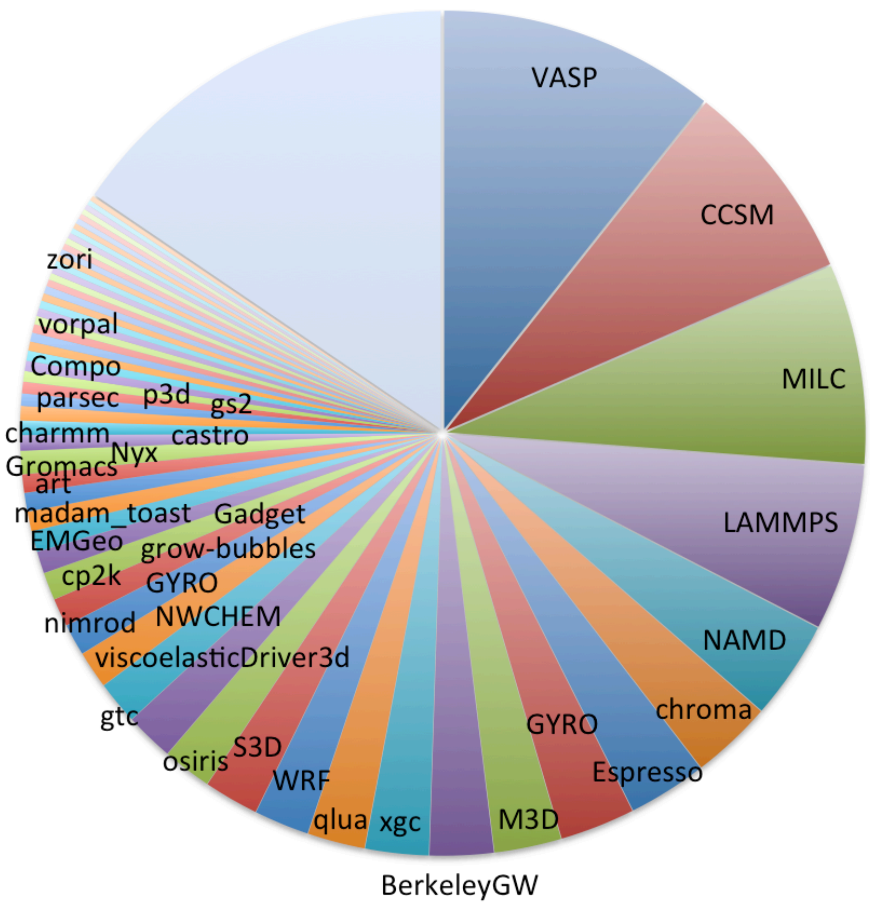
Tier-1, 2 Code

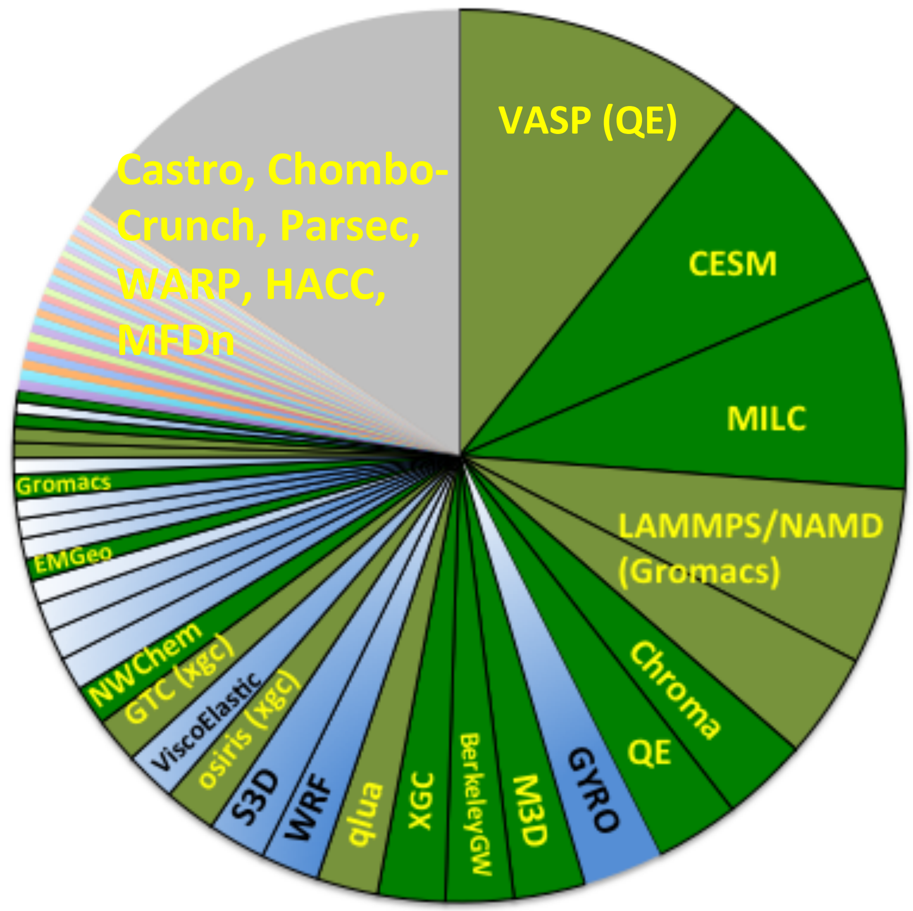
Tier-3

Tier-1, 2 Proxy Code

Other Codes

Breakdown of Application Hours on Hopper and Edison 2013





The Selected Codes are Diverse in Several Dimensions

Almost all others

production codes *vs.* up and coming

MPAS-O
ACME
PARSEC

HACC, Chroma, MILC, BerkeleyGW, XGC

Higher *vs.* lower readiness

M3D, CESM, EMgeo, WARP & Synergia, Crunch

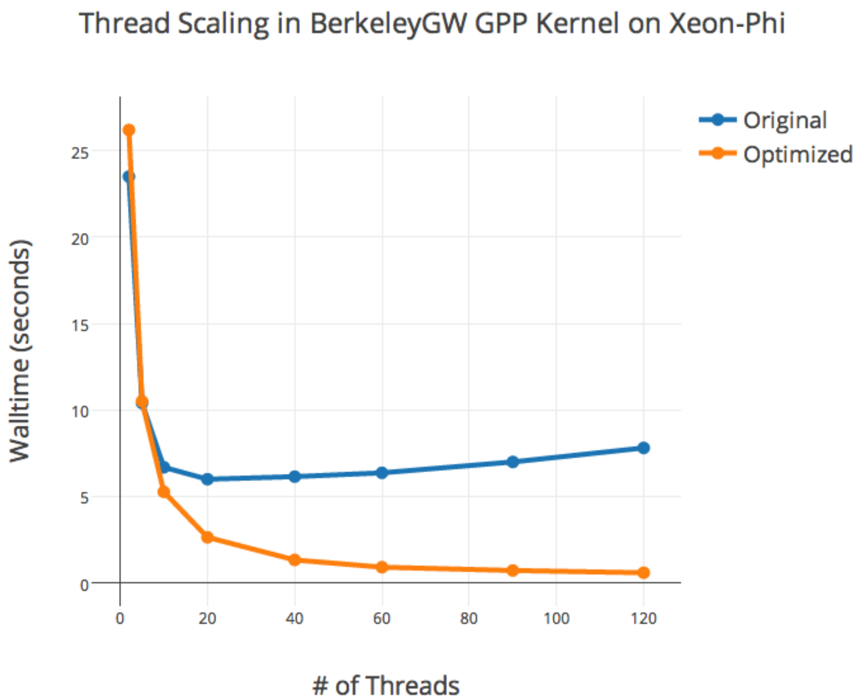
EMGeo, WARP, Meraculous, Chombo-crunch, BoxLib

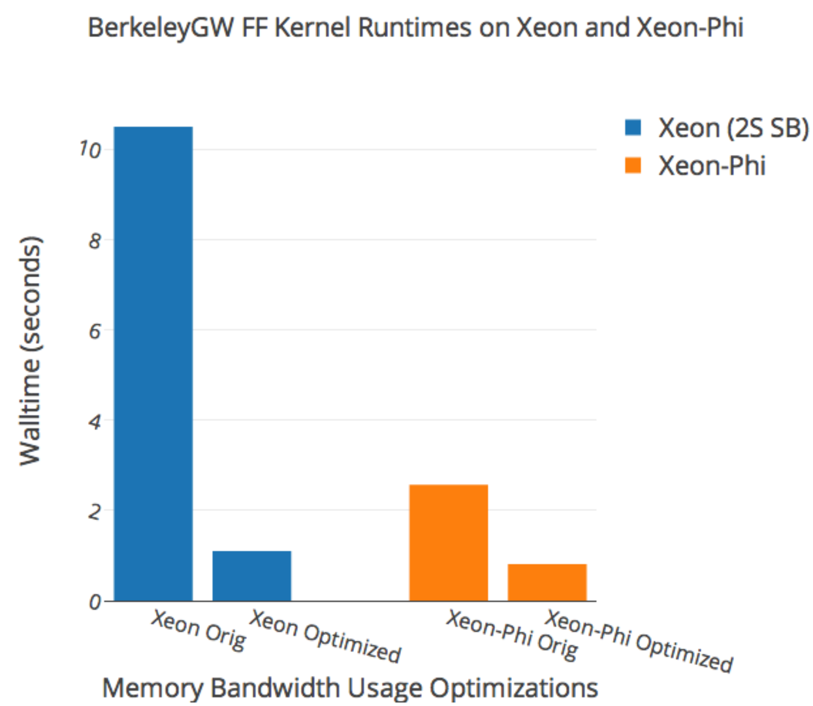
Smaller community apps vs Overlap with established ALCF/ OLCF readiness codes

MILC, CESM, Chroma, HISQ/DWF, Gromacs, QE, NWChem

NESAP has already received recognition

NESAP Intel and Cray collaboration improved BerkeleyGW performance

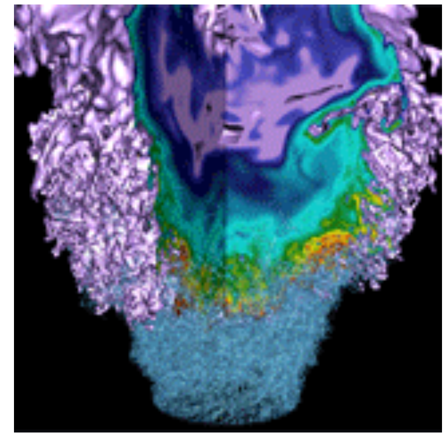


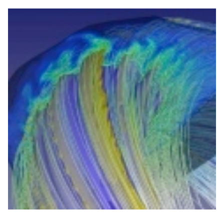


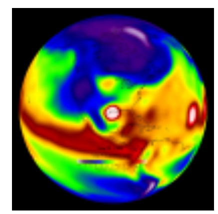
NESAP (Cray COE and Intel Dungeon Session) advances for BerkeleyGW kernels include:

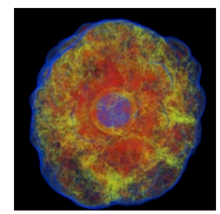
1. Thread scaling improvements beyond 10 threads
2. Addition of cache-blocking in bandwidth-bound kernels
3. Improved vectorization in kernels (including small matmuls)

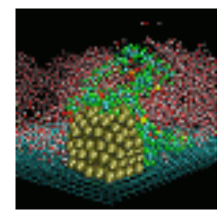
Extreme Data Science



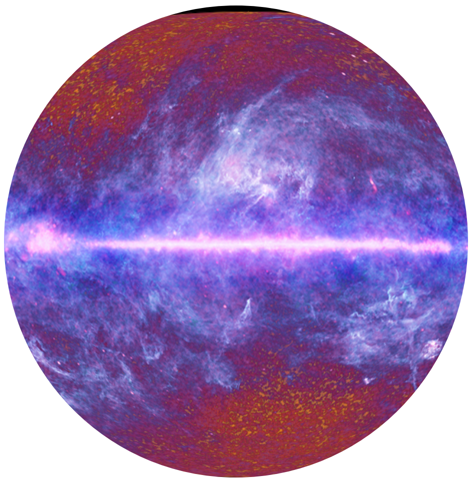




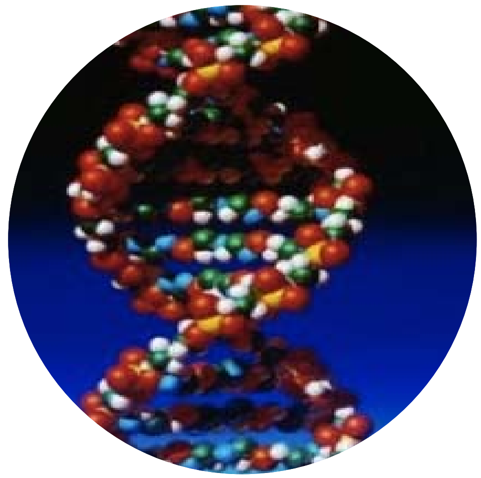




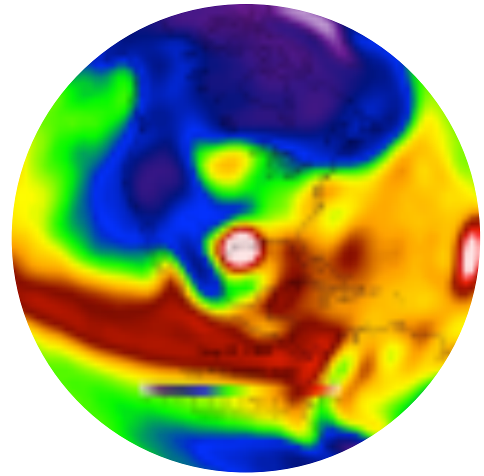
DOE Facilities are Facing a Data Deluge



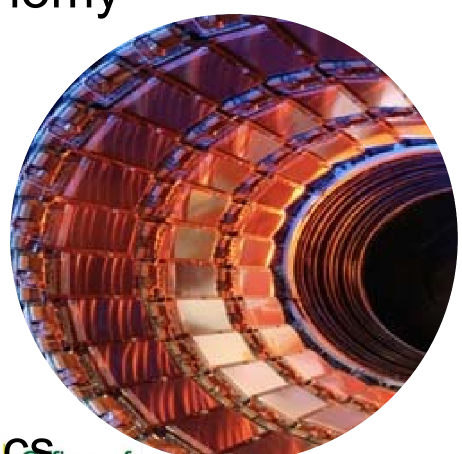
Astronomy



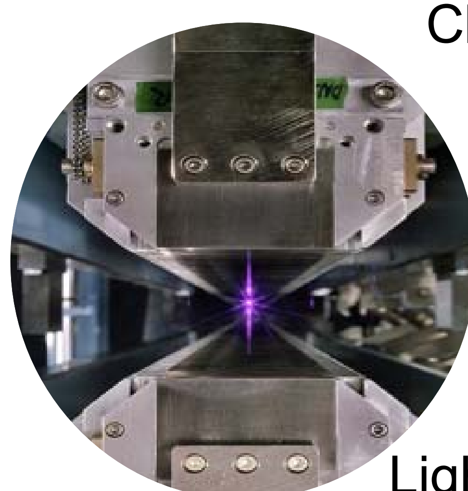
Genomics



Climate

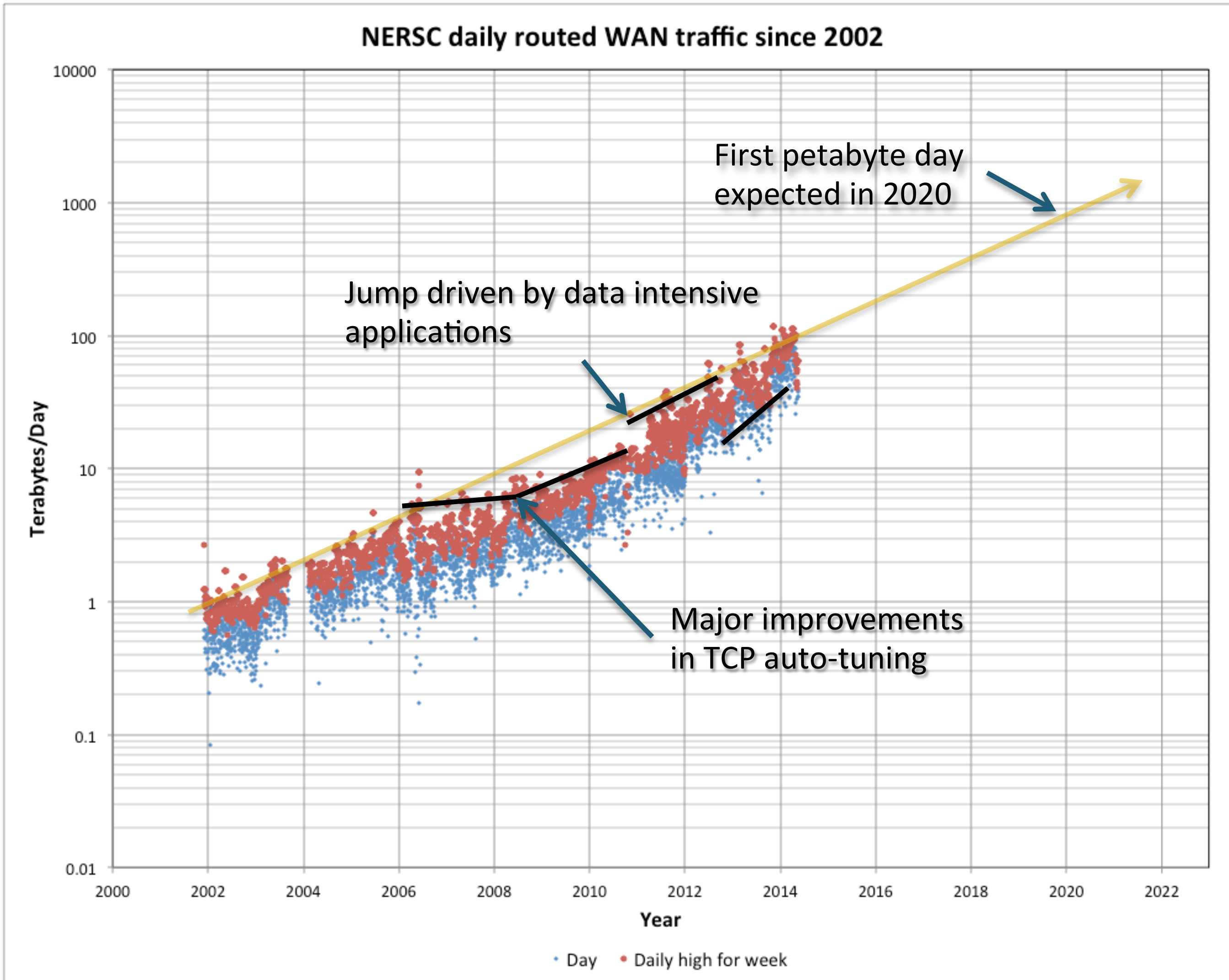


Physics

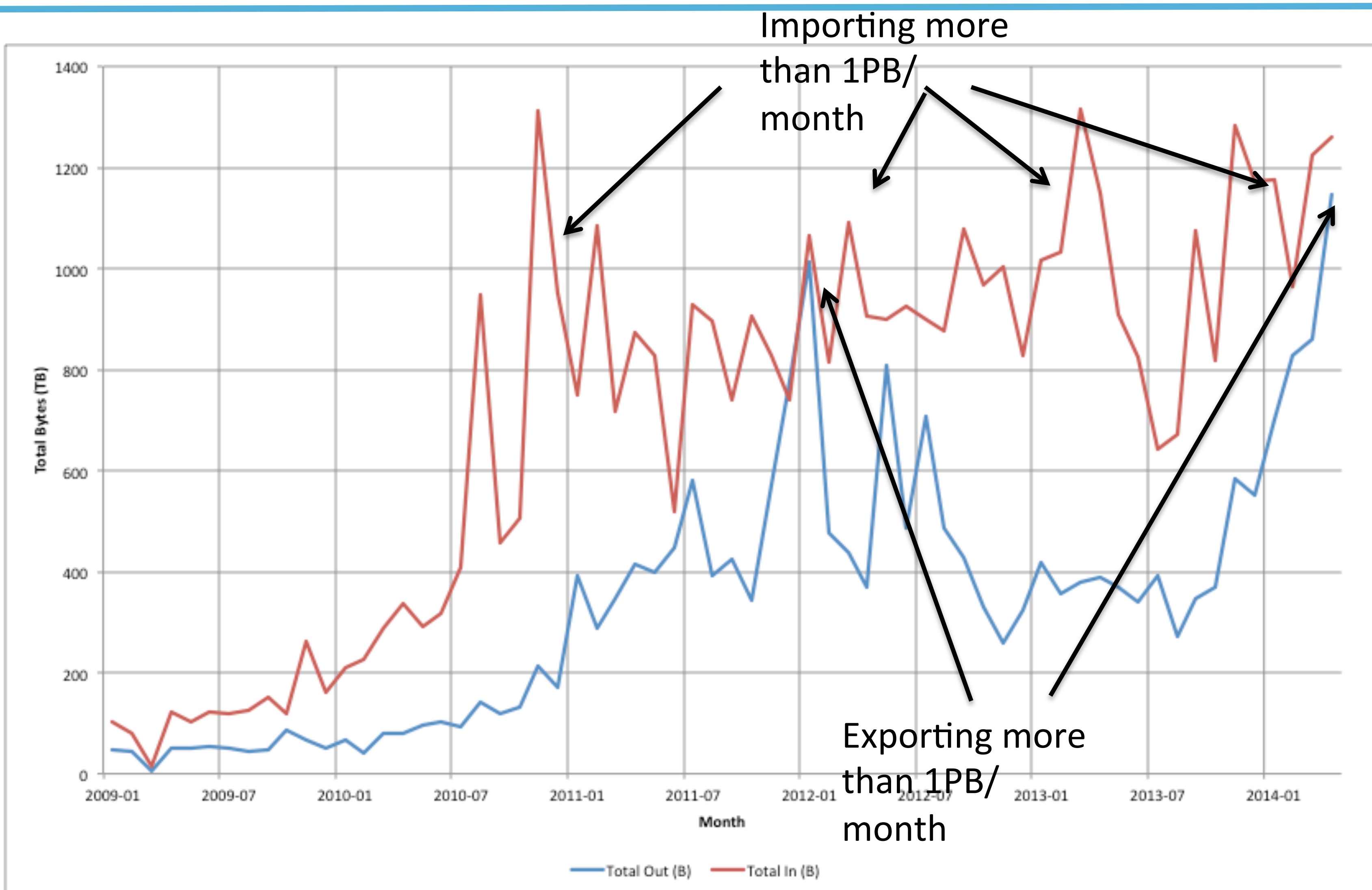


Light Sources

Exponentially increasing data traffic

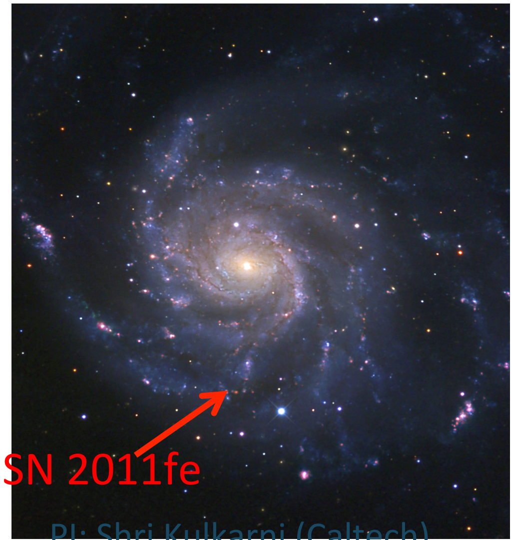


NERSC users import more data than they export!



Extreme Data Science is Playing a Key Role in Scientific Discovery

- Measurement of the important "θ_{13}" neutrino parameter. One of Science Magazine's Top-Ten Breakthroughs of 2012.
  - Last and most elusive piece of a longstanding puzzle: why neutrinos appear to vanish as they travel
- The Palomar Transient Factory Discovered over 2000 supernovae in the last 5 years, including the youngest and closest Type Ia supernova in past 40 years
- Trillions of measurements by the Planck satellite led to the most detailed maps ever of cosmic microwave background (One of Physics Today's Top 10 breakthroughs of 2013)
- Materials project has over 5000 users and was featured on the cover of Scientific American



PI: Shri Kulkarni (Caltech)

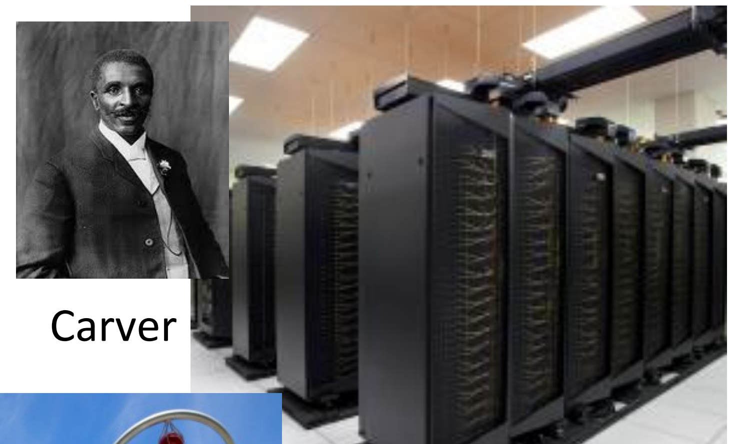
We currently deploy separate Compute Intensive and Data Intensive Systems

NERSC 40 YEARS at the FOREFRONT 1974-2014

Compute Intensive	*Data Intensive*

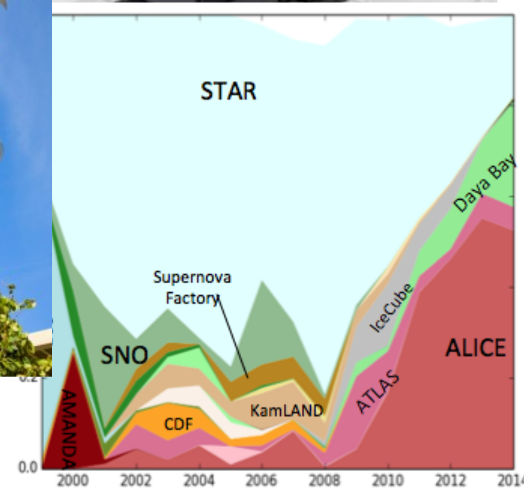
Compute Intensive

Data Intensive



Carver

Genepool



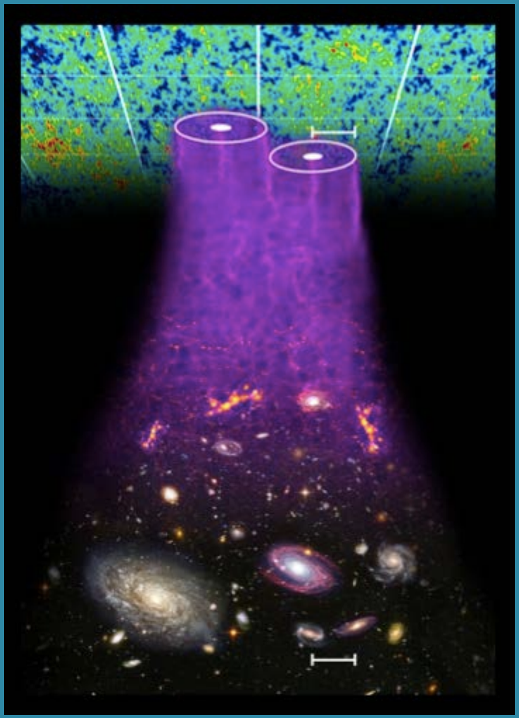
PDSF

The Need for Data Intensive Systems

- Communicate with databases / host databases
- Complex workflows (including High Throughput Computing - HTC)
- Policy flexibility
- Local disk
- Very large memory
- Massive serial jobs (~100K)
- Easy to customize environment and the environment is familiar

- Dramatically growing data sets require Petascale+ computing for analysis
- In addition, we increasingly need to couple large-scale simulations and data analysis

Baryon Acoustic Oscillations (BAO):

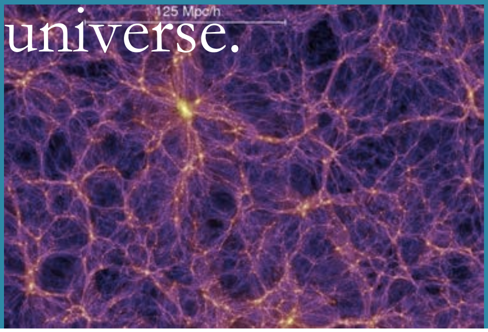


Large quantities of data need to be analyzed.

Imaging survey in 2005: 20 TB
in 2025 60 PB

Statistical analyses need MCMC for cross-correlation of the millions of galaxies
-- collapsing the problem to just 2-point statistics.

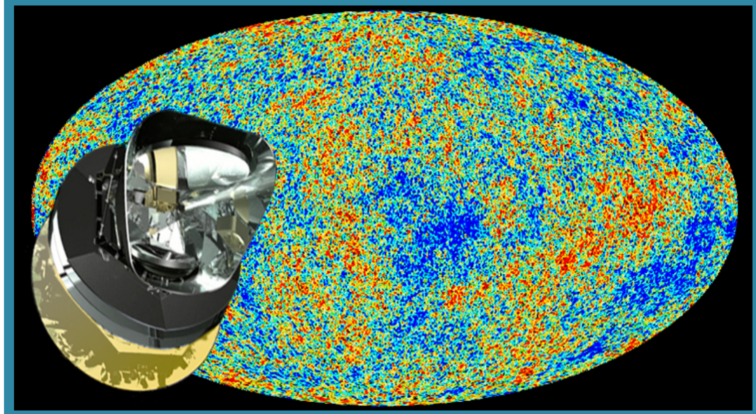
All data analysis dependent on comparisons to supercomputer-based N-body simulations of the evolution of matter in the universe.



Current state of art: 2048^3 – 4096^3 “particles.”
Need an order of magnitude more.

Cosmic Microwave Background (CMB):

Exponentially growing data chasing fainter echos:



- BOOMERanG: 10^9 samples in 2000
- Planck: 10^{12} samples in 2013 (0.5 PB)
- CMBpol: 10^{15} samples in 2025

Uncertainty quantification through Monte Carlos

- Simulate 10^4 realizations of the entire mission
- Control both systematics and statistics

Mission-class science relies on HPC evolution.

Cori Data Enhancements

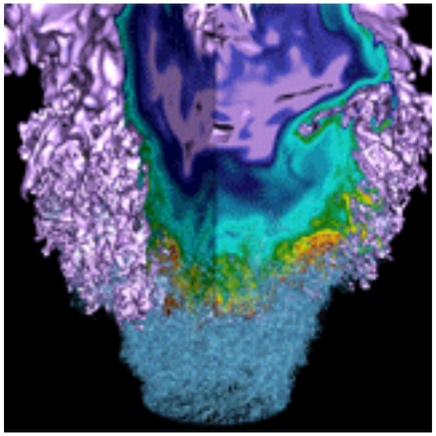
- **Data partition with large memory nodes and throughput optimized processors**
- **Burst buffer -- NVRAM nodes on the interconnect fabric for IO caching**
- **Larger disk system**

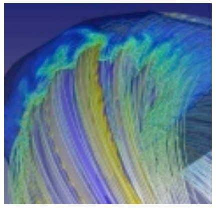
Goals are to enable the analysis of large experimental data sets and in-situ analysis coupled to Petascale simulations

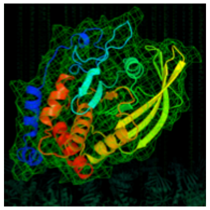
Popular features of a data intensive system can be supported on Cori

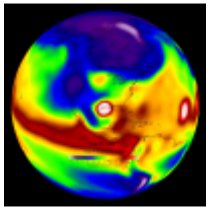
Data Intensive Workload Need	Cori Solution
Local Disk	NVRAM 'burst buffer'
Large memory nodes	128 GB/node on Haswell; Option to purchase fat (1TB) login node
Massive serial jobs	NERSC serial queue prototype on Edison; MAMU
Complex workflows	More (14) external login nodes; CCM mode for now
Communicate with databases from compute nodes	***Compute Gateway Node***
Stream Data from observational facilities	***Compute Gateway Node***
Easy to customize environment	***User Defined Images***
Policy Flexibility	Improvements coming with Cori: Rolling upgrades, CCM, MAMU, above COEs would also contribute

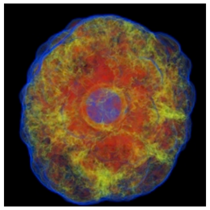
Conclusions

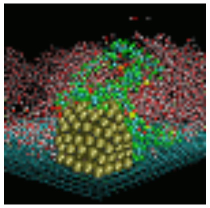










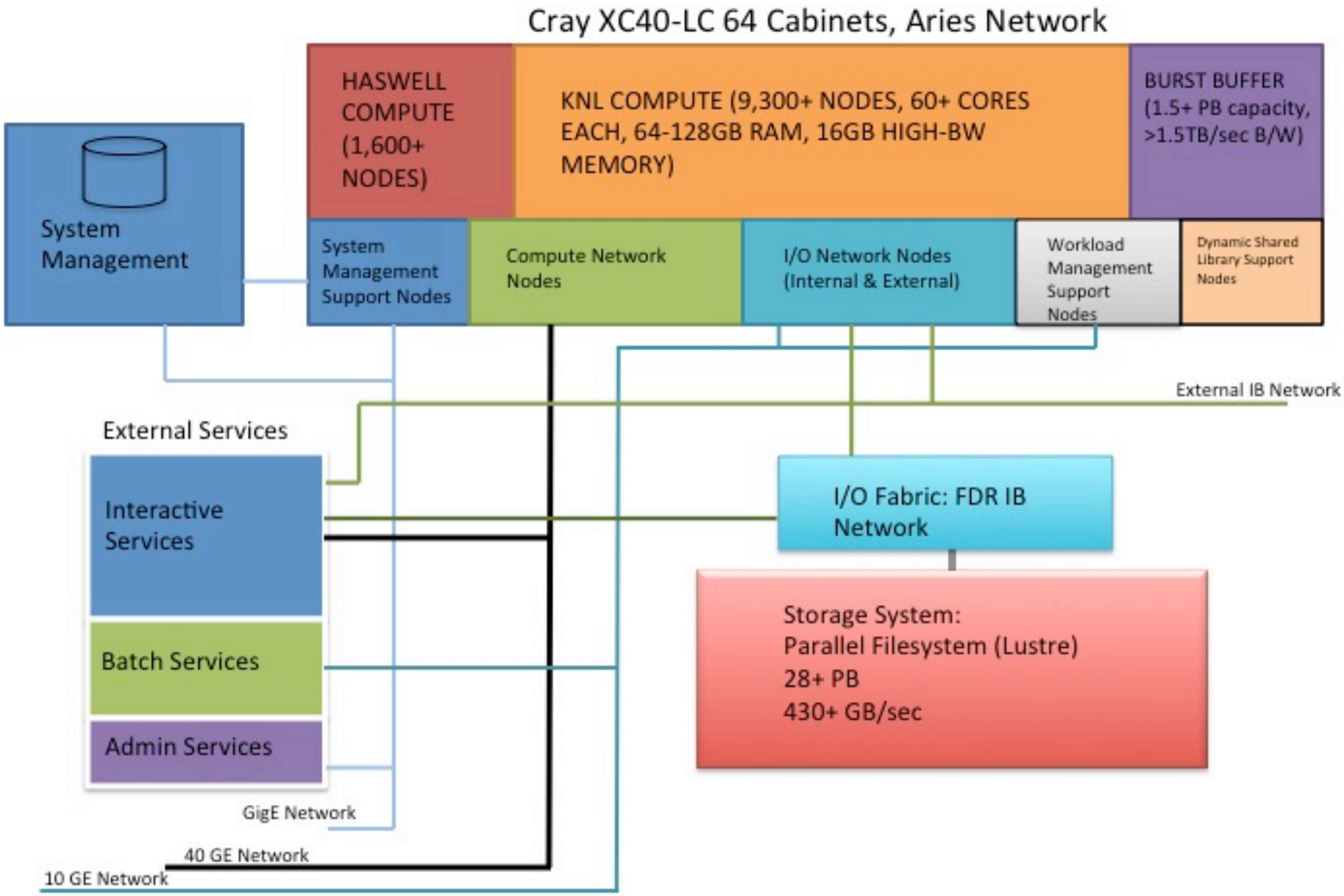


NERSC 40 YEARS at the FOREFRONT 1974-2014

U.S. DEPARTMENT OF ENERGY | Office of Science

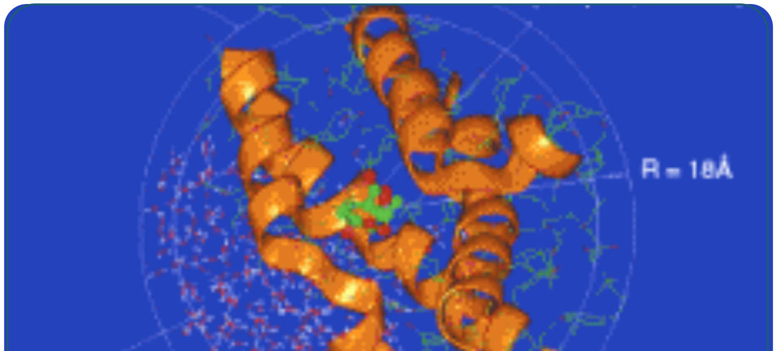
BERKELEY LAB

The Cori System



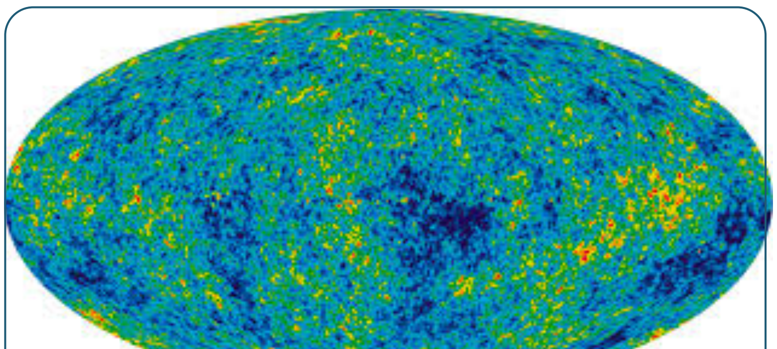
Our goal is to enable science that can’t be done on today’s supercomputers

NERSC 40 YEARS at the FOREFRONT 1974-2014



Martin Karplus

Saul Perlmutter



George Smoot

Warren Washington