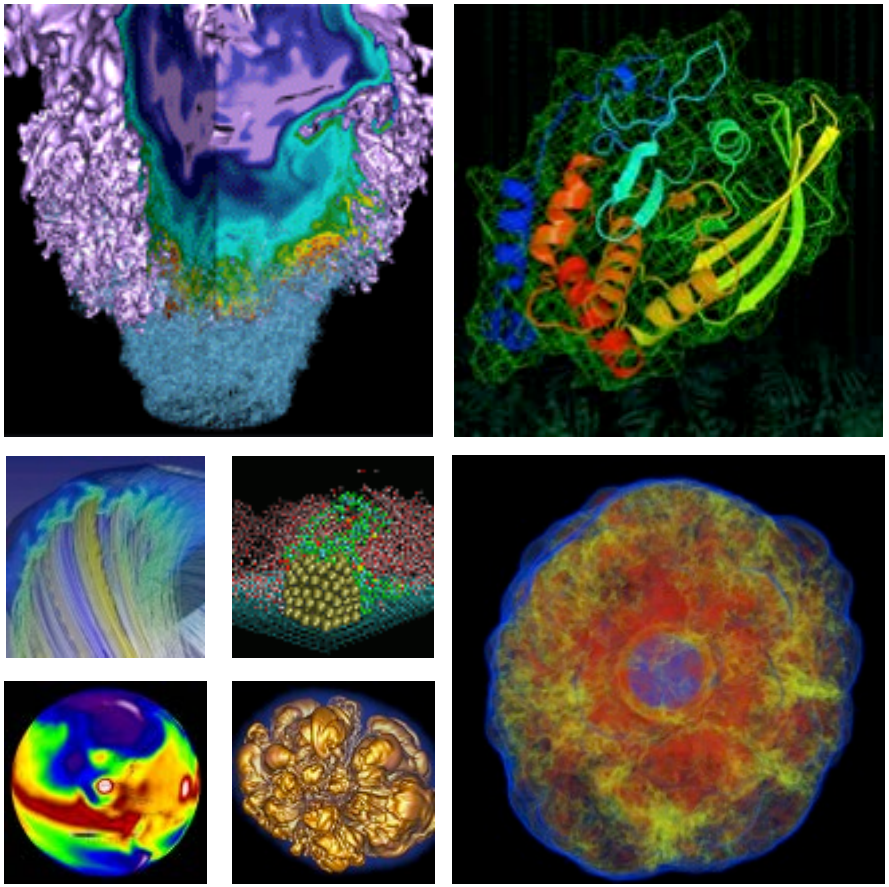
Cori (NERSC-8)



Sudip Dosanjh
Director

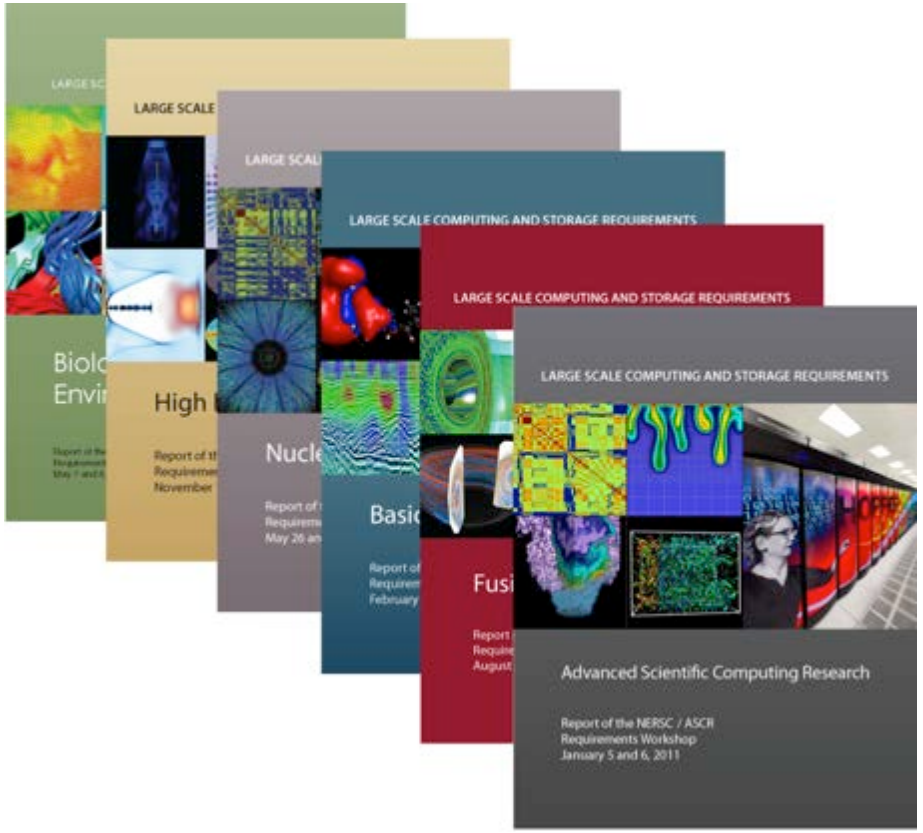
August 2, 2014

NERSC currently deploys two Petascale systems

System attributes	Hopper	Edison
	Hopper	**Edison**
System peak	1.3 PF	2.6PF
System memory	0.21 PB	0.35 PB
Node performance	202GF	460 GF
Node memory BW	50 GB/s	100 GB/s
Node concurrency	24 AMD Magnycours cores	24 Intel Ivy Bridge Cores
System size (nodes)	6,384 nodes	5,576 nodes
MPI Node Interconnect BW	~3 GB/s	~9GB/s

Requirements with six program offices

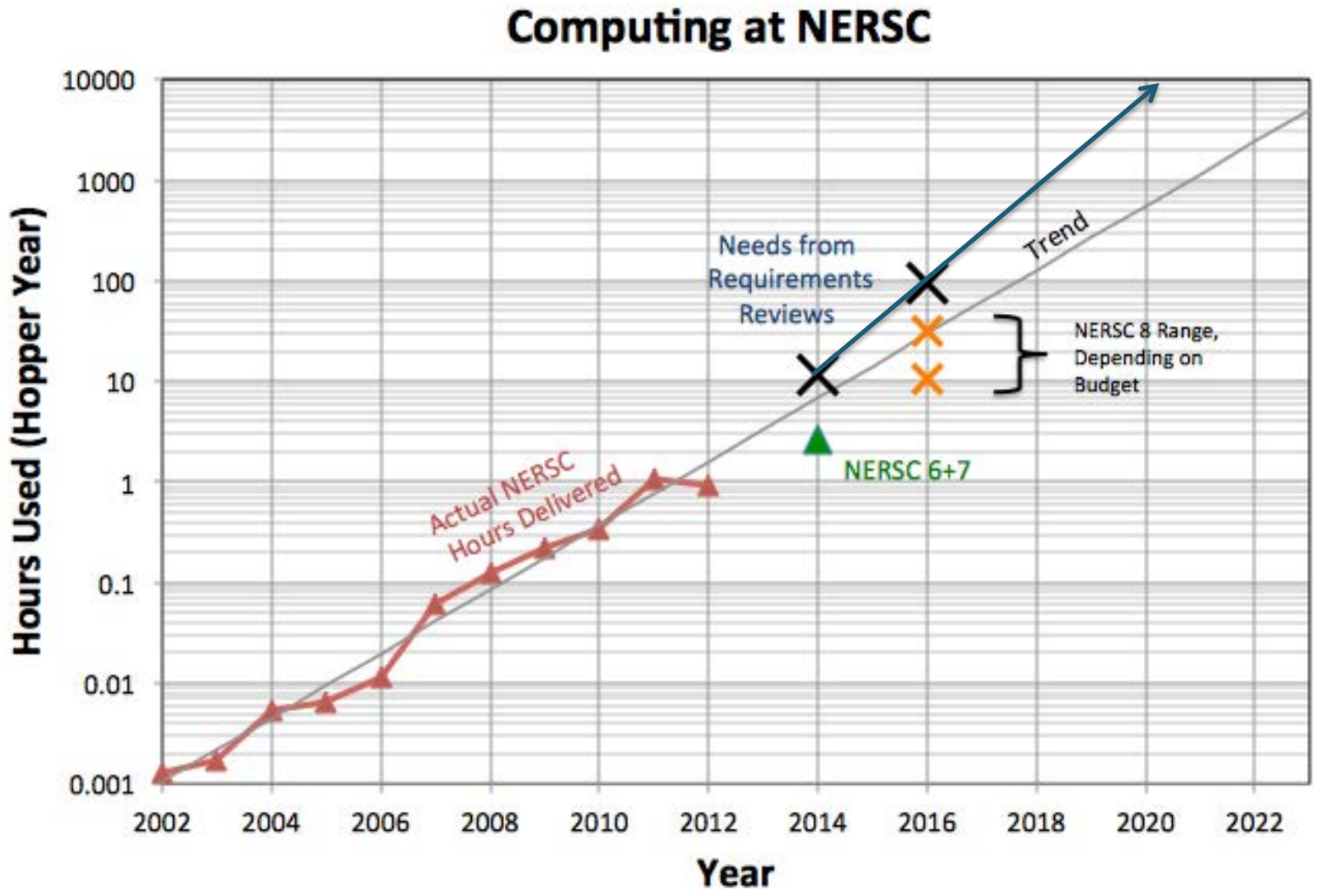
- Reviews with six program offices every three years
- Program managers invite representative set of users (typically represent >50% of usage)
- Identify science goals and representative use cases
- Based on use cases, work with users to estimate requirements
- Re-scale estimates to account for users not at the meeting (based on current usage)
- Aggregate results across the six offices
- Validate against information from in-depth collaborations, NERSC User Group meetings, user surveys



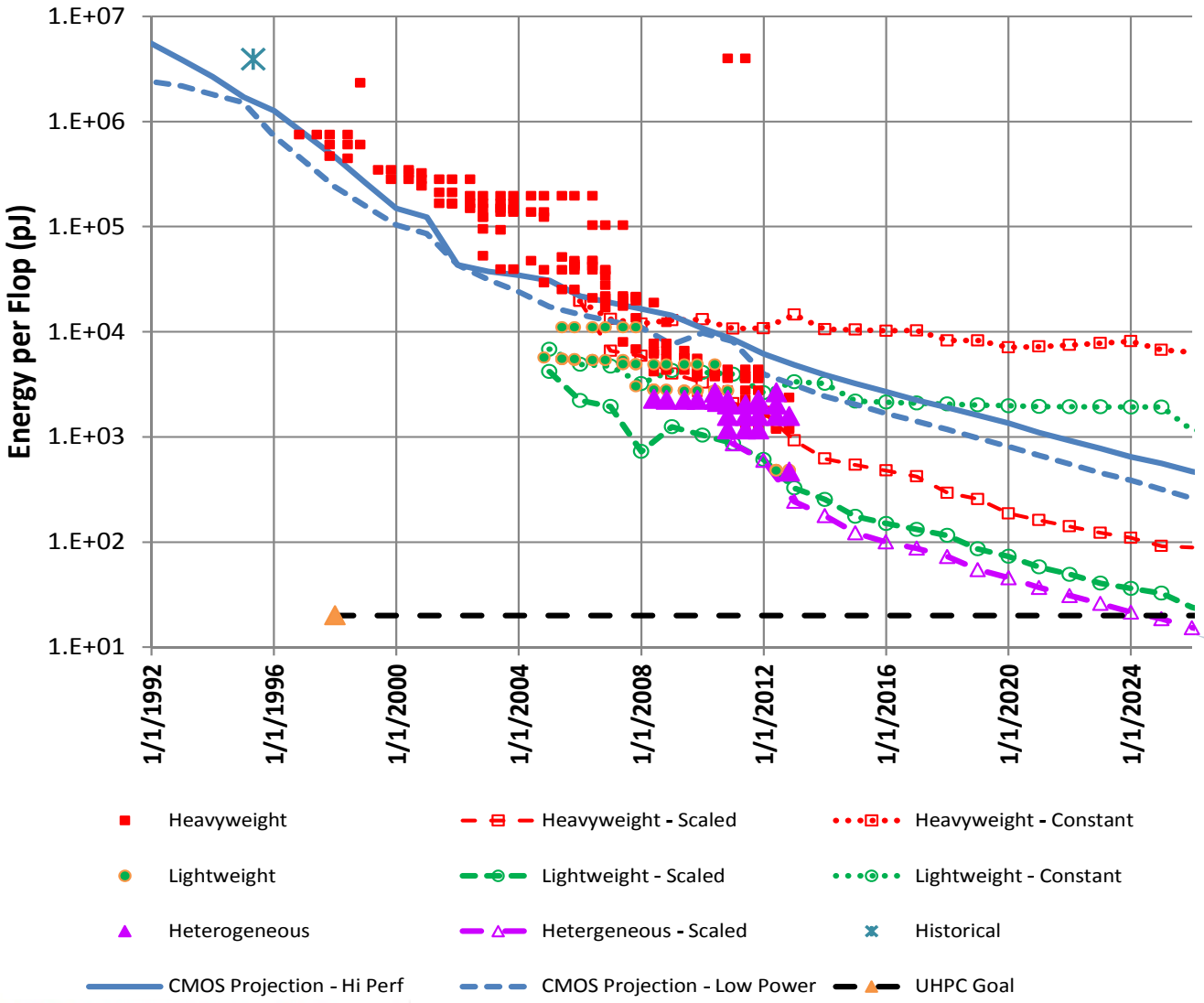
Tends to underestimate need because we are missing future users

http://www.nersc.gov/science/requirements-reviews/final-reports/

Keeping up with user needs will be a challenge



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

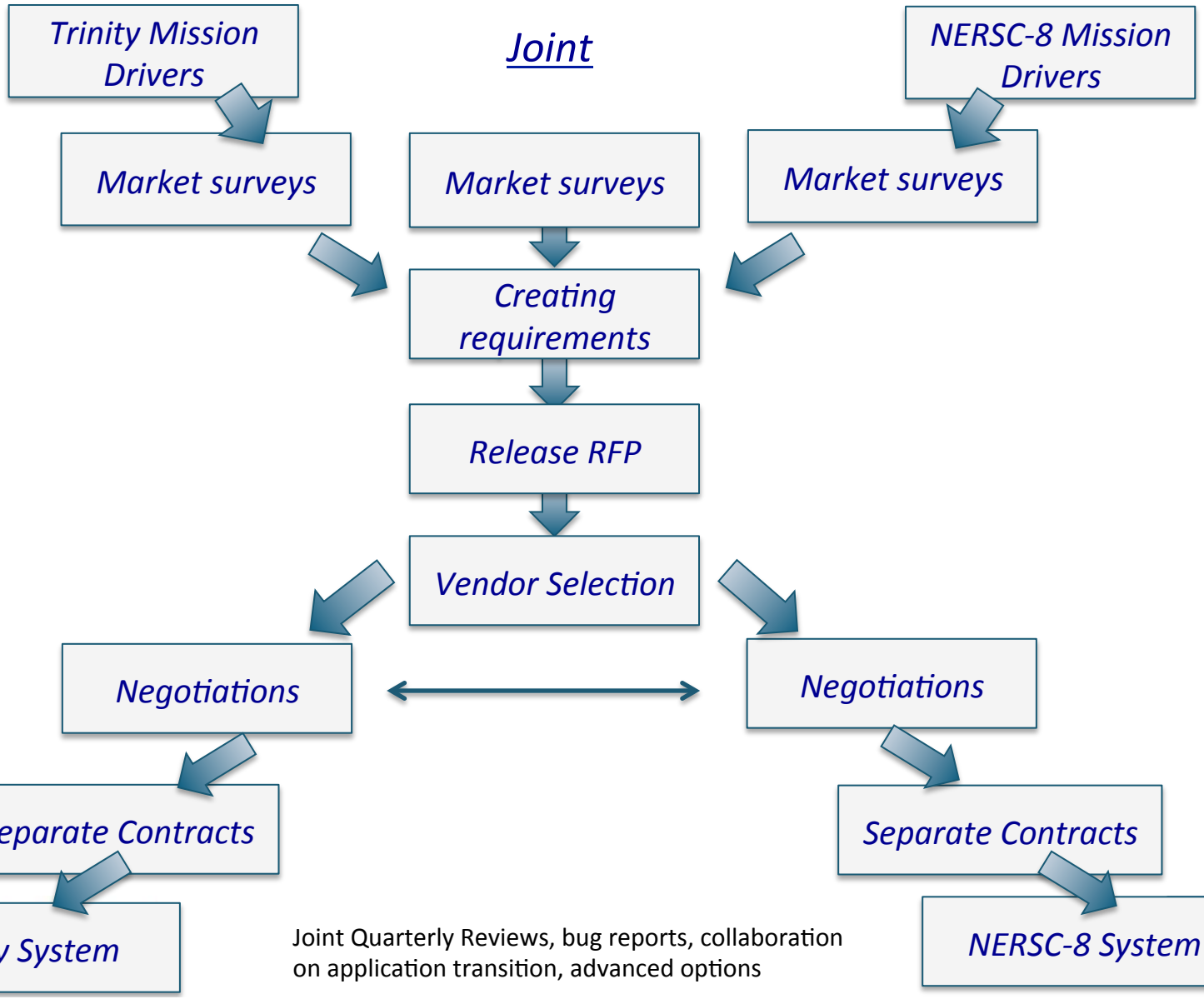
The Department of Energy Office of Science requires an HPC system to support the rapidly increasing computational demands of the entire spectrum of DOE SC computational research.

- 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

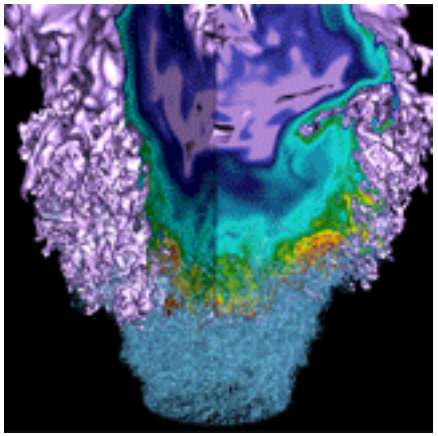
- Visible collaboration between ASCR and ASC
- Strengthen impact on industry
- Address challenges transitioning applications to advanced manycore architectures with a broader coalition
- Act as a risk mitigation strategy for NERSC-8 and Trinity systems by having a partner to work with on technical challenges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

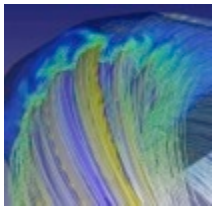
This was a collaboration of two separate projects

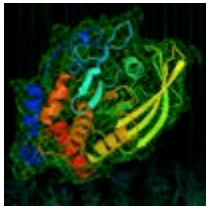


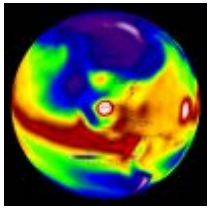
Joint Quarterly Reviews, bug reports, collaboration on application transition, advanced options

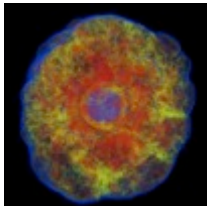
The Cori system

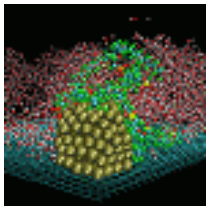












NERSC 40 YEARS at the FOREFRONT 1974-2014

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

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

Programming Model Considerations

- **Knight's Landing is a self-hosted part**
  - Users can focus on adding parallelism to their applications without concerning themselves with PCI-bus transfers
- **MPI + OpenMP preferred programming model**
  - Should enable NERSC users to make robust code changes
- **MPI-only will work – performance may not be optimal**
- **On package MCDRAM**
  - How to optimally use ?
    - Explicitly or implicitly ??

Knights Landing Integrated On-Package Memory

Cache Model

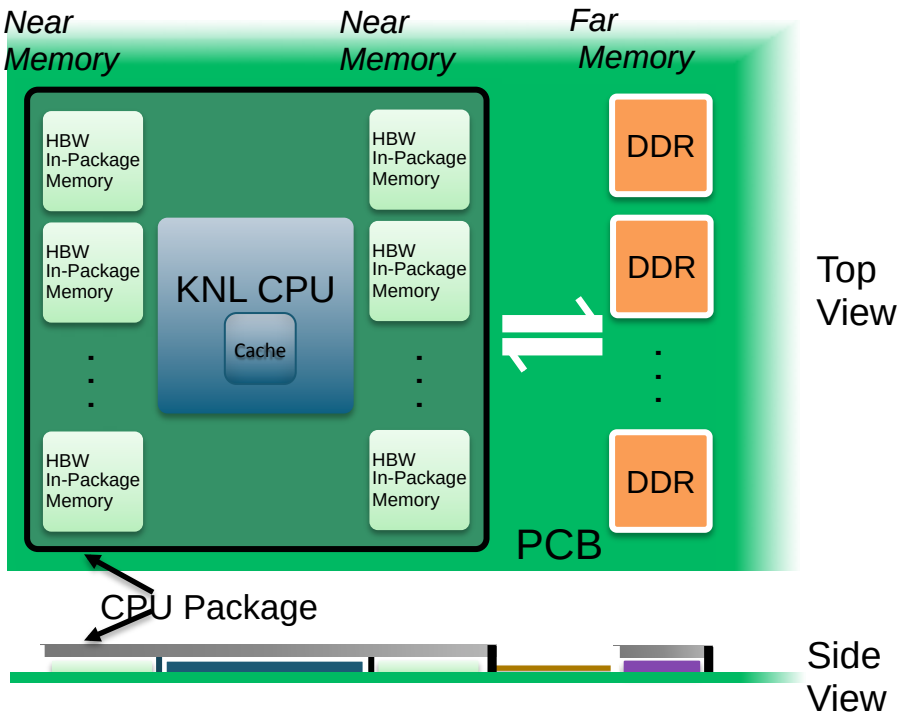
Let the hardware automatically manage the integrated on-package memory as an "L3" cache between KNL CPU and external DDR

Flat Model

Manually manage how your application uses the integrated on-package memory and external DDR for peak performance

Hybrid Model

Harness the benefits of both cache and flat models by segmenting the integrated on-package memory



Maximum performance through higher memory bandwidth and flexibility

Cori's Programming Environment

- **Key point: Cori will look basically just like NERSC-5/ N6/N7 to users**
- **Cori is not a heterogeneous or accelerator system; programmed via Fortran/C/C++ plus MPI+OpenMP**
- **Intel / Cray/ GNU programming environments and commitment from Cray and Intel for optimized math and I/O libraries**
- **Multiple vendor profiling tools: CrayPAT and Vtune**
  - Interest from 3rd-party suppliers, too
- **DDT and Totalview support + Cray debugging tools**

Cori's Programming Environment

- **Advanced features: OpenMP4.0, OpenACC, Intel KNL High-Bandwidth Memory Placement Extensions, UPC, Fortran CoArrays, CoArray C++, Chapel**
- **Intel: Parallel Studio XE/Composer/Advisor/ Inspector**
- **Use of fast on-package memory: cache, directly addressable memory, or combination; configured as boot option with each job launch**

Running on Cori

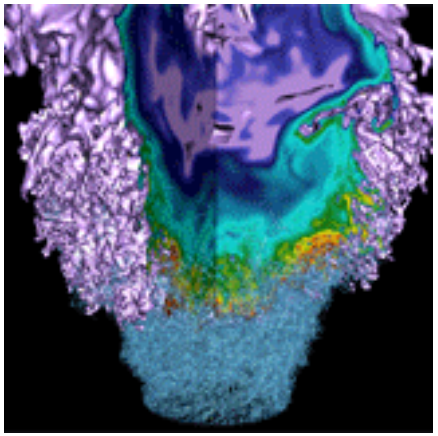
- Codes will probably run on NERSC-8 without any changes.
- To take advantage of the Knights Landing architecture, applications must
  - Exploit more parallelism
  - Express thread-level parallelism
  - Exploit data level parallelism
  - Manage data placement and movement
  - Accommodate less memory per process space

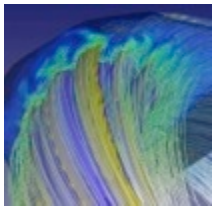
All of these changes will also help with portability

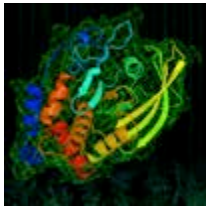
Cori will be installed in the Computational Research and Theory (CRT)

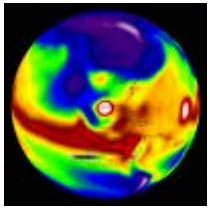
- **Four story, 140,000 GSF**
  - 300 offices on two floors
  - 20K -> 29Ksf HPC floor
  - 12.5MW -> 42 MW to building
- **Located for collaboration**
  - CRD and ESnet
  - UC Berkeley
- **Exceptional energy efficiency**
  - Natural air and water cooling
  - Heat recovery
  - PUE < 1.1
  - LEED gold design
- **Initial occupancy Fall 2014**

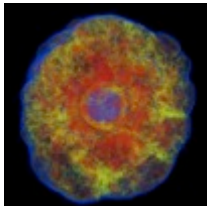
Application Readiness -- Challenges and Strategy

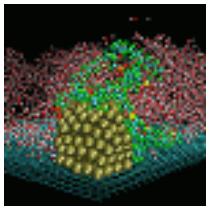












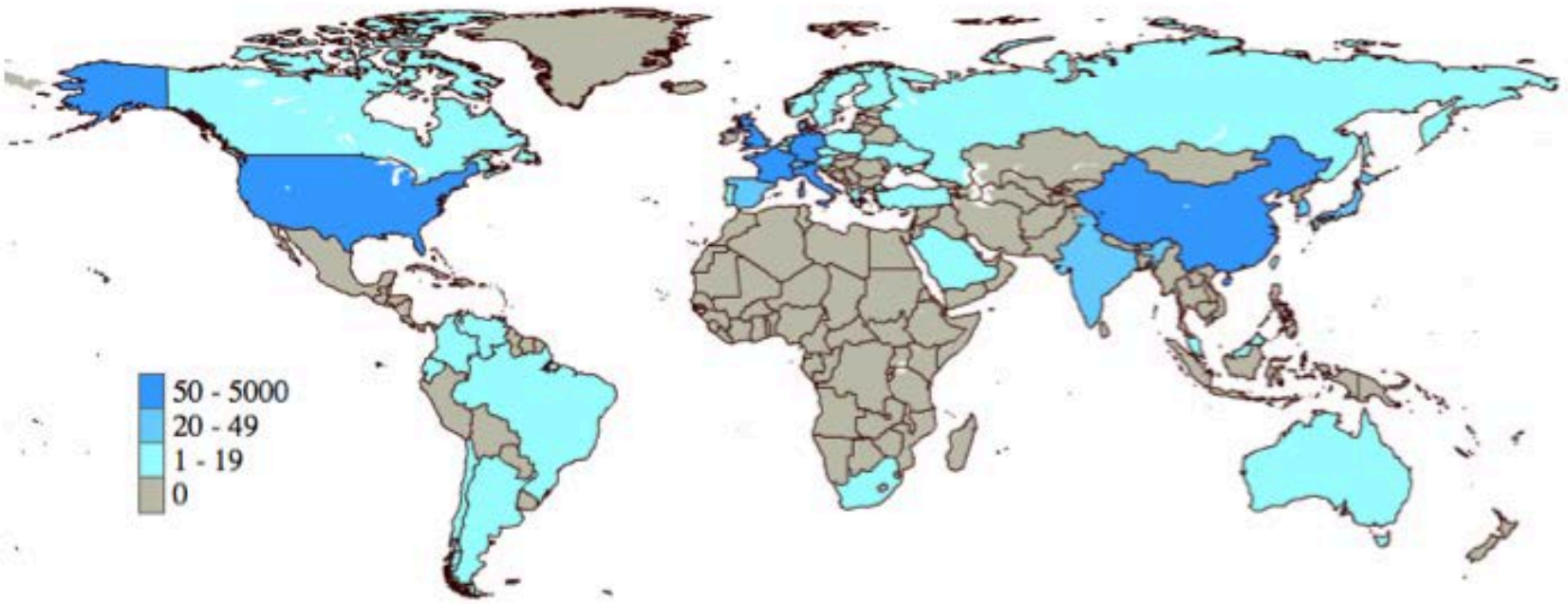
NERSC 40 YEARS at the FOREFRONT 1974-2014

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

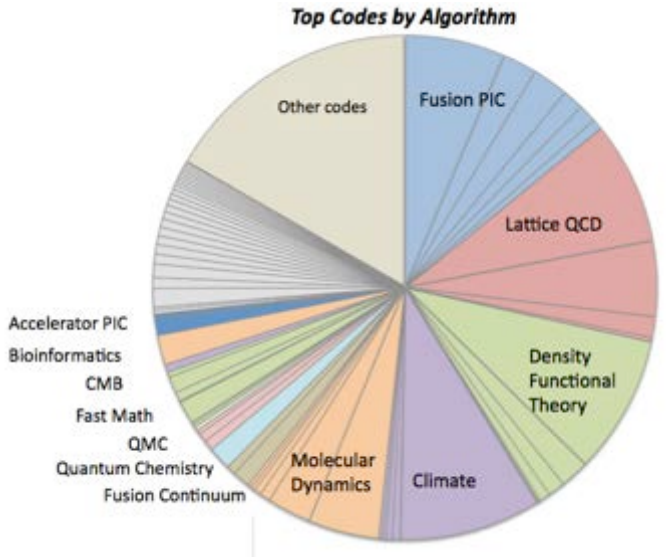
We support a broad user base

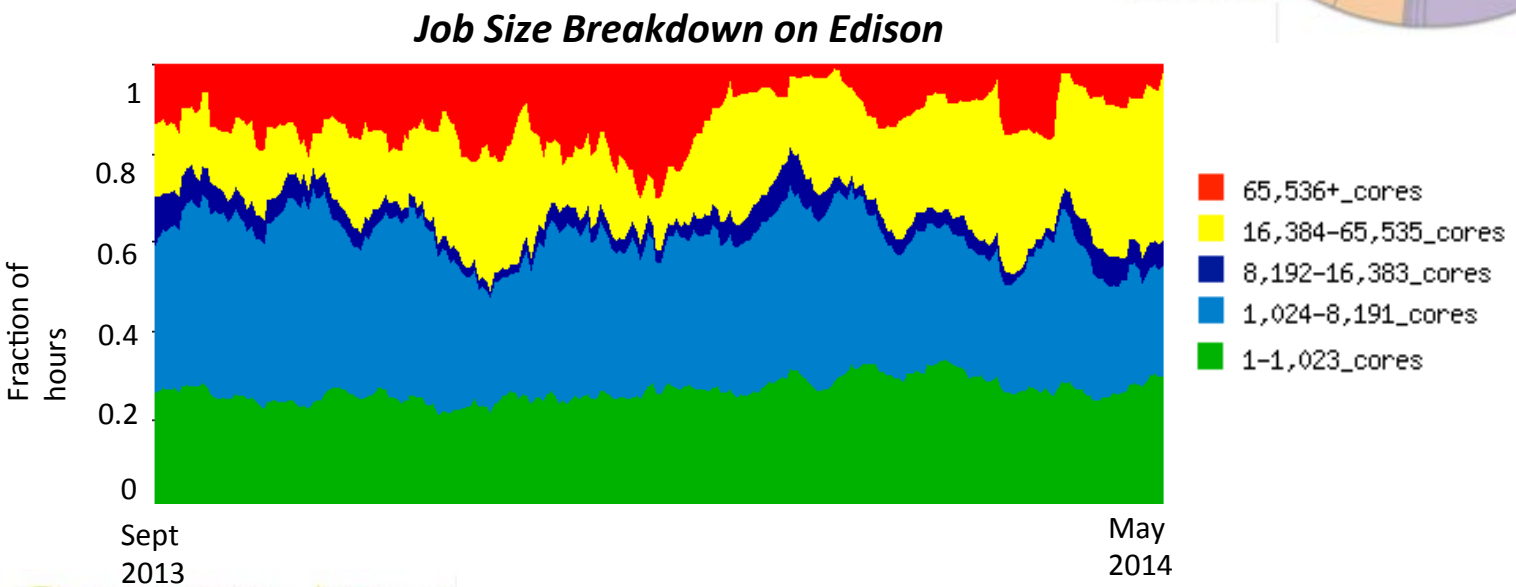
- **5000 users, and we typically add 350 per year**
- **Geographically distributed: 47 states as well as multinational projects**



We support a diverse workload

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

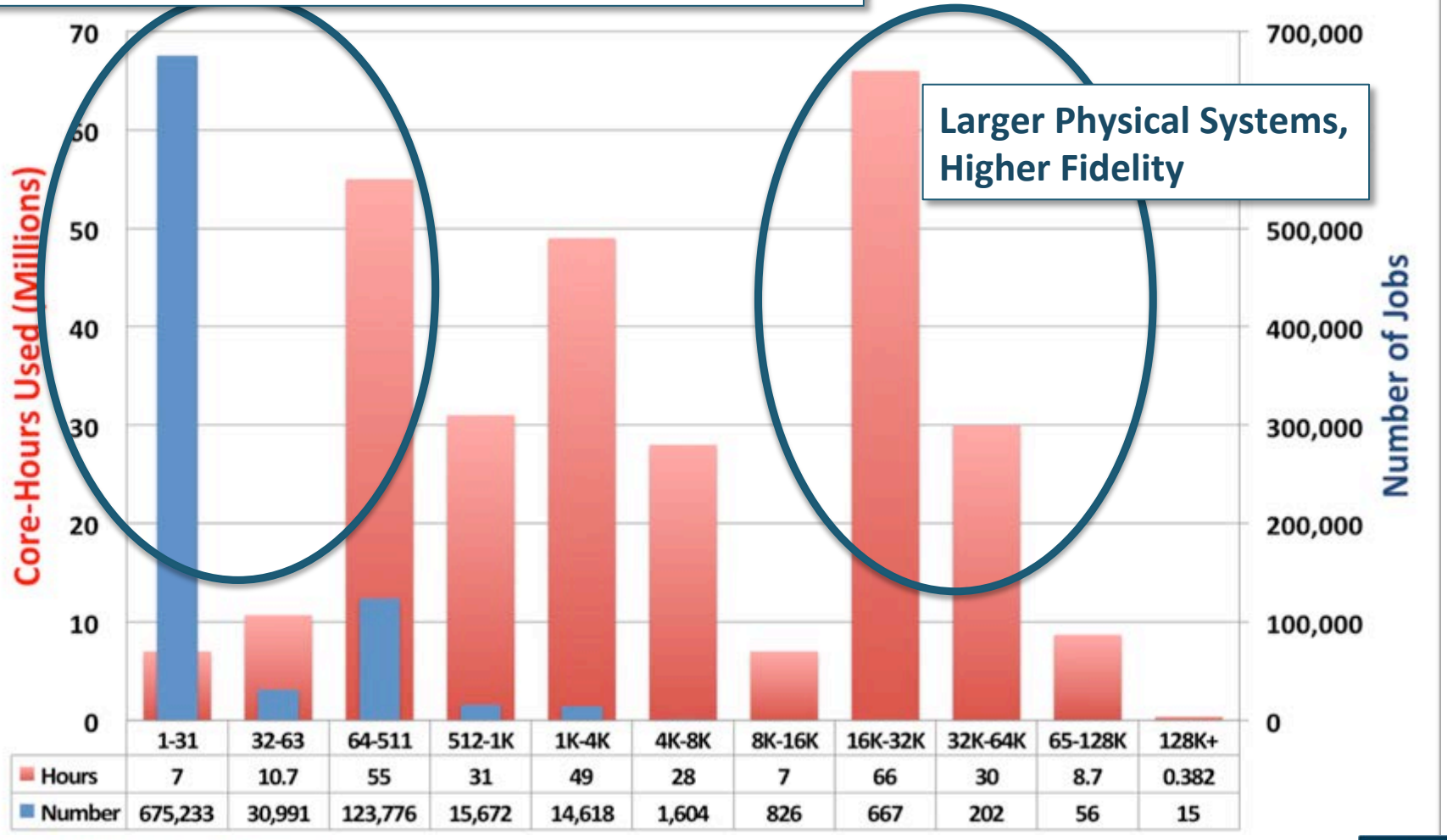




NERSC Supports Science Needs at Many Difference Scales and Sizes

High Throughput: Statistics, Systematics, Analysis, UQ

Larger Physical Systems, Higher Fidelity

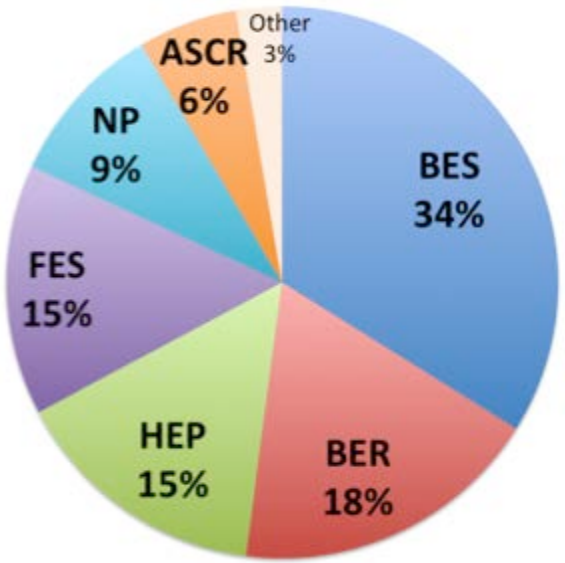


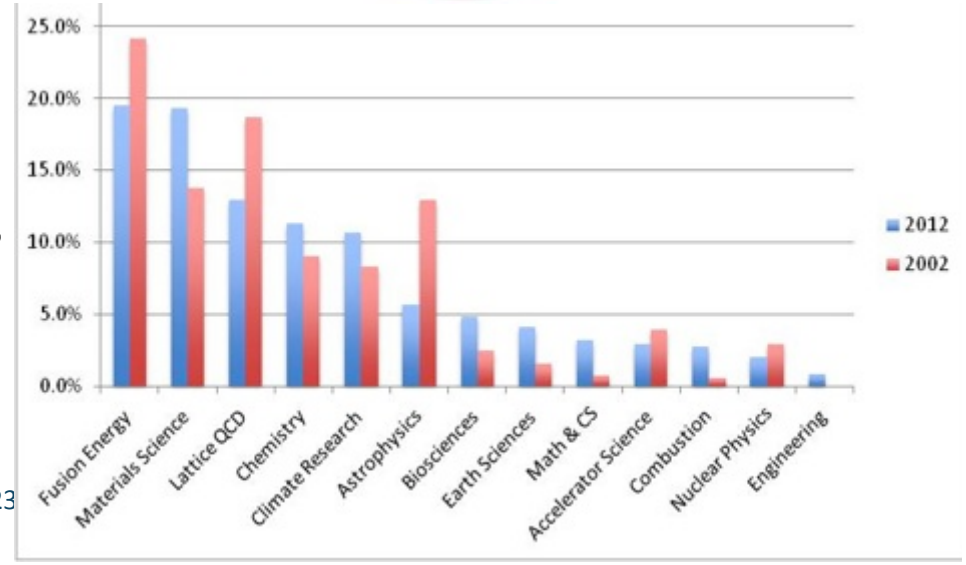
	1-31	32-63	64-511	512-1K	1K-4K	4K-8K	8K-16K	16K-32K	32K-64K	65-128K	128K+
Hours	7	10.7	55	31	49	28	7	66	30	8.7	0.382
Number	675,233	30,991	123,776	15,672	14,618	1,604	826	667	202	56	15

We 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



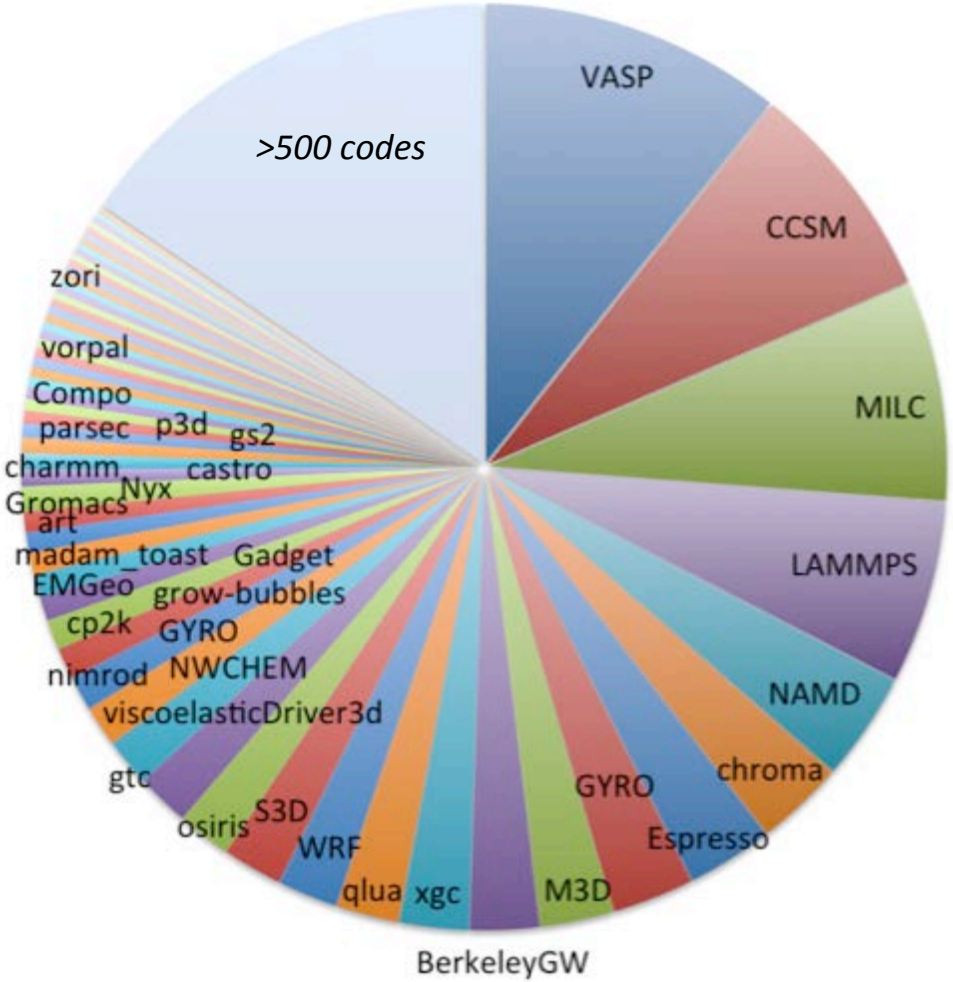


Disruptions in programming models are a challenge for NERSC

- Many users
- Many codes
- We don't select our users

We will initially focus on 20 codes

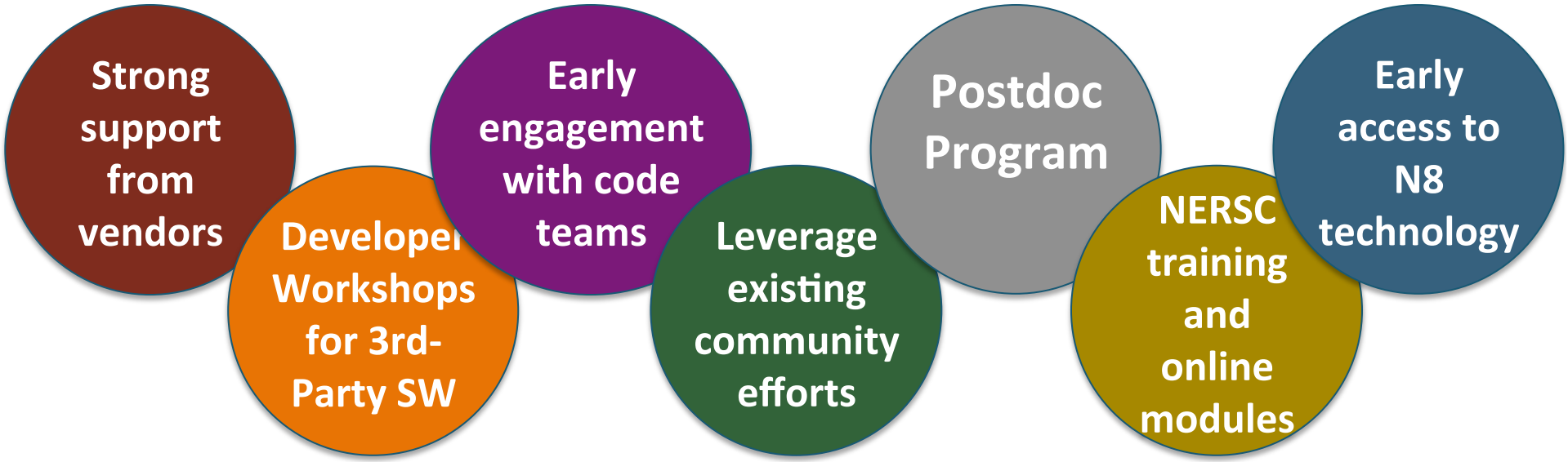
Breakdown of Application Hours on Hopper and Edison 2013



- **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:**



SC Lab coordination

- **Tri-facility app readiness meeting with LCFs, 3/25/2014 at Berkeley Lab**
  - Compilation of Office of Science applications
- **Ongoing discussion of best practices**
- **LCFs representatives helped evaluate NESAP proposals**
  - Aware of initial scoring of proposals
  - We will communicate our final choices with them

OLCF
Oak Ridge Leadership Computing Facility

CENTER FOR ACCELERATED APPLICATION READINESS (CAAR)

Argonne Leadership Computing Facility

Early Science Program (ESP)

NESAP Resources

- **DOE/NERSC**
  - 8 Post-docs
- **Cray**
  - 5 FTE years of application and optimization support
  - User training
- **Intel**
  - Quarterly Dungeon sessions – 16 in total
  - Remote access to an early KNL system
  - KNL white boxes @ NERSC before arrival of N8
  - 4 Training sessions – 2 per year
  - Intel associate on-site 1 week/month for 4 years

Because of the uneven amount of resources, we are planning a 3 tiered program

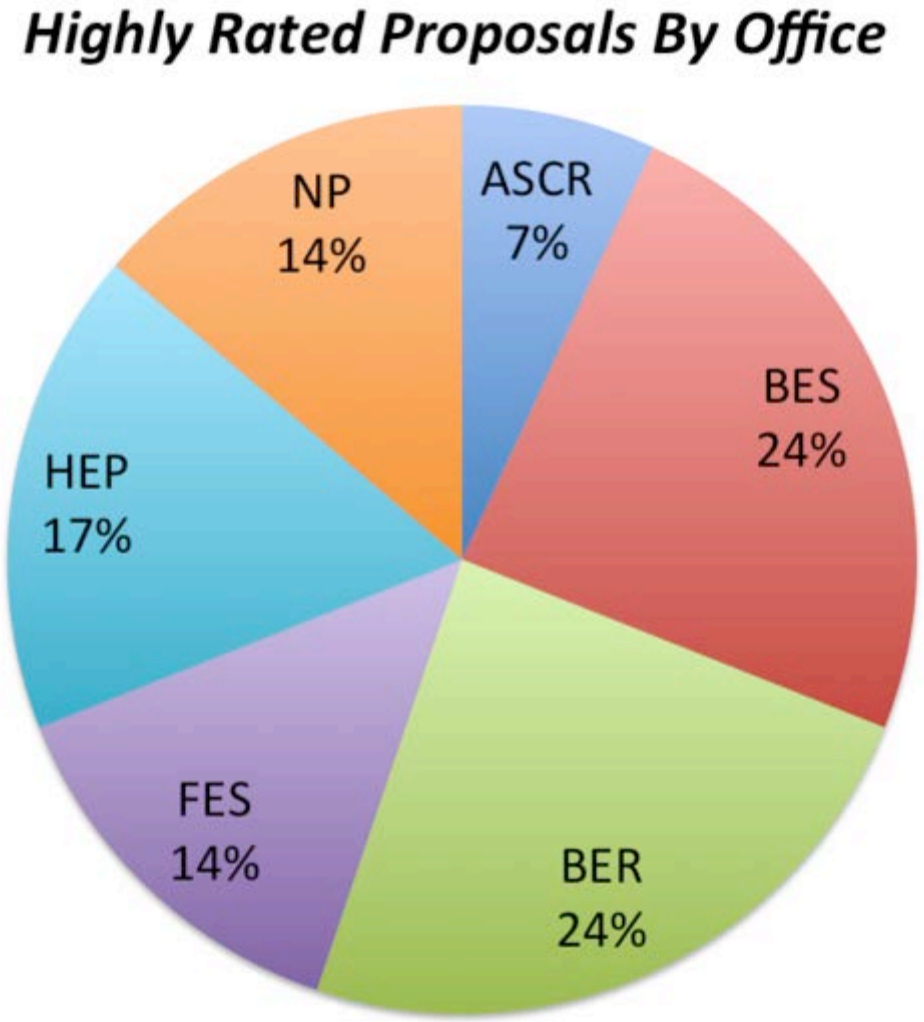
- **Tier 1: 8 Application teams**
  - Each team will have a embedded post-doc
  - Access to an Intel dungeon session
  - Support from NERSC Application Readiness and Cray COE staff
  - Early access to KNL testbeds and Cori system
  - User training sessions from Intel, Cray and NERSC staff
- **Tier 2: 12 Application teams**
  - All the resources of the Tier 1 teams except for an embedded post-doc
- **Tier 3: ~10-20 Application teams**
  - Access to KNL testbeds, Cori system and user trainings
    - Could be very advanced user groups who don't need much help
    - Or code teams where there were many submissions from the same area (example: QCD)
    - Motivated code teams that didn't make the cut to tier 1 or tier 2

Key point for Tier 3 applications teams: Encourage capable teams to prepare for Cori! Early access and training sessions can be provided to more users at minimal cost.

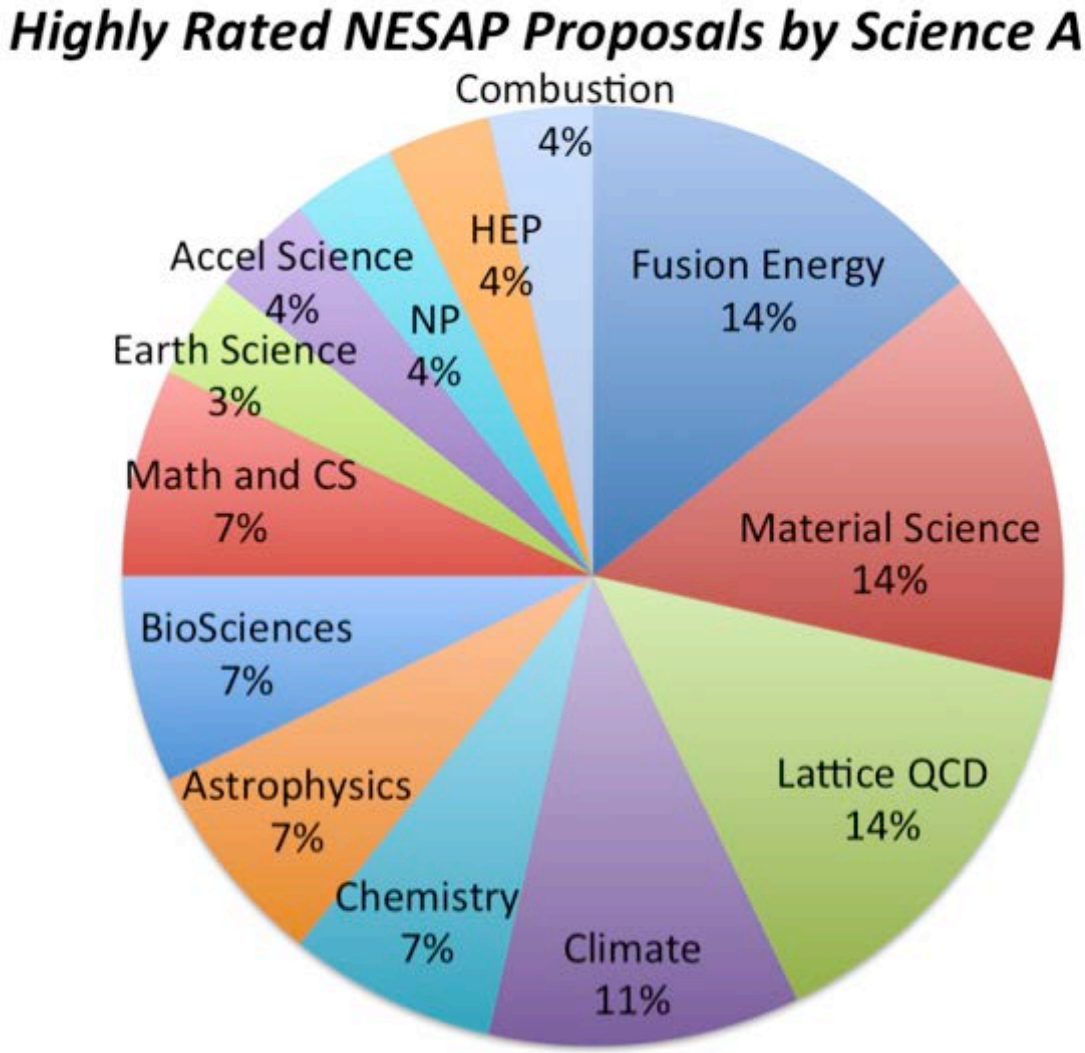
We used 4 criteria for the preliminary evaluations of the 50+ proposals

- An Application's usage within the DOE Office of Science
- Ability for proposal/application team to produce scientific advancements
- Ability for code development and optimizations to be transferred to the broader community through libraries, algorithms, kernels or community codes
- Resources available from the application team to match NERSC/Vendor resources

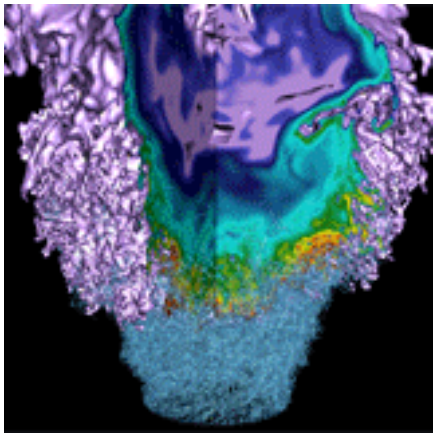
All Offices are represented by the highly rated proposals

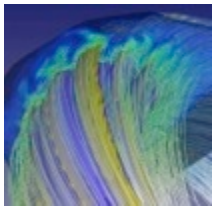


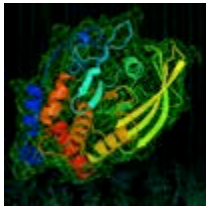
Science areas are also well represented

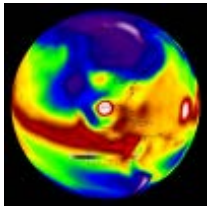


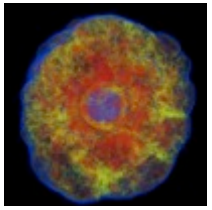
Extreme Data Science

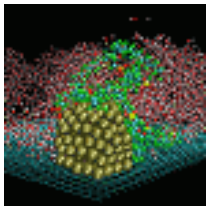










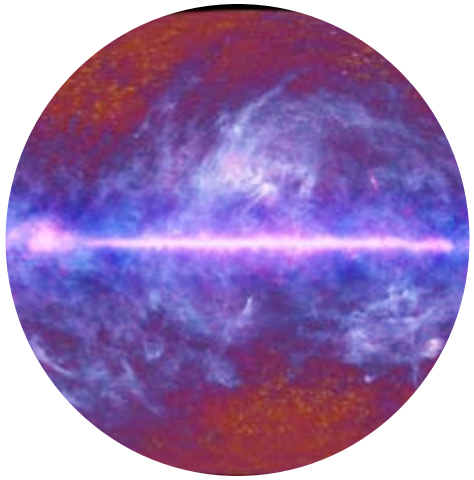


NERSC 40 YEARS at the FOREFRONT 1974-2014

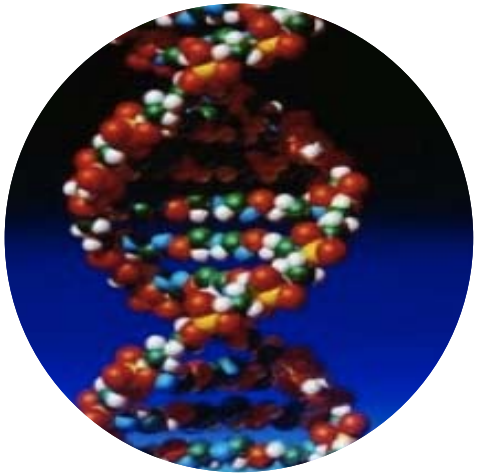
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

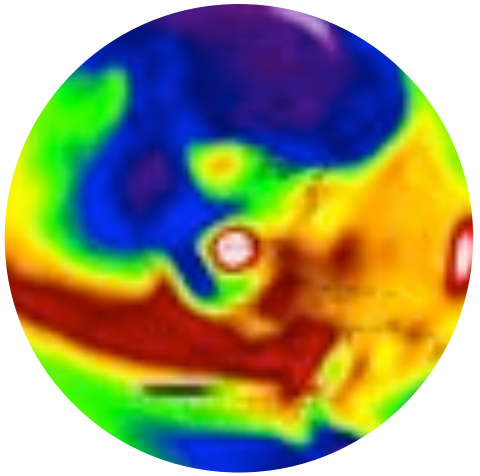
DOE Facilities are Facing a Data Deluge



Astronomy

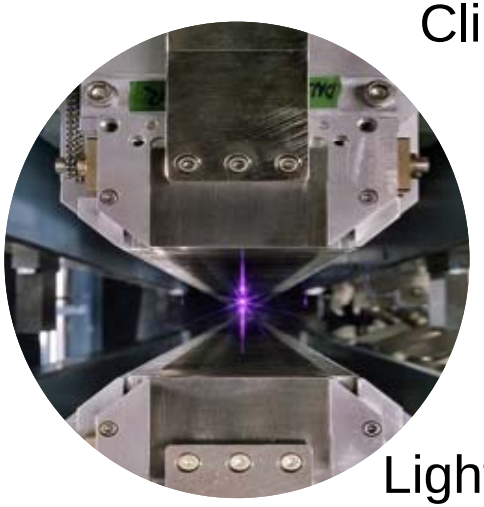


Genomics



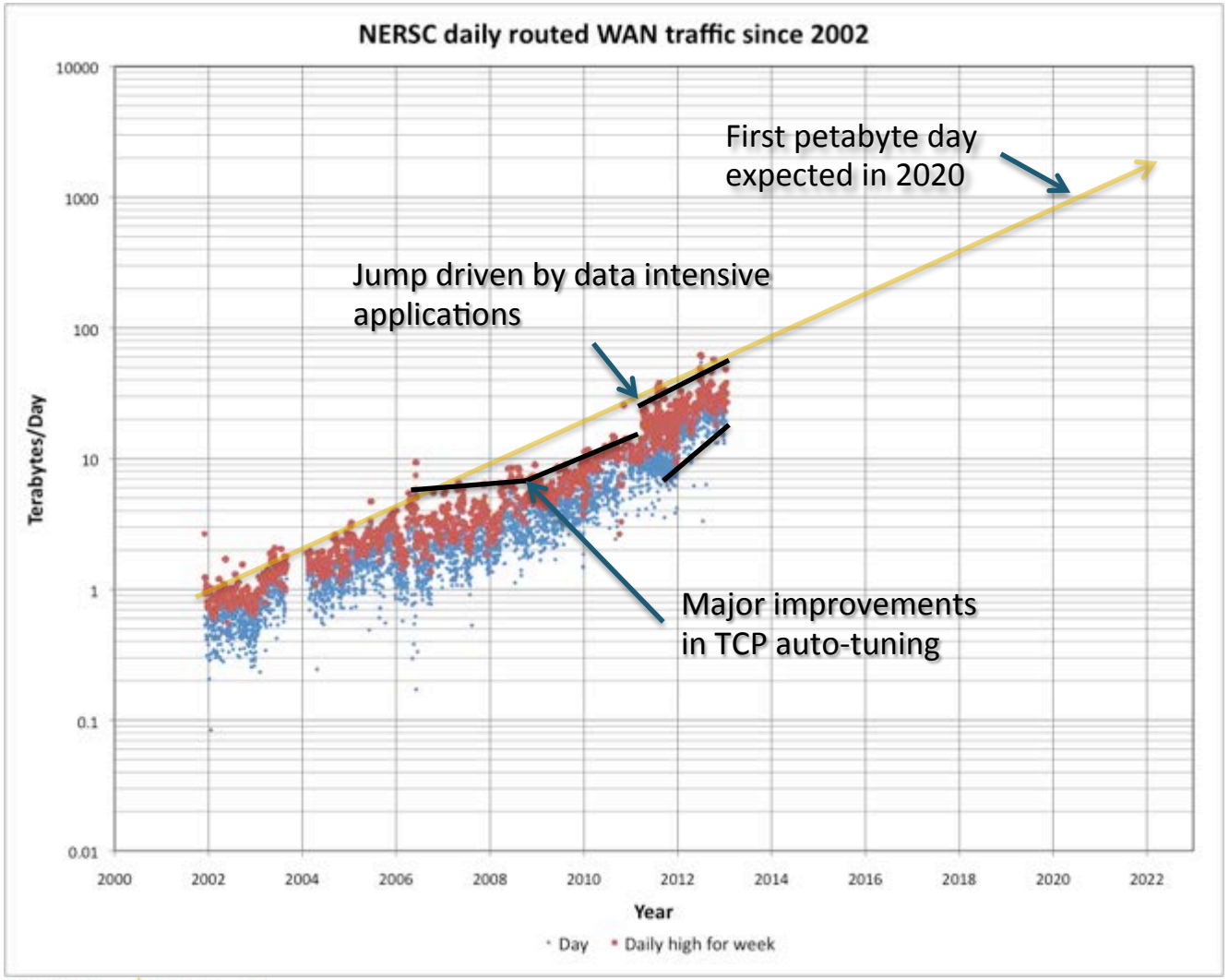
Climate

Physics



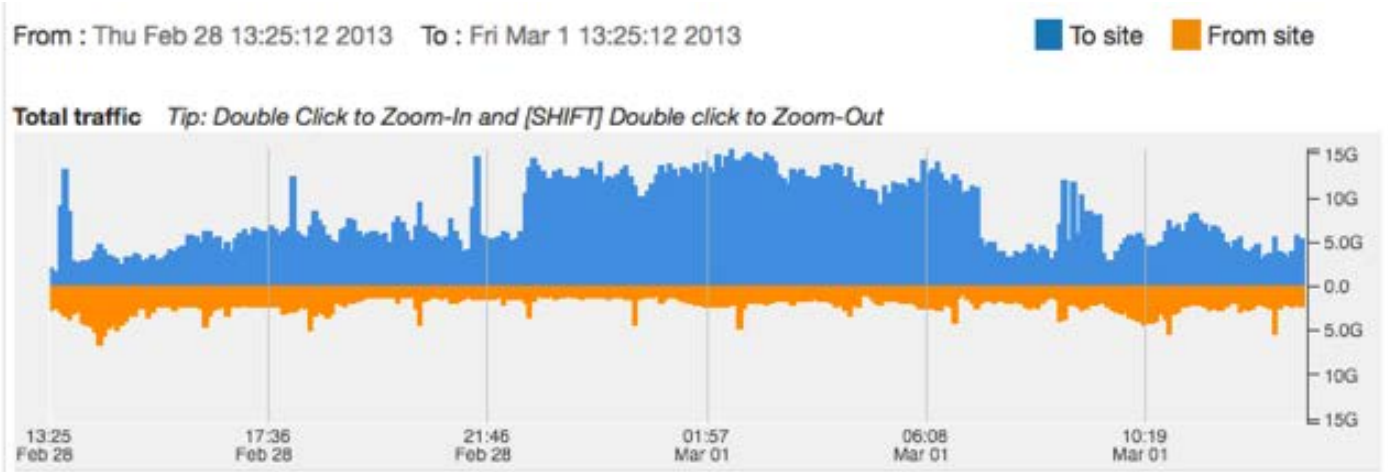
Light Sources

Exponentially increasing data traffic



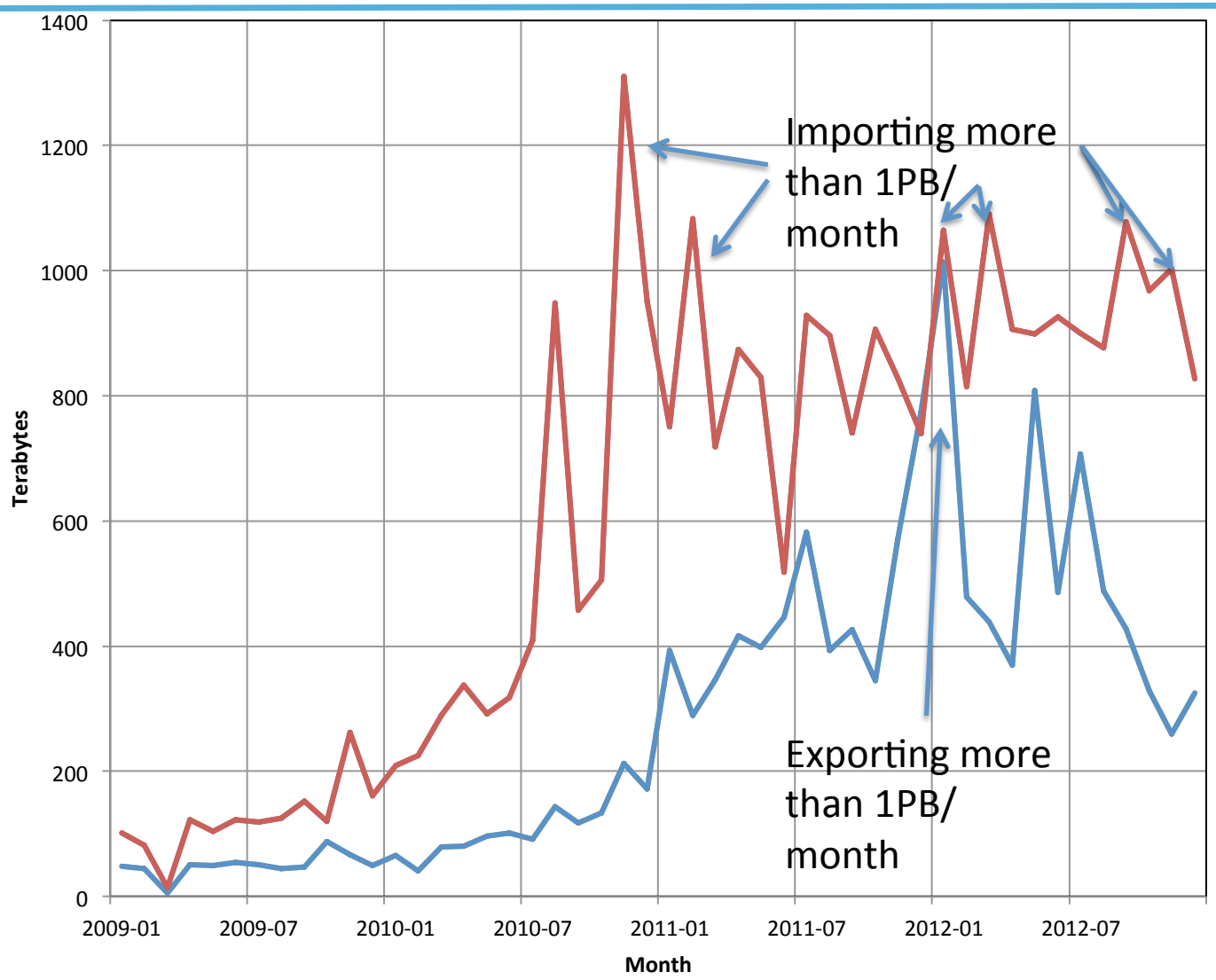
Cross Bay Data Transfer

All NERSC Traffic



Photosystem II X-Ray Study

NERSC users import more data than they export!



Extreme Data Science is Playing a Key Role in Scientific Discovery

- Discovery of the Higgs Boson
- 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
- Four of Science Magazines breakthroughs of the last decade were in Genomics
- Materials project has over 5000 users and was featured on the cover of Scientific American

PI: Shri Kulkarni (Caltech)

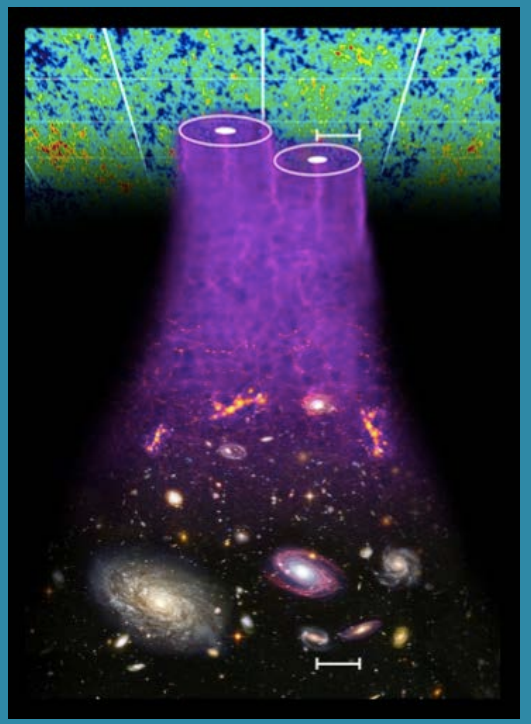
We currently deploy separate HPC systems and Data Intensive Systems

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

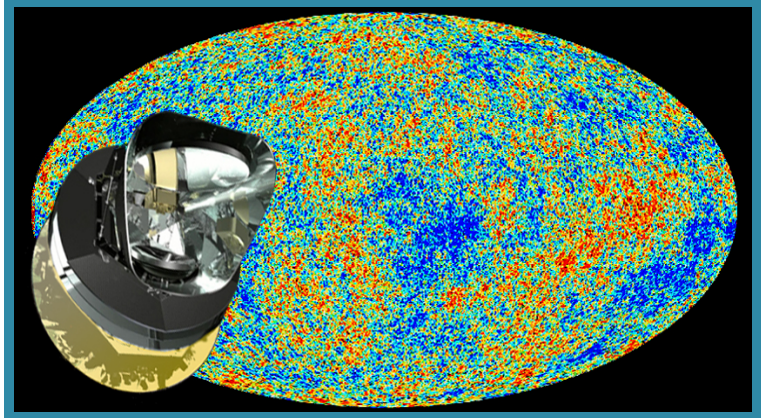
Statistical analyses need MCMC for cross-correlation of the millions of galaxies
-- collapsing the problem to just 2-point statistics.

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

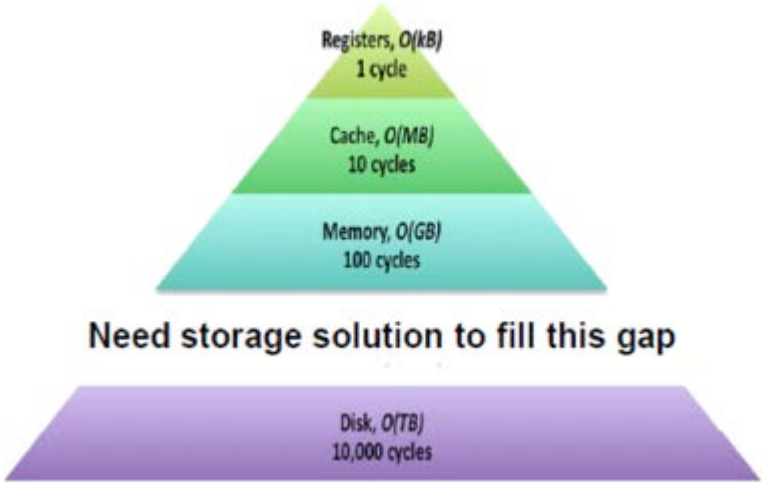
Parallel file system comparison

	Cori	Hopper (2 filesystems aggregate)
Bandwidth	432 GB/s	70 GB/s (35+35)
Metadata ops (creates/s)	77 K/s	34 K/s (17+17)
Capacity	28.5 PB	2.2 PB (1.1 +1.1)
Delta-PFS*	29 min	44 min

Delta-PFS: Time to write 80% of memory to the Parallel File System

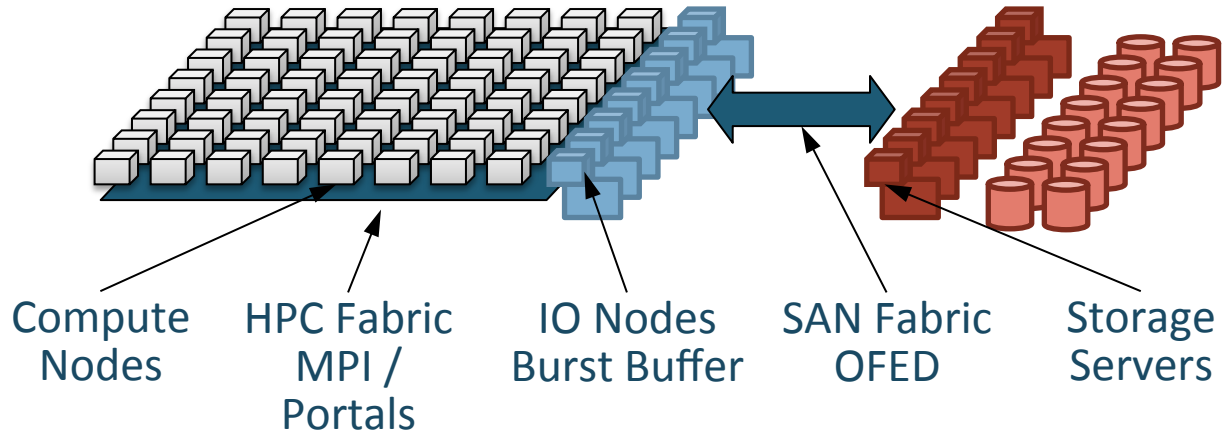
Burst Buffer

- **Flash storage which would act as a cache to improve peak performance of the PFS.**



- **Flash is currently as little as 1/6 the cost of disk per GB/s bandwidth and has better random access characteristics (no seek penalty).**

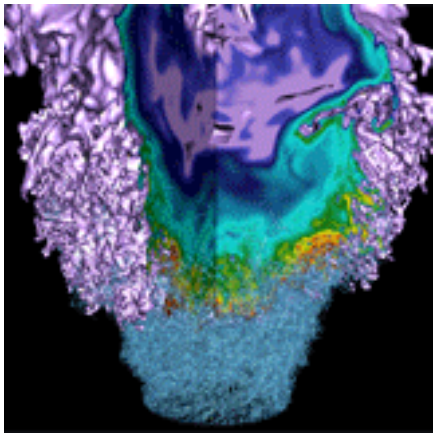
Burst Buffer Software NRE Efforts

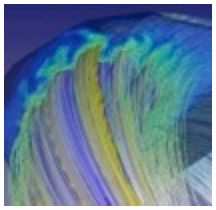


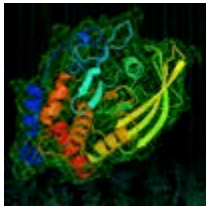
Create Software to enhance usability and to meet the needs of all NERSC users

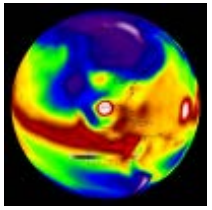
- Scheduler enhancements
  - Automatic migration of data to/from flash
  - Dedicated provisioning of flash resources
  - Persistent reservations of flash storage
- Enable In-transit analysis
  - Data processing or filtering on the BB nodes – model for exascale
- Caching mode – data transparently captured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

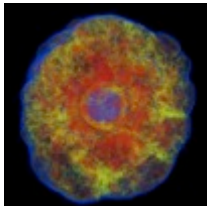
Conclusions

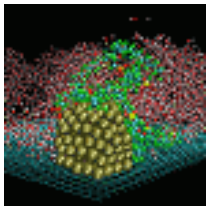










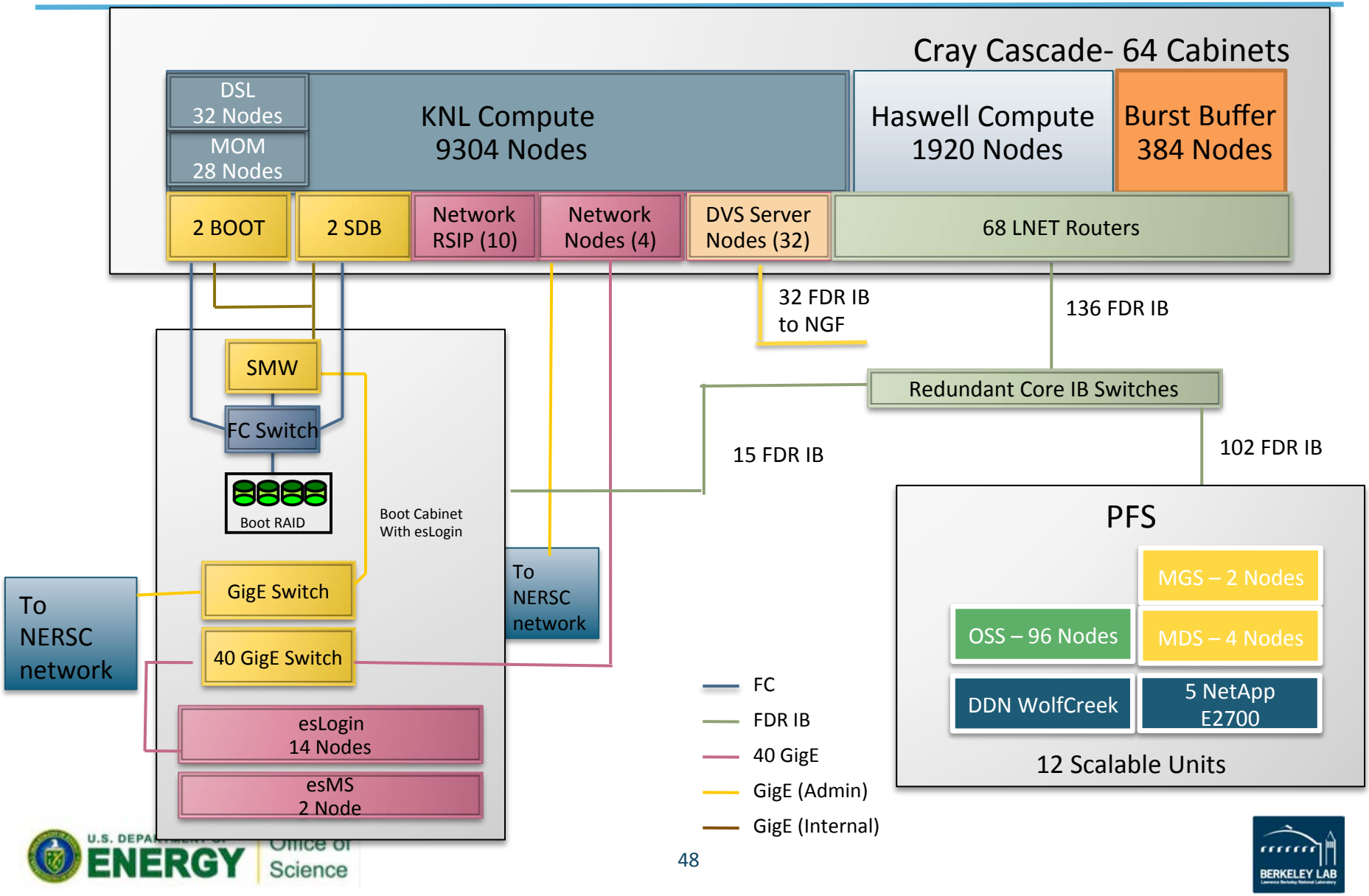


NERSC 40 YEARS at the FOREFRONT 1974-2014

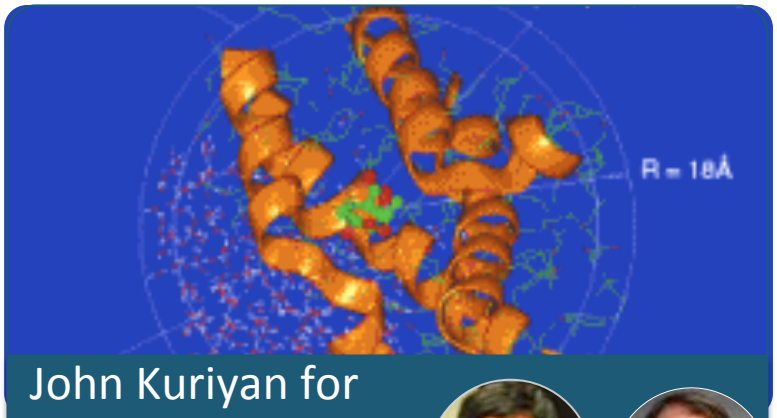
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

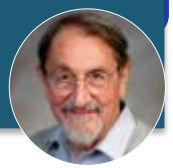
The Cori System



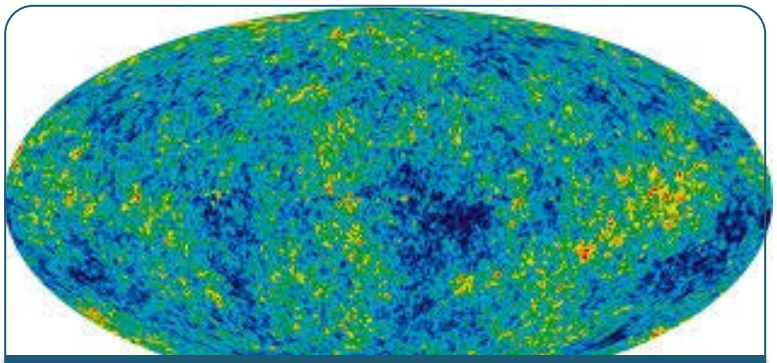
Our goal is to enable science that can't be done on today's supercomputers



John Kuriyan for Martin Karplus



Saul Perlmutter



George Smoot

Warren Washington