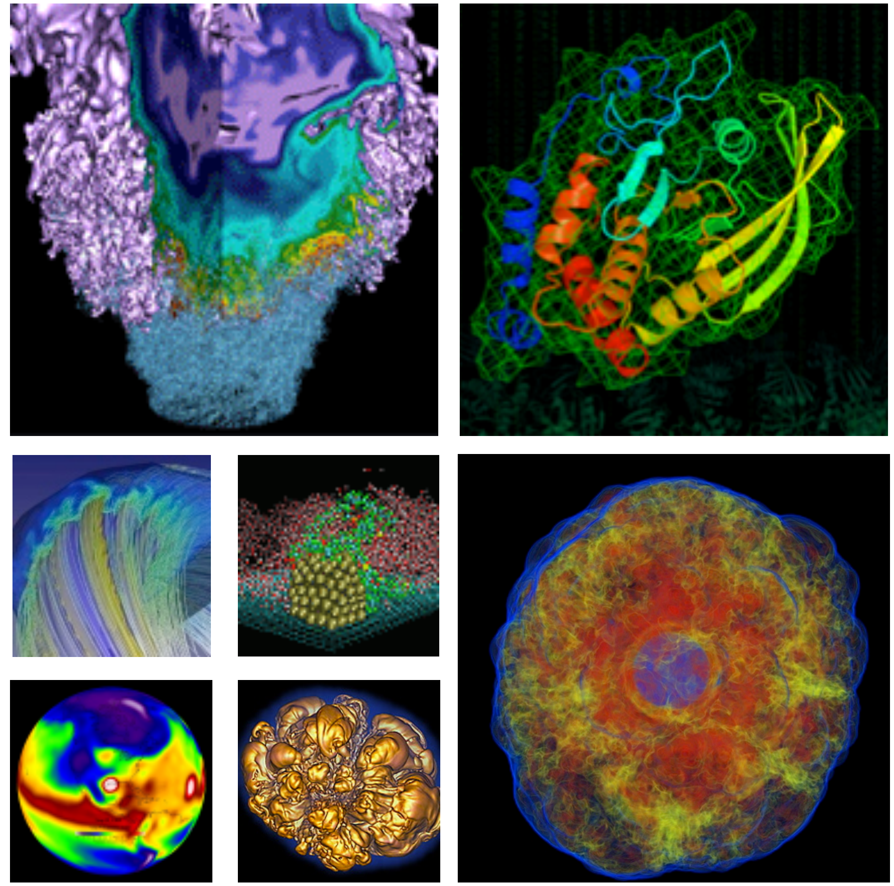
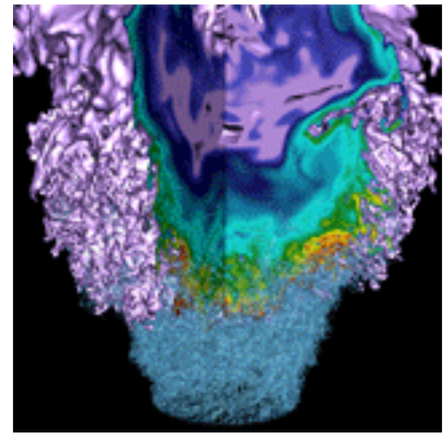
Big Computing, Big Data, Big Science

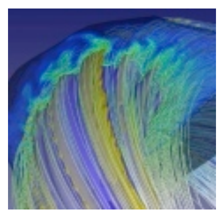


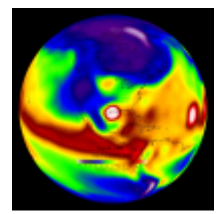
Sudip Dosanjh
Director

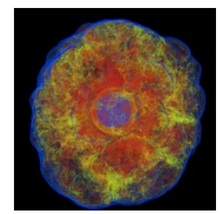
July 8, 2014

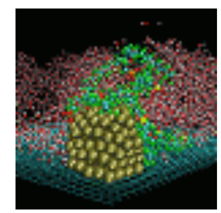
Exascale at NERSC











NERSC 40 YEARS at the FOREFRONT 1974-2014

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Exascale Strategic Objective

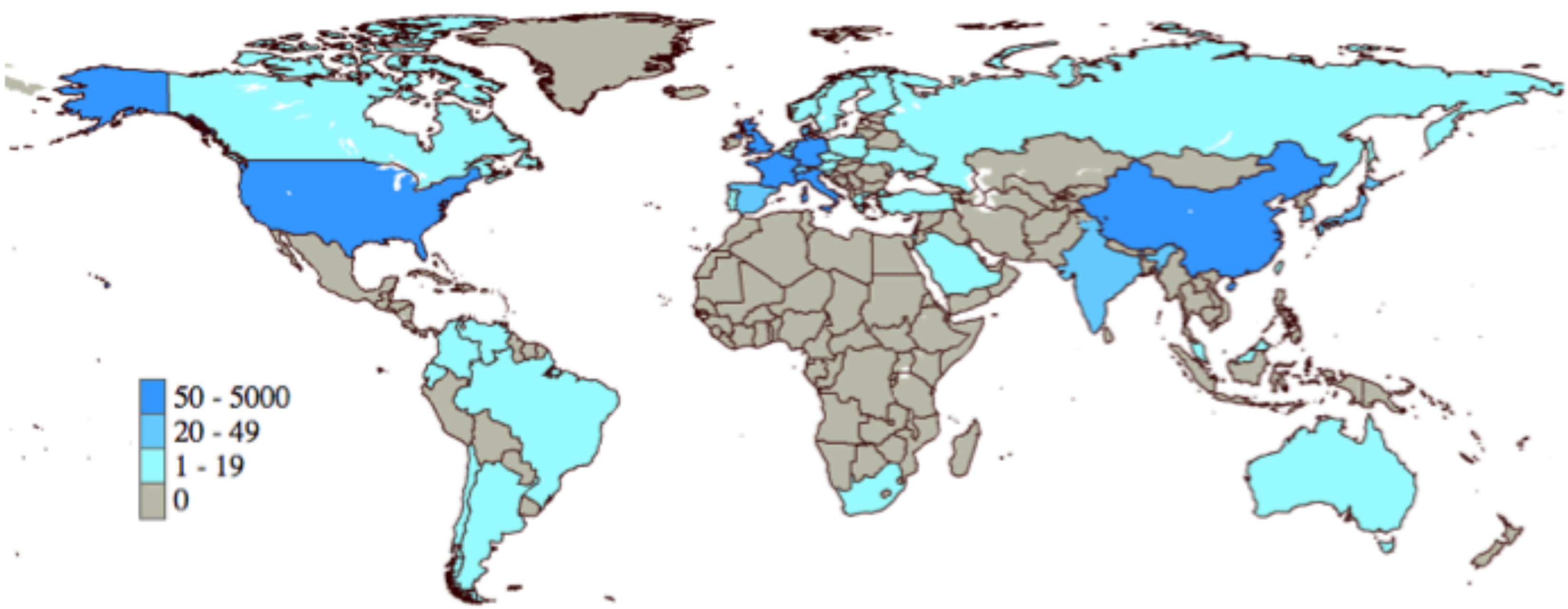
- **Meet the ever-growing computing and data needs of our users by**
  - providing usable exascale computing and storage systems
  - transitioning SC codes to execute effectively on manycore architectures
  - influencing the computer industry to ensure that future systems meet the mission needs of SC

NERSC collaborates with computer companies to deploy advanced HPC and data resources

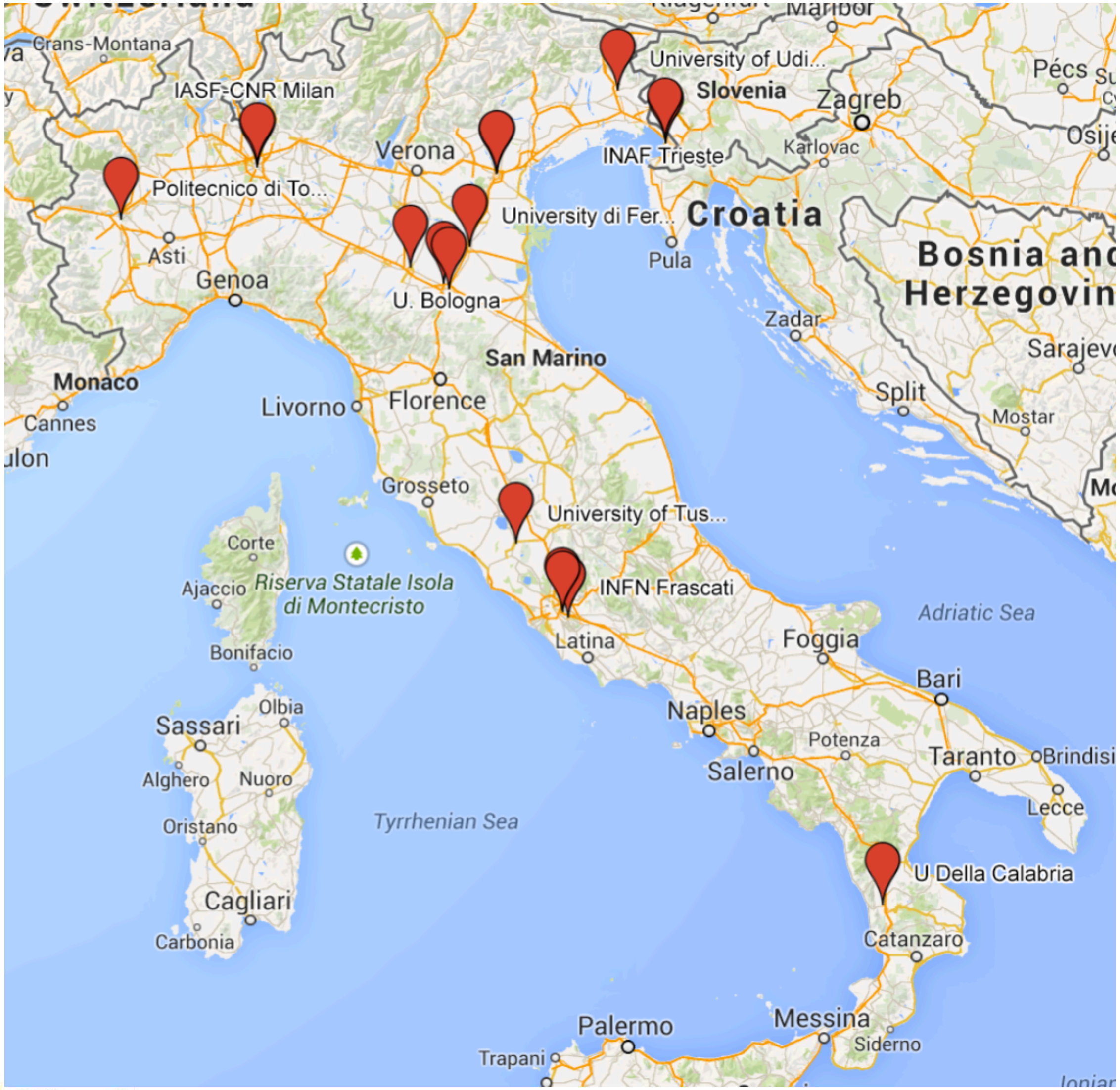
- **Hopper (N6) and Cielo (ACES) were the first Cray petascale systems with a Gemini interconnect**
- **Architected and deployed data platforms including the largest DOE system focused on genomics**
- **Edison (N7) is the first Cray petascale system with Intel processors, Aries interconnect and Dragonfly topology (serial #1)**
- **Cori (N8) will be one of the first large Intel KNL systems and will have unique data capabilities**

We support a broad user base

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



Users in Italy

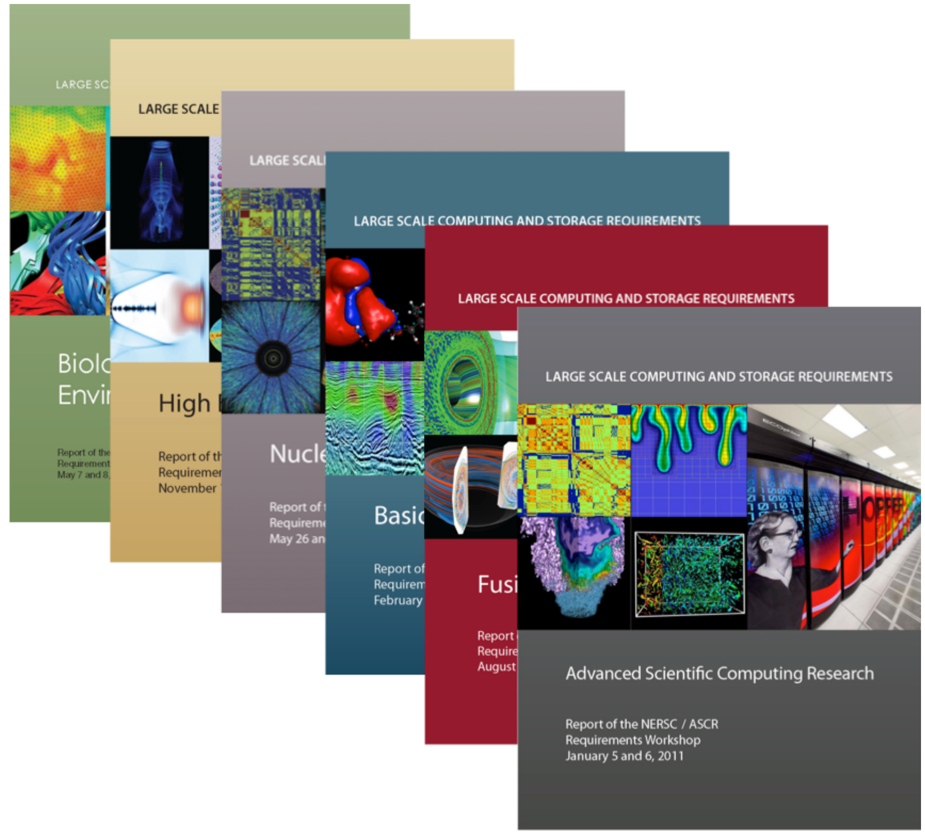


Disruptions in programming models are a challenge for NERSC

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

Requirements with six program offices

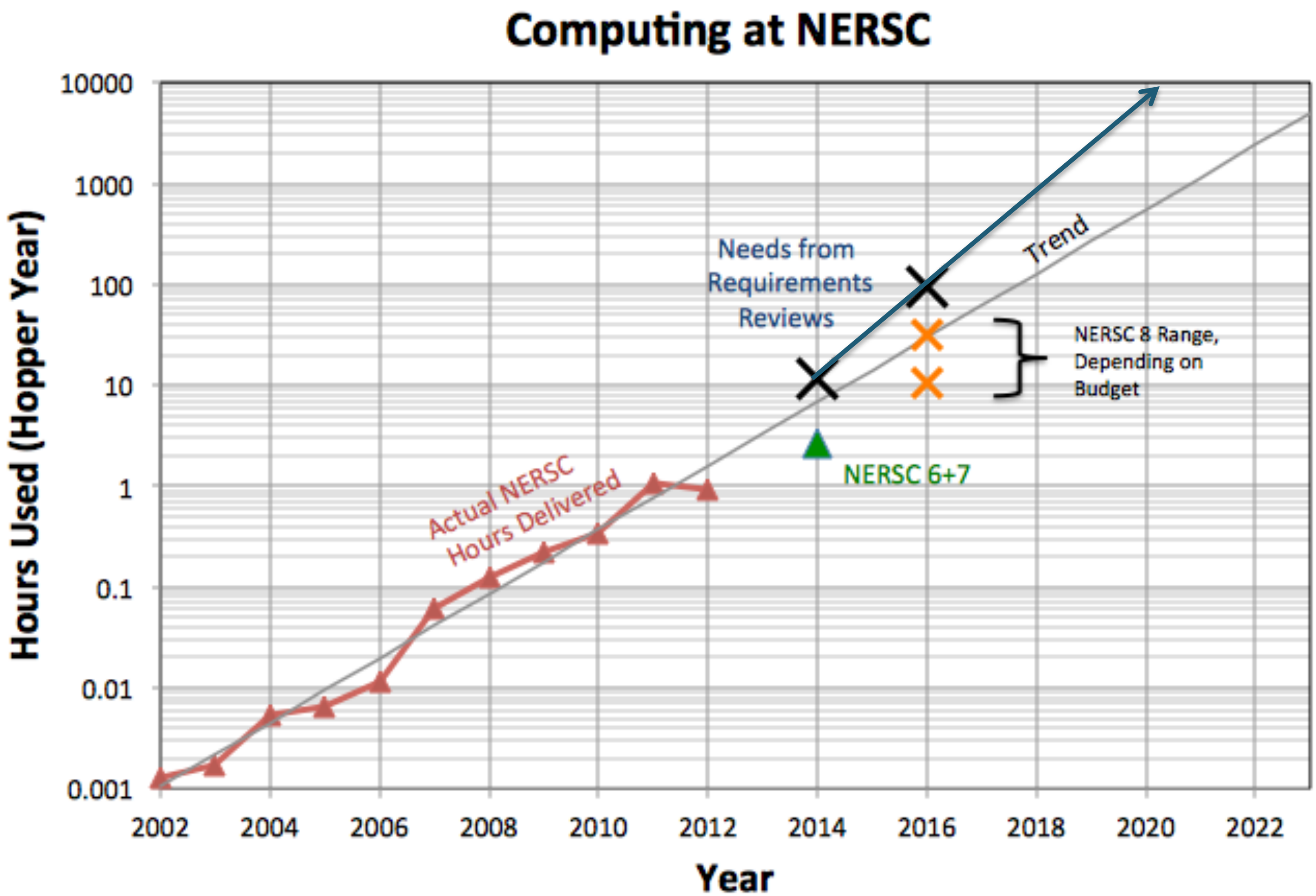
- Reviews with six program offices every three years
- Program managers invite representative set of users (typically represent >50% of usage)
- Identify science goals and representative use cases
- Based on use cases, work with users to estimate requirements
- Re-scale estimates to account for users 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-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.

Cori Configuration

- **64 Cabinets of Cray XC System**
  - Over 9,300 'Knights Landing' compute nodes
    - Self-hosted (not an accelerator)
    - Greater than 60 cores per node with multiple hardware threads each
    - 64-128 GB memory per node
    - High bandwidth on-package memory
  - 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

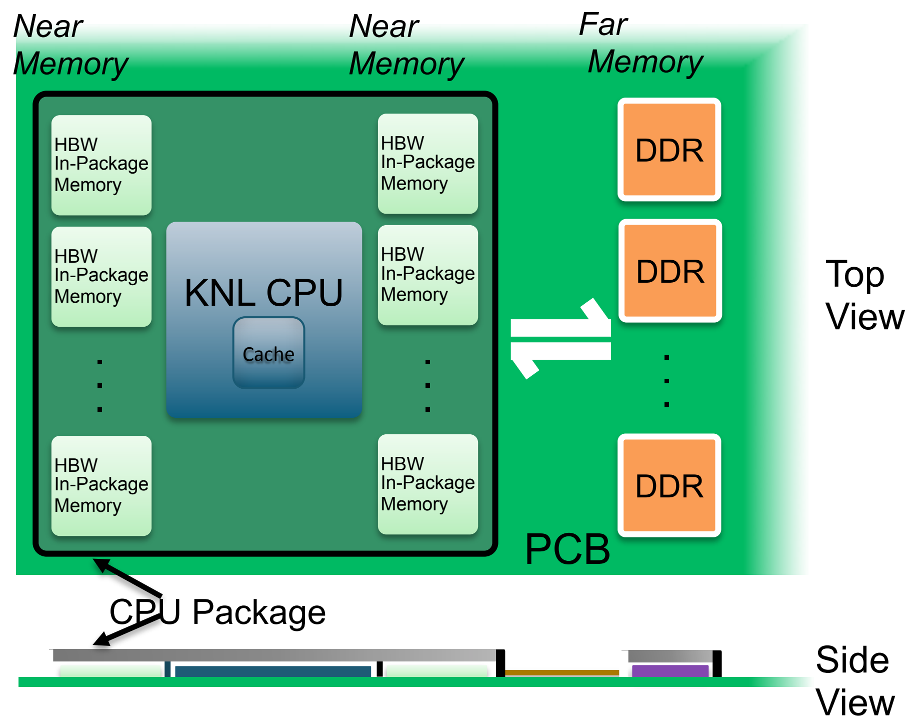
- **Knight's Landing is a self-hosted part**
  – Users can focus on adding parallelism to their applications without concerning themselves with PCI-bus transfers
- **MPI + OpenMP preferred programming model**
  – Should enable NERSC users to make robust code changes
- **MPI-only will work – performance 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



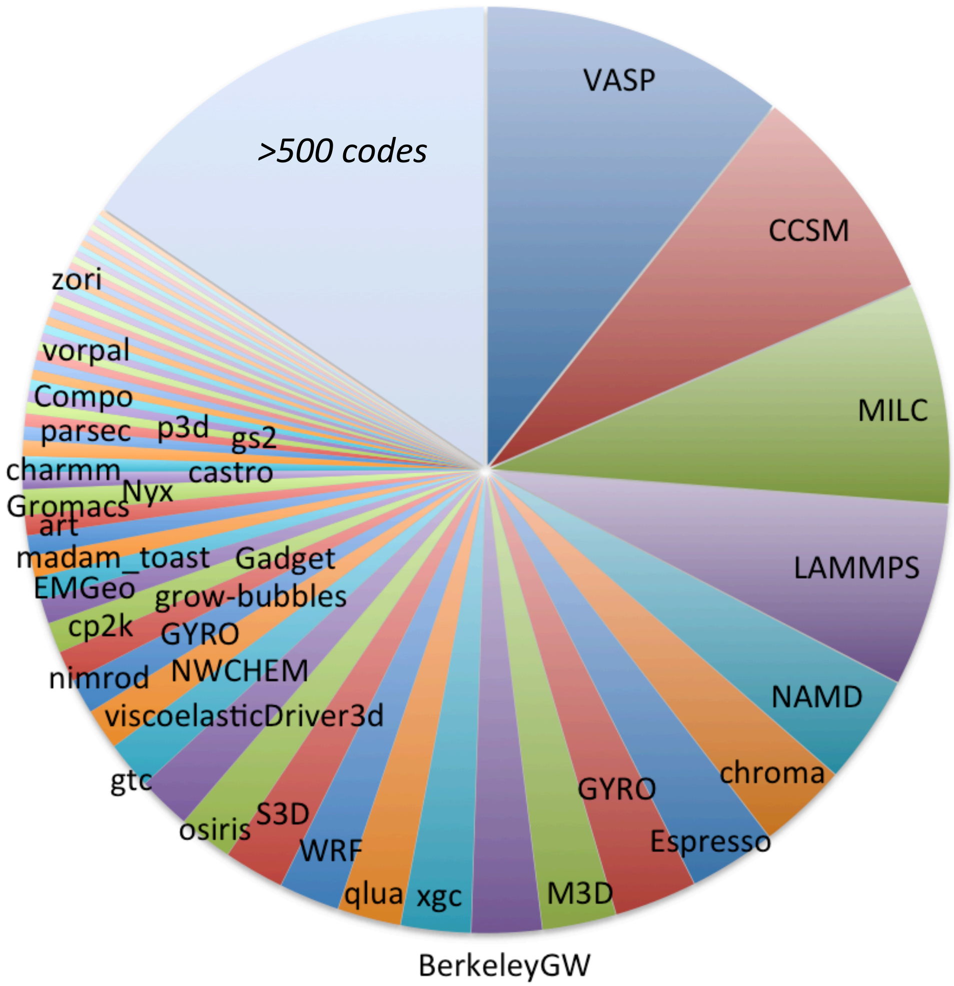
Maximum performance through higher memory bandwidth and flexibility

NERSC's Key Challenges

- **Application Readiness**
  - We must prepare the broad user community for manycore architectures, not just a few codes
  - Will require deep collaboration with select code teams
  - Finding additional application parallelism is the main challenge
  - Unclear how to use on-package memory, as explicit memory or cache
- **Burst Buffer**
  - How to integrate and monitor in a production environment?
  - Which applications are best suited to use the Burst Buffer?
  - How to make the Burst Buffer user friendly
- **Integration into NERSC environment in CRT**
  - Mounting NERSC-8 file system across other systems, (Edison)
  - Integration into a new facility

NERSC's workload is highly concentrated and unequally distributed

<u>***Breakdown of Application Hours on Hopper and Edison 2013***</u>



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

We will partner closely with Cray and Intel

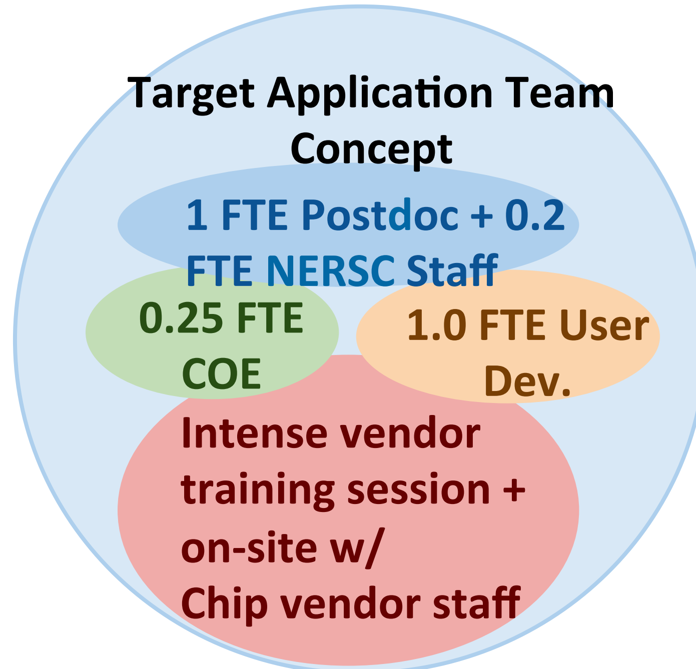
- **Cray**
  - 5 FTE years of application and optimization support
- **Intel**
  - Remote access to an early KNL system
  - KNL white boxes @ NERSC before arrival of N8
  - 4 Training sessions – 2 per year
  - Quarterly Dungeon sessions – 16 in total
  - Intel associate on-site 1 week/month for 4 years

We Launched the NERSC Exascale Science Application Program

- **Umbrella program for all NERSC Application Readiness Activities**
- **Approximately 20 application teams will be accepted into NESAP**
- **Each application team will be partnered with a member of NERSC's App Readiness team who will assist with code profiling and scaling analyses**
- **Through this program NERSC will allocate resources from Cray and Intel**
- **8 application teams will receive NERSC funded Post-docs**
- **Partnership with ALCF, OLCF and the DOE HPC community is a key**

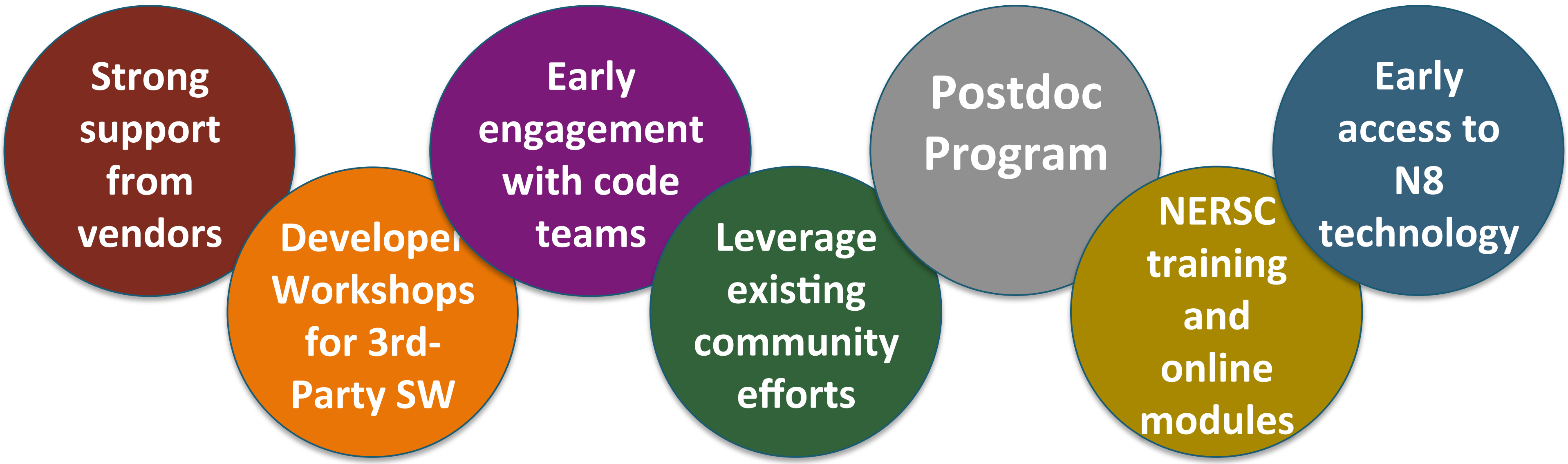
A Portion of the NESAP Projects will have Postdocs Assigned

- **Postdocs conduct applied R&D in energy-efficient HPC to enable new, pathbreaking science with Cori supercomputer.**
- **Opportunity to work at the forefront of HPC, ensuring that Cori pushes the limit of what can be done; successful only if codes are state of the art**
- **Ensure that methods feedback to other postdocs, NERSC staff, vendors, and NERSC users**
  - Cross pollinate good solutions from different communities
- **Publication in journals/conferences**
- **NERSC will advertise, begin hiring early 2015 after NESAP projects selected**



NERSC Exascale Science Applications Program (NESAP)

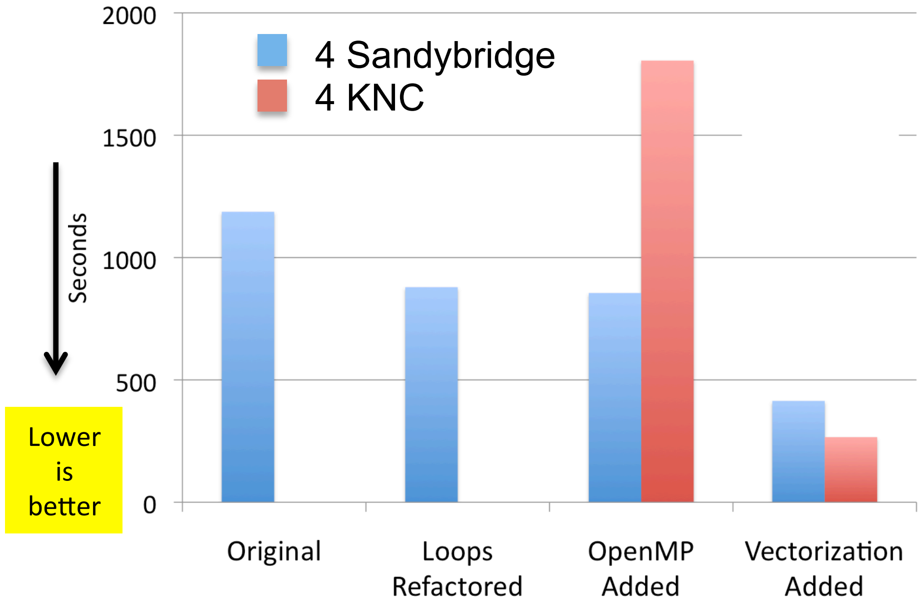
- **NESAP components:**



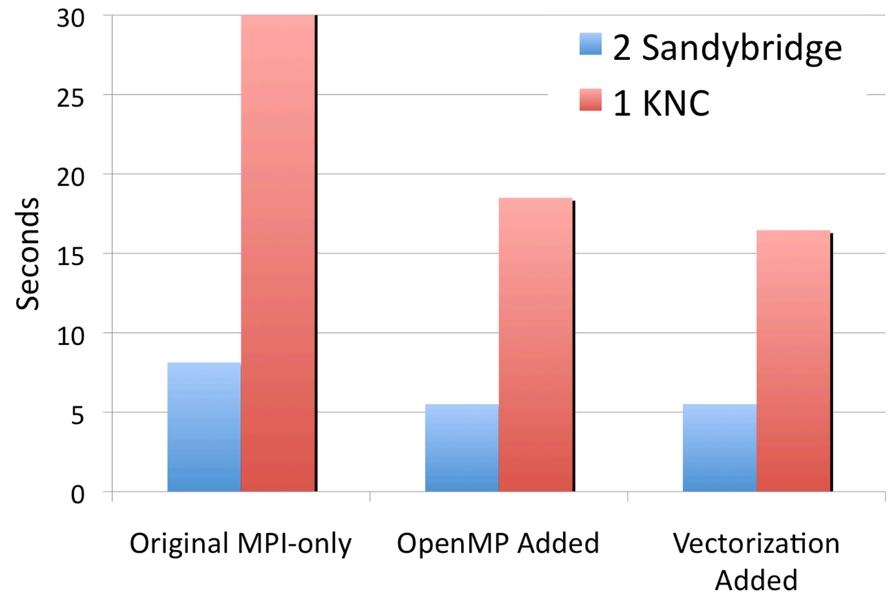
Application Readiness team is examining KNC (and GPUs)

Some applications are well suited to the Knight's architecture, while others will need significant changes to achieve good performance.

Berkeley GW Kernel Performance on Knight's Corner (KNC)



CSU Atmospheric Model Multigrid Solver on Knight's Corner (KNC)



- BerkeleyGW kernel is example of code that can benefit from manycore architecture.
- Early prototype KNC hardware roughly equals performance of Sandybridge processor
- Optimizations for KNC improve performance on Sandybridge

- Despite improvements from adding OpenMP and vectorization, this multigrid solver will need further restructuring to run on optimally on KNC

The Computational Research and Theory (CRT) building will be the home for Edison and Cori

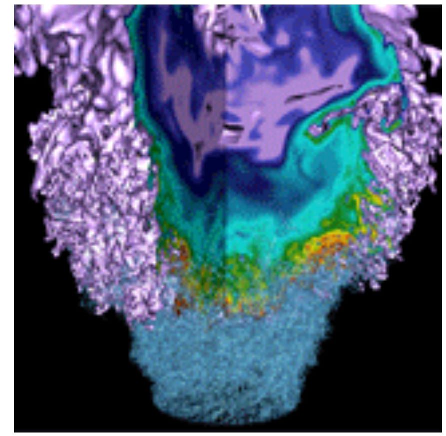
- **Four story, 140,000 GSF**
  - 300 offices on two floors
  - 20K -> 29Ksf HPC floor
  - 12.5MW -> 42 MW to building
- **Located for collaboration**
  - CRD and ESnet
  - UC Berkeley
- **Exceptional energy efficiency**
  - Natural air and water cooling
  - Heat recovery
  - PUE < 1.1
  - LEED gold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

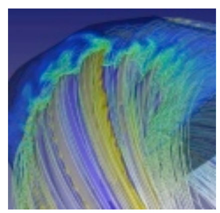
NERSC 40 YEARS at the FOREFRONT 1974-2014

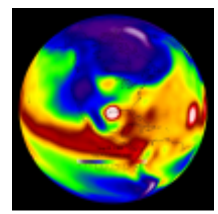
U.S. DEPAR
ENE

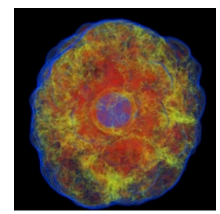
BERKELEY LAB
Lawrence Berkeley National Laboratory

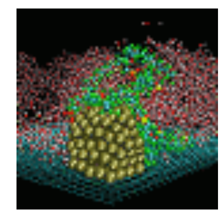
Extreme Data Science









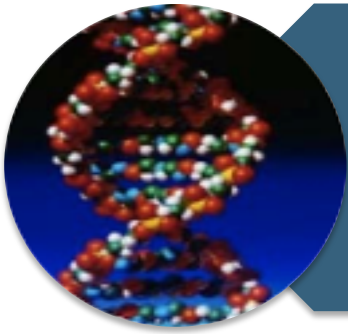


Data Strategic Objective

- **Increase the productivity, usability, and impact of DOE's user facilities by providing comprehensive data systems and services to store, analyze, manage, and share data from those facilities**

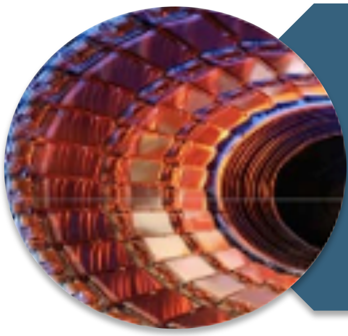
DOE "Big Data" Challenges
Volume, velocity, variety, and veracity

NERSC 40 YEARS at the FOREFRONT 1974-2014



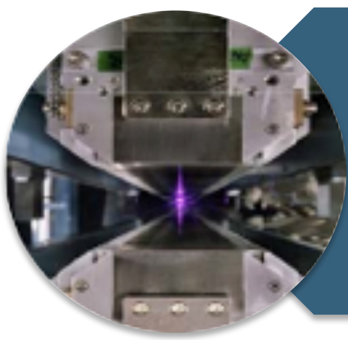
Biology

- *Volume:* Petabytes now; computation-limited
- *Variety*: multi-modal analysis on bioimages



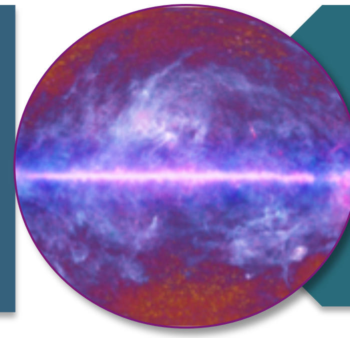
High Energy Physics

- *Volume*: 3-5x in 5 years
- *Velocity*: real-time filtering adapts to intended observation



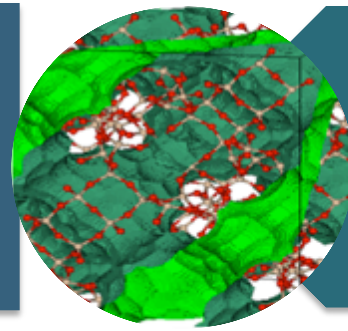
Light Sources

- *Velocity:* CCDs outpacing Moore's Law
- *Veracity:* noisy data for 3D reconstruction



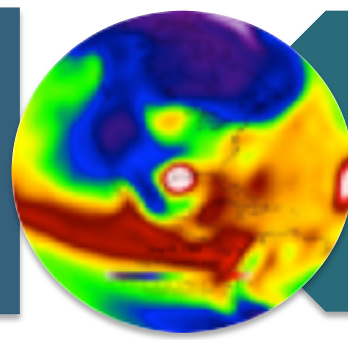
Cosmology & Astronomy:

- *Volume:* 1000x increase every 15 years
- *Variety:* combine data sources for accuracy



Materials:

- *Variety:* multiple models and experimental data
- *Veracity:* quality and resolution of simulations



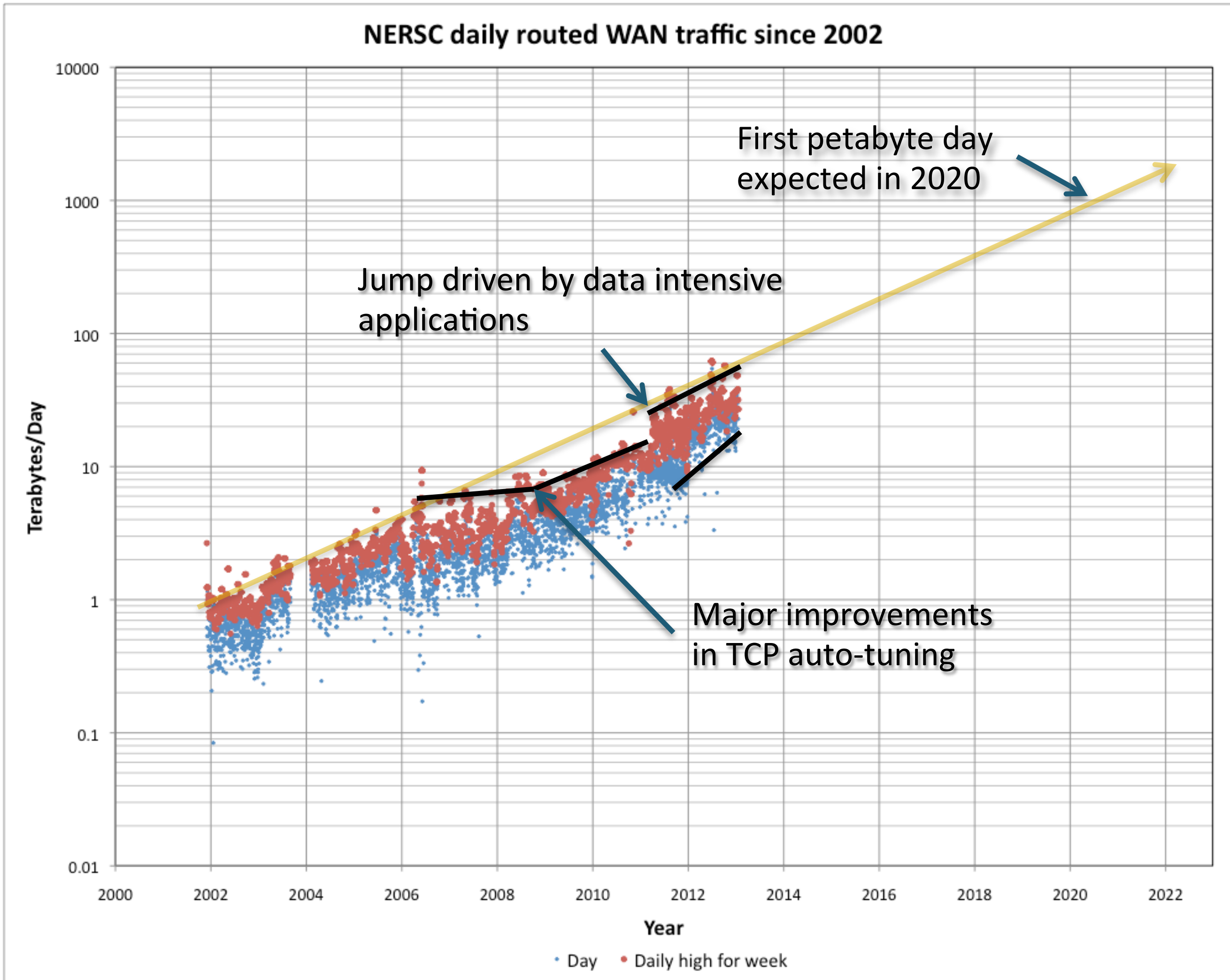
Climate

- *Volume:* Hundreds of exabytes by 2020
- *Veracity:* Reanalysis of 100-year-old sparse data

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Exponentially increasing data traffic



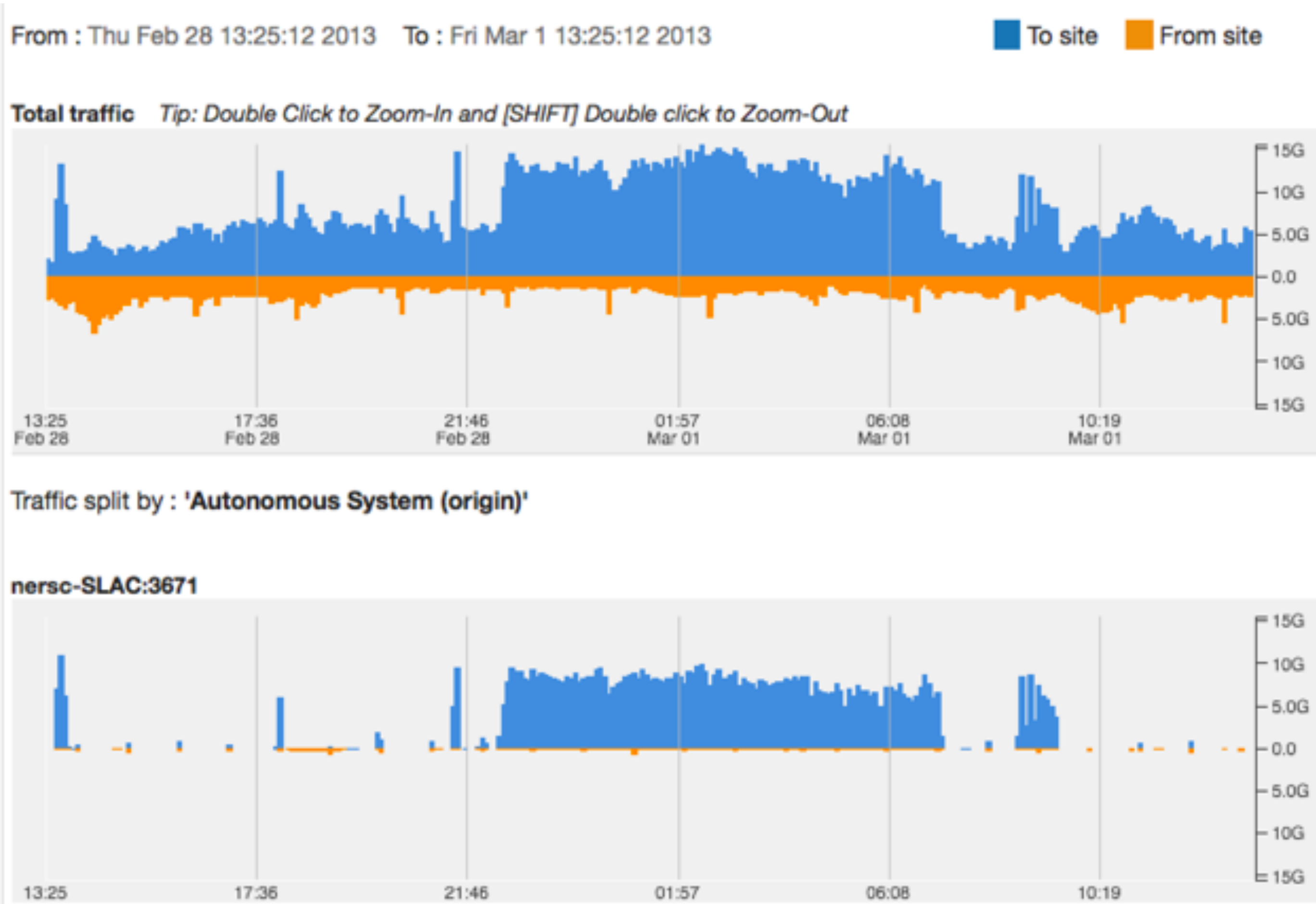
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

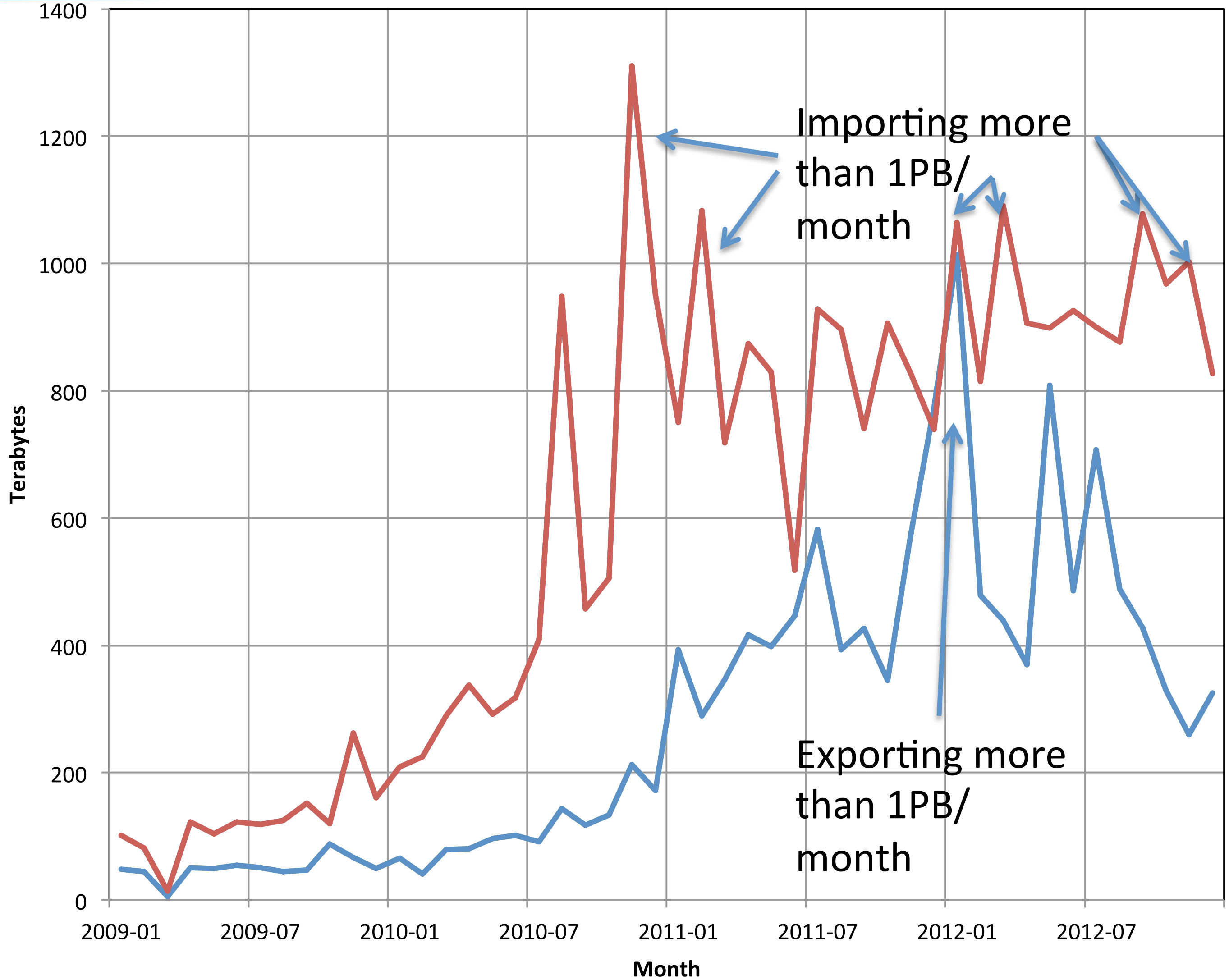
Cross Bay Data Transfer

All NERSC Traffic

Photosystem II X-Ray Study



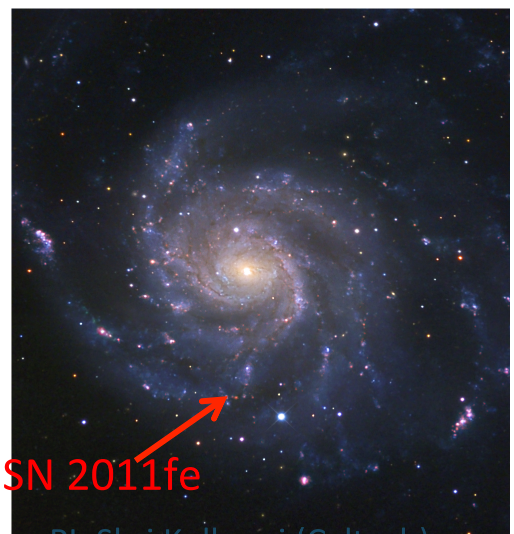
NERSC users import more data than they export!



Total Out (TB) Total In (TB) Total Out (TB) Total In (TB)

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

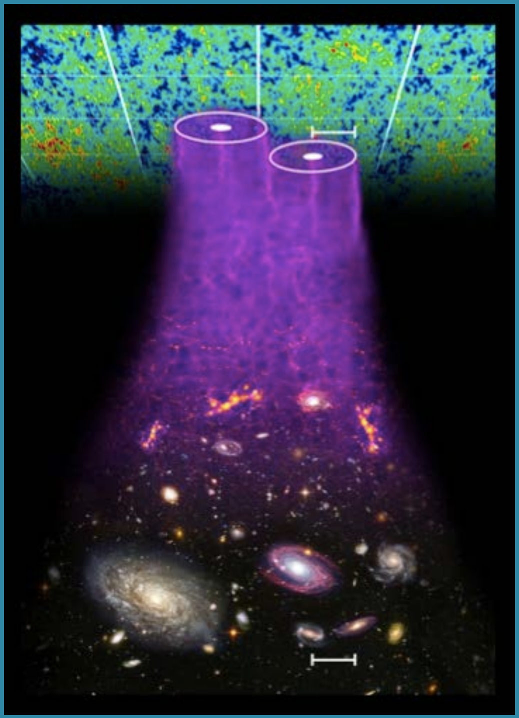
The Need for Data Intensive Systems

- **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

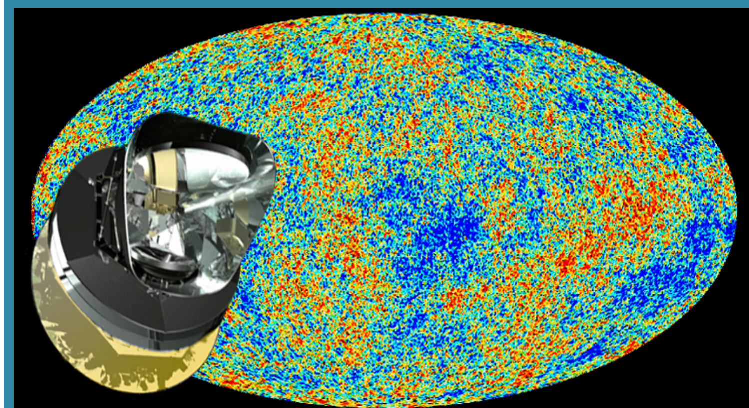
Statistical analyses need MCMC for cross-correlation of the millions of galaxies
-- collapsing the problem to just 2-point statistics.

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.

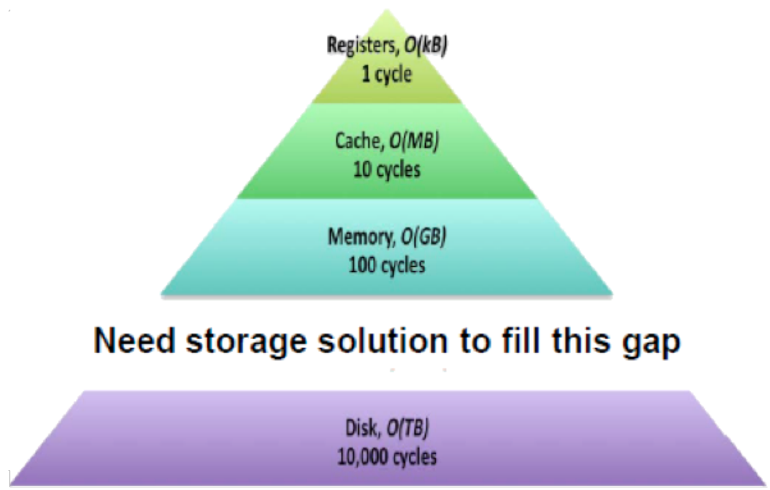
Parallel file system comparison

	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

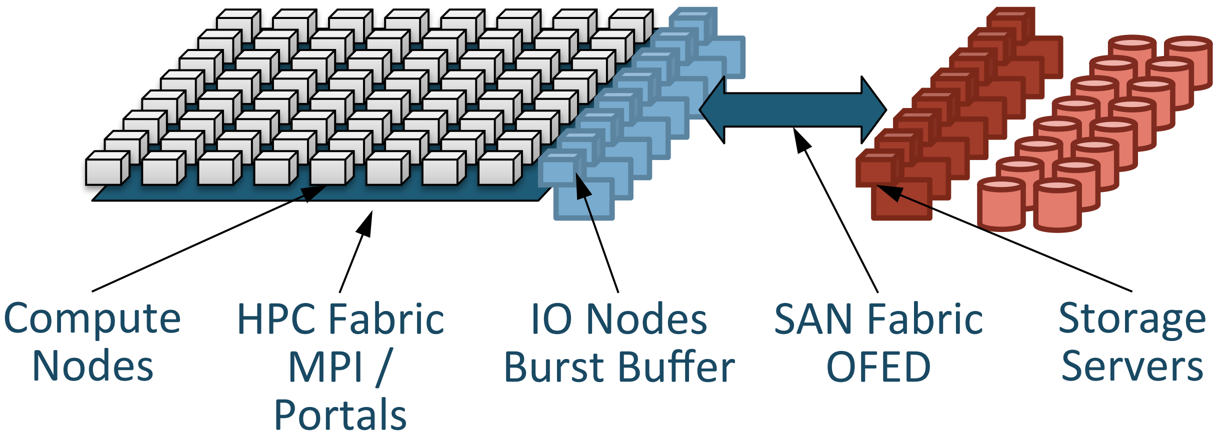
Burst Buffer

- **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).**

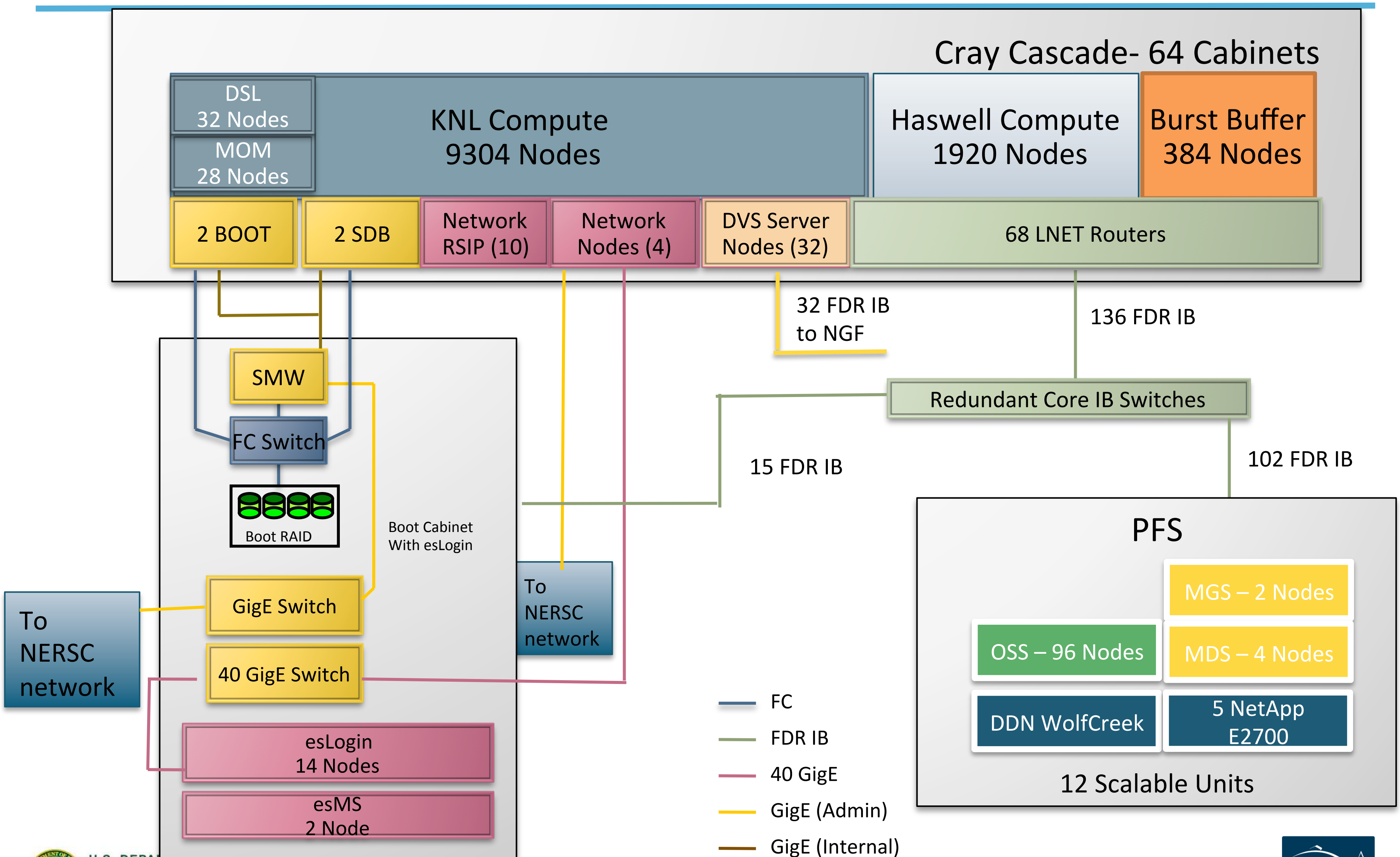
Burst Buffer Software NRE Efforts



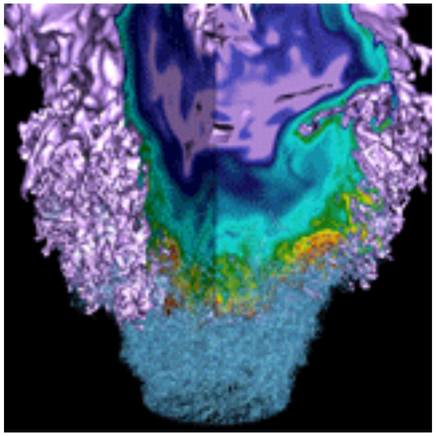
Create Software to enhance usability and to meet the needs of all NERSC users

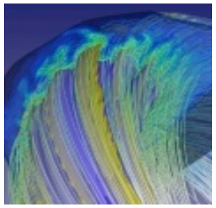
- Scheduler enhancements
  - Automatic migration of data to/from flash
  - Dedicated provisioning of flash resources
  - Persistent reservations of flash storage
- Enable In-transit analysis
  - Data processing or filtering on the BB nodes – model for exascale
- Caching mode – data transparently captured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

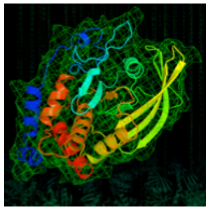
The Cori System

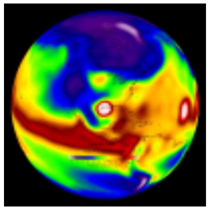


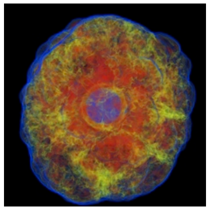
NERSC System Plan

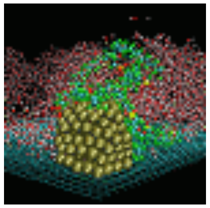












NERSC 40 YEARS at the FOREFRONT 1974-2014

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Major Technology Changes That Will Improve Usability

- **2015-16 NERSC-8/Trinity**
  - High-bandwidth on-package memory
  - "Burst Buffers" – NVRAM enhanced I/O
- **2017-18 CORAL**
  - On-die NIC – lower latency
  - On-node NVRAM
- **2019-20 NERSC-9/ATS-3**
  - P0 exascale processor
  - Emerging Exascale Programming Model
  - Object-based storage
  - Advanced memory technologies
  - Processing Near Memory (processing data where it is located)
  - Advanced power management technology
  - Coherence domains & fine-grained interprocessor communication
- **2021-22 CORAL+1**
  - P1 exascale processor
  - .....

All of these can be enhanced with judicious NRE investments

Holy Grail: Can a single computer system meet the needs of Data and Simulation?

NERSC 40 YEARS at the FOREFRONT 1974-2014

	Compute Intensive Arch	Data Intensive Arch
Compute On-Package DRAM Capacity Memory On-node-Storage In-Rack Storage Interconnect Global Shared Disk Off-System Network	**Goal:** ***Maximum computational density and local bandwidth*** *for given power/cost constraint.* Maximizes bandwidth density near compute	**Goal:** ***Maximum data capacity and global bandwidth*** *for given power/cost constraint.* Bring more storage capacity near compute (or conversely embed more compute into the storage). *Requires software and programming environment support for such a paradigm shift*

Direct from each node

NERSC Upgrades

System attributes	NERSC-6	NERSC-7	*NERSC-8 (proposed)*	*NERSC-9 (Proposed)*
	Hopper	**Edison**		
System peak	1.3 PF	2.6PF	*20-40PF*	*200-300 PF*
Power	2.9 MW (Peak) 2.2MW (Typical)	2.3 MW (Peak) 1.6 MW (Typical)	<5 MW (Peak)	*< 15 MW (peak)*
System memory	0.21 PB	0.35 PB	*1-2 PB*	*~10 PB (128 GB on package, 512-1024 GB DRAM)*
Node performance	202GF	460 GF	*2-3.5TF*	*~10 TF*
Node memory BW	50 GB/s	90 GB/s	*100-500 GB/s*	*~200 GB/s ? 2-4 TB/s on package*
Node concurrency	24 AMD Magnycours cores	24 Intel Ivy Bridge Cores	*up to 300*	*Up to 2048*
System size (nodes)	6,384 nodes	5,576 nodes	8,000-12,000 nodes	O(10,000)
MPI Node Interconnect BW	~3 GB/s	~9GB/s	~9 GB/s	*Up to 50 GB/s*