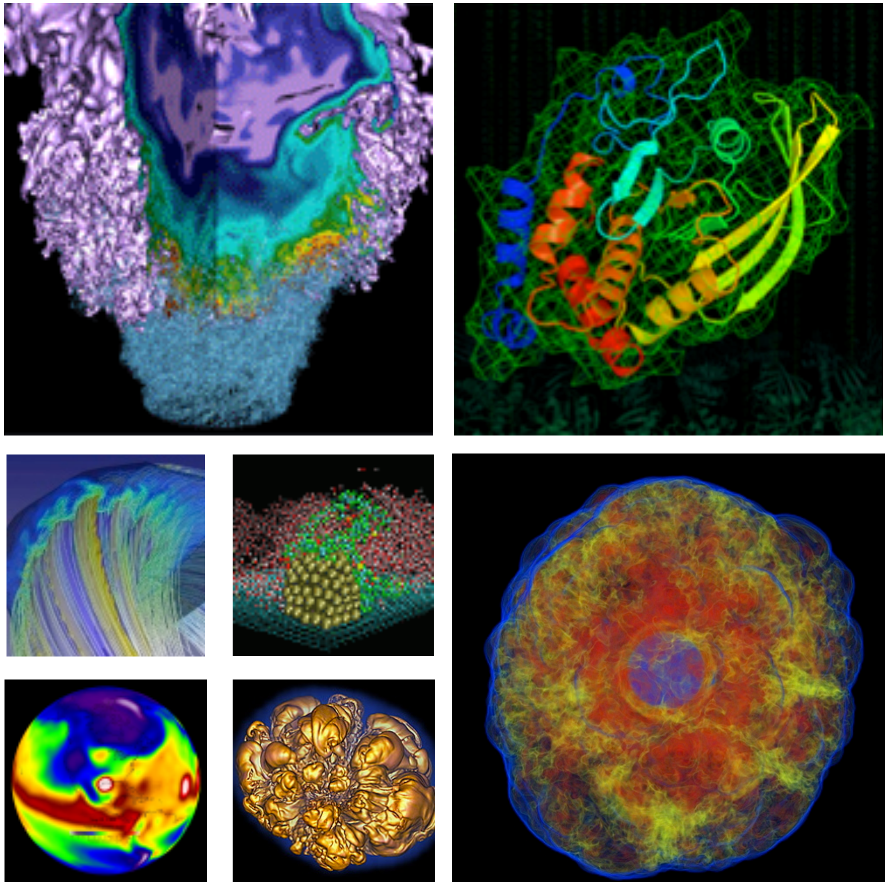
NERSC Today and over the next Ten Years



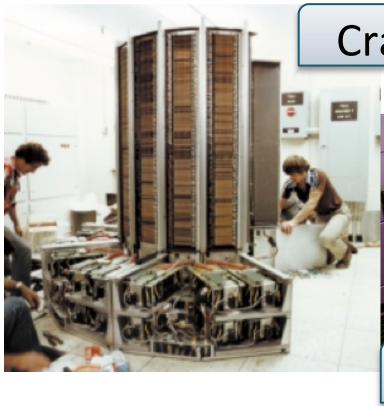
Sudip Dosanjh
Director

February 13, 2013

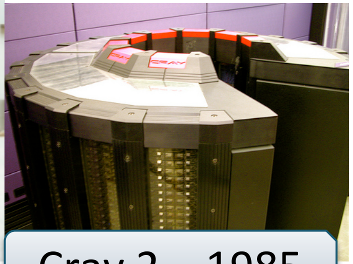
NERSC's Mission

- Accelerate scientific discovery at the DOE Office of Science through high performance computing and extreme data analysis

NERSC History



Cray 1 - 1978



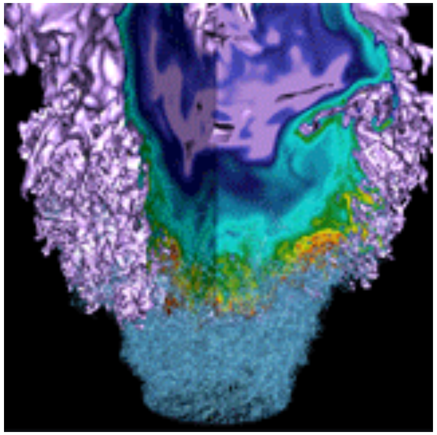
Cray 2 – 1985

Cray T3E Mcurie - 1996

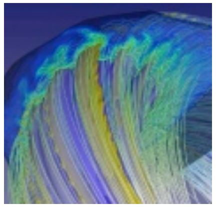
IBM Power3 Seaborg - 2001

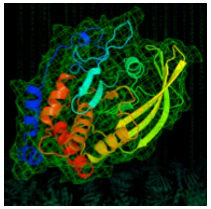
1974	Founded at Livermore to support fusion research with a CDC system
1978	Cray 1 installed
1983	Expanded to support today’s DOE Office of Science
1986	ESnet established at NERSC
1994	Cray T3D MPP testbed
1994 - 2000	Transitioned users from vector processing to MPP
1996	Moved to Berkeley Lab
1996	PDSF data intensive computing system for nuclear and high energy physics
1999	HPSS becomes mass storage platform
2005	Facility wide filesystem
2010	Collaboration with JGI

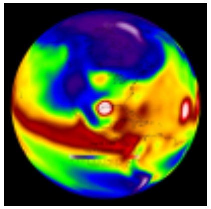
NERSC Today

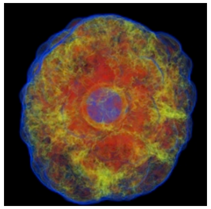


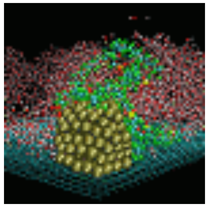
NERSC











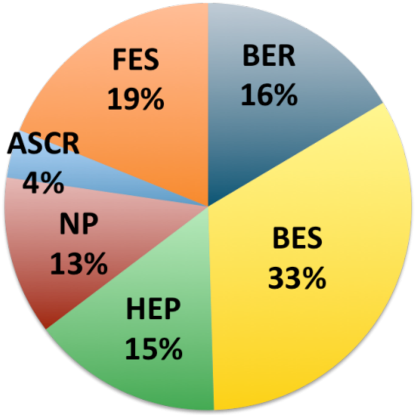
NERSC collaborates with computer companies to deploy advanced HPC and data resources

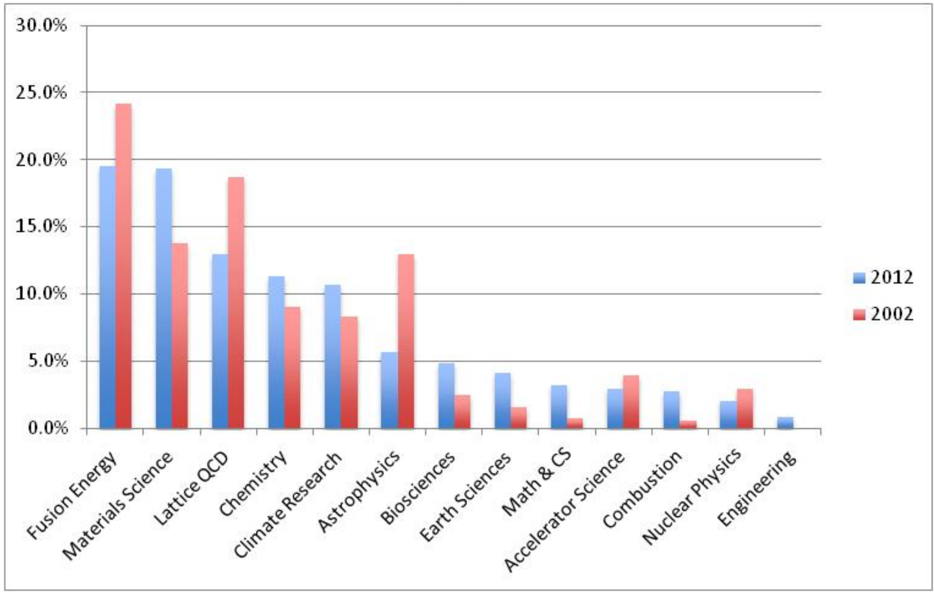
- Hopper (N6) and Cielo (ACES) were the first Cray petascale systems with a Gemini interconnect
- Edison (N7) will be the first Cray petascale system with Intel processors, Aries interconnect and Dragonfly topology (serial #1)
- N8 and Trinity (ACES) are being jointly designed as on-ramps to Exascale
- Architected and deployed data platforms including the largest DOE system focused on genomics
- One of the first facility wide filesystems

We employ experts in high performance computing, computer systems engineering, data, storage and networking

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**





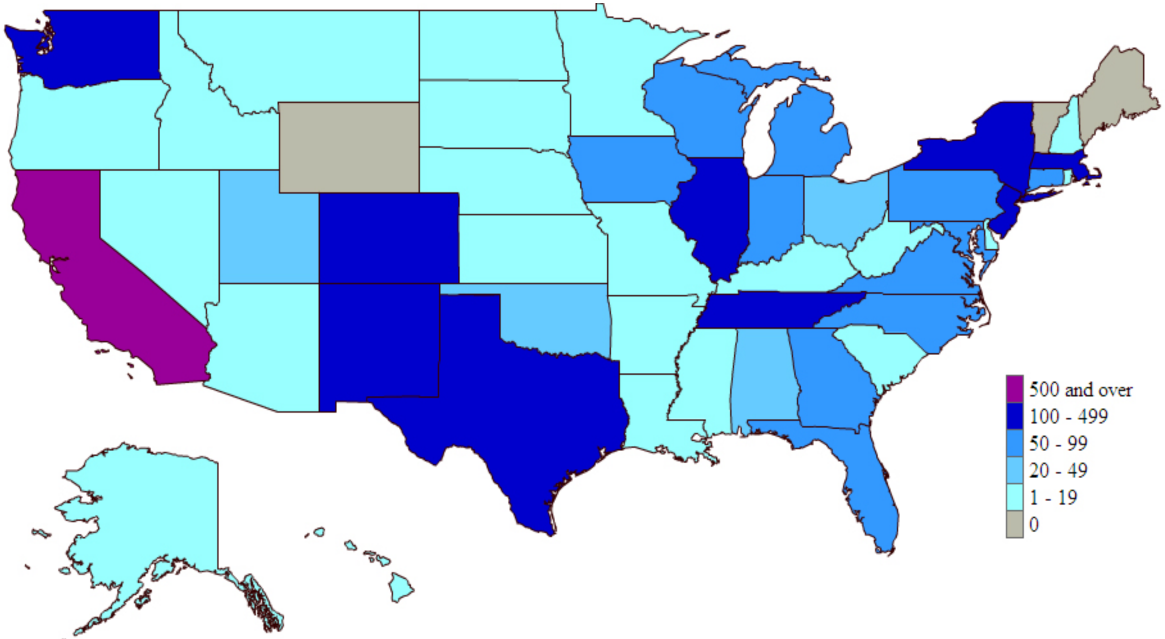
We focus on the scientific impact of our users

- 1500 journal publications per year
- 10 journal cover stories per year on average
- Simulations at NERSC were key to 2 Nobel Prizes (2007 and 2011)
- Supernova 2011fe was caught within hours of its explosion in 2011 and telescopes from around the world were redirected to it the same night
- Data resources and services at NERSC played important roles in two of Science Magazine's Top Ten Breakthroughs of 2012 — the discovery of the Higgs boson and the measurement of the Θ_{13} neutrino weak mixing angle
- MIT researchers developed a new approach for desalinating sea water using sheets of graphene, a one-atom-thick form of the element carbon. Smithsonian Magazine's fifth "Surprising Scientific Milestone of 2012."
- Four of Science Magazine's insights of the last decade (3 in genomics, 1 related to cosmic microwave background)

13 Journal Covers in 2012

We support a broad user base

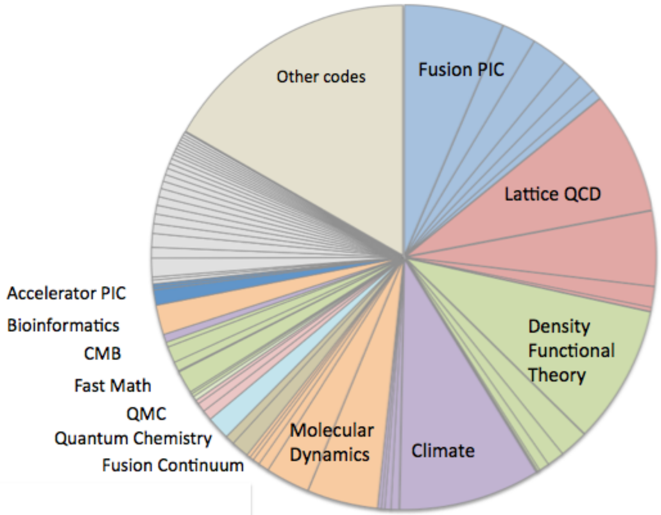
- 4500 users and we typically add 350 per year
- Geographically distributed: 47 states as well as multinational projects



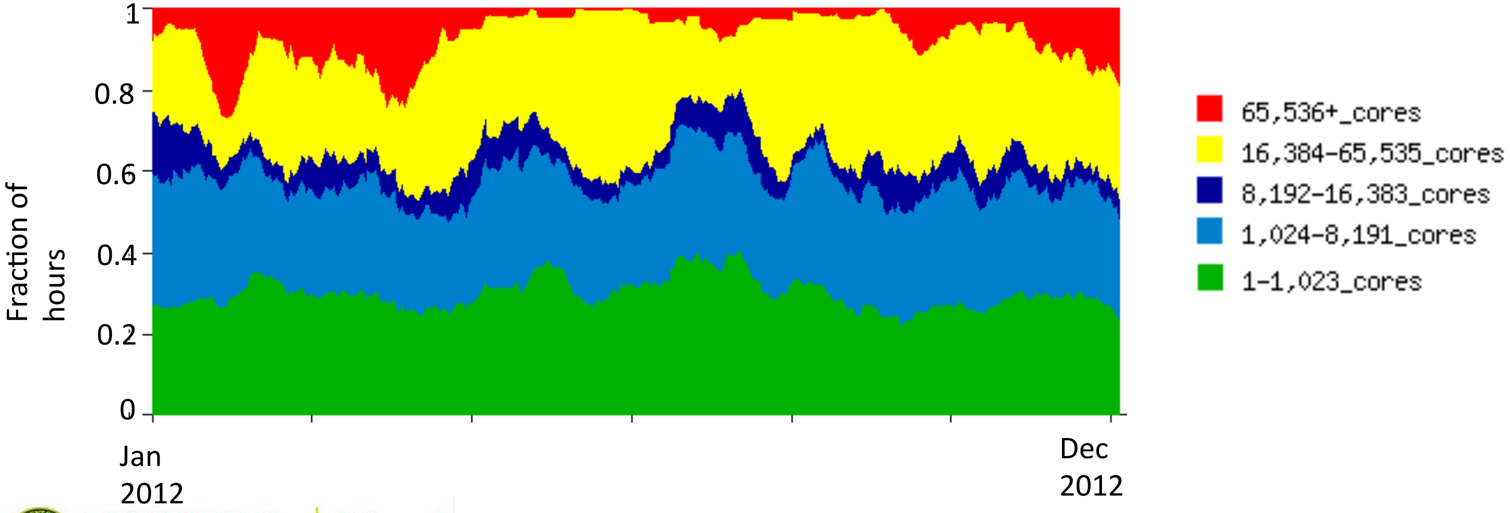
We support a diverse workload

- Many codes (600+) and algorithms
- Computing at Scale and at High Volume

Top Codes by Algorithm

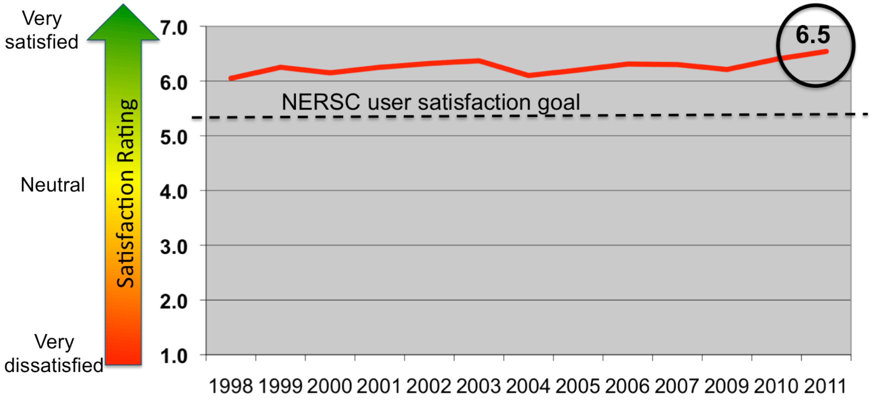


2012 Job Size Breakdown on Hopper

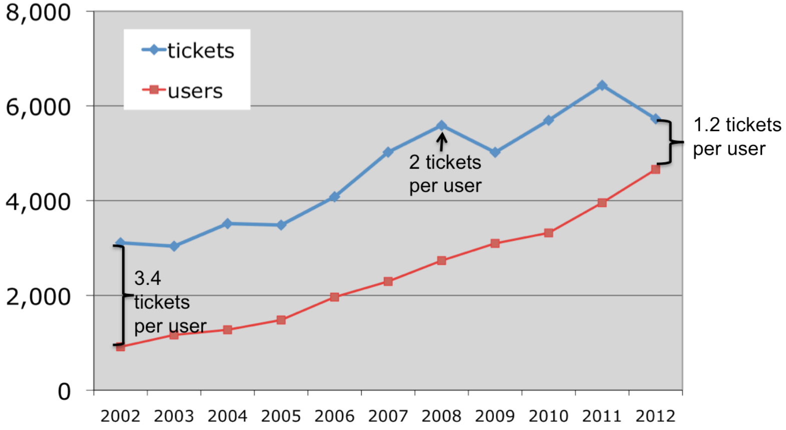


Our operational priority is providing highly available HPC resources backed by exceptional user support

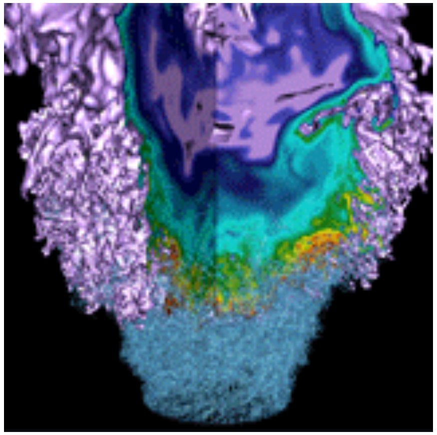
- **We maintain a very high availability of resources (>90%)**
  - One large HPC system is available at all times to run large-scale simulations and solve high throughput problems
- **Our goal is to maximize the productivity of our users**
  - One-on-one consulting
  - Training (e.g., webinars)
  - Extensive use of web pages
  - We solve or have a path to solve 80% of user tickets within 3 business days

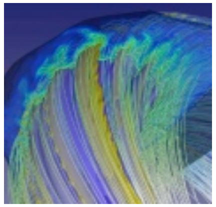


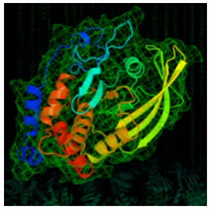
Number of NERSC Users and User Tickets Created per Year

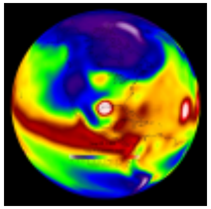


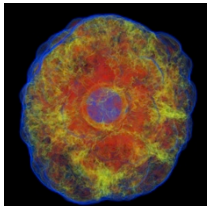
Future Needs and Challenges

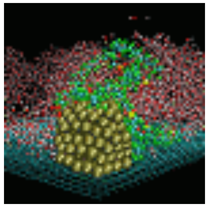






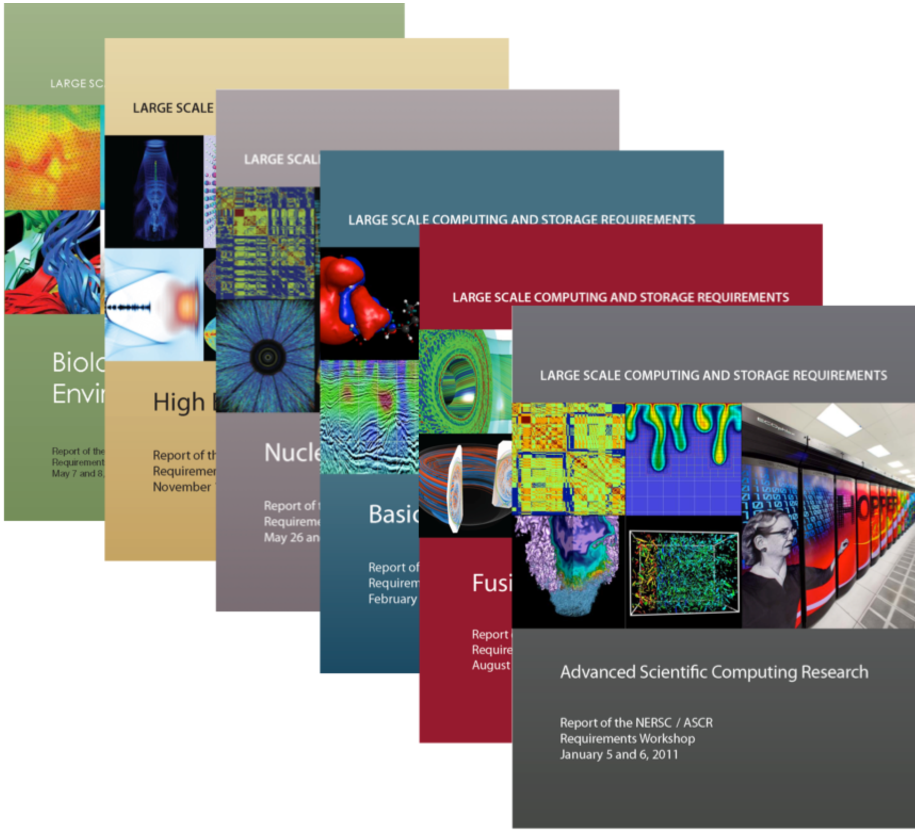






Requirements with 6 program offices

- Reviews with 6 program offices every 3 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 6 offices
- Validate against information from in-depth collaborations, NERSC User Group meetings, user surveys



LARGE SCALE COMPUTING AND STORAGE REQUIREMENTS

Advanced Scientific Computing Research

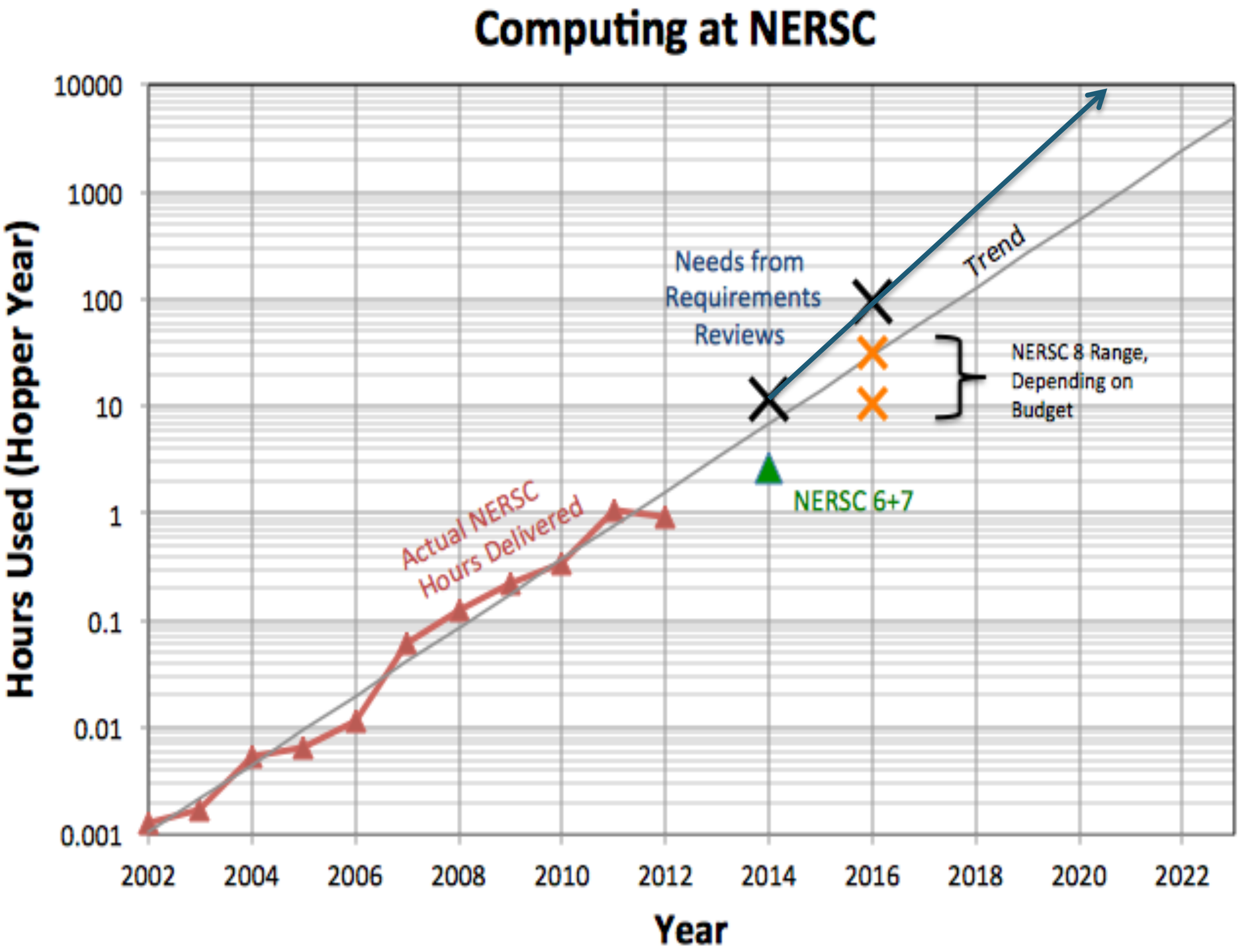
Report of the NERSC / ASCR
Requirements Workshop
January 5 and 6, 2011

LARGE SC
Biolo
Envir
LARGE SCALE
High
LARGE SCAL
Nucle
LARGE SCALE COMPUTING AND STORAGE REQUIREMENTS
Basic
LARGE SCALE COMPUTING AND STORAGE REQUIREMENTS
Fusi
HOPPER

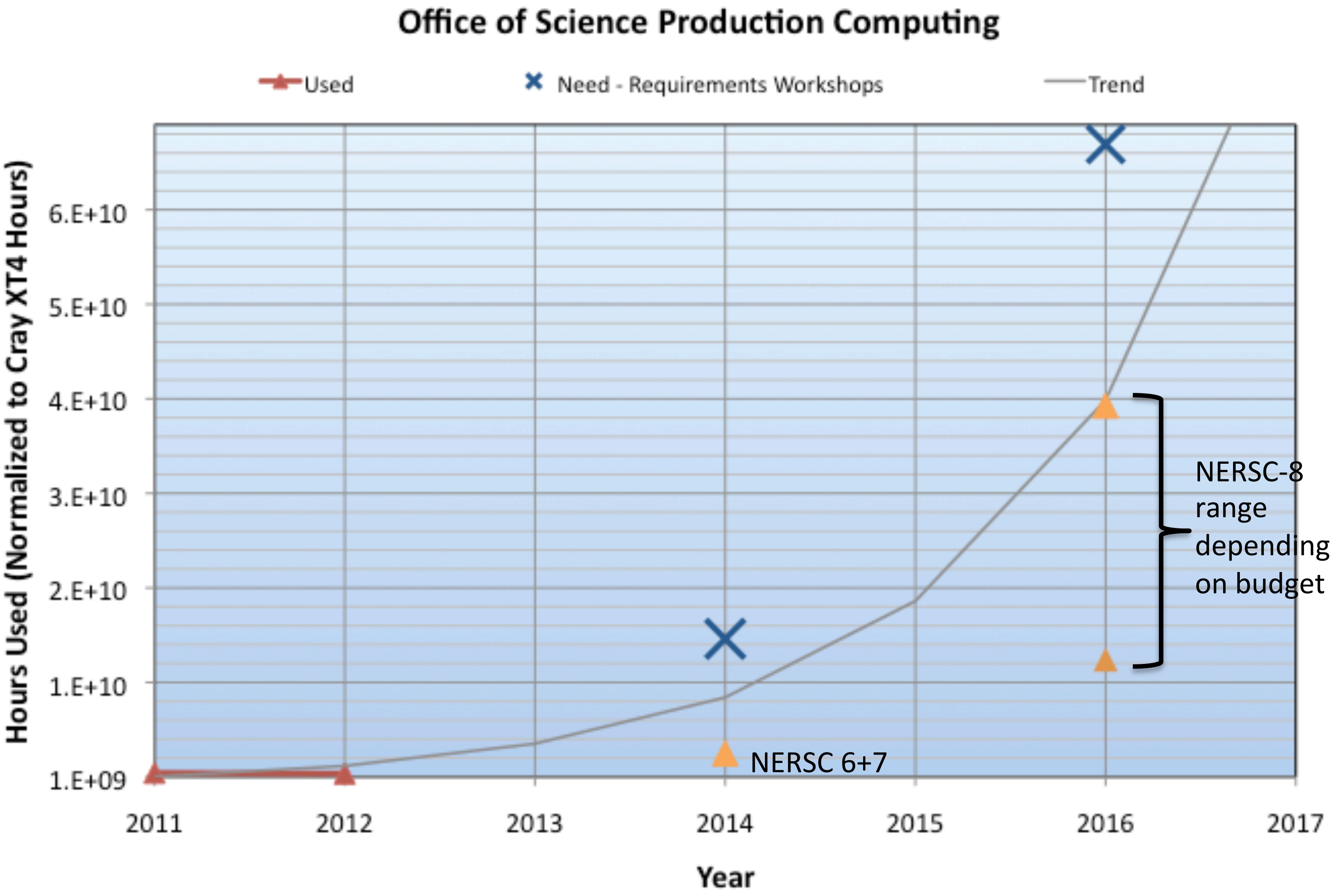
Tends to underestimate need because we are missing future users

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

Keeping up with user needs will be a challenge



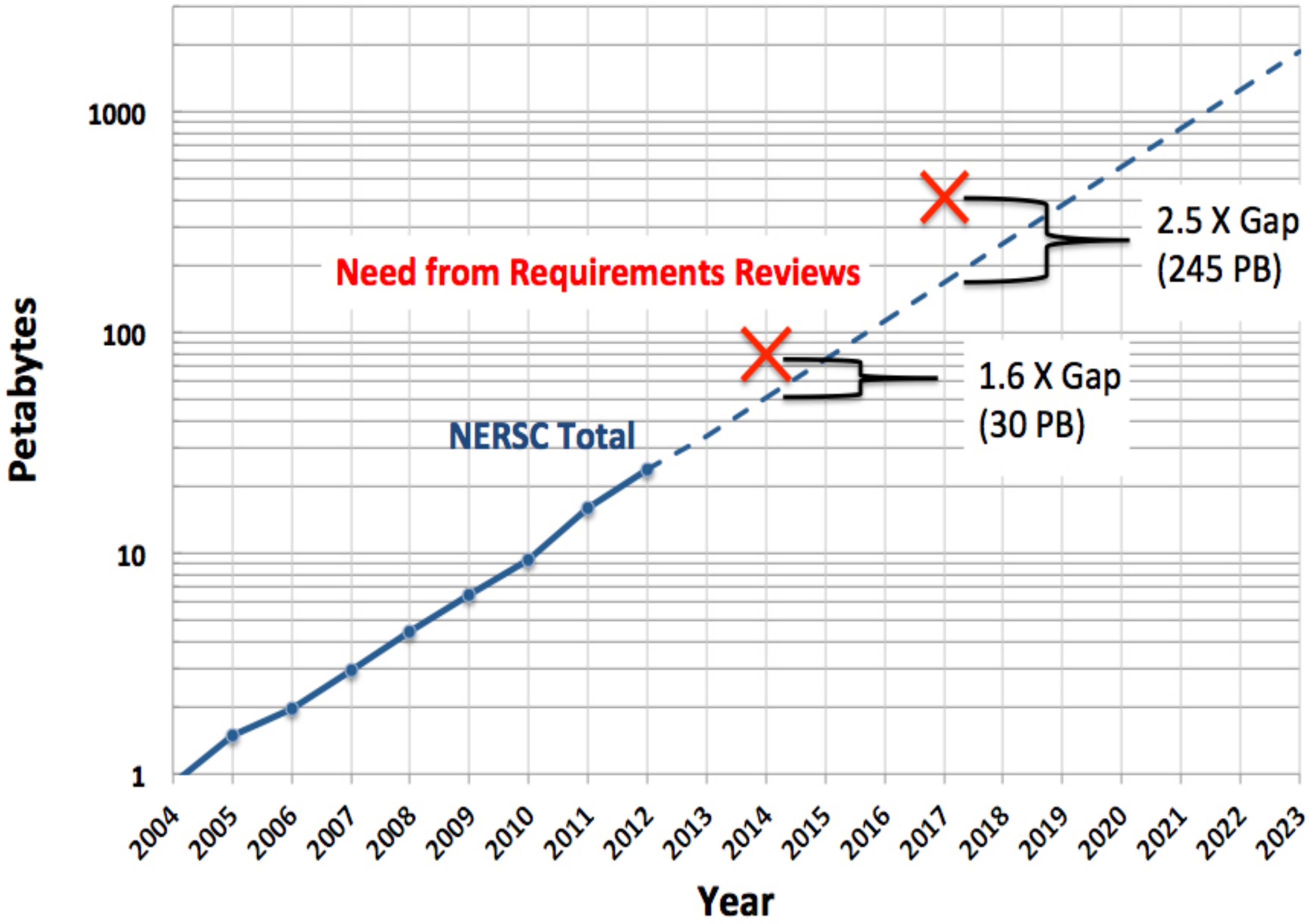
Keeping up with user needs will be a challenge (cont.)



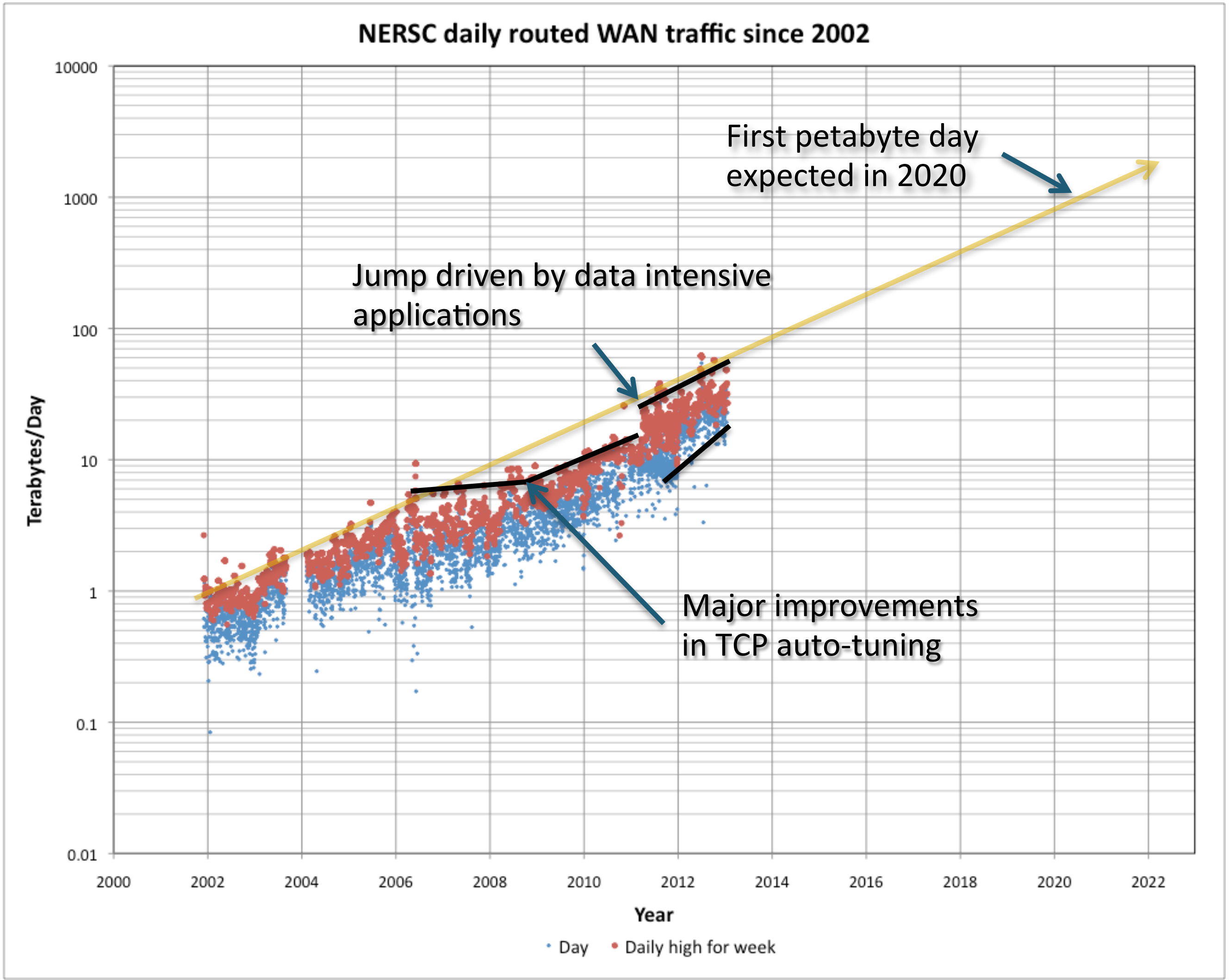
NERSC 6+7

NERSC-8 range depending on budget

Future archival storage needs



Exponentially increasing data traffic

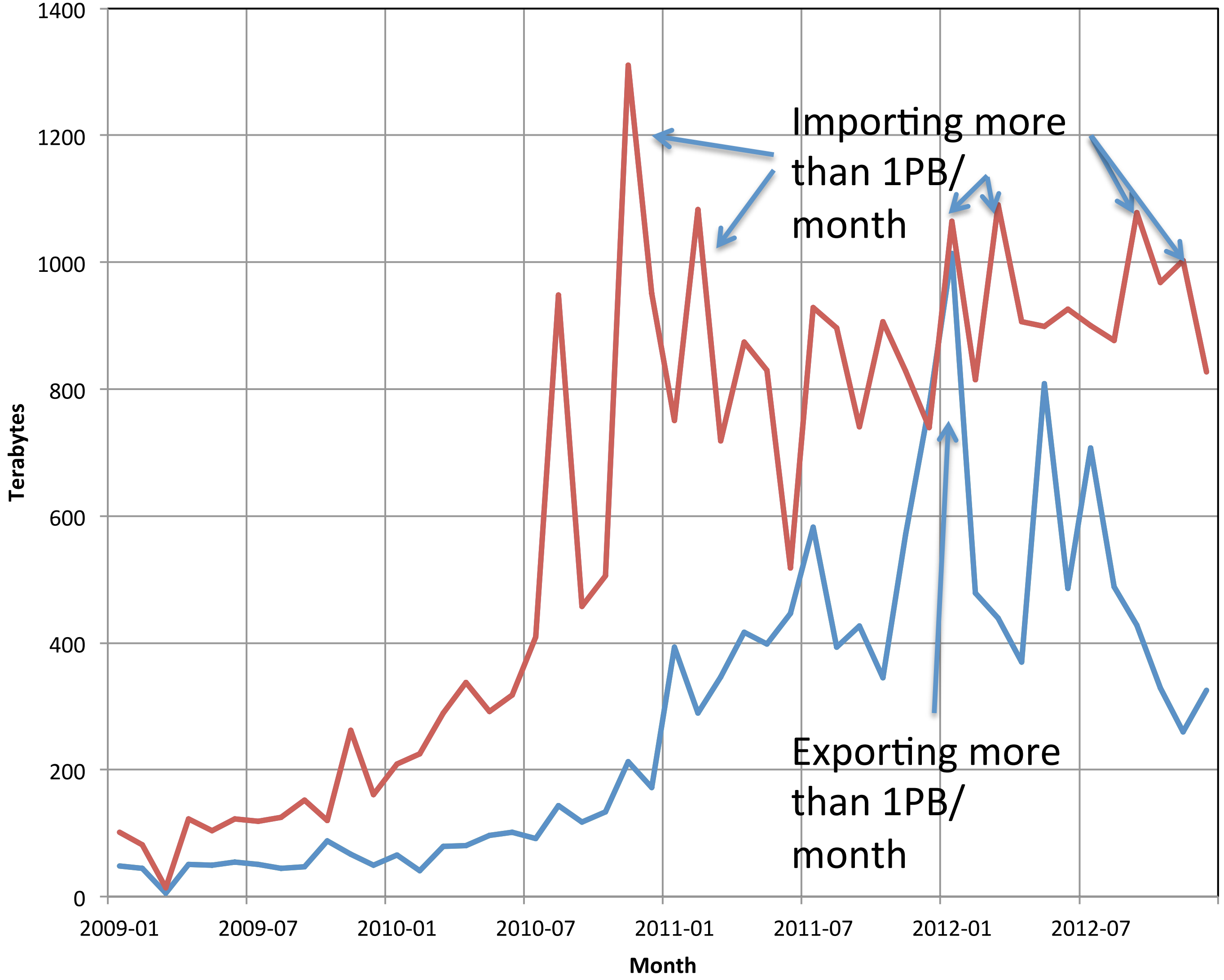


First petabyte day expected in 2020

Jump driven by data intensive applications

Major improvements in TCP auto-tuning

NERSC users import more data than they export!



Importing more than 1PB/ month

Exporting more than 1PB/ month

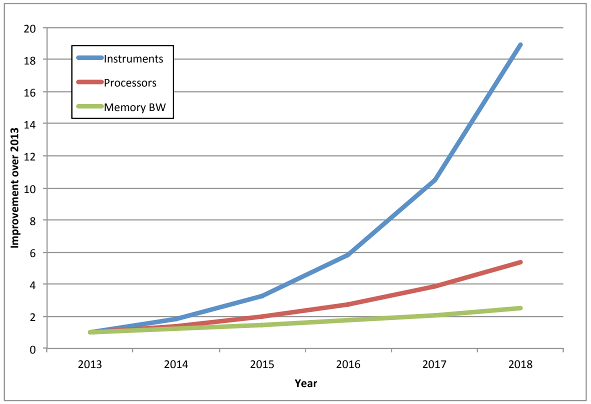
Increased data emphasis in requirements reviews

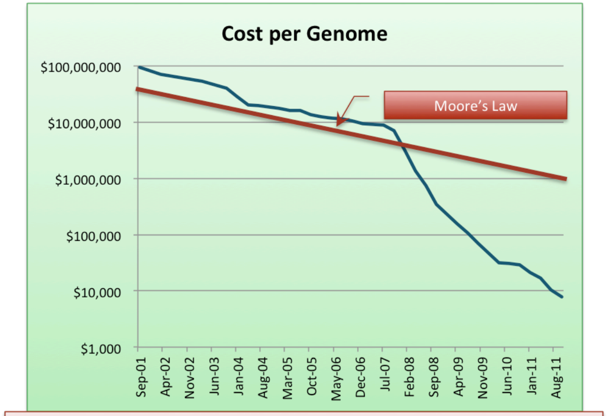
- **BER (2017 draft): "***Access to more computational and storage resources … and the ability to access, read, and write data at a rate far beyond that available today***"**
- **HEP (2017 pre-draft):** *"Need for more computing cycles and fast-access storage; support for data-intensive science, including*
  - *Improvements to archival storage*
  - *Analytics (parallel, DBs, services, gateways etc.)*
  - *Sharing, curation, provenance of data*
- **ASCR (2014): "***Applications will need to be able to read, write, and store 100s of terabytes of data for each simulation run. Many petabytes of long-term storage will be required to store and share data with the scientific community***."**
- **BES (2014): "***[There is a need to support] … huge volumes of data from the ramp-up of the SLAC LINAC Coherent Light Source (LCLS) [and other experimental facilities in BES]."*
- **FES (2014): "***[Researchers need] data storage systems that can support high-volume/high-throughput I/O.***"**
- **NP (2014):** *Needs include*
  - *"Useable methods for cross-correlating across large databases …"*
  - *"[…] grid infrastructure, including the Open Science Grid (OSG) interface […]. "*
  - *[…] The increased capacity afforded by GPUs has resulted in […] a significant increase in IO demands in both intermediate and long term storage. "*

DOE experimental facilities are also facing extreme data challenges

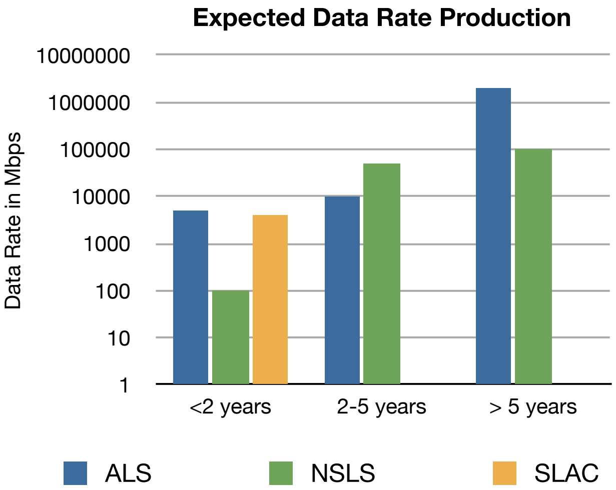
- **The observational dataset for the Large Synoptic Survey Telescope will be ~100 PB**
- **The Daya Bay project will require simulations which will use over 128 PB of aggregate memory**

- **By 2017 ATLAS/CMS will have generated 190 PB**
- **Light Source Data Projections:**
  - 2009: 65 TB/yr
  - 2011: 312 TB/yr
  - 2013: 1.9 PB /yr
  - EB in 2021?
  - NGLS is expected to generate data at a Terabit per second





Source: National Human Genome Research Institute

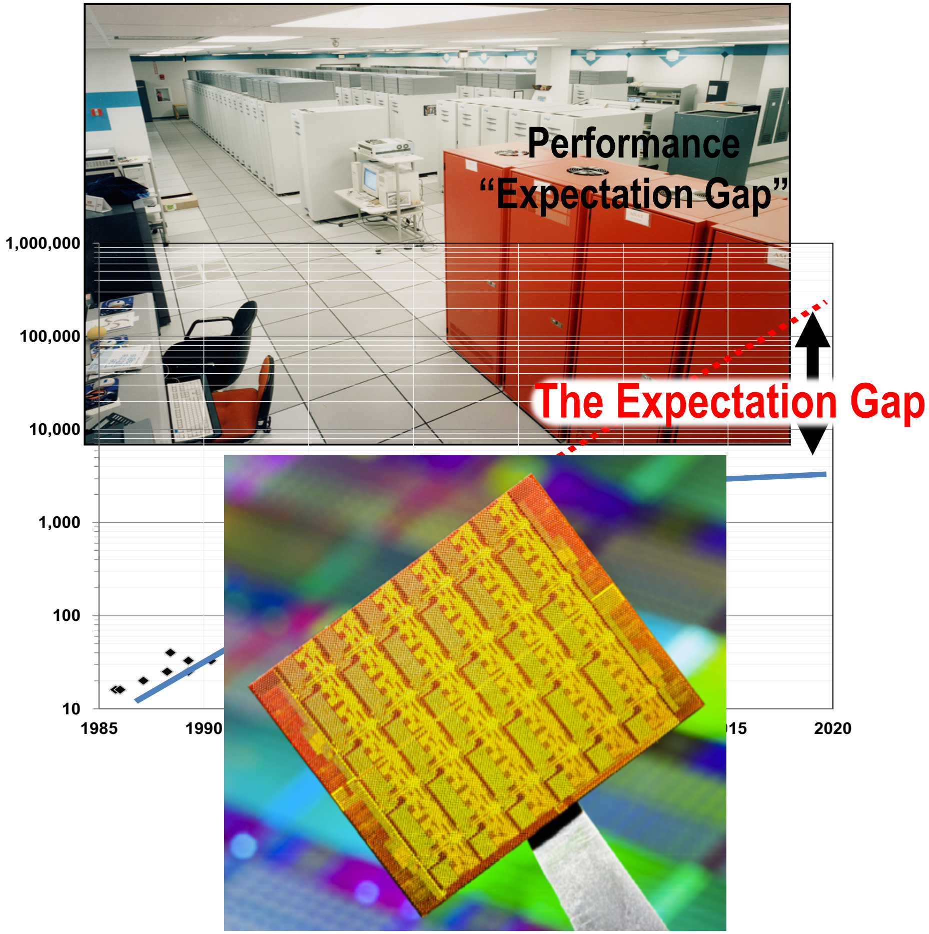


Computer industry roadmaps will not meet DOE mission needs

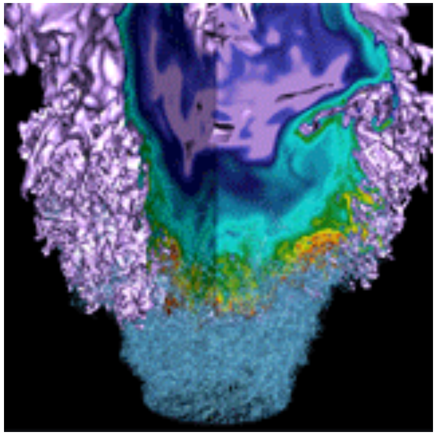
- Technology disruption is underway at the processor and memory level. Computing challenges include:
  - Energy efficiency
  - Concurrency
  - Data movement
  - Programmability
  - Resilience

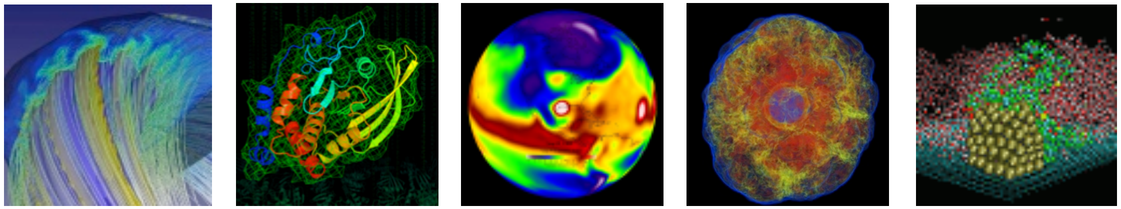
These will impact all scales of computing

- We can only meet these challenges through both hardware and software innovation
  - Rewrite application codes
  - Influence computer industry



NERSC Strategy



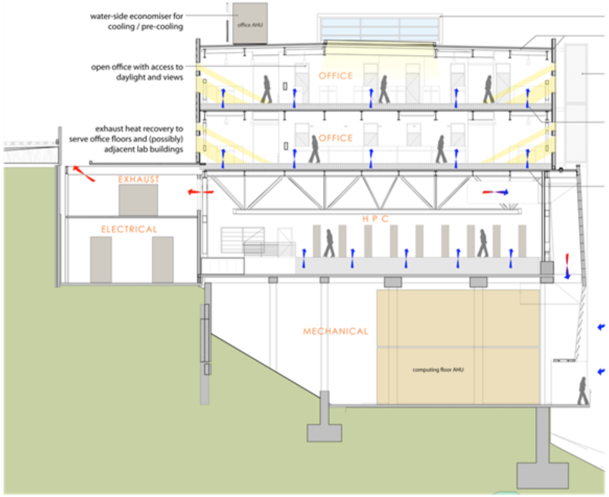


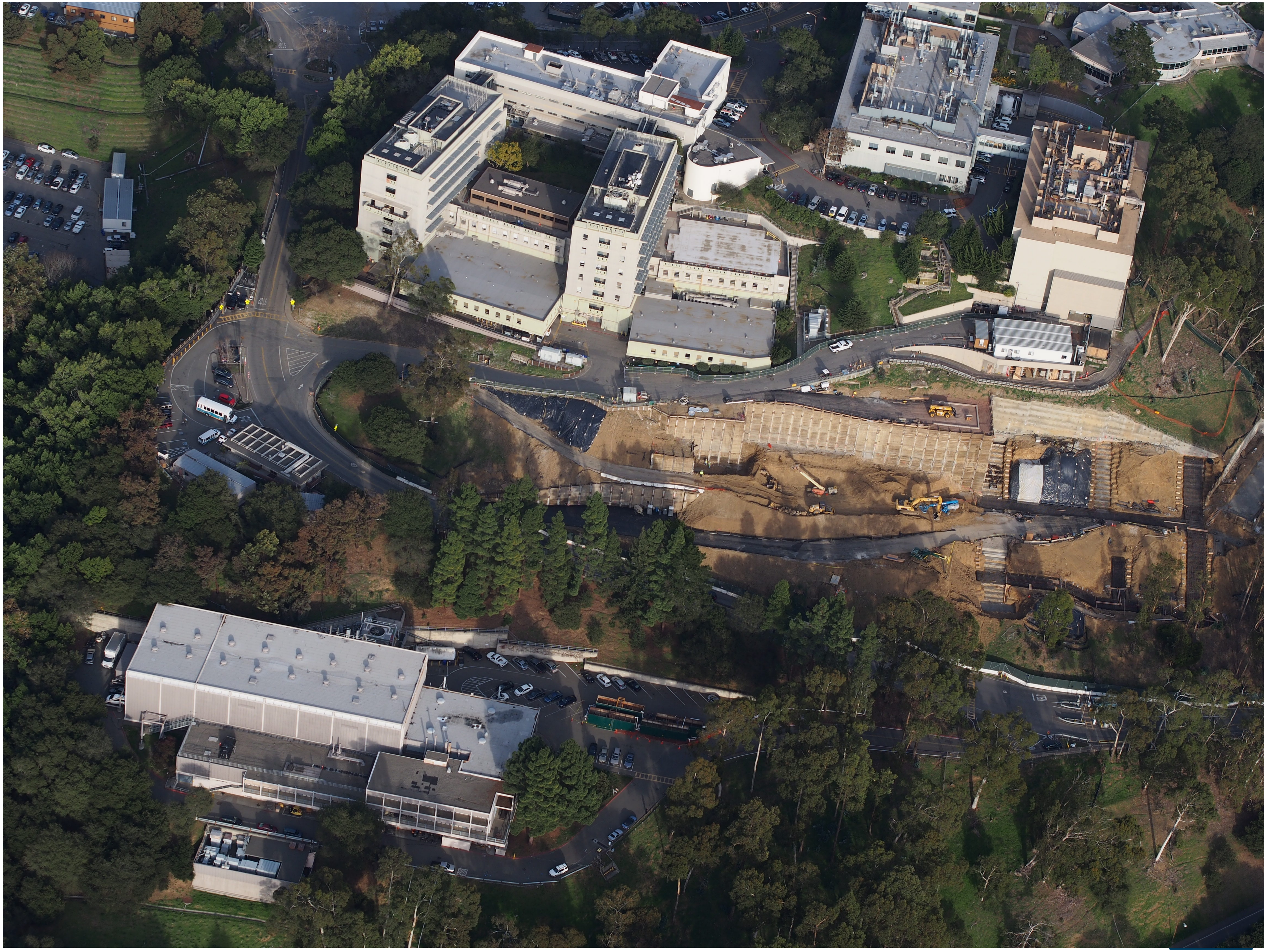
Strategic Objectives

- **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 many core architectures
  - influencing the computer industry to ensure that future systems meet the mission needs of SC
- **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**

We are deploying the CRT facility to meet the ever growing computing and data needs of our users

- **Four story, 140,000 GSF**
  - Two 20Ksf office floors, 300 offices
  - 20K -> 29Ksf HPC floor
  - Mechanical floor
- **42MW to building**
  - 12.5MW initially provisioned
  - WAPA power: Green hydro
- **Energy efficient**
  - Year-round free air and water cooling
  - PUE < 1.1
  - LEED Gold
- **Occupancy Early 2015**


water-side economiser for cooling / pre-cooling
office AHU
open office with access to daylight and views
OFFICE
exhaust heat recovery to serve office floors and (possibly) adjacent lab buildings
OFFICE
EXHAUST
ELECTRICAL
HPC
MECHANICAL
computing floor AHU



Providing usable Exascale computing and storage systems

- **We made NERSC-7 a x86-based system because our broad user base wasn't ready in 2013 for GPUs, accelerators or greatly increased threading**
- **We will deploy pre-Exascale systems in 2015 (NERSC-8) and 2019 (NERSC-9), and an Exascale system in 2023. Our strategy is:**
  - Open competition for best solutions
  - Focus on the performance of a broad range of applications, not synthetic benchmarks
  - General-purpose architectures are needed in order to support a wide range of applications, both large-scale simulations and high volumes of smaller simulations
  - Earlier procurements to influence designs
  - Leverage Fast Forward and Design Forward
  - Engage co-design efforts
  - Transition users to a new programming model

NEW

Programming Models Strategy

- **Our near-term strategy is**
  - Smooth progression to exascale from a user's point of view
  - Support for legacy code, albeit at less than optimal performance
  - Reasonable performance with MPI+OpenMP
  - Support for a variety of programming models
  - Support optimized libraries
- **Longer term, Berkeley Lab is willing to lead a multinational effort to converge on the next programming model**
  - Leverage research efforts (XStack, XTune, DEGAS) for advanced programming constructs
  - Assess existing models through participation in standards bodies (OMP, MPI, Fortran, C/C++) and assess emerging languages
  - Engage co-design community of vendors & HPC for cross-cutting solutions
  - Share results and build consensus

Strategy for transitioning the SC Workload to Energy Efficient Architectures

- We will deploy testbeds to gain experience with new technologies and to better understand emerging programming models and potential tradeoffs.
- We will have in-depth collaborations with selected users and application teams to begin transitioning their codes to our testbeds and to NERSC-8
- We will develop training and online resources to help the rest of our users based on our in-depth collaborations, as well as on results from co-design centers and ASCR research
- We will add consultants with an algorithms background who can help users when they have questions about improving the performance of key code kernels

It is important to note that all users will be impacted by technology changes because the disruption is at the processor and memory level.

Strategy for ensuring that future systems meet SC mission requirements

- Partner with Los Alamos and Sandia on procurements in 2015 and 2019. The larger size of these procurements will give us greater leverage with industry
- Provide industry with greater information on NERSC's workload through new and innovative instrumentation, measurement, and analysis
- Actively engage with industry through DOE's Fast Forward and Design Forward programs
- Leverage the Berkeley/Sandia Computer Architecture Laboratory (CAL) that has been established by ASCR
- Serve as a conduit for information flow between computer companies and our user community

Extreme Data Strategy

- Partner with DOE experimental facilities to identify requirements and create early success
- Develop and deploy new data resources and capabilities
- Provide new classes of HPC expertise required for data-intensive workloads
- Leverage ESnet and ASCR research to create end-to-end solutions

Unique data-centric resources will be needed

Compute

On-Package DRAM

Capacity Memory

On-node-Storage

In-Rack Storage

Interconnect

Global Shared Disk

Off-System Network

Compute Intensive Arch

Goal: *Maximum Computational Density and local bandwidth for given power/cost constraint.*

Maximizes bandwidth density near compute

Data Intensive Arch

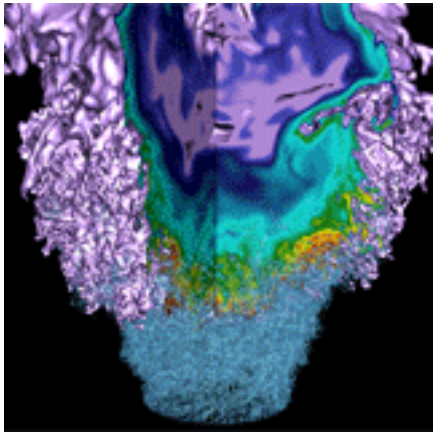
Goal: *Maximum Data Capacity and global bandwidth for given power/cost constraint.*

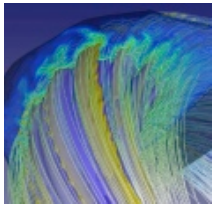
Bring more storage capacity near compute (or conversely embed more compute into the storage).

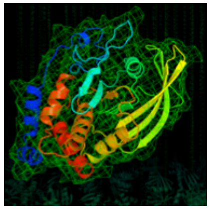
Requires software and programming environment support for such a paradigm shift

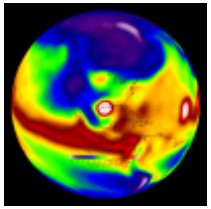
Direct from each node

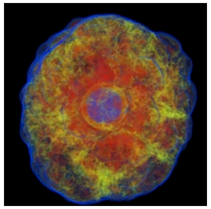
NERSC System Plan

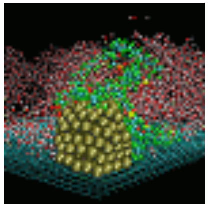




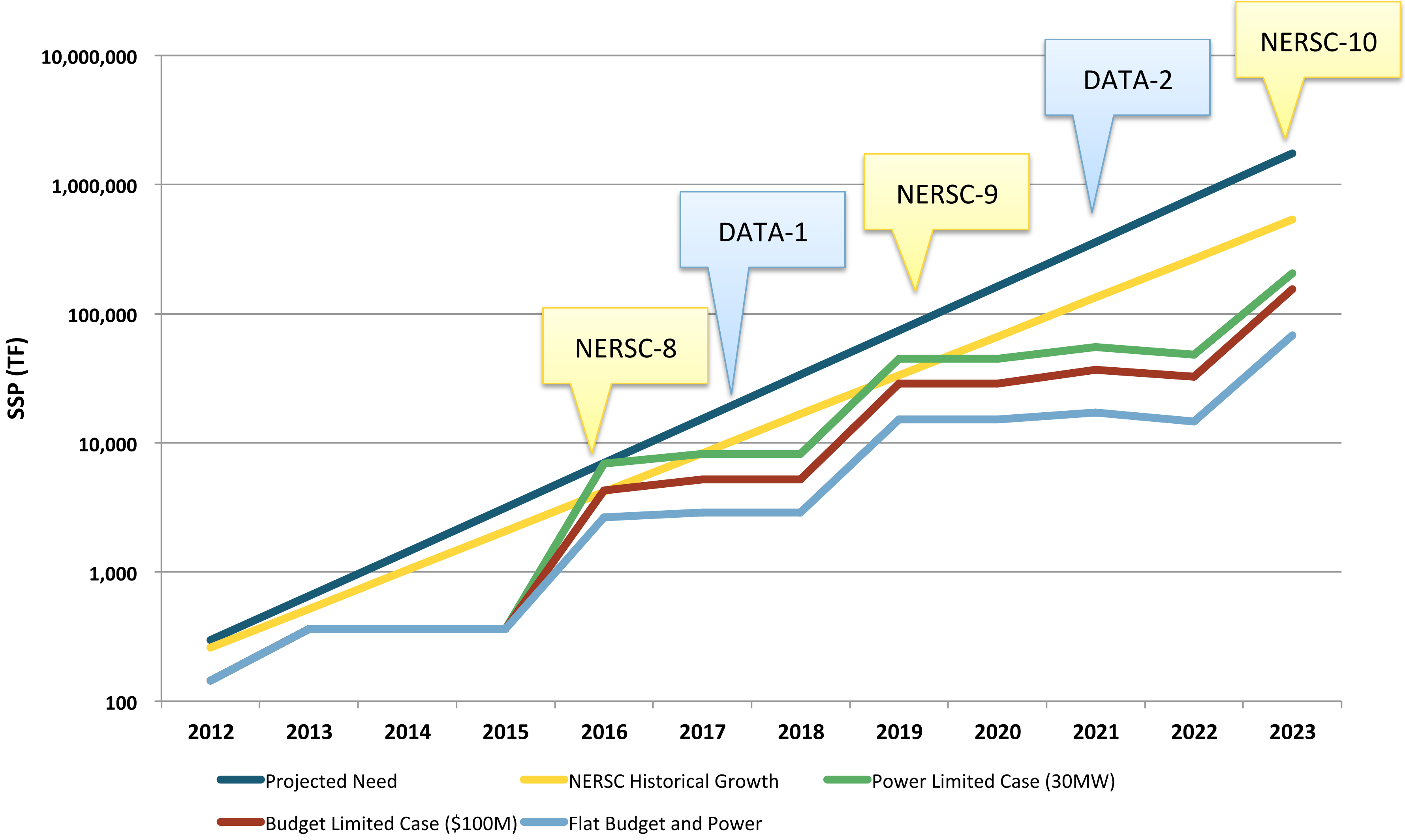








Projections of Installed Capacity

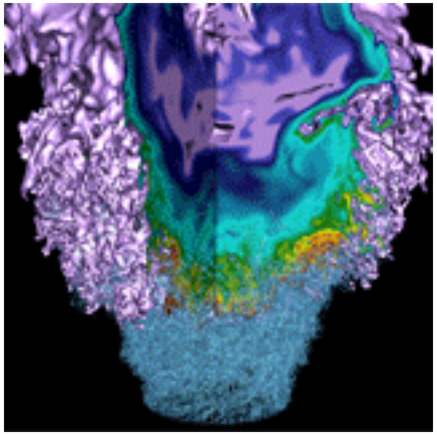


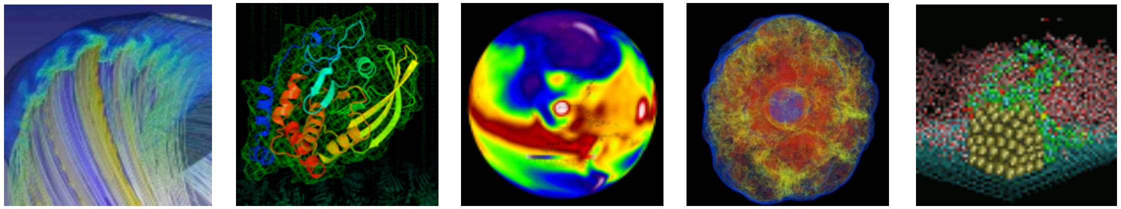
Conclusions

- **NERSC has a strategy and a plan for meeting the ever growing computing and storage needs of our users**
  - We need to overcome key exascale challenges

- **We want to enable science teams with the nation's largest data-intensive challenges to rely on NERSC to the same degree they already do for modeling and simulation**

Backup Slides

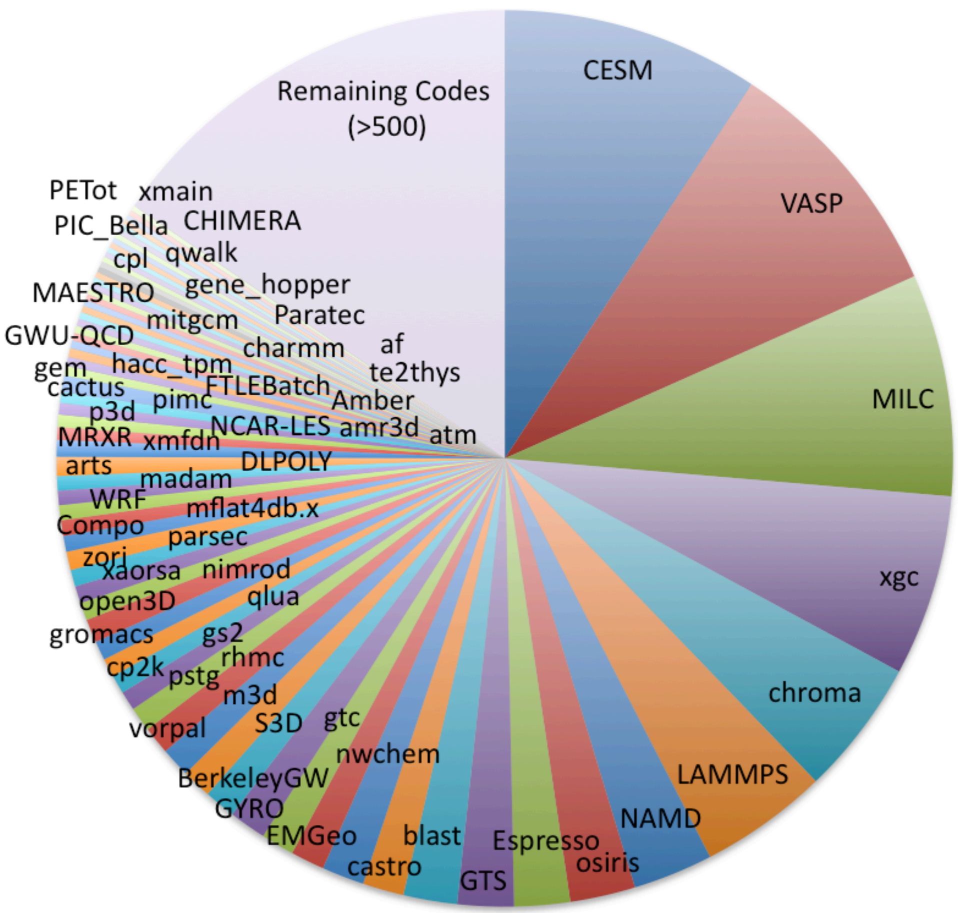




Although NERSC has a broad user base, the workload is highly concentrated and unevenly distributed

Top Application Codes

Jan – Nov 2012



- **10 codes make up 50% of workload**
- **25 codes make up 66% of workload**
- **75 codes make up 85% of workload**
- **remaining codes make up bottom 15% of workload**

Approximately 80% of the workload needs to transfer to NERSC-8
(20% can remain on Edison for the next few years)

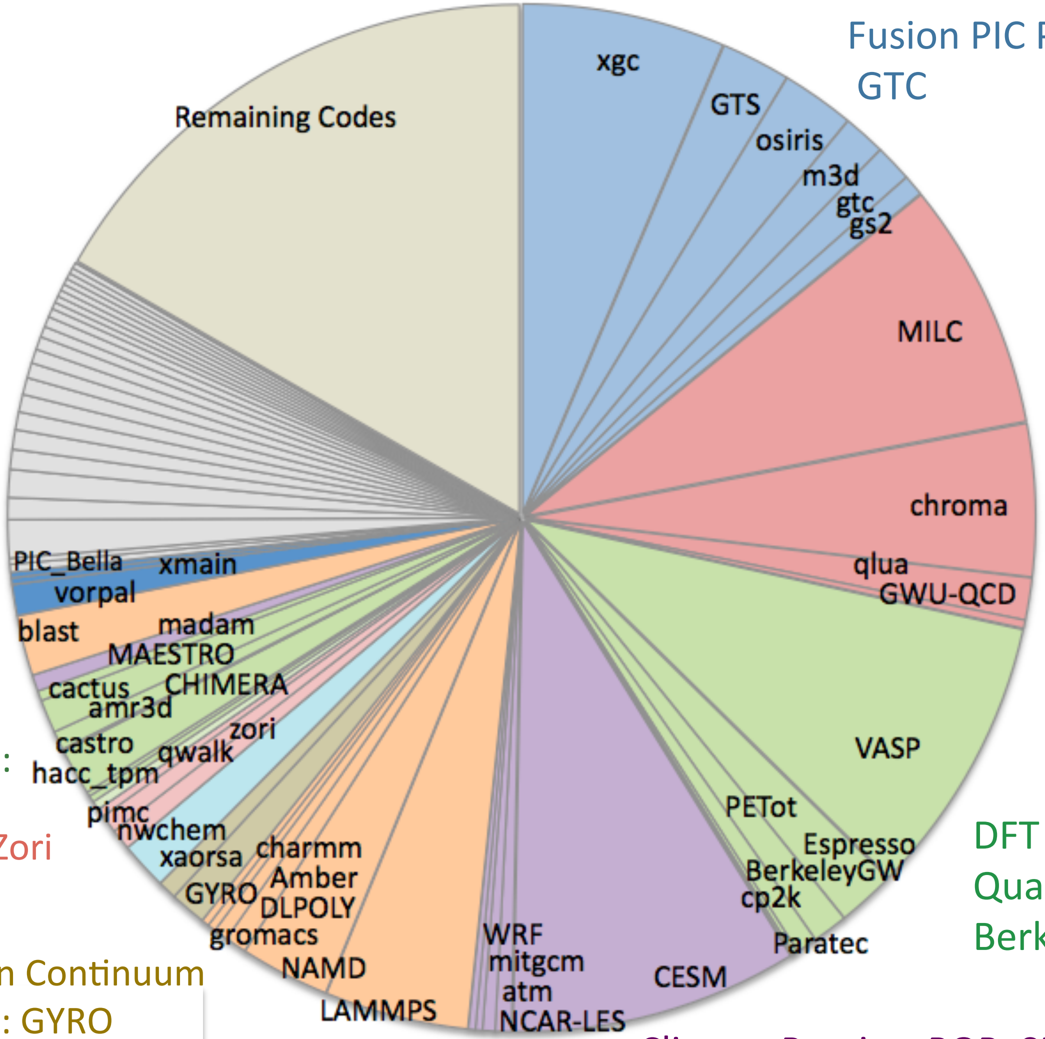
Codes, percent usage, algorithm

Top 30 codes run at NERSC 2012

CESM	9.29%	CLIMATE
VASP	9.05%	PWDFT
MILC	8.01%	LQCD
xgc	6.53%	FUSIONPIC
chroma	4.91%	LQCD
LAMMPS	4.63%	MD
NAMD	2.88%	MD
osiris	2.35%	FUSIONPIC
Espresso	2.01%	PWDFT
GTS	2.03%	FUSIONPIC
blast	1.94%	BIOINFORMATICS
castro	1.49%	FASTMATH
nwchem	1.53%	ABINITIO
EMGeo	1.22%	GEOPhysics
gtc	1.19%	FUSIONPIC
GYRO	1.34%	FUSIONGRID
BerkeleyGW	1.13%	PWDFT
m3d	1.07%	FUSIONPIC
S3D	1.07%	FASTMATH
vorpal	1.00%	ACCELERATORPIC
rhmc	0.89%	QCD - monte carlo
pstg	0.69%	FUSION
cp2k	0.67%	PWDFT
gs2	0.67%	FUSIONPIC
gromacs	0.69%	MD
qlua	0.84%	LQCD
open3D	0.63%	VIS
nimrod	0.62%	MHD
xaorsa	0.60%	FUSIONGRID
zori	0.66%	QMC

The App Readiness team proxies cover almost 75% of the workload

Top Codes by Algorithm and Application Readiness Coverage



Fusion PIC Proxy: GTC

Lattice QCD Proxy: MILC

DFT Proxies: Quantum Espresso Berkeley GW

Climate Proxies: POP, SE-CAM, MPAS, WRF

Molecular Dynamics Proxies: NAMD, Amber

Fusion Continuum Proxy: GYRO

Quantum Chemistry Proxy: NWChem

QMC Proxy: Zori

Fast Math Proxies: FLASH, MAESTRO

CMB Proxy: MADAM

Bioinformatics Proxy: BLAST

Accelerator PIC: Impact

Staffing Gaps For Transitioning Codes

Roles	Effort (FTEs)	Added Capability
Application consultants	3	Assistance with transitioning users to new architectures; exploration of performance issues with new architectures.
Postdoctoral associates	6	New Exascale Postdoc Program. Application-specific expertise with areas to be transitioned to using new architectures.
Systems engineer	1	Deployment of testbed systems; exploration of issues with putting new architectures into production in a reliable and stable fashion.