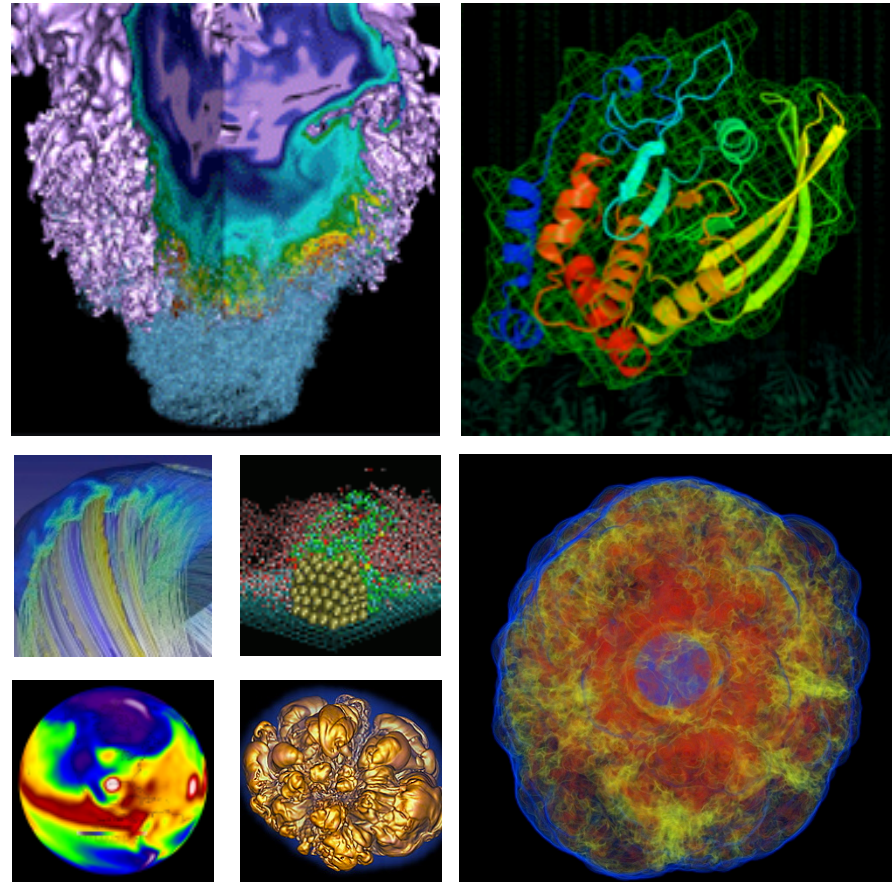
High Performance Computing and Big Data at NERSC



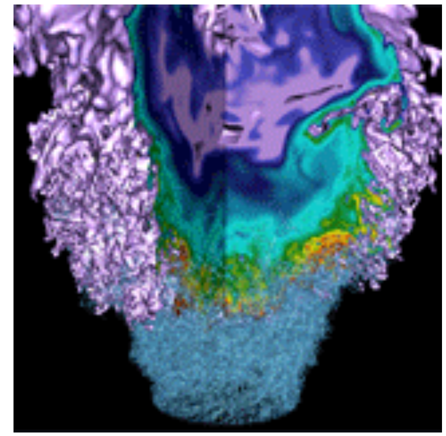
Richard Gerber, Ph.D.
NERSC User Services

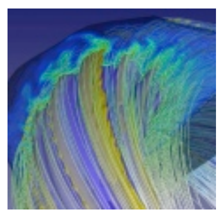
April 3, 2013

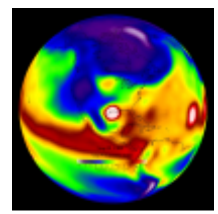
Outline

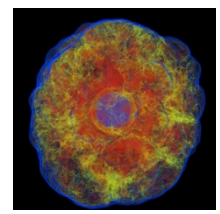
- **What is NERSC?**
- **Who runs at NERSC?**
- **Can you use/do you need High Performance Computing and Data Services?**

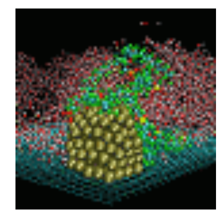
What is NERSC?











NERSC is the High Performance Production Computing Center for the DOE Office of Science

13 Journal Covers in 2012

NERSC's focus is on enabling scientific productivity

- *1,500 refereed publications per year*
- *~10 major journal covers per year*
- *Key contributor to 2 Nobel Prizes (2007 & 2011)*
- *Data services contributed to 2 of Science Magazine's Top 10 breakthroughs of 2012*

Large and varied user community

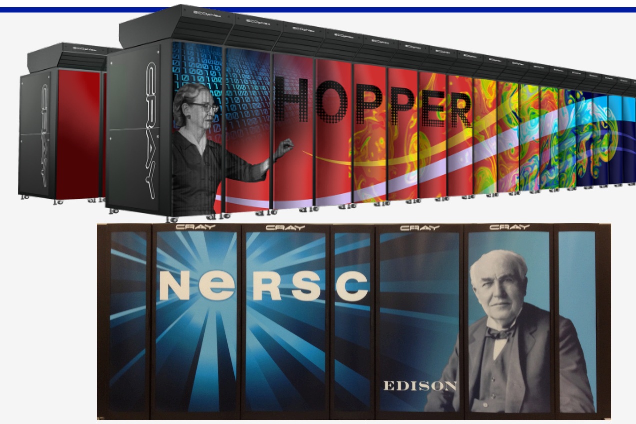
- *5,500 users, 600 projects*
- *From 48 states; 65% from universities*
- *Hundreds of users each day*

Science-driven systems and services

- *World-class computers, storage systems, & networks based on needs of scientific applications*
- *Services designed to optimize science*

Current NERSC Systems

World-Class Supercomputers



Hopper: Cray XE6

- 6,384 compute nodes, 153,216 cores
- 144 Tflop/s on applications; 1.3 Pflop/s peak

Edison: Cray XC30 (Cascade)

- Phase I (10K processors), Phase II in 2013 (~120K)
- Over 200 Tflop/s on applications, 2 Pflop/s peak

Midrange

140 Tflops total

Carver

- IBM iDataplex cluster
- 9884 cores; 106TF

PDSF (HEP/NP)

- ~2K core cluster

GenePool (JGI Genomics)

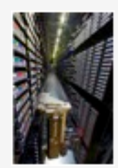
- ~5K core cluster
- 2.1 PB Isilon File System

NERSC Global Filesystem (NGF)

Uses IBM's GPFS

- 8.5 PB capacity
- 15 GB/s of bandwidth

HPSS Archival Storage



- 240 PB capacity
- 5 Tape libraries
- 200 TB disk cache

Analytics & Testbeds

Dirac 48 Fermi GPU nodes

NERSC PDSF

- Funded and used by High Energy Physics, Nuclear Physics
- Networked distributed commodity Linux cluster in continuous operations since 1996
- Detector simulation & data analysis
- Data intensive, high throughput workflows
- Grid Support
  - OSG, WLCG stacks
  - Compute and storage elements for OSG, ALICE
  - Storage elements for ATLAS

PDSF Quick Facts

- 2300 cores
- 1 PB globally accessible disk
- Interconnect 1GigE, 10 GigE, IB
- SGE batch system

PDSF Usage

- **PDSF is an essential resource for a number of groups such as STAR (Tier 1), KamLAND, ATLAS (Tier 3), ALICE (Tier 2), DayaBay (Tier 1), IceCube, CUORE, etc.**
- **Groups such as SNO, SNFactory, CDF, BaBar, LUMI, Planck have used PDSF in the past.**

NERSC Ecosystem

NERSC

Hopper: 1.3PF, 212 TB RAM

Cray XE6, 150,000 Cores

2.2 PB Local Scratch 70 GB/s

16 x QDR IB

Edison: >2PF, 333 TB RAM

Cray XC30

6.4 PB Local Scratch 140 GB/s

16 x FDR IB

Production Clusters
Carver, PDSF, JGI,KBASE,HEP,BES
14x QDR, 2GB/s, 15 TB

Vis & Analytics Data Transfer Nodes
Adv. Arch. Testbeds Science Gateways

Ethernet &
IB Fabric
Science Friendly Security
Production Monitoring
Power Efficiency
WAN

15 GB/s
Global Scratch
1.1 PB
5 x DDN9900

36 GB/s
/project
3.8 PB
4 xDDN9900
NexSAN

5GB/s
/home
250 TB
2 x NetApp 5460

12 GB/s
HPSS
40 PB stored, 240 PB capacity, 30 years of community data

2 x 10 Gb

1 x 100 Gb

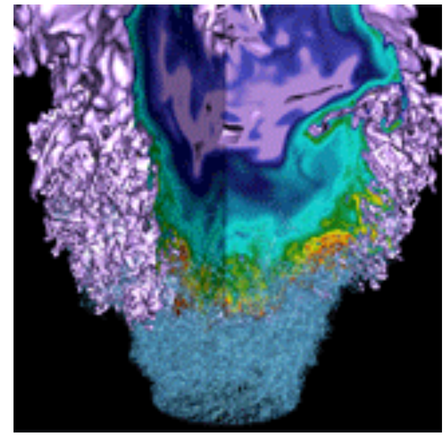
Software Defined Networking

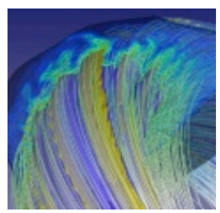
ESnet
Energy Sciences Network

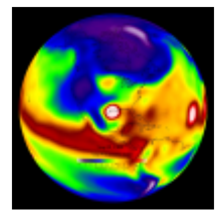
U.S. DEPARTMENT OF ENERGY | Office of Science

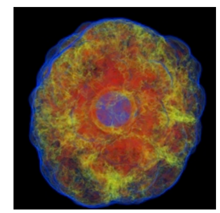
BERKELEY LAB
Lawrence Berkeley National Laboratory

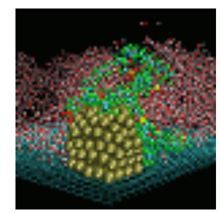
Computing & Storage History and Trends



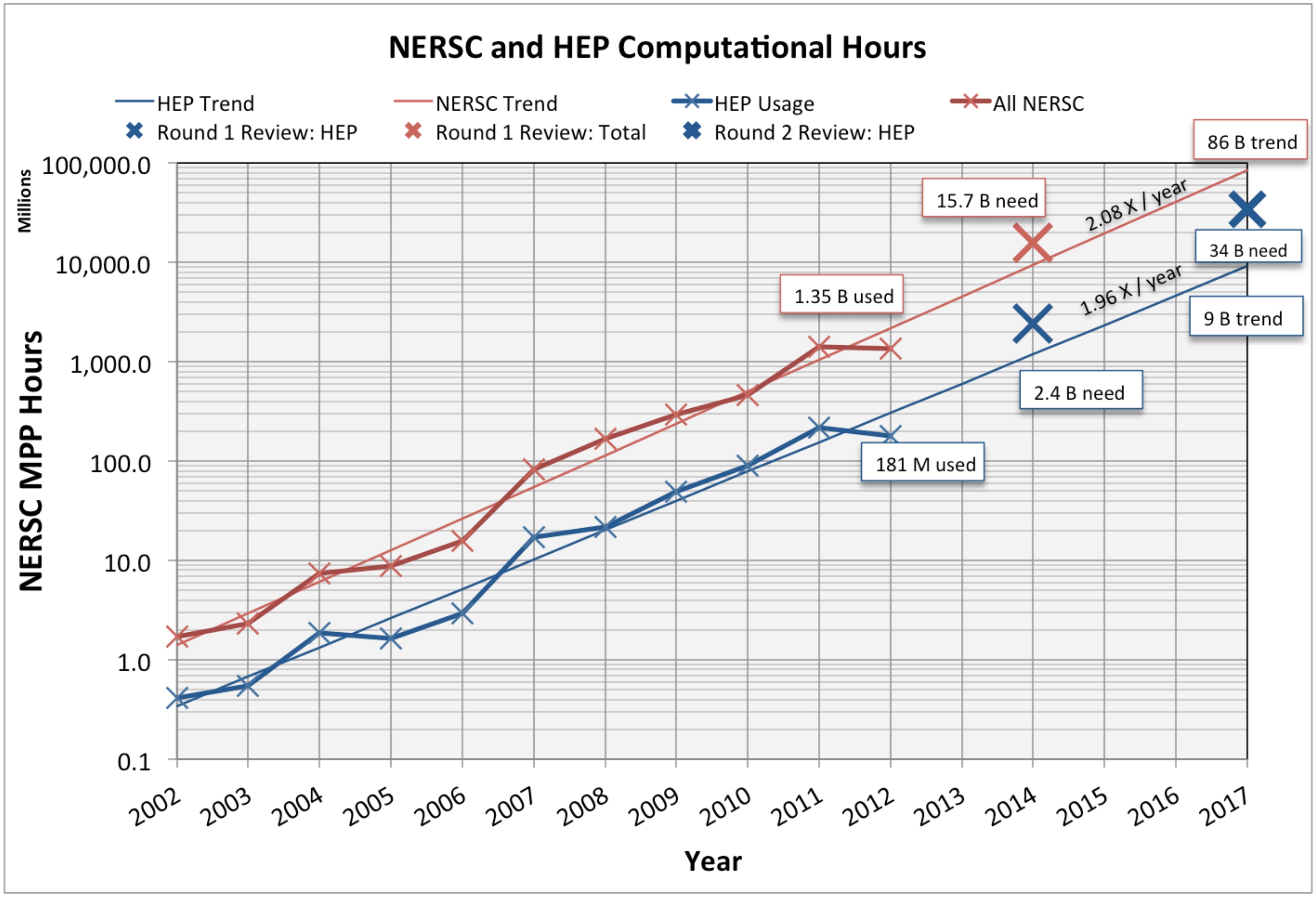




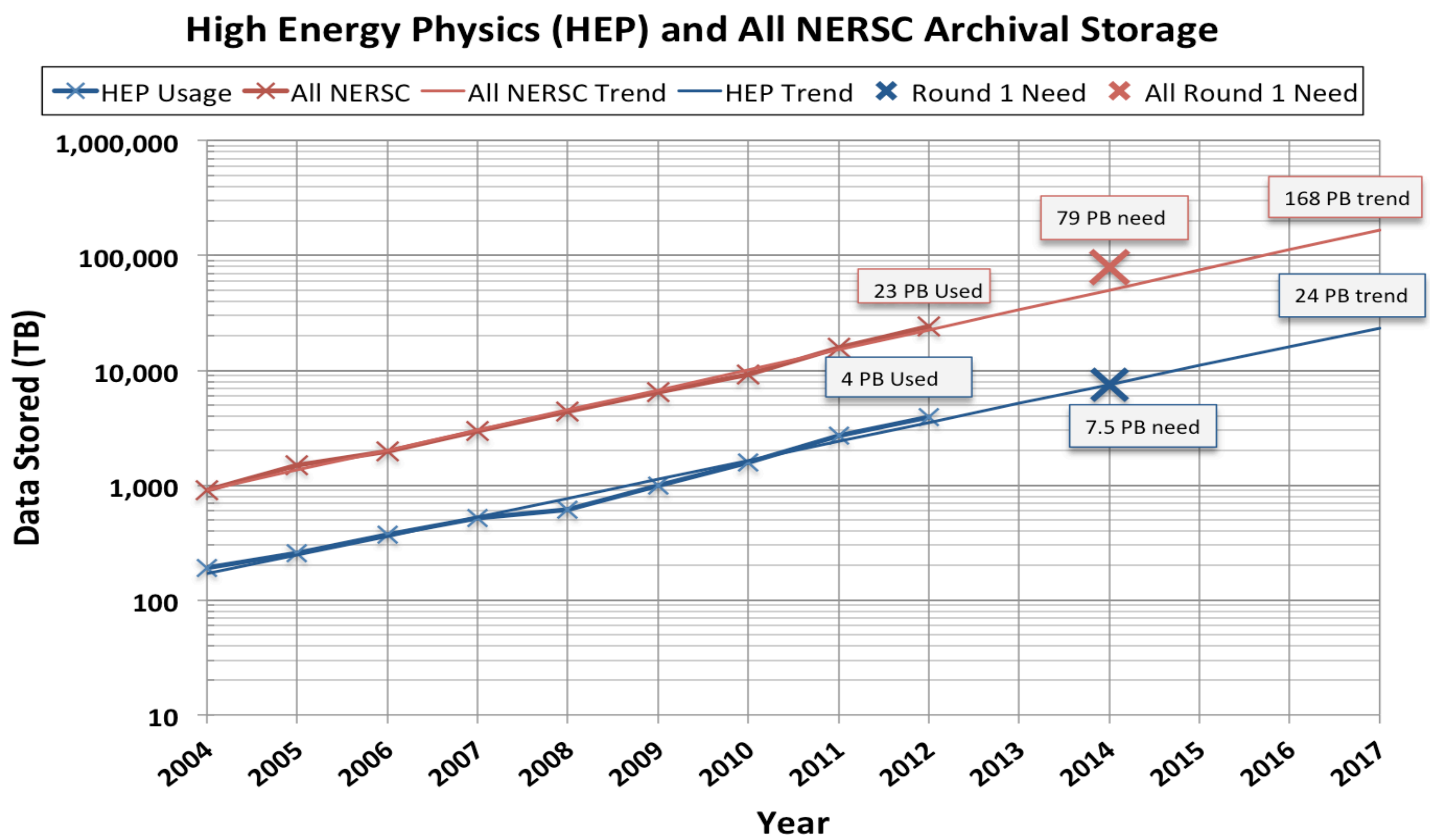




Computational Hours

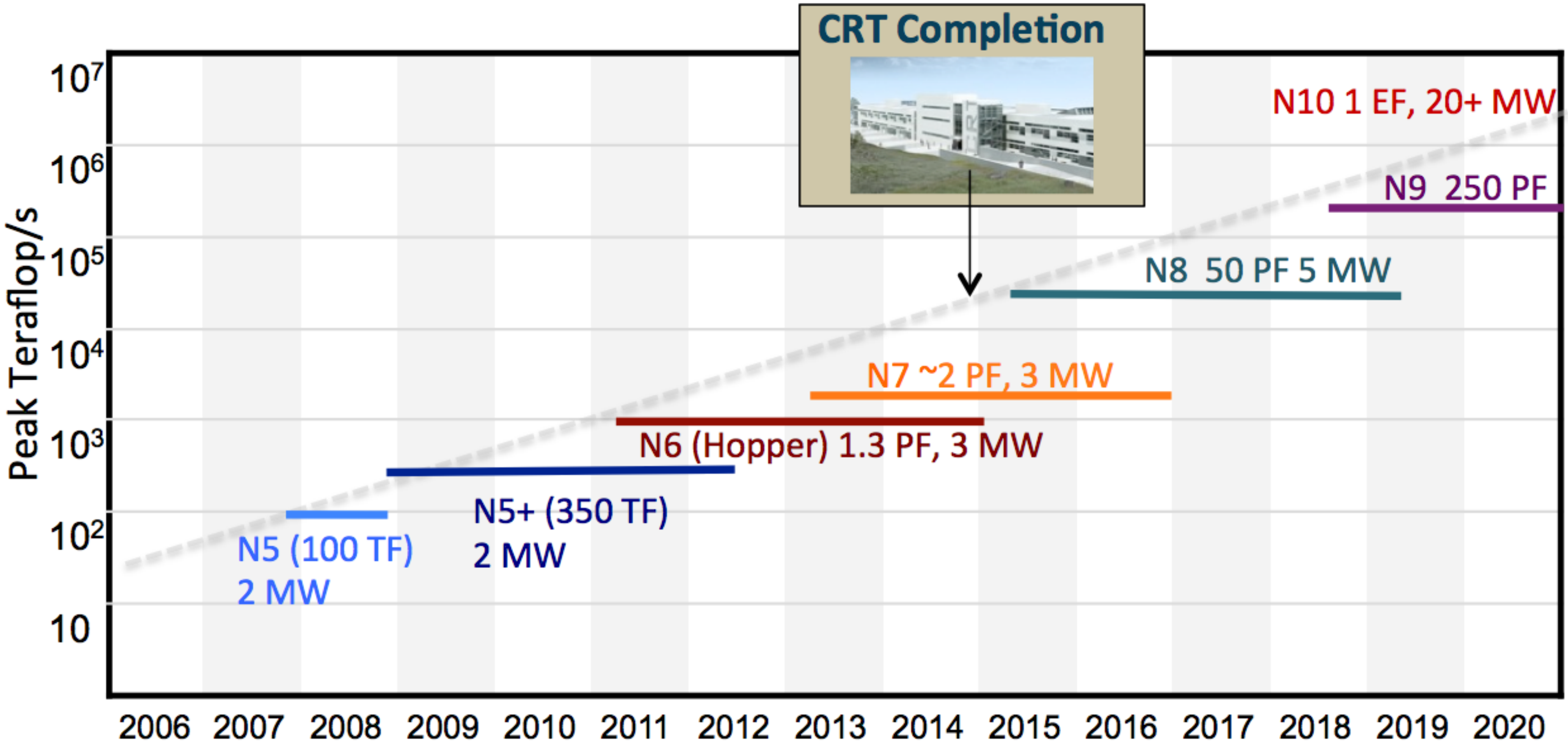


Archival Data



NERSC is Planning for Future Growth

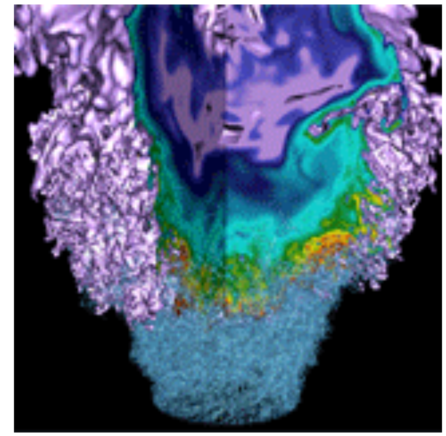
NERSC Systems Roadmap



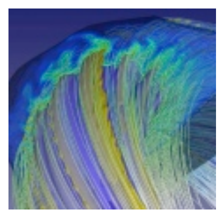
Extreme Data Strategy

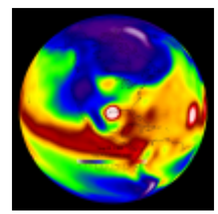
- **Partner with DOE experimental facilities to identify requirements and create early success**
  - NERSC pilot projects have shown great success with automated data pipelining, indexing, searching, archiving, sharing, and distributing end users via the web
- **Develop and deploy new data resources and capabilities**
  - Accelerate NERSC's traditional storage growth rate to meet rapidly increasing requirements for capacity and bandwidth.
  - We are proposing to enhance the data processing capabilities of Edison in 2014 by adding large memory visualization/analysis nodes, adding a flash-based burst buffer or node local storage, and deploying a throughput partition for fast turnaround of jobs.
- **Provide the expertise required to run data-intensive workloads**
  - Develop sophisticated web-based gateways to interact with and leverage data
  - Support database-driven workflows and storage
  - Use scalable structured and unstructured object stores
  - Provide search and analysis software for massive data
  - Provide comprehensive scientific data curation
- **Partner with ESnet for advanced networking capabilities**

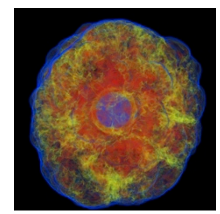
Who runs at NERSC?

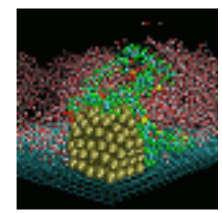


NERSC





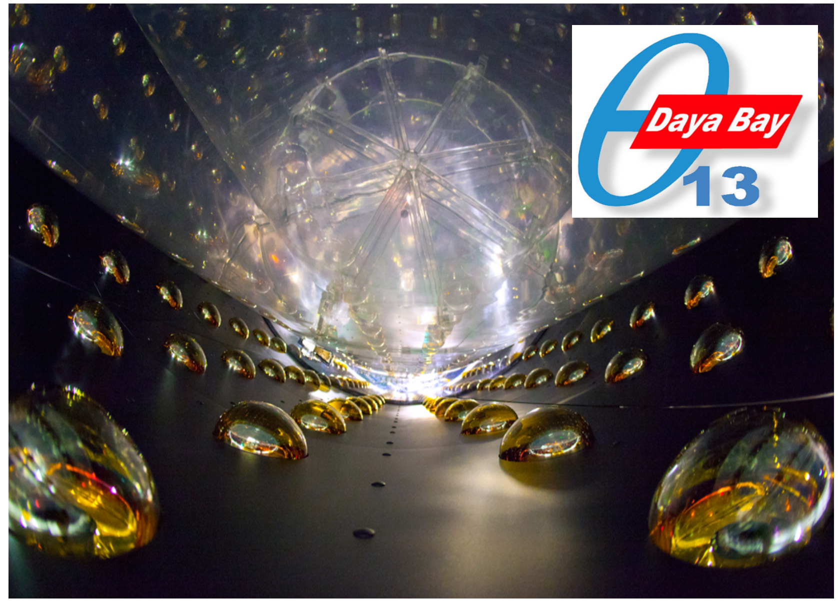




One of *Science* Magazine's Top 10 Breakthroughs of 2012

Discovery of θ_{13} weak mixing angle

- The last and most elusive piece of a longstanding puzzle: How can neutrinos appear to vanish as they travel?
- The answer – a new, large type of neutrino oscillation
  - Affords new understanding of fundamental physics
  - May help solve the riddle of matter-antimatter asymmetry in the universe.



Detectors count antineutrinos near the Daya Bay nuclear reactor in Japan. By calculating how many would be seen if there were no oscillation and comparing to measurements, a 6.0% rate deficit provides clear evidence of the new transformation.

Experiment Could Not Have Been Done Without NERSC and ESNet

- PDSF for simulation and analysis
- HPSS for archiving and ingesting data
- ESNet for data transfer into NERSC
- NERSC Global File System & Science Gateways for distributing results
- NERSC is the *only* US site where all raw, simulated, and derived data are analyzed and archived

PI: Kam-Biu Luk (LBNL)

NERSC Played Key Role in Nobel Prize-Winning Discovery

Physics

NERSC

Accelerating Expansion of the Universe Subject of 2011 Prize

Type Ia supernovae are used as "standard candles" to measure the distance to remote galaxies.

Simulations run at NERSC modeled how Type Ia supernovae should appear from Earth.

This provided the crucial calibration needed to enable the Nobel Prize-winning discovery.

When NERSC moved to Berkeley 1996, this project's work was one of the first funded in a new computational science program created to encourage collaborations between physical and computer scientists.

Berkeley Lab's Saul Perlmutter was awarded the 2011 Nobel Prize in Physics along with two others for their discovery.

It implies the existence of so-called dark energy, a mysterious force that acts to oppose gravity.

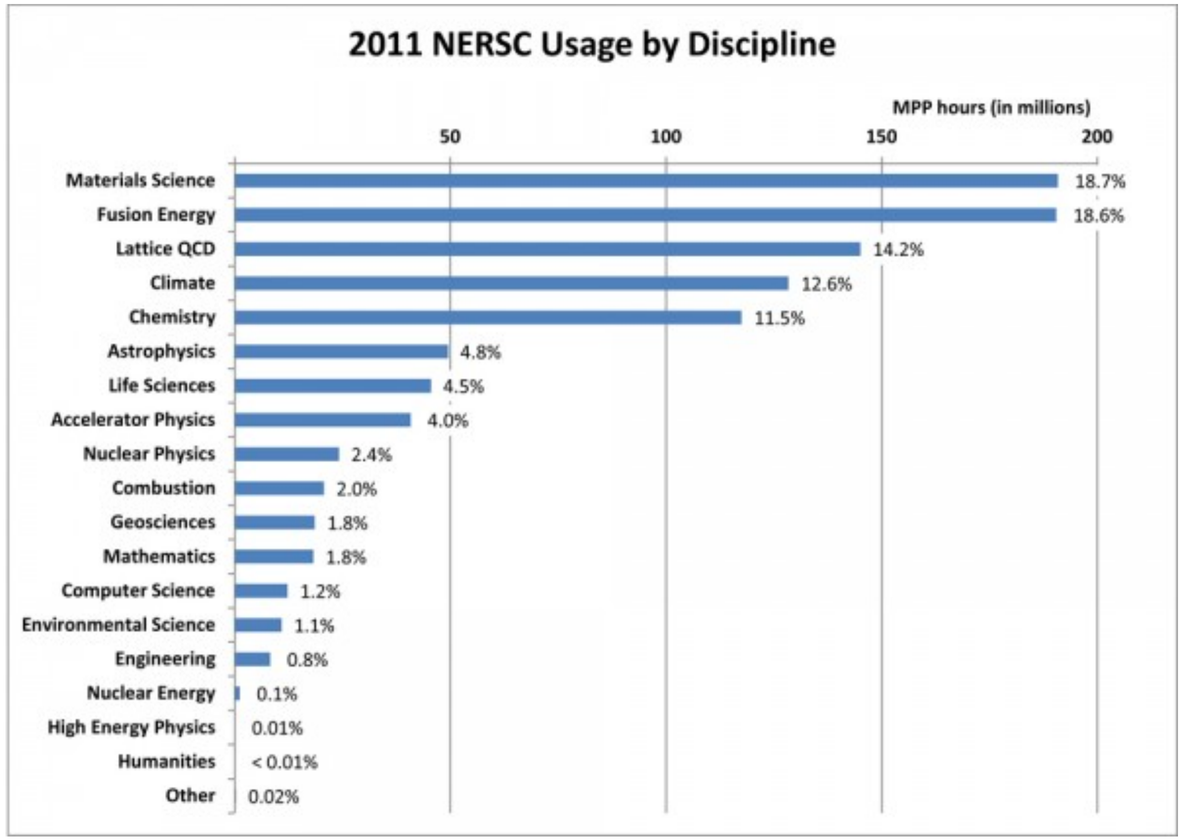
The nature of dark energy is unknown and has been termed the most important problem facing 21st century physics.

U.S. DEPARTMENT OF ENERGY | Office of Science

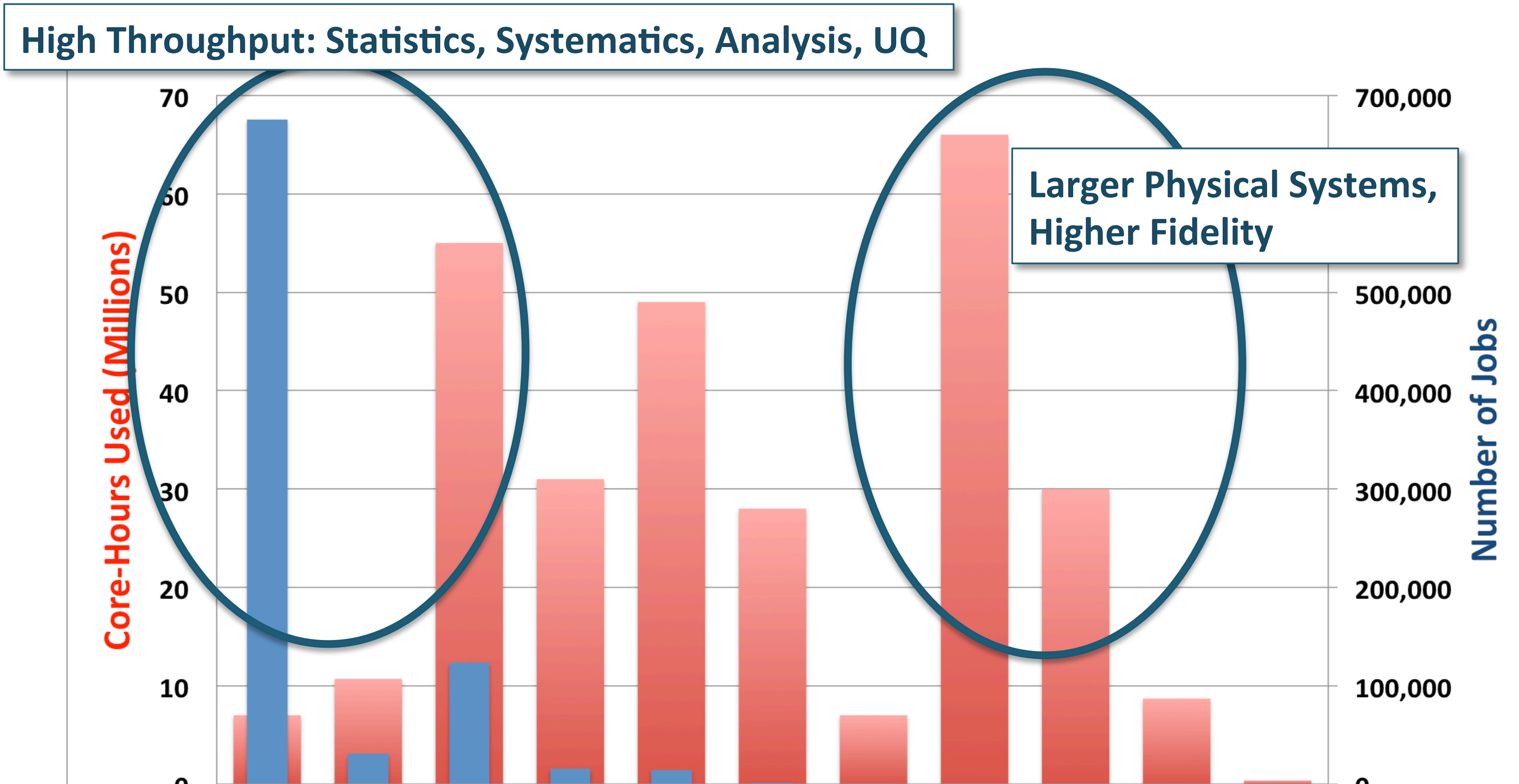
BERKELEY LAB
Lawrence Berkeley National Laboratory

NERSC Supports DOE Open Science

- Scientists from all Office of Science offices rely on NERSC
- Universities (54%), DOE Labs (39%)
- U.S. and International
- Individuals, teams of all sizes



NERSC Supports Jobs of all Kinds 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

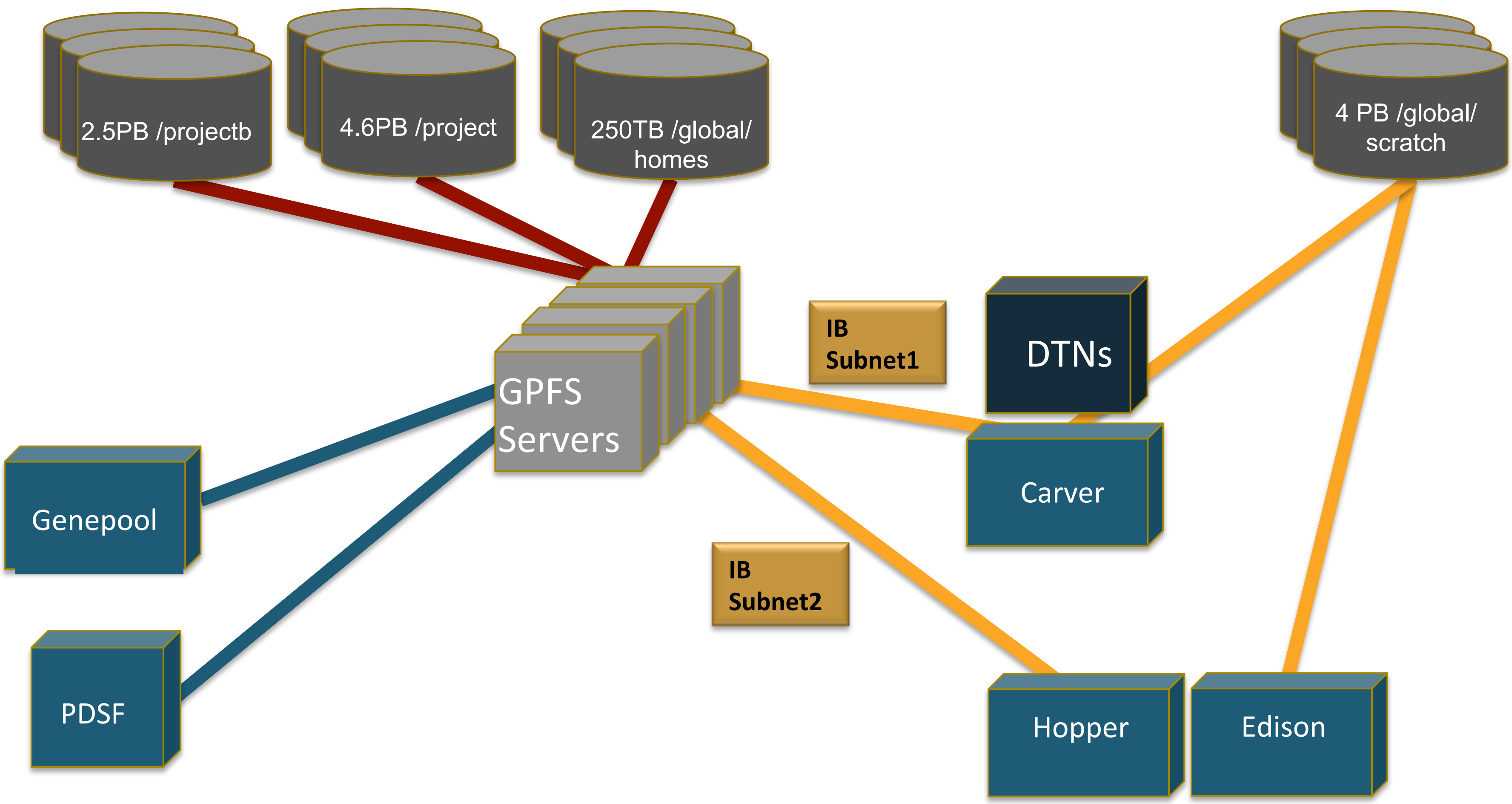
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

NERSC Data Resources

- **Global shared filesystems (aka NGF)**
  - Connected to all NERSC computational systems
  - Large, fast, permanent data storage
  - Intended for data sharing within and among projects
  - Many PBs
  - Default quotas ~ 5-10 TB, but often increased
- **Hopper and Edison have dedicated "local" scratch systems**
  - 2 PB & 6.4 PB, respectively
- **Archival storage system**
  - HPSS tape-backed storage
  - Permanent, many 10s of PB
  - No quotas per se, current 240 PB capacity
- **Grid enabled for fast and easy transfers**
- **Dedicated data transfer nodes**

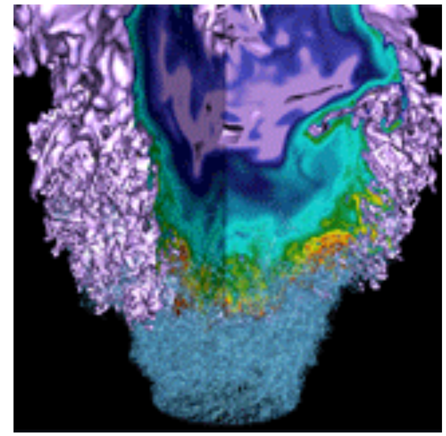
NERSC Global (GPFS) File Storage 2013



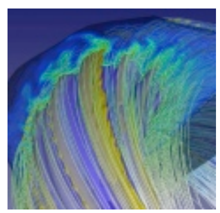
HPSS resources

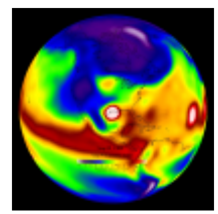
- **User Archive System**
  - As of Feb 2012, contains 24 PB of scientific data:
    - Dating back to 1979
    - Largest file is 38 TB
  - 240 TB disk cache
  - More 5TB enterprise tape drives to improve ingest and read capability
- **Backup System**
  - Contains 14 PB of various backup data
    - ~50% is NGF/GPFS file system backups
  - 60 TB disk cache
  - 4TB enterprise tape drives to handle increase in backup/restore demand
  - Perform a user requested restore operation every other week (single file to several TBs)

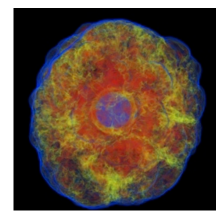
Do you need High Performance Computing (and Data)?

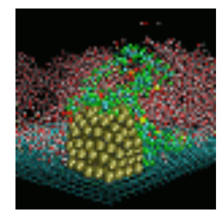


NERSC





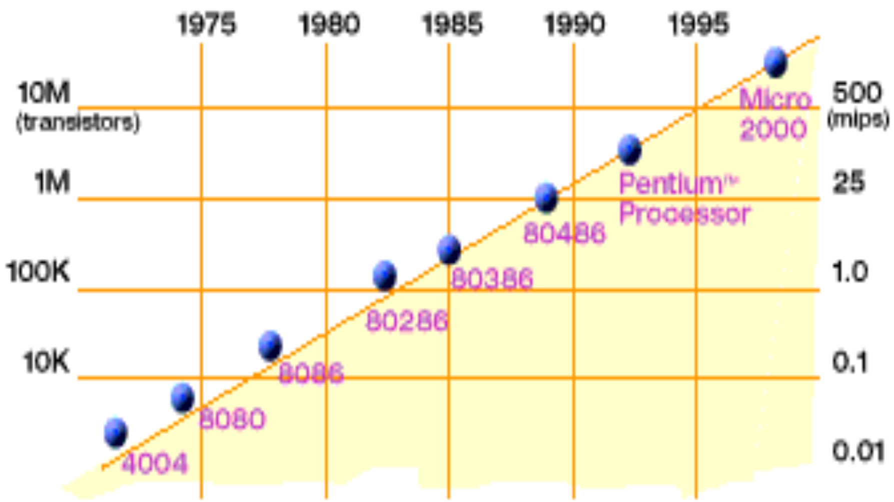




Some Advantages

- Access to large, high performing computing and data systems
- Access to consulting, well-maintained and secure software environment
- Ability to complete simulations and/or analysis thousands up to 100s of thousands times faster than with a single processor
- Ability to run bigger simulations
- Ability to easily share data and codes
- Easy to use account management tools
- Access to huge, permanent archival data storage
- Ability to build web-based science gateways

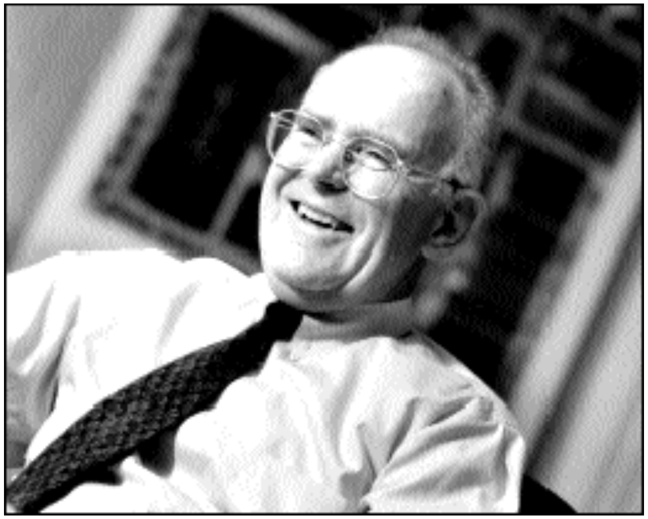
Why You Need Parallel Computing: The End of Moore's Law?



2X transistors/Chip Every 1.5 years

Called "[Moore's Law]"

Microprocessors have become smaller, denser, and more powerful.

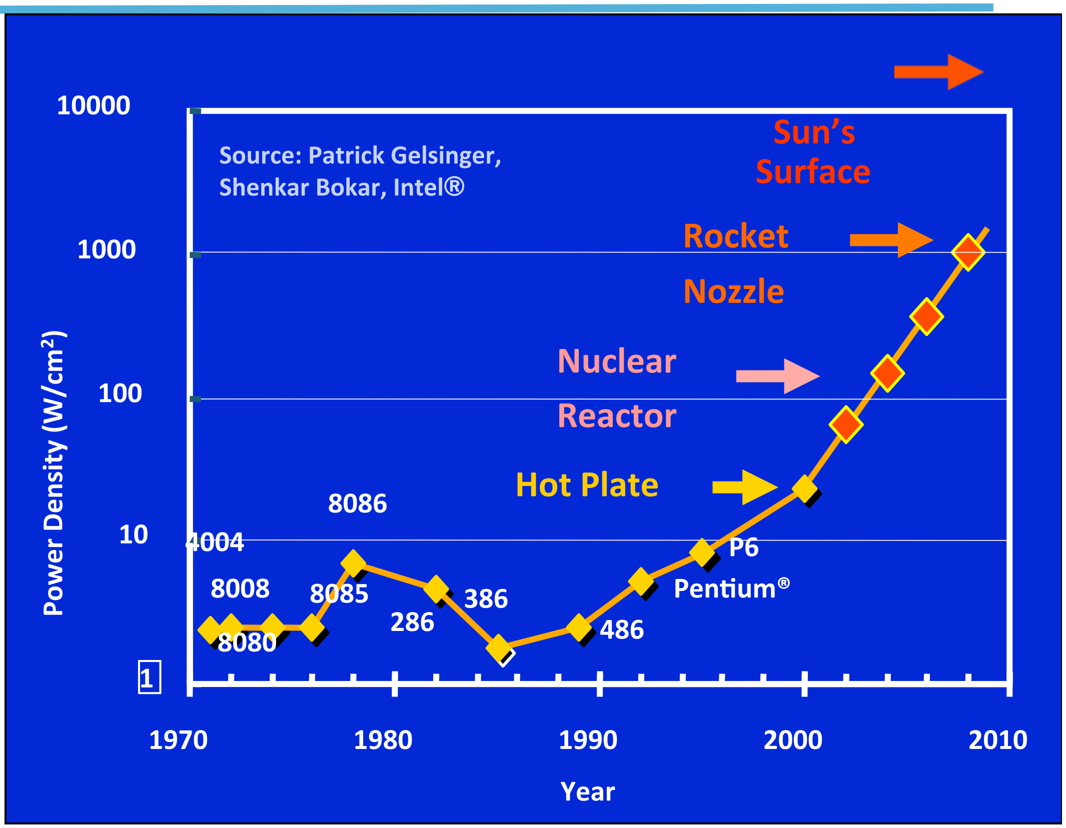


Gordon Moore (co-founder of Intel) predicted in 1965 that the transistor density of semiconductor chips would double roughly every 18 months.

Slide source: Jack Dongarra

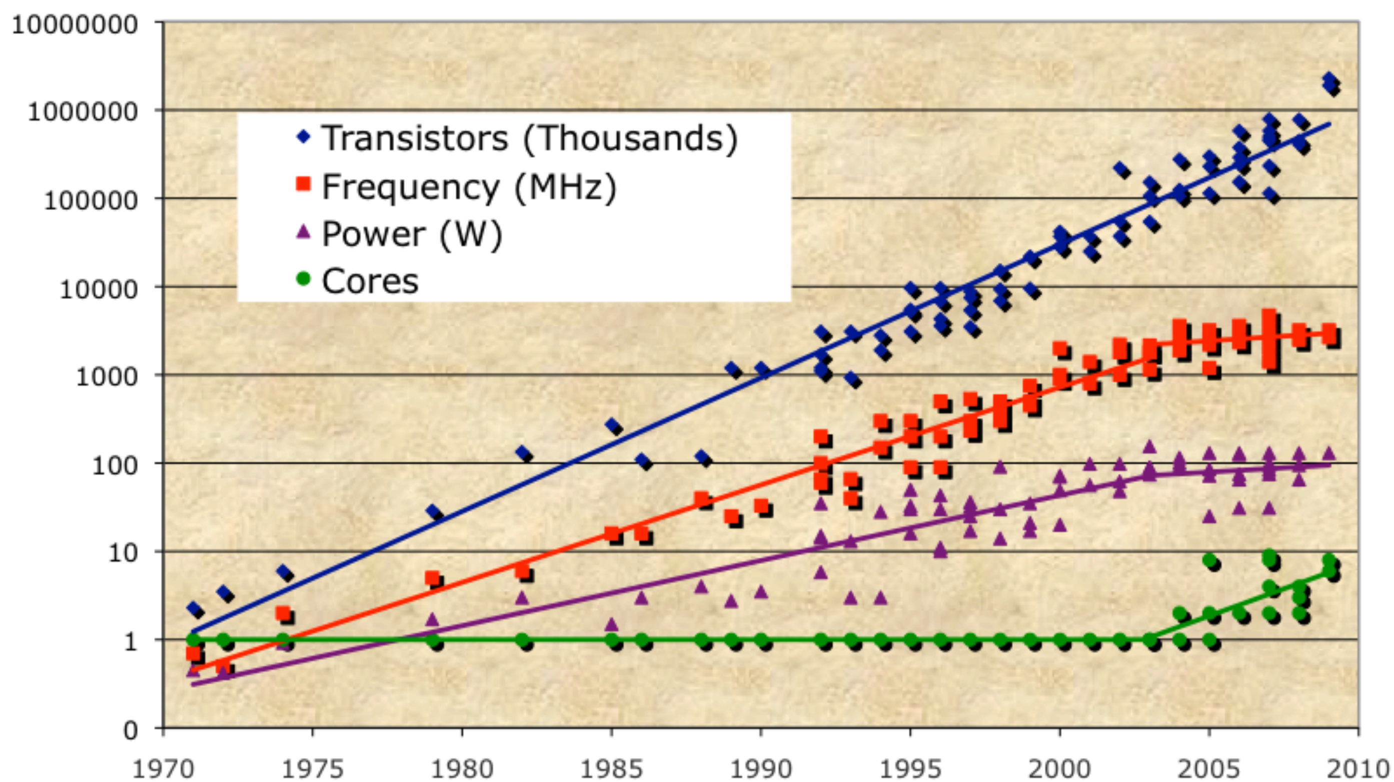
Power Density Limits Serial Performance

- Concurrent systems are more power efficient
  - Dynamic power is proportional to V^2fC
  - Increasing frequency (f) also increases supply voltage (V) → cubic effect
  - Increasing cores increases capacitance (C) but only linearly
  - Save power by lowering clock speed



- High performance serial processors waste power
  - Speculation, dynamic dependence checking, etc. burn power
  - Implicit parallelism discovery
- More transistors, but not faster serial processors

Revolution in Processors

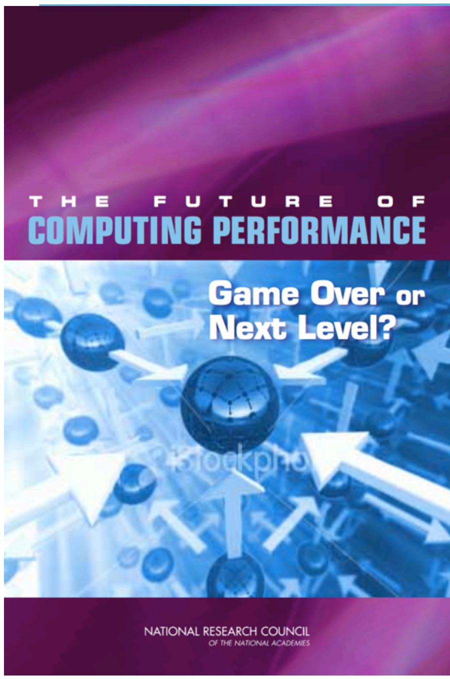


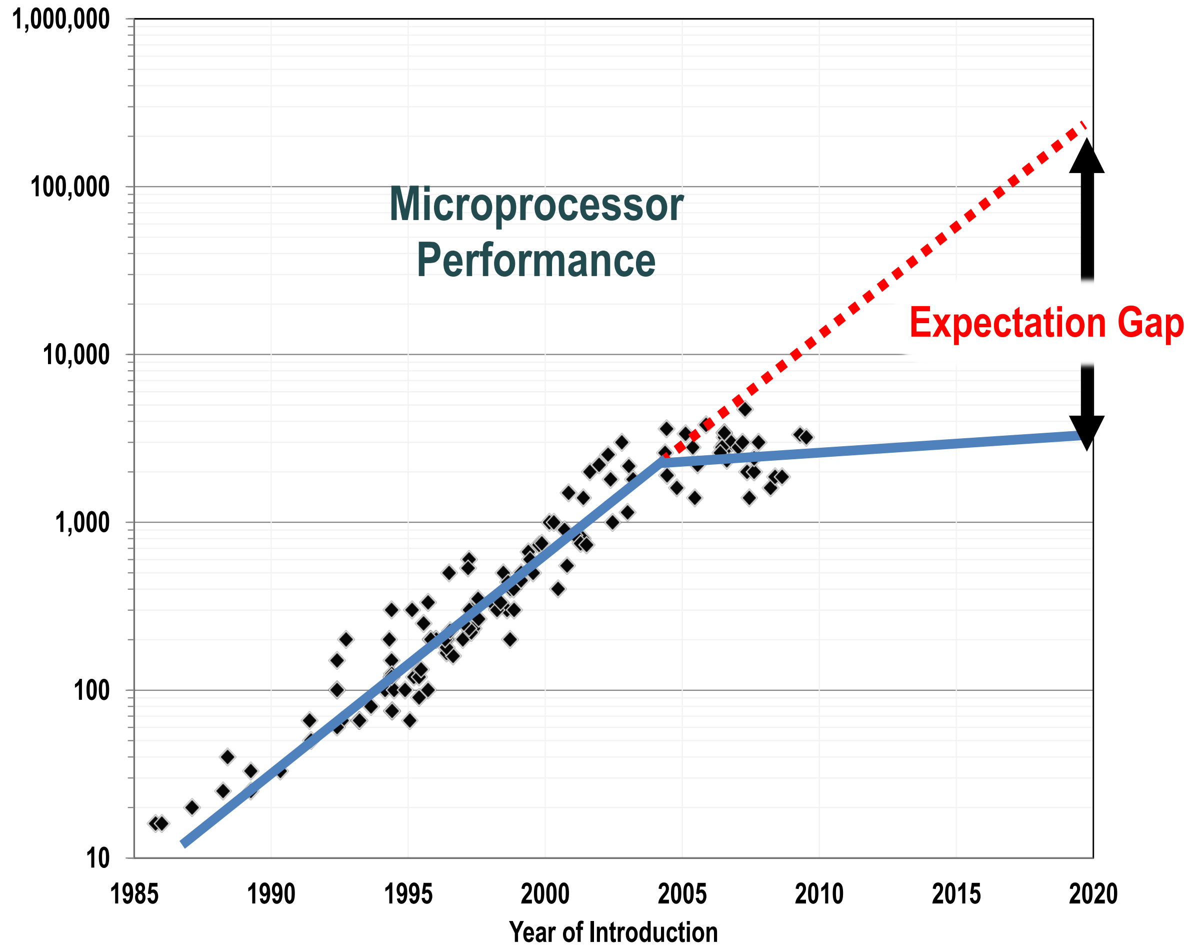
- **Chip density is continuing increase ~2x every 2 years**
- **Clock speed is not**
- **Number of processor cores may double instead**
- **Power is under control, no longer growing**

Moore's Law Reinterpreted

- Number of cores per chip will increase
- Clock speed will not increase (~~possibly~~ probably decrease)
- Need to deal with systems with millions of concurrent threads
- Need to deal with inter-chip parallelism (OpenMP threads) as well as intra-chip parallelism (MPI)
- Any performance gains are going to be the result of increased parallelism, not faster processors

Serial Processing = Left Behind





Can Accelerators The Answer?

- **GPUs show promise for some applications**
  - Many small, energy-efficient cores (GPUs)
  - Accelerators are theoretically very fast
  - Much better theoretical Flop/Watt
- **Challenges are considerable**
  - GPU have private memory space
  - Attached to motherboard via PCI interface currently
  - Need one fat core (at least) for running the OS
  - Data movement from main memory to GPU memory kills performance
  - Programmability is very poor
  - Most codes will require extensive overhauls

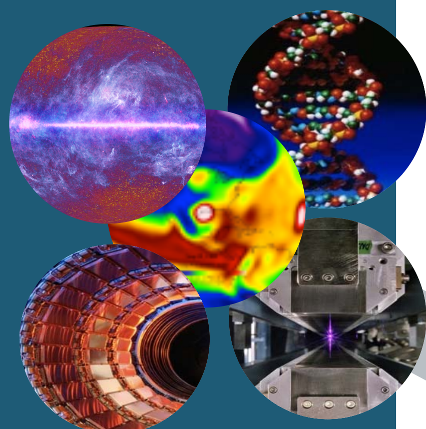
My Personal Opinion

- “Many core” is here to stay
- You will have to find fine-grained parallelism in your code or you will be left behind
- OpenMP or a similar threading model is the most likely viable long-term (5-10 years) programming model
- GPU accelerators have a lot of momentum in the short term and can be useful for certain applications
- Simulation and data analysis will become even more intertwined and will need to share close data spaces

NERSC Is the Place for Integration of Data, Large Simulation, & High Throughput Computing

Big Data

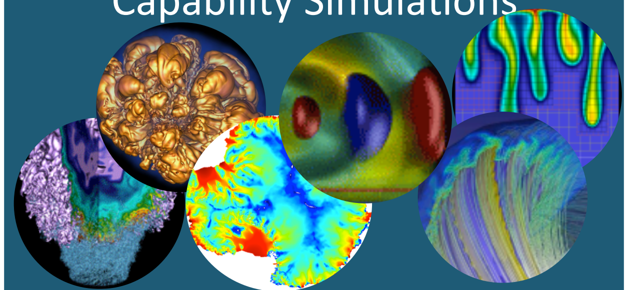
From Experiments and Simulation



NERSC ingests, stores and analyzes data from Telescopes, Sequencers, Light sources, Particle Accelerators (LHC), climate, and environment

Large Scale

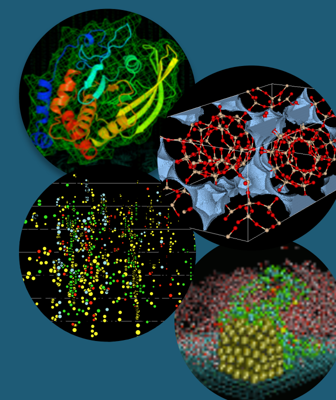
Capability Simulations



Petascale systems run simulations in Physics, Chemistry, Biology, Materials, Environment and Energy at NERSC

High Volume

Job Throughput



NERSC computer, storage and web systems support complex workflows that run thousands of simulations to screen materials ("Materials Genome"), proteins, structures and more; the results are shared with academics and industry through a web interface

NERSC

Petascale Computing, Petabyte Storage, and Expert Scientific Consulting

NERSC

National Energy Research Scientific Computing Center