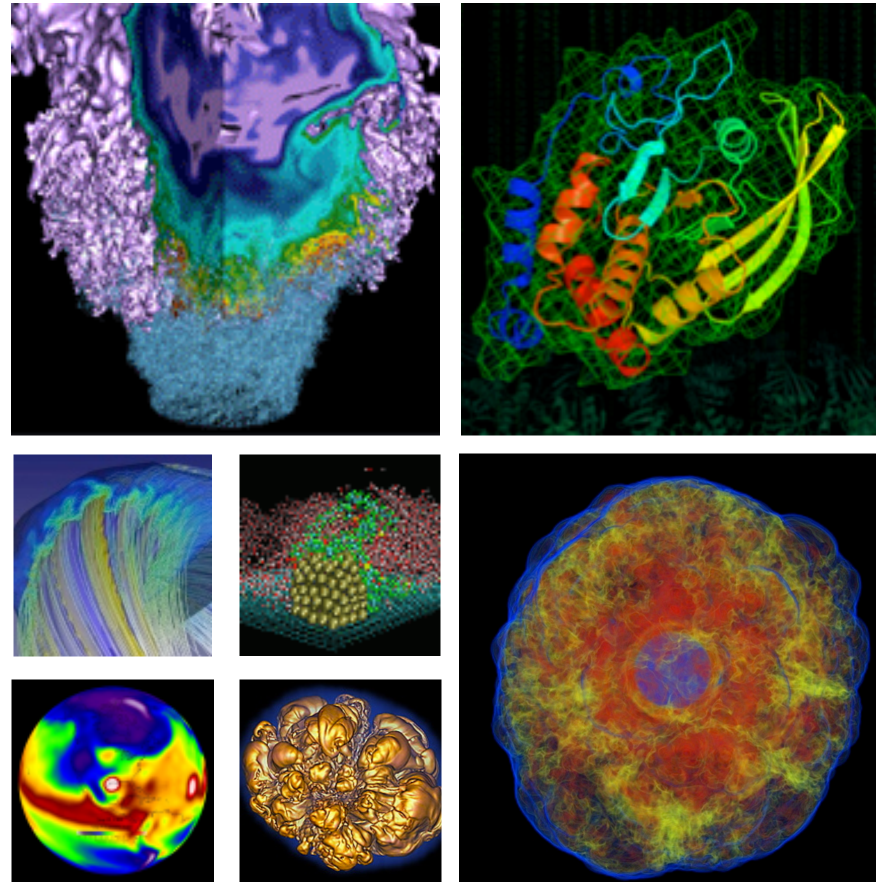
Exascale Computing, Big Data, and World-Class Science at NERSC



NERSC 40 YEARS at the FOREFRONT

Richard Gerber
NERSC Senior Science Advisor
User Services Group Lead
NERSC @ Berkeley Lab

December 18, 2014

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

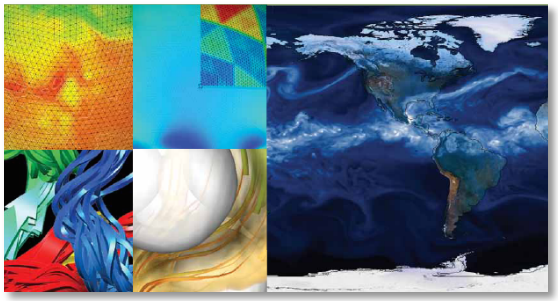
Outline

- What is NERSC
- Science at NERSC
- NERSC Resources
- Who uses NERSC?
- How to get access to NERSC Resources
- Exascale and Cori

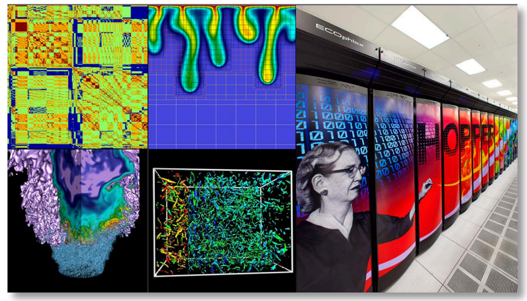
NERSC is the Supercomputing & Data Facility for DOE Office of Science

U.S. DEPARTMENT OF ENERGY | Office of Science

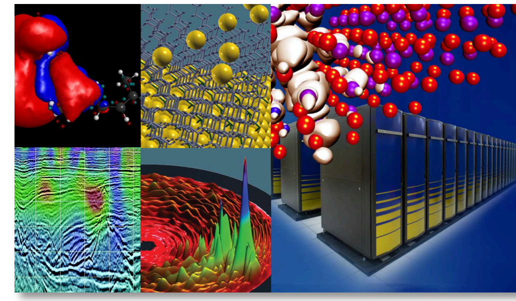
Largest funder of physical science research in U.S.



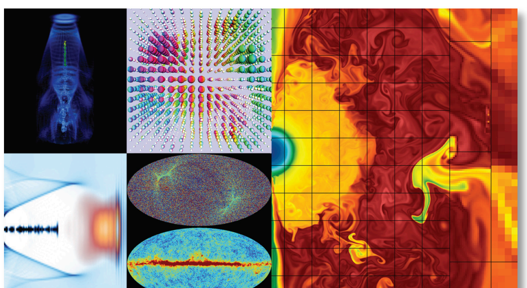
Bio Energy, Environment



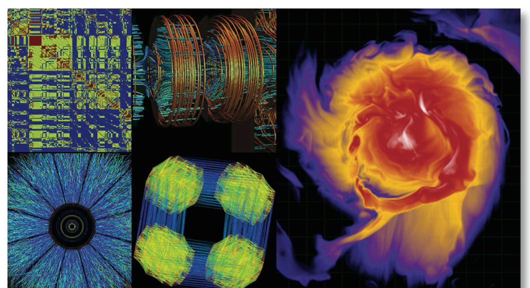
Computing



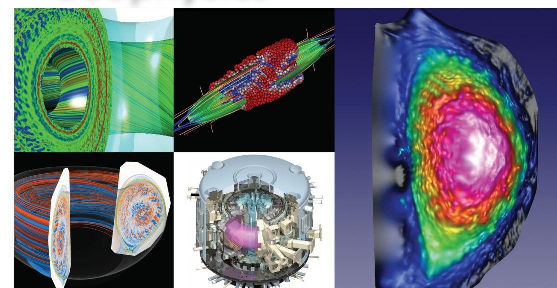
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



Nuclear Physics



Fusion Energy, Plasma Physics

National Energy Research Scientific Computing Center (NERSC)

- **NERSC is a national supercomputer center funded by the U.S. Department of Energy Office of Science (SC)**
  - The SC HPC center that supports its research mission
  - Part of Berkeley Lab, currently located in downtown Oakland
- **NERSC Provides**
  - Reliable HPC supercomputers and data storage systems
  - Expert consulting services
- **If you are a researchers with funding from SC, you can apply to use NERSC**
  - Other researchers are eligible if research is in SC mission
- **NERSC supports 5,000 users, 700 projects**
- **From 47 states; 65% from universities**
- **Hundreds of users log on each day**


BERKELEY LAB
OAKLAND
SCIENTIFIC

What Makes NERSC Special

- Large, state-of-the-art supercomputing and data systems
- Consulting, system admin, 24x7 operations support
- Well maintained software environment, prebuilt optimized applications & libraries
- Designed for massive parallelism, but support all scales
- Easy to share data and codes
- Easy to use account management
- Large permanent archival data storage
- Ongoing technology refreshes
- Word-class cybersecurity
- Open science environment
- Web-based science /data gateways

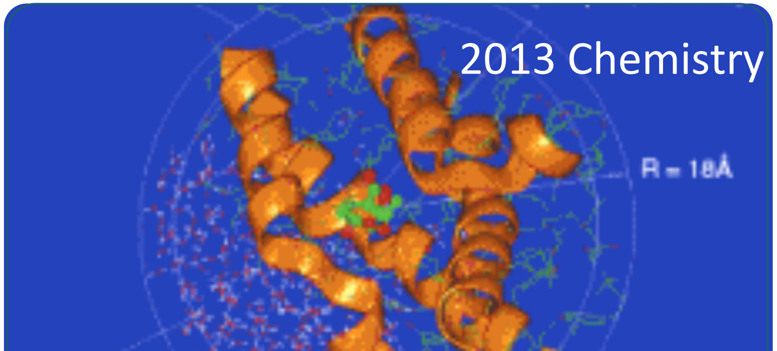
NERSC: Science First

- 1,977 refereed journal publications in 2013
- 5-20 journal covers per year
- 4 Nobel Prize winners (5 if count Higgs)
- 3 of *Physics World's* Top 10 Breakthroughs of 2013 (IceCube, Planck, B-Mode CMB polarization)
- Finding that Earth-like planets are not uncommon in our galaxy recognized as a top 2013 discovery by Wired Magazine and covered in The New York Times.
- 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."
- One of Science Magazine's Top 10 Breakthroughs of 2012: Measurement of θ_{13} ν mixing angle at Daya Bay, China

18 Journal Covers in 2013

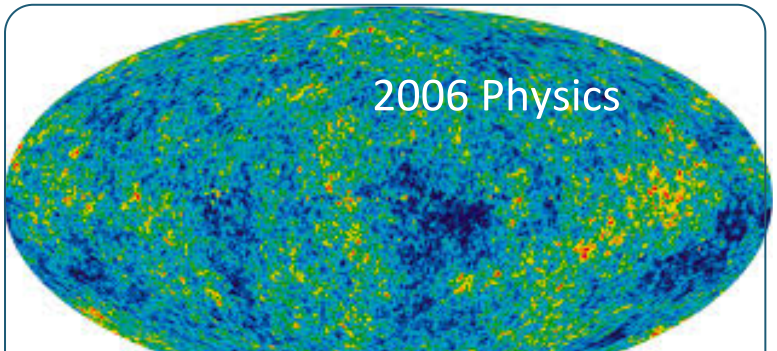
U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC Nobel Prize Winners



John Kuriyan for Martin Karplus

Saul Perlmutter

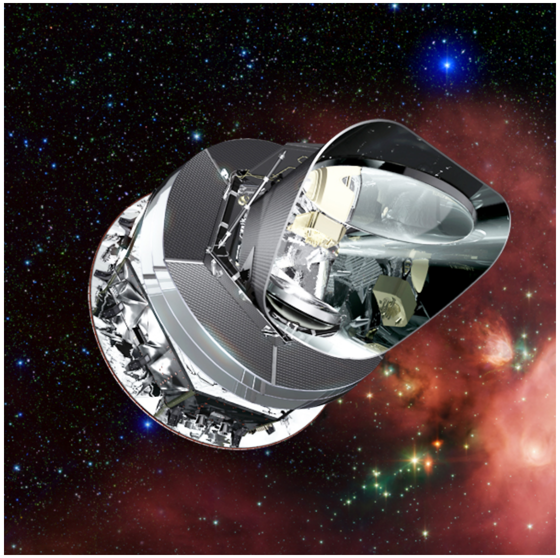


George Smoot

Warren Washington

The Planck Mission

- **A European Space Agency (+NASA) satellite mission to measure the temperature and polarization of the Cosmic Microwave Background.**
  - The echo of the Big Bang: primordial photons have seen it all.
  - Fluctuations encode all of fundamental physics & cosmology.
  - Planck results assumed by all Dark Energy experiments.

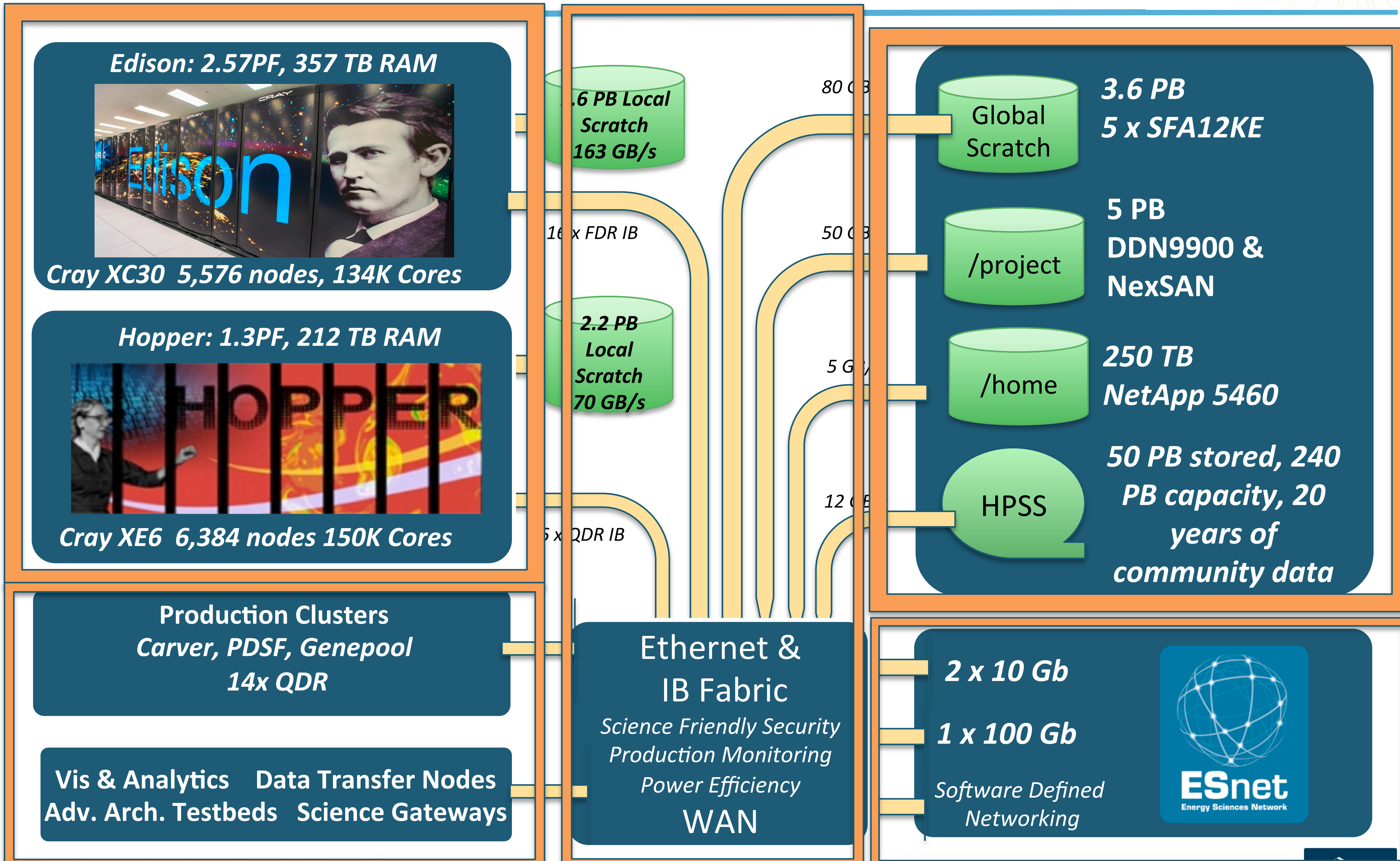


- **Realizing the full scientific potential of Planck requires very significant computing resources**
  - Tiny signal (μK - nK) requires huge data volume for sufficient S/N
  - 72 detectors sampling at 30-180Hz for 2.5 years => 10^{12} samples.
  - Analysis depends critically on Monte Carlo methods
    - Simulate and analyze 10^4 realizations of the entire mission!
- **One of Physics World's Top 10 Breakthroughs of 2013**

Astronomy & Astrophysics Projects at NERSC

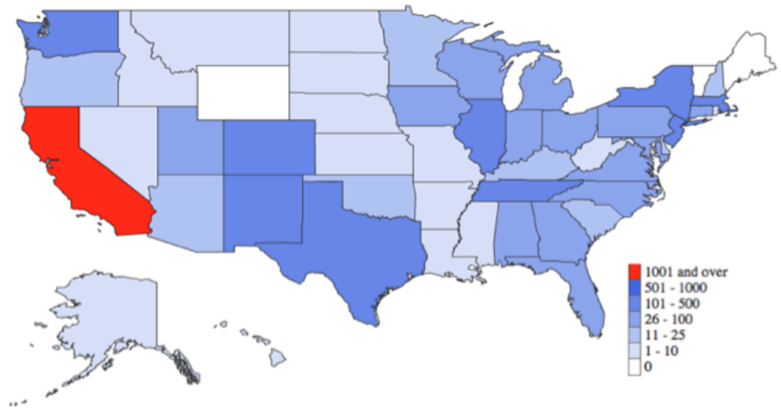
- **55 Projects in 2014**
  - 250 Million hours of compute time allocated
  - 6.5 PB of archival data currently stored
  - 1 PB on permanent spinning disk shared among project members (/project)
- **Science Emphasis**
  - Planck data analysis and synthetic observations/maps
  - Supernova & transients searches
  - Cosmological simulations
  - Supernova simulations
  - Other: Neutrino astrophysics, radio astronomy data analysis, galaxy formation, X-ray bursts, MHD, …

NERSC Systems Today

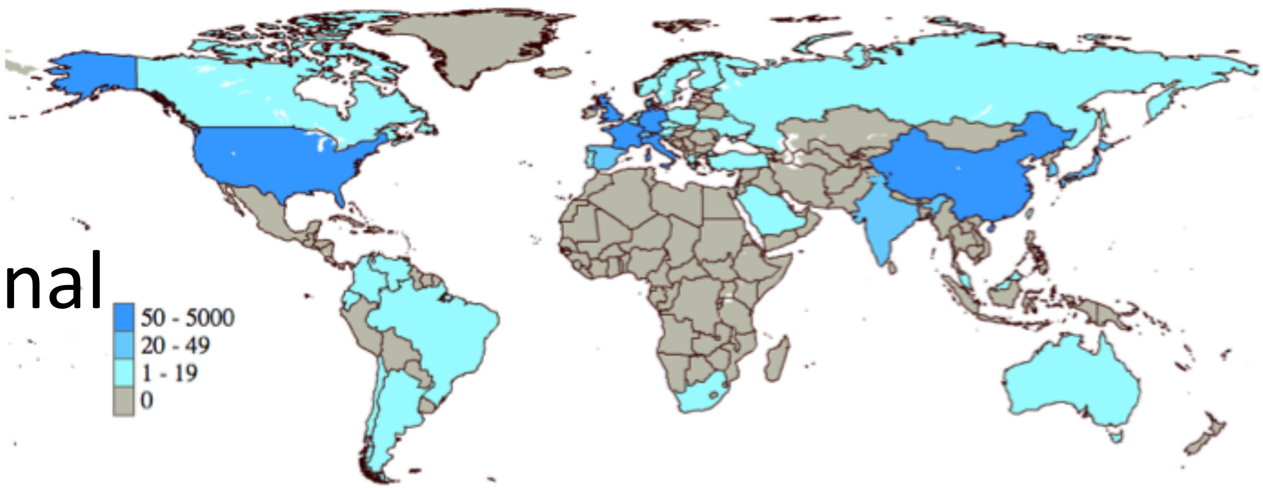


NERSC Demographics

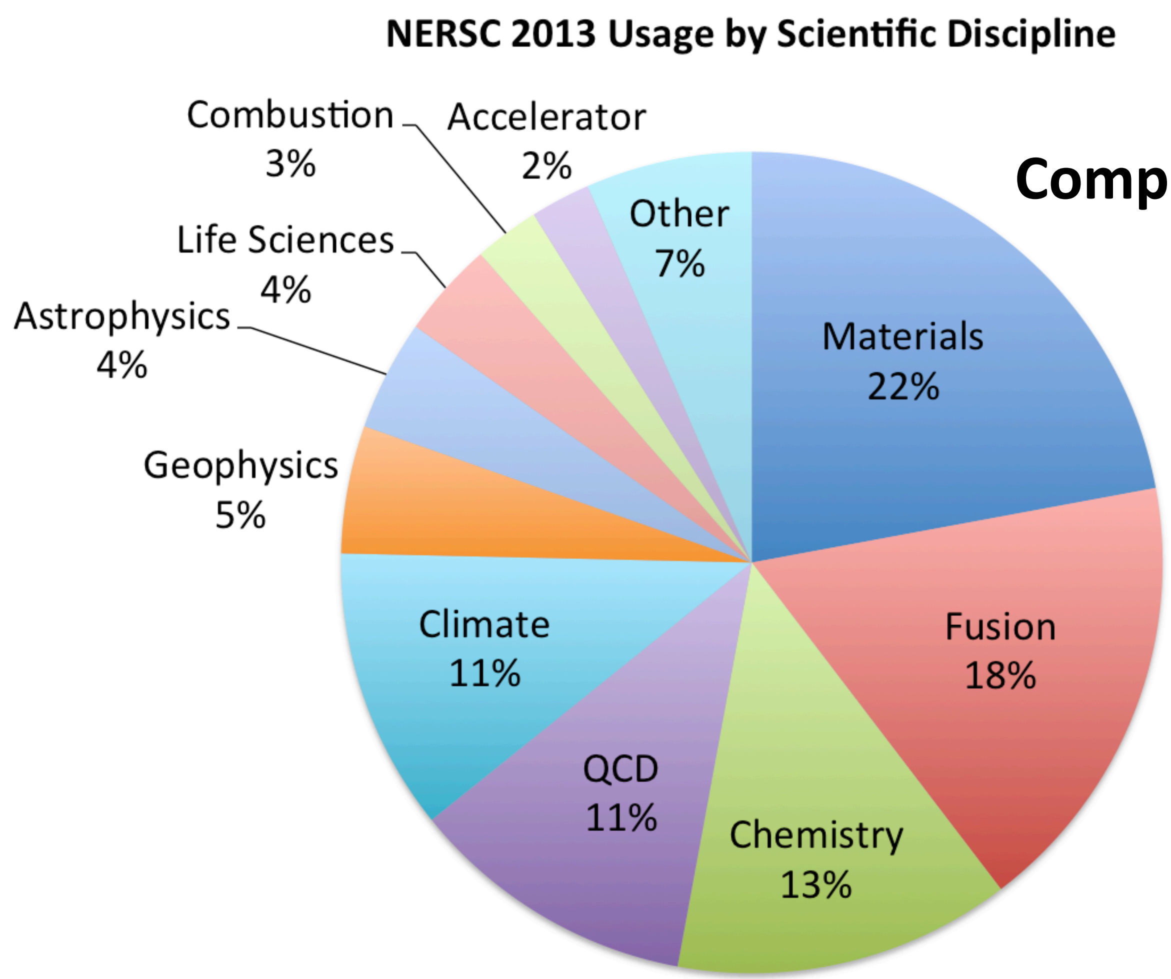
5,000 users from 47 U.S. states



642 international users

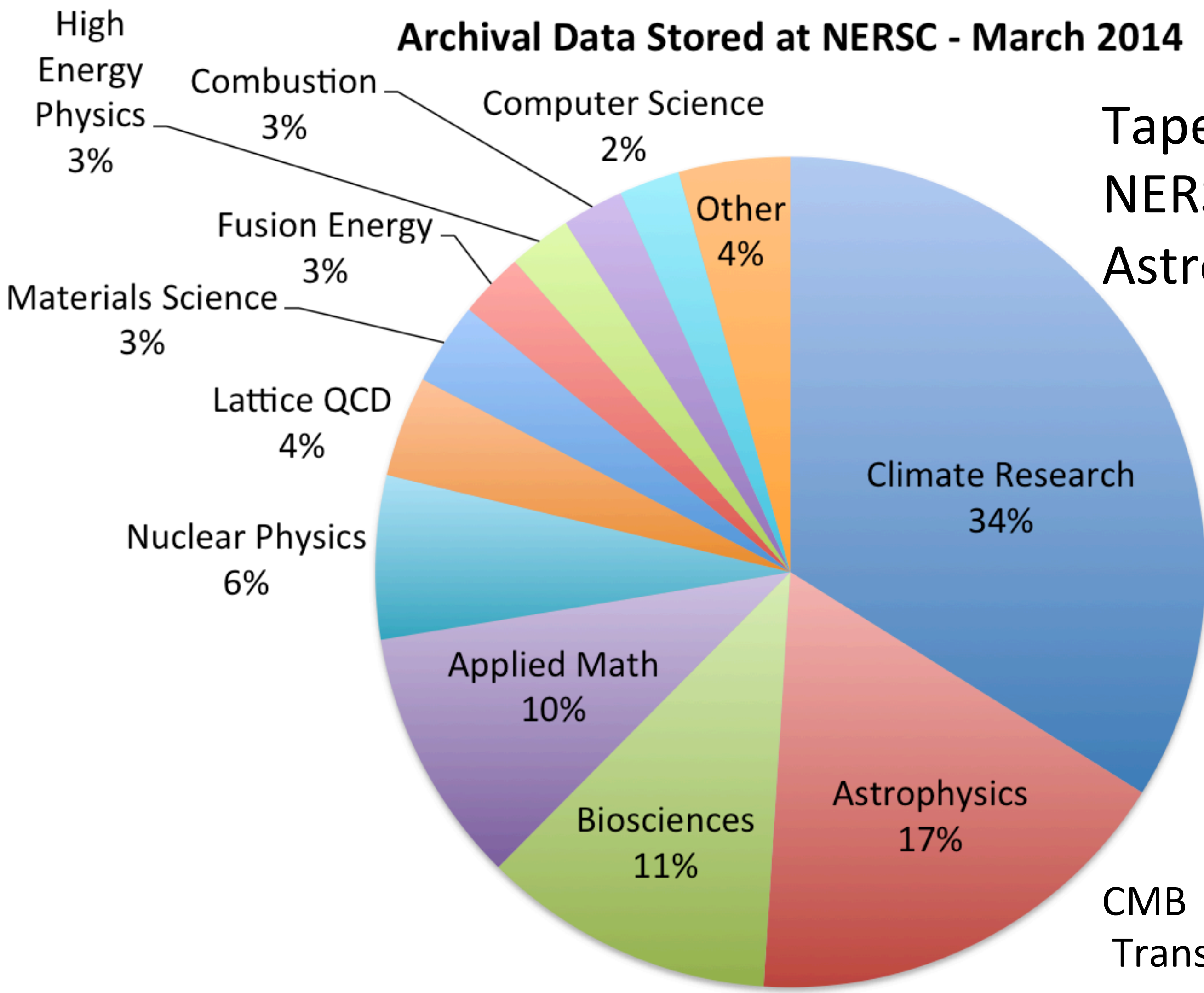


Who Uses NERSC? : Compute Hours



Compute Core Hours

Who Uses NERSC? – Archival Storage



Tape-Backed Storage

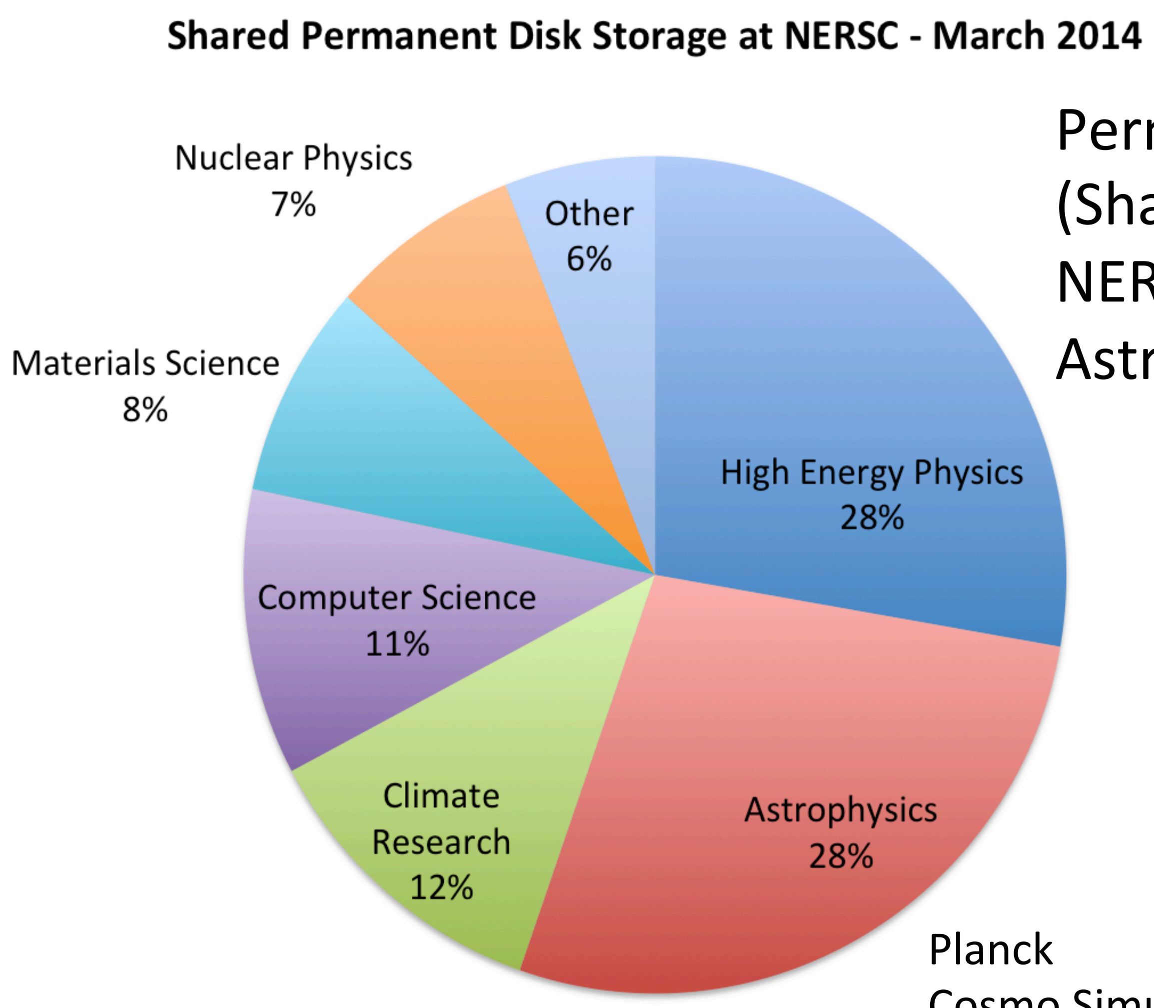
NERSC Total: 38 PB

Astro: 6.5 PB

CMB

Transient/SN searches

Who Uses NERSC? – Permanent Disk



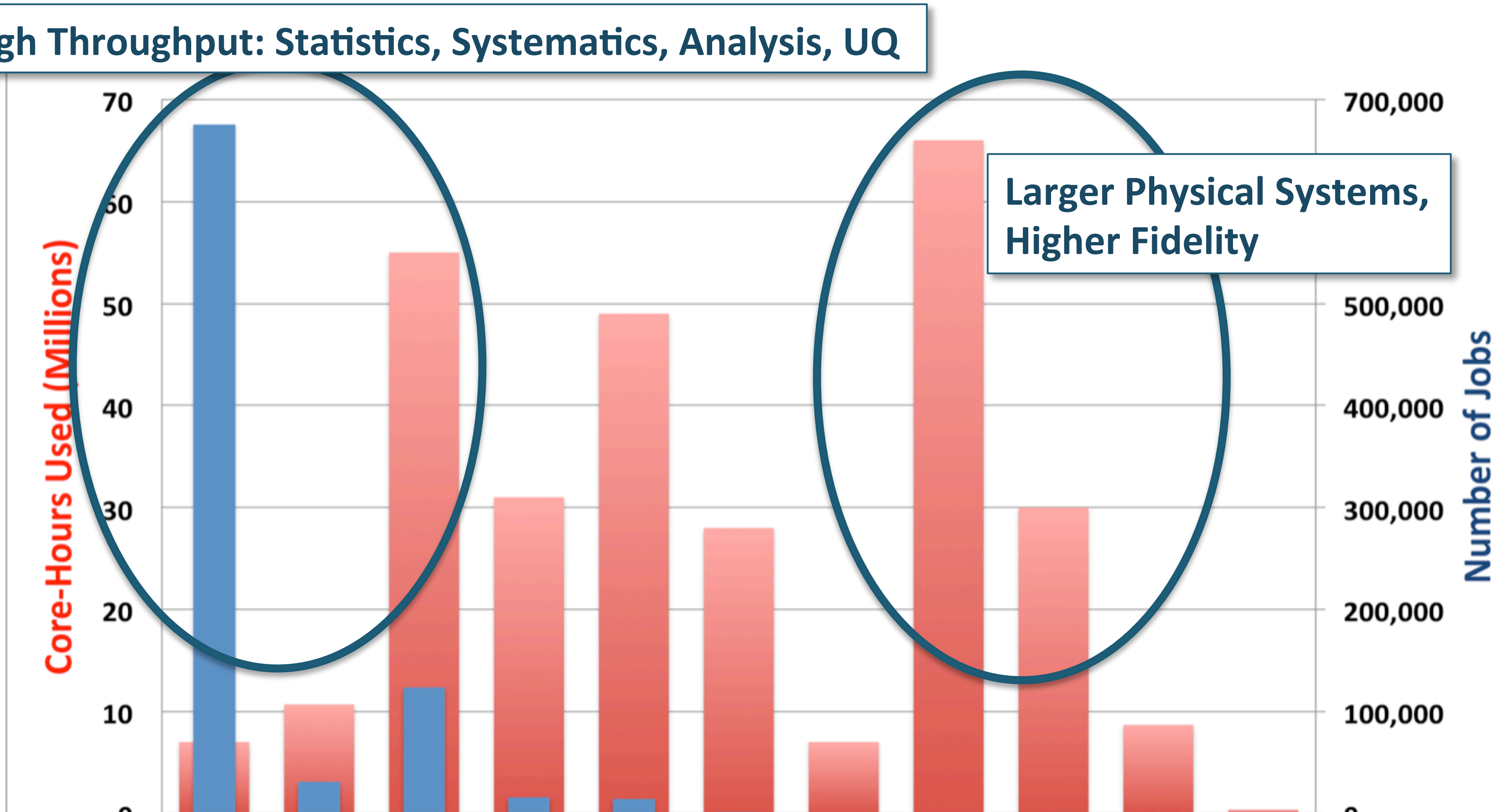
Permanent Spinning Disk (Shared)

NERSC Total: 3.25 PB

Astro: 0.9 PB

Planck
Cosmo Simulation Data

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

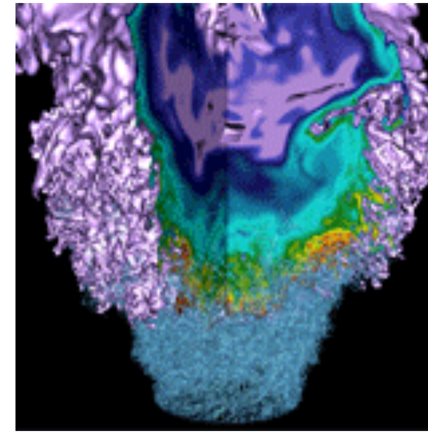
How to Get Access to NERSC Resources

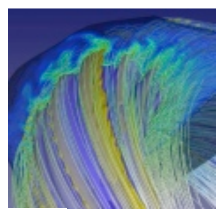
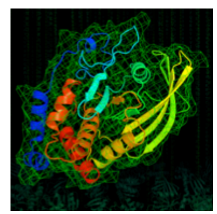
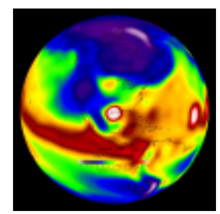
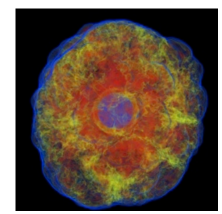
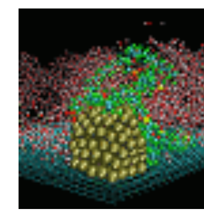
- **"ERCAP" allocations process**
  - 80% of compute hours allocated by DOE program managers to projects doing research within the DOE mission
  - 10% allocated through ALCC (high-risk, high-payoff)
  - Archival storage (tape) also allocated
  - Project funding from DOE not required; at discretion of program managers
- **NERSC Director's Reserve for strategic projects**
  - 10% of computer time (300 M hours in 2015)
- **Startup Projects**
  - At NERSC's discretion
  - Up to 50 K hours for 18 months

How to Get Access to NERSC II

- **Buy-in model for hardware and support**
  - PDSF cluster: Nuclear and High Energy Physics
  - Genepool cluster & file systems: Joint Genome Institute
- **A La Carte resources run by NERSC**
  - Planck bought a rack of a compute cluster
  - Fixed cost for 5 years of spinning disk in NERSC Global File System
  - The Materials Project has dedicated nodes
  - Science Gateways

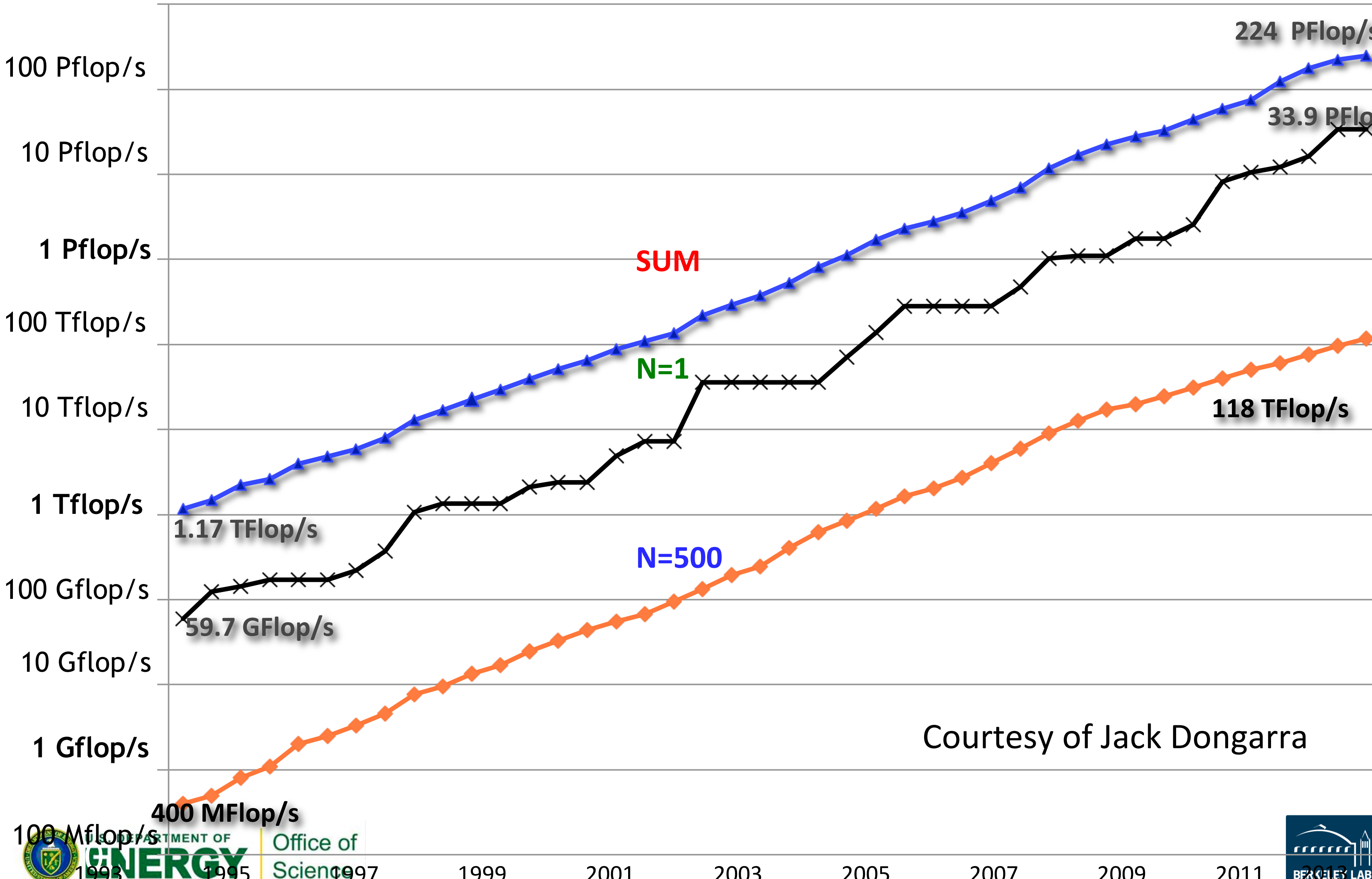
The Road Toward Exascale



NERSC 40 YEARS at the FOREFRONT

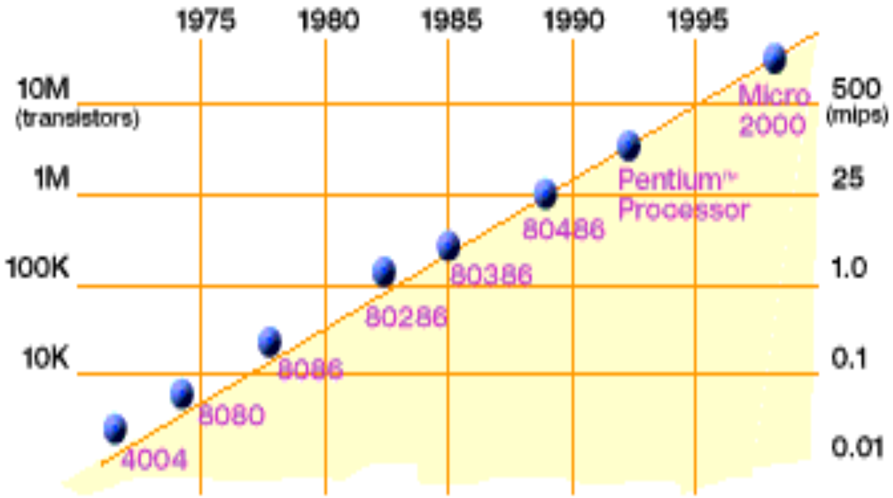
Performance Development of HPC Over the Last 20 Years From Top 500



U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

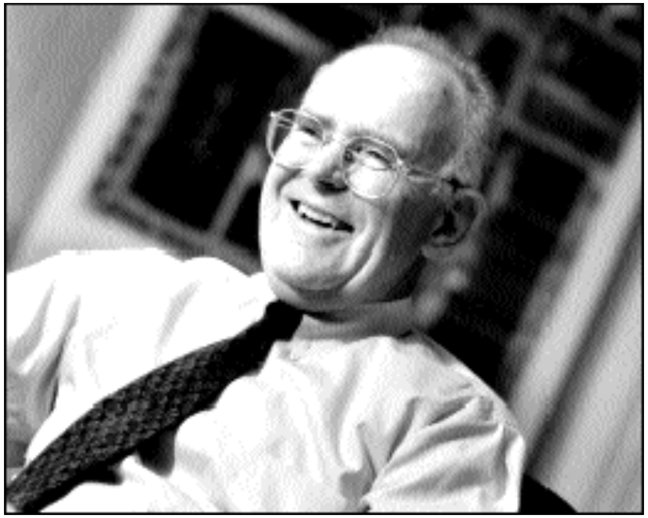
Moore's Law



2X transistors/Chip Every 1.5 years

Called "Moore's Law"

Microprocessors have become smaller, denser, and more powerful.

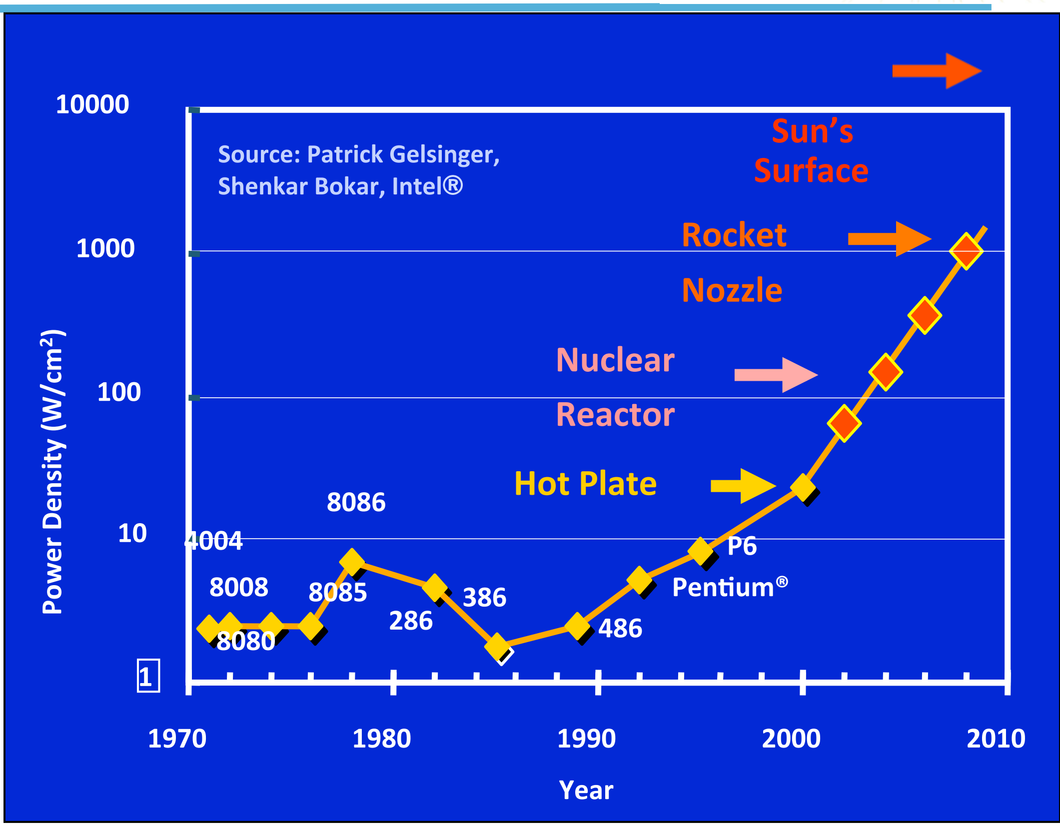


Gordon Moore (co-founder of Intel) predicted in 1965 that the transistor density of semiconductor chips would double roughly every 18 months.

Slide source: Jack Dongarra

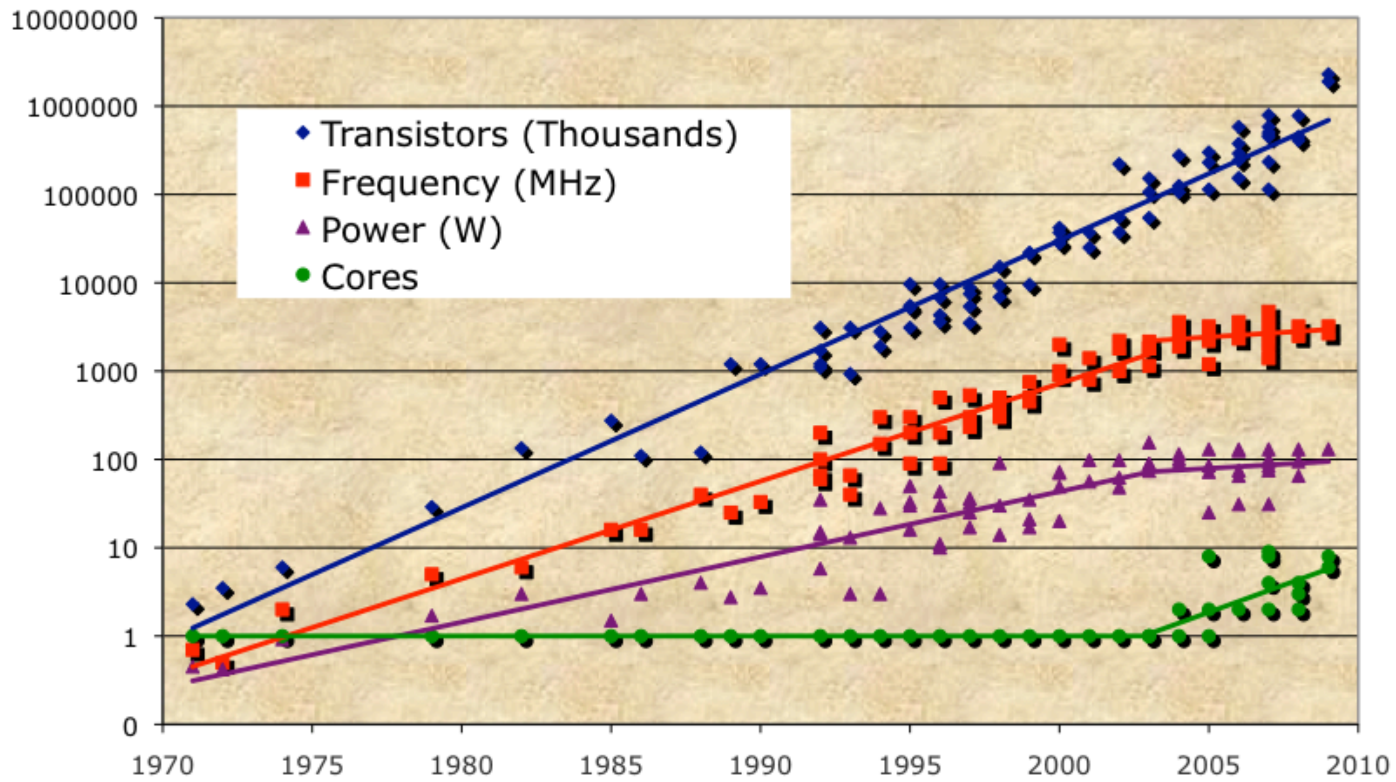
Power Density Limits Serial Performance

- 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 as fast

Moore's Law Reinterpreted

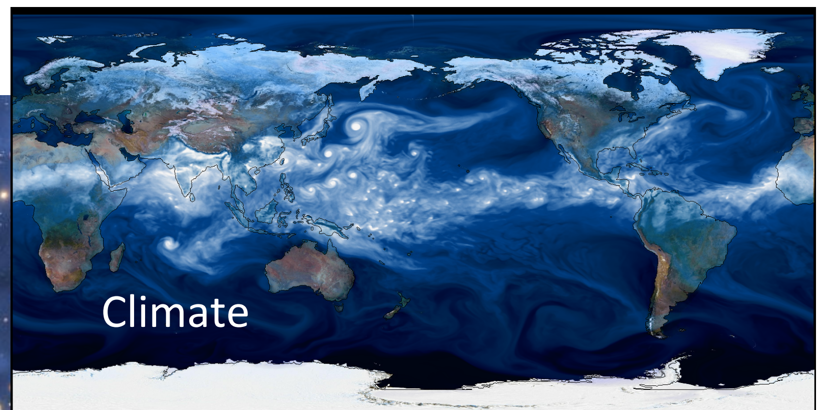
- Number of cores per chip will increase
- Clock speed will not increase (~~possibly~~ probably decrease)
- Need to deal with systems with millions of concurrent threads
- Need to deal with inter-chip parallelism (OpenMP threads) as well as intra-chip parallelism (MPI)
- Any performance gains are going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

This is a Pain!

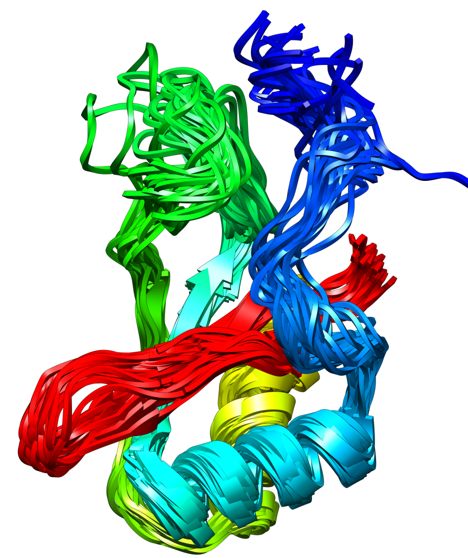
Legacy programs will have to be rewritten to run well (or at all) on next-generation “manycore” processors

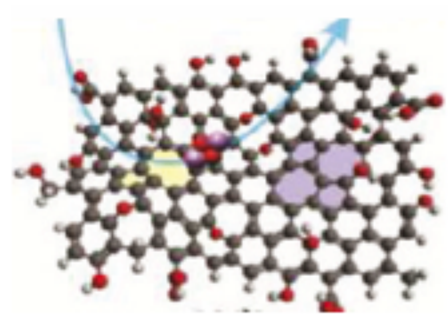
But We Have Exascale Science Problems

The science community has to transition to manycore computing systems



Understanding How Proteins Work





Designing Better Batteries

Requirements Reviews

1½-day reviews with each Program Office

Computing and storage requirements for next 5 years

Adv. Comp. Science Research Jan. 2014

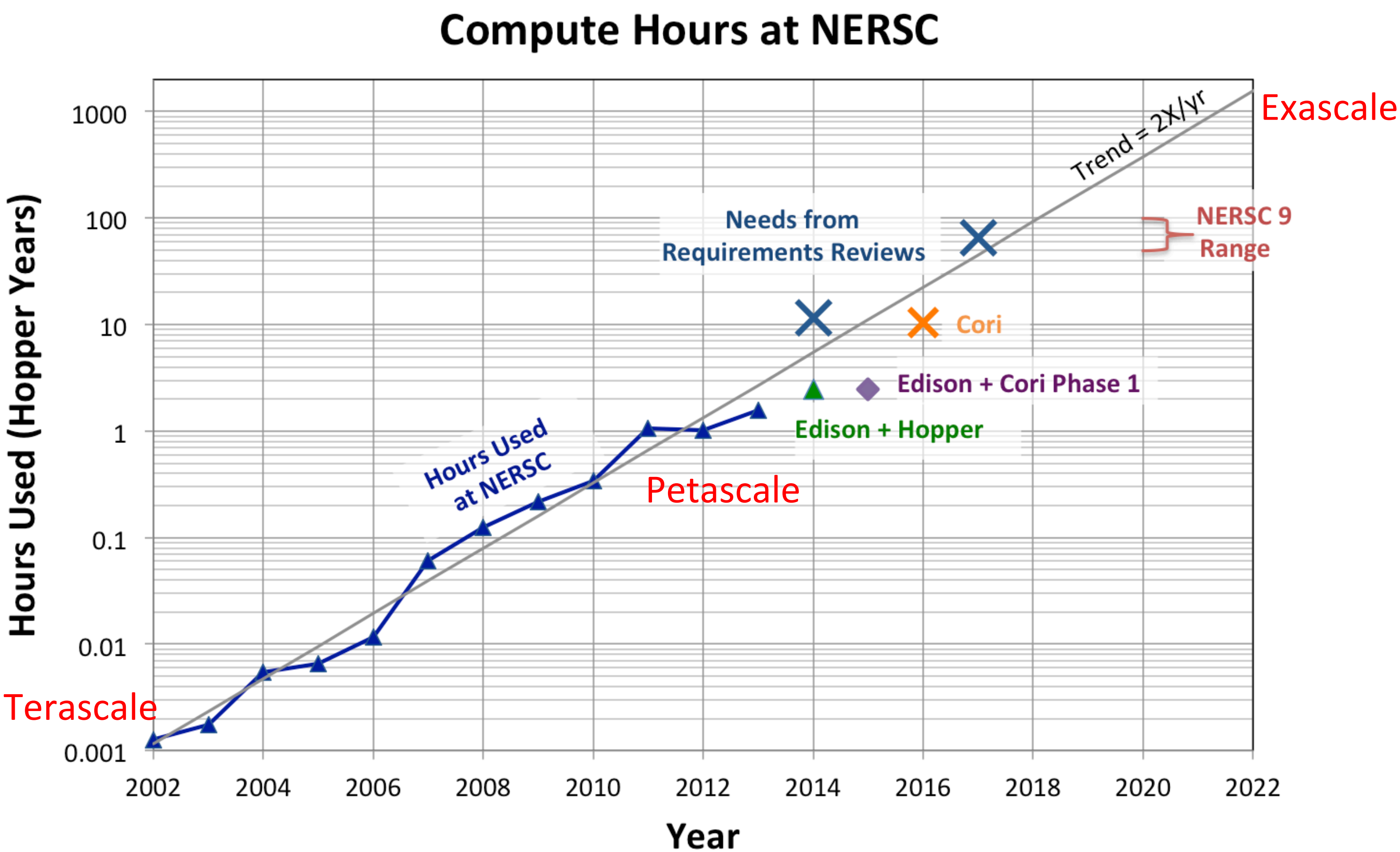
- Participants
  - DOE ADs & Program Managers
  - Leading NERSC users & key potential users
  - NERSC staff & CS Experts

Scientific Objectives

↓

Computing, Storage, Software, Services Requirements

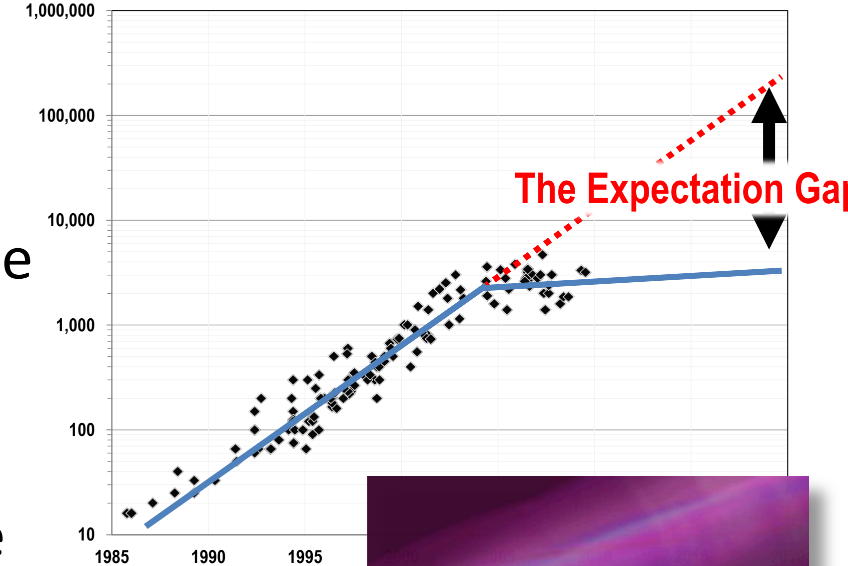
Keeping up with user needs will be a challenge



NERSC 40 YEARS at the FOREFRONT

Challenges to Exascale

1) **Power** is the primary constraint
2) **Parallelism** (1000x today)
3) **Processor architecture** will change
4) **Data movement** dominates
5) **Memory** growth will not keep up
6) **Programming models** will change
7) **Algorithms** to minimize data movement
8) **I/O** performance will not keep up
9) **Resilience** will be critical at this scale
10) **Interconnect bisection** must scale



Exascale Strategic Objective

- **Meet the ever-growing computing 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

The Cori System: Pre-exascale computing and big data capabilities

- **Cori will support the broad Office of Science research community transition to advanced architectures**
- **Cray XC system with over 9300 Intel Knights Landing compute nodes (Intel Xeon Phi or MIC)**
  - Self-hosted, (not an accelerator) manycore processor with over 60 cores per node
  - On-package high-bandwidth memory
- **Data Partition**
  - NVRAM Burst Buffer to accelerate data intensive applications
  - 28 PB of disk, 432 GB/sec I/O bandwidth
- **NERSC Exascale Science Applications Program will prepare users for Cori**
  - Outreach and training for user community
  - Application deep dives with Intel and Cray
  - 8 post-docs integrated with key application teams

Image source: Wikipedia

System named after Gerty Cori, Biochemist and first American woman to receive the Nobel prize in science.

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

NERSC 40 YEARS at the FOREFRONT

- **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**

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™
- 512b vector units (32 flops/clock – AVX 512)
- 3X single-thread performance over current generation Xeon Phi co-processor (KNC)
- 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 ??

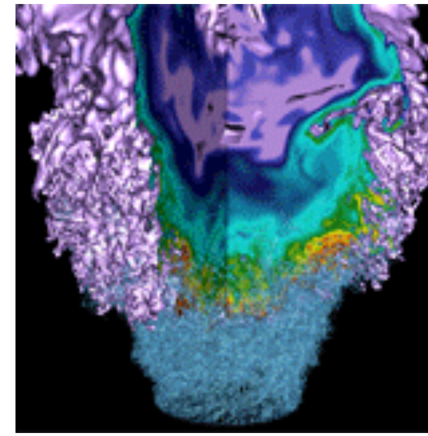
We will partner closely with Cray and Intel

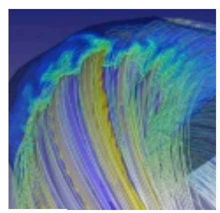
- **Cray**
  - 5 FTE years of application and optimization support
- **Intel**
  - Remote access to an early KNL system
  - KNL white boxes @ NERSC before arrival of N8
  - 4 Training sessions – 2 per year
  - Quarterly Dungeon sessions – 16 in total
  - Intel associate on-site 1 week/month for 4 years

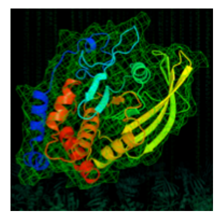
We Launched the NERSC Exascale Science Application Program

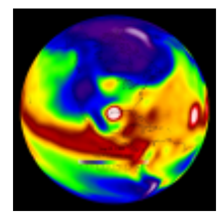
- Umbrella program for all NERSC Application Readiness Activities
- 20 application teams were accepted into NESAP
- Each application team is partnered with a member of NERSC's App Readiness team who will assist with code profiling and scaling analyses
- Through this program NERSC will allocate resources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

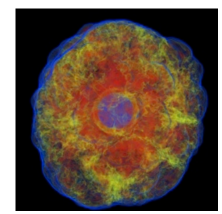
Extreme Data Science

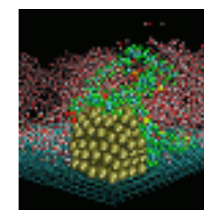








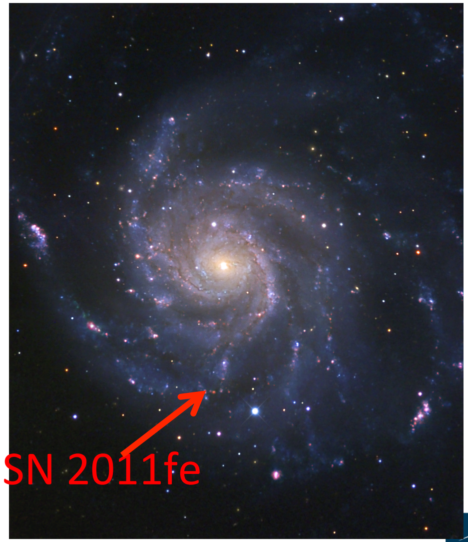




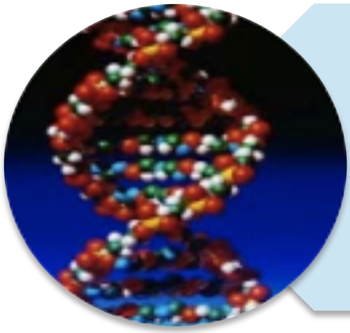
NERSC 40 YEARS at the FOREFRONT

Data is Playing a Key Role in Scientific Discovery

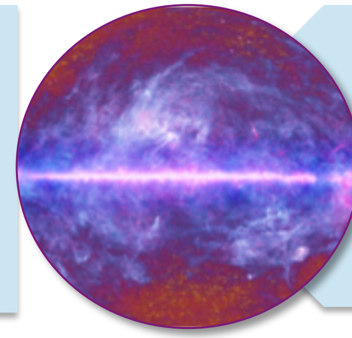
- Discovery of the Higgs Boson
- Measurement of the important "θ_{13}" neutrino parameter
- The Palomar Transient Factory discovered over 2000 supernovae in the last 5 years, including the youngest and closest Type Ia supernova in past 40 years
- Trillions of measurements by the Planck satellite led to the most detailed maps ever of cosmic microwave background
- Four of Science Magazine's breakthroughs of the last decade were in genomics
- Materials project has over 5,000 users and was featured on the cover of Scientific American



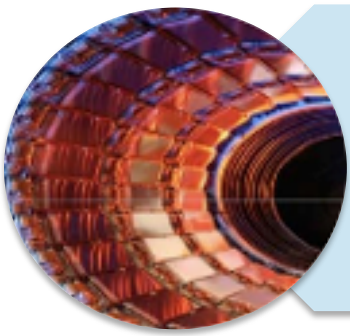
DOE "Big Data" Challenges
Volume, velocity, variety, and veracity



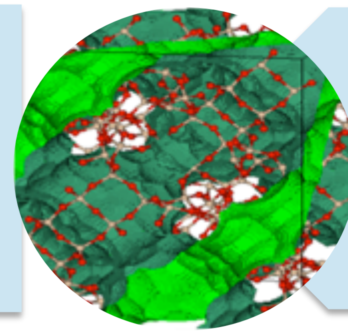
Biology
- *Volume:* Petabytes now; computation-limited
- *Variety*: multi-modal analysis on bioimages



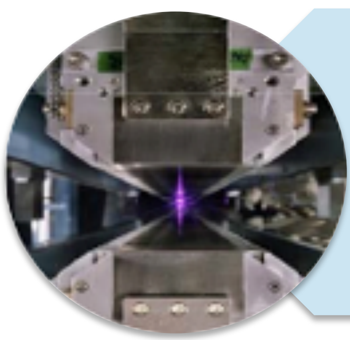
Cosmology & Astronomy:
- *Volume:* 1000x increase every 15 years
- *Variety:* combine data sources for accuracy



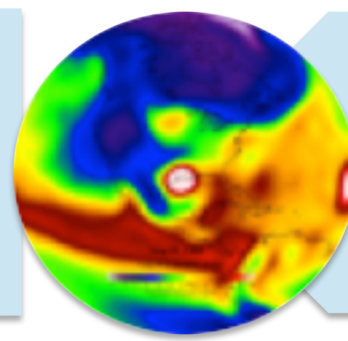
High Energy Physics
- *Volume*: 3-5x in 5 years
- *Velocity*: real-time filtering adapts to intended observation



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

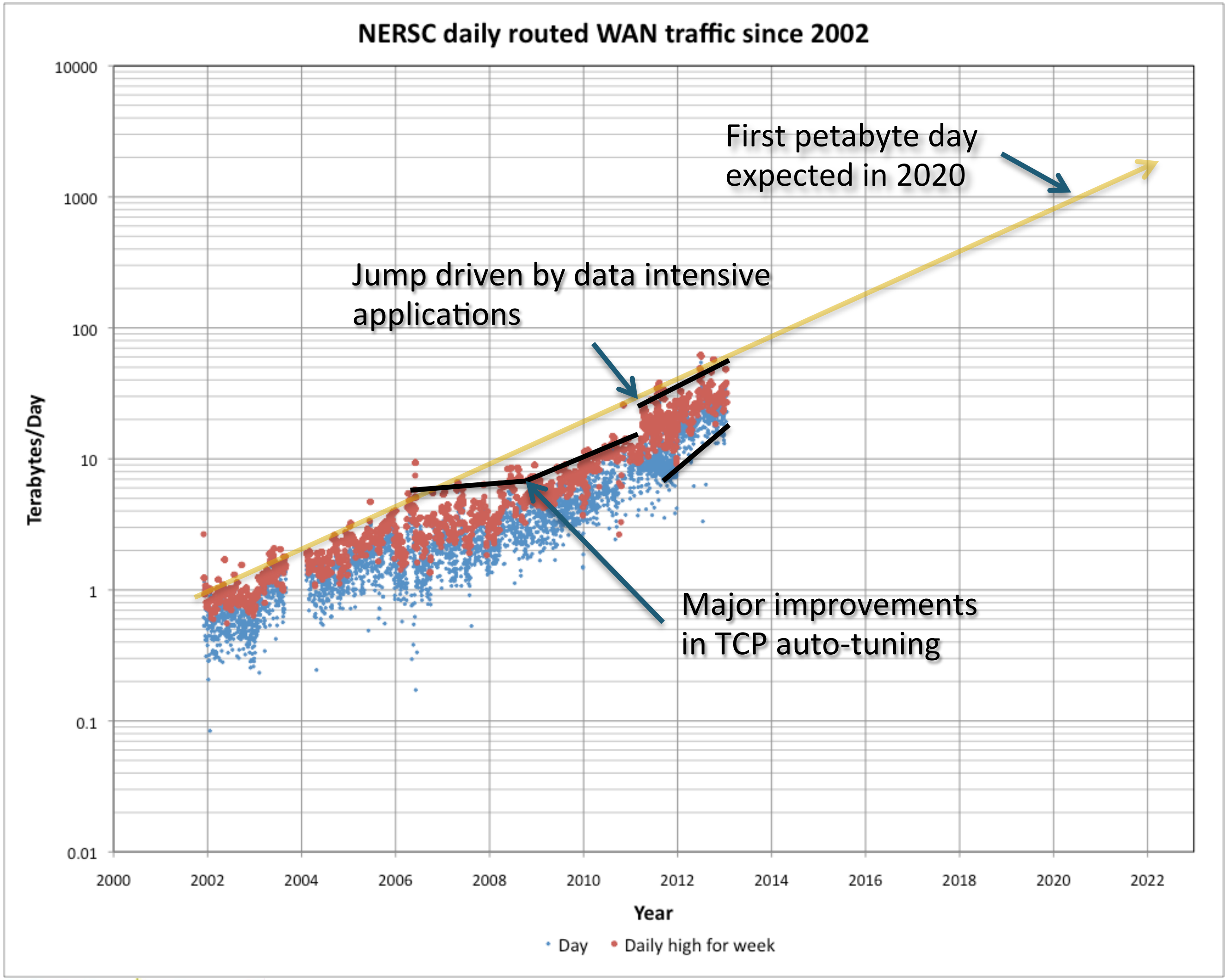


Light Sources
- *Velocity:* CCDs outpacing Moore's Law
- *Veracity:* noisy data for 3D reconstruction

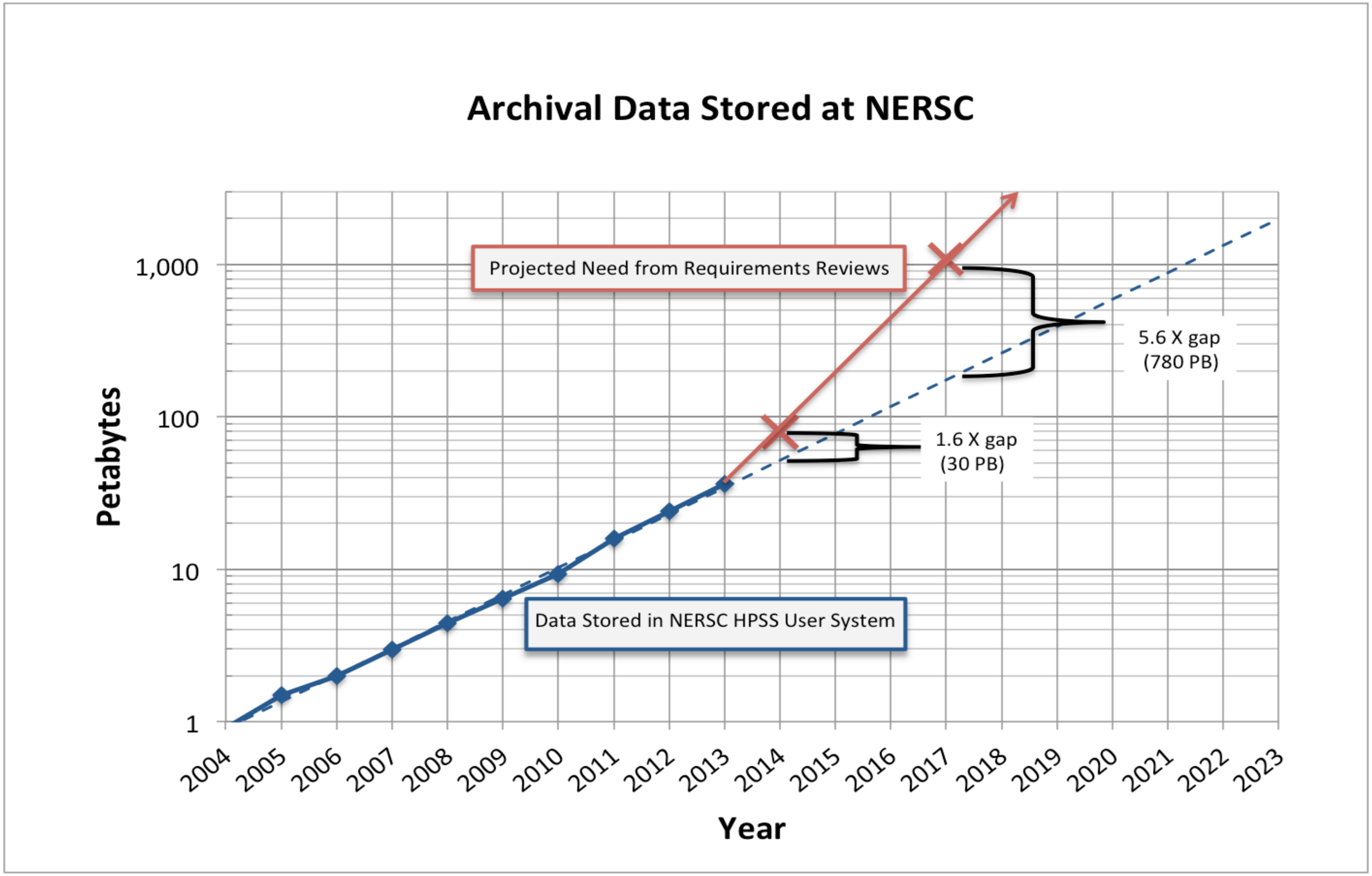


Climate
- *Volume:* Hundreds of exabytes by 2020
- *Veracity:* Reanalysis of 100-year-old sparse data

Exponentially increasing data traffic



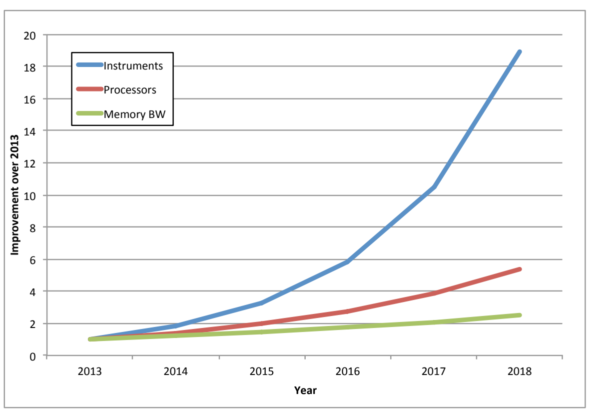
Future archival storage needs

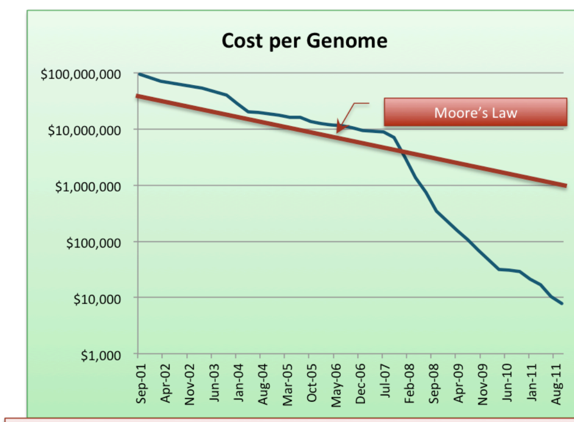


The data deluge will only get worse

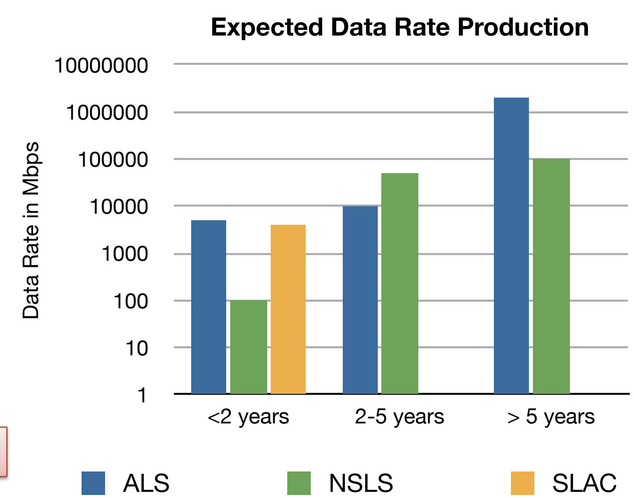
- The observational dataset for the Large Synoptic Survey Telescope will be ~100 PB
- The Daya Bay project will require simulations which will use over 128 PB of aggregate memory

- 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



Extreme Data Strategy

- Develop and deploy new data resources and capabilities
- Partner with DOE experimental facilities and projects to identify requirements and create early success
- Provide expertise and services for extreme data
- Leverage ESnet and DOE Advanced Scientific Computing Research to create end-to-end solutions

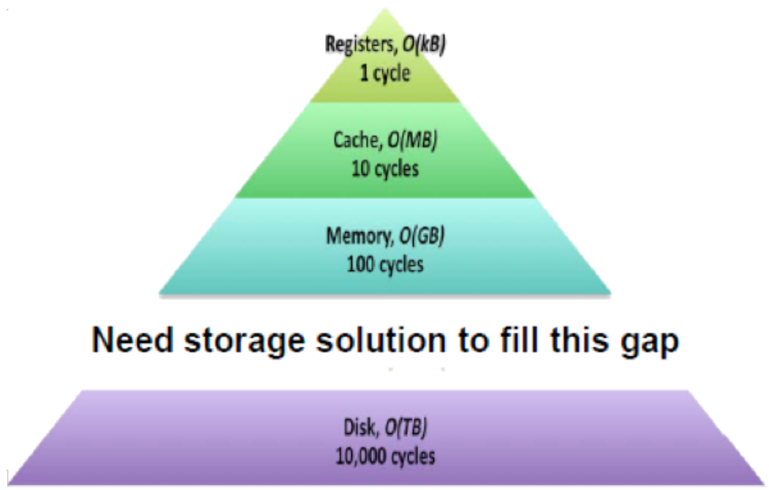
Cori Data Enhancements

- **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.

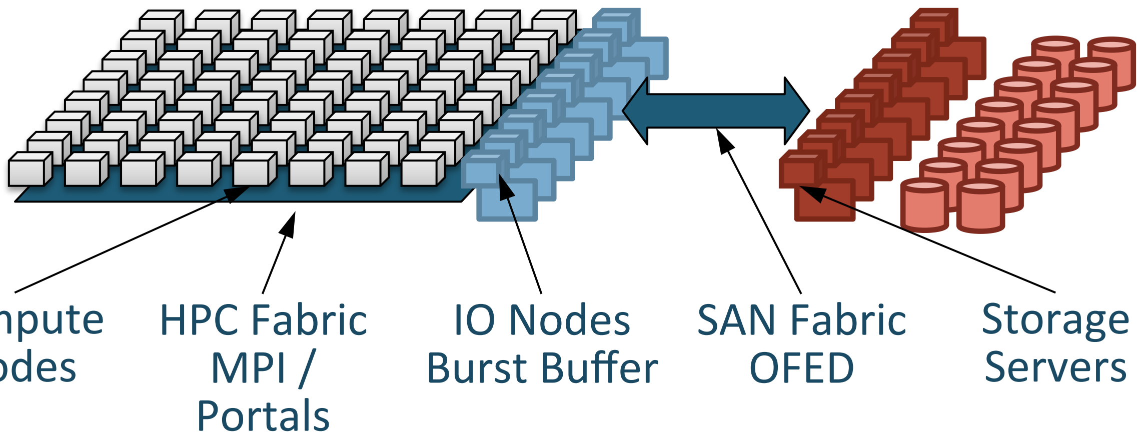
Burst Buffer

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



- **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).**

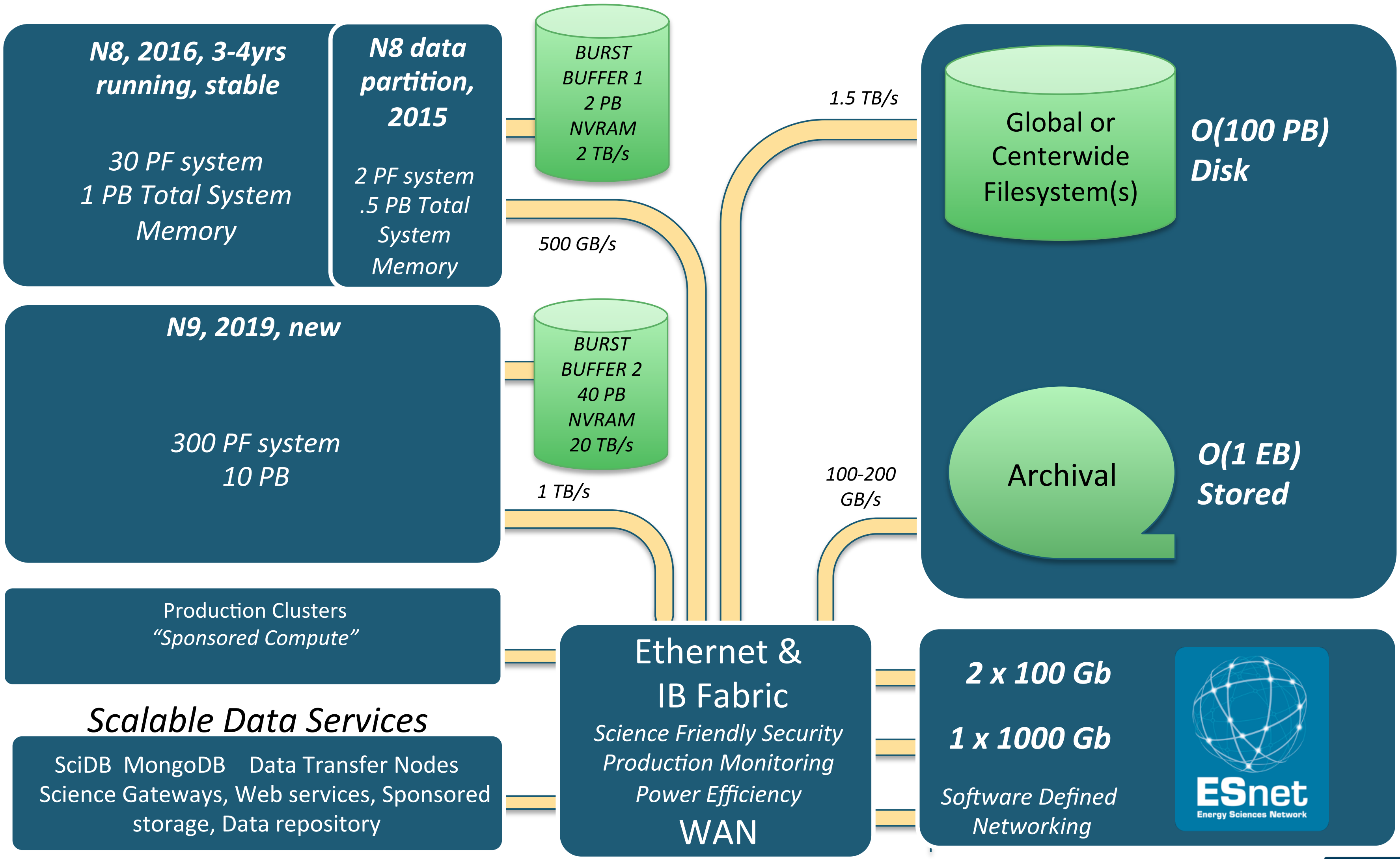
Burst Buffer Software NRE Efforts



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

NERSC 2020



N8, 2016, 3-4yrs running, stable
30 PF system
1 PB Total System Memory
N8 data partition, 2015
2 PF system
.5 PB Total System Memory
BURST BUFFER 1
2 PB
NVRAM
2 TB/s
500 GB/s
1.5 TB/s
Global or Centerwide Filesystem(s)
O(100 PB) Disk
N9, 2019, new
300 PF system
10 PB
BURST BUFFER 2
40 PB
NVRAM
20 TB/s
1 TB/s
100-200 GB/s
Archival
O(1 EB) Stored
Production Clusters
"Sponsored Compute"
Scalable Data Services
SciDB MongoDB Data Transfer Nodes
Science Gateways, Web services, Sponsored storage, Data repository
Ethernet &
IB Fabric
Science Friendly Security
Production Monitoring
Power Efficiency
WAN
2 x 100 Gb
1 x 1000 Gb
Software Defined Networking
ESnet
Energy Sciences Network

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

Thank you.