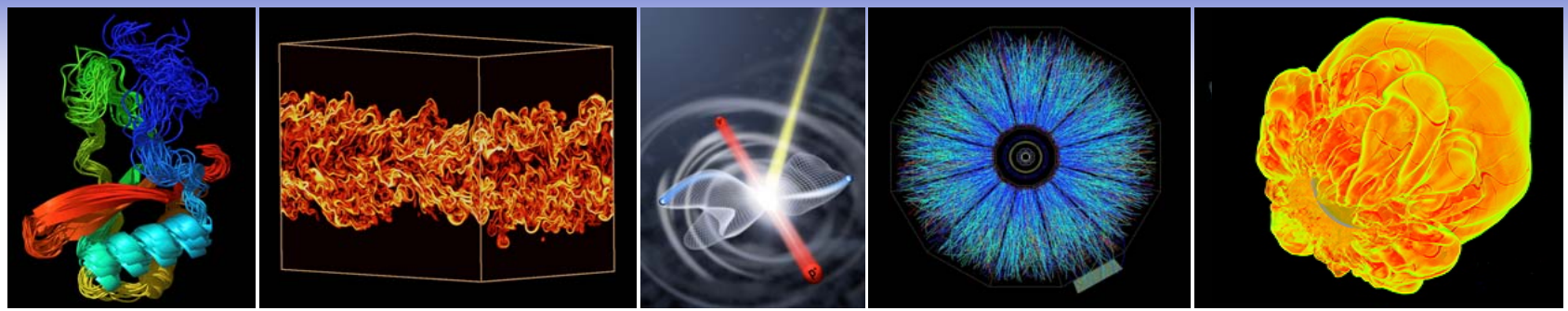


Science at NERSC

Katherine Yelick
NERSC Director

Office of Science
U.S. DEPARTMENT OF ENERGY

NERSC

BERKELEY LAB

NERSC Mission

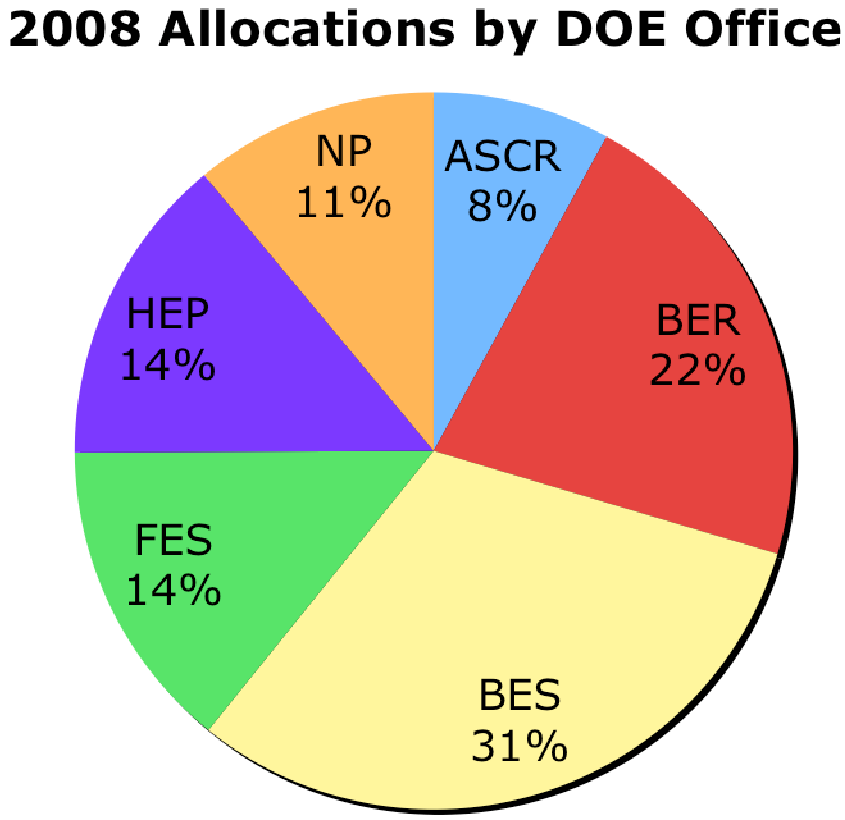
The mission of the National Energy Research Scientific Computing Center (NERSC) is to *accelerate the pace of scientific discovery* by providing high performance computing, information, data, and communications services for *all DOE Office of Science (SC) research.*

NERSC is the Production Facility for DOE SC

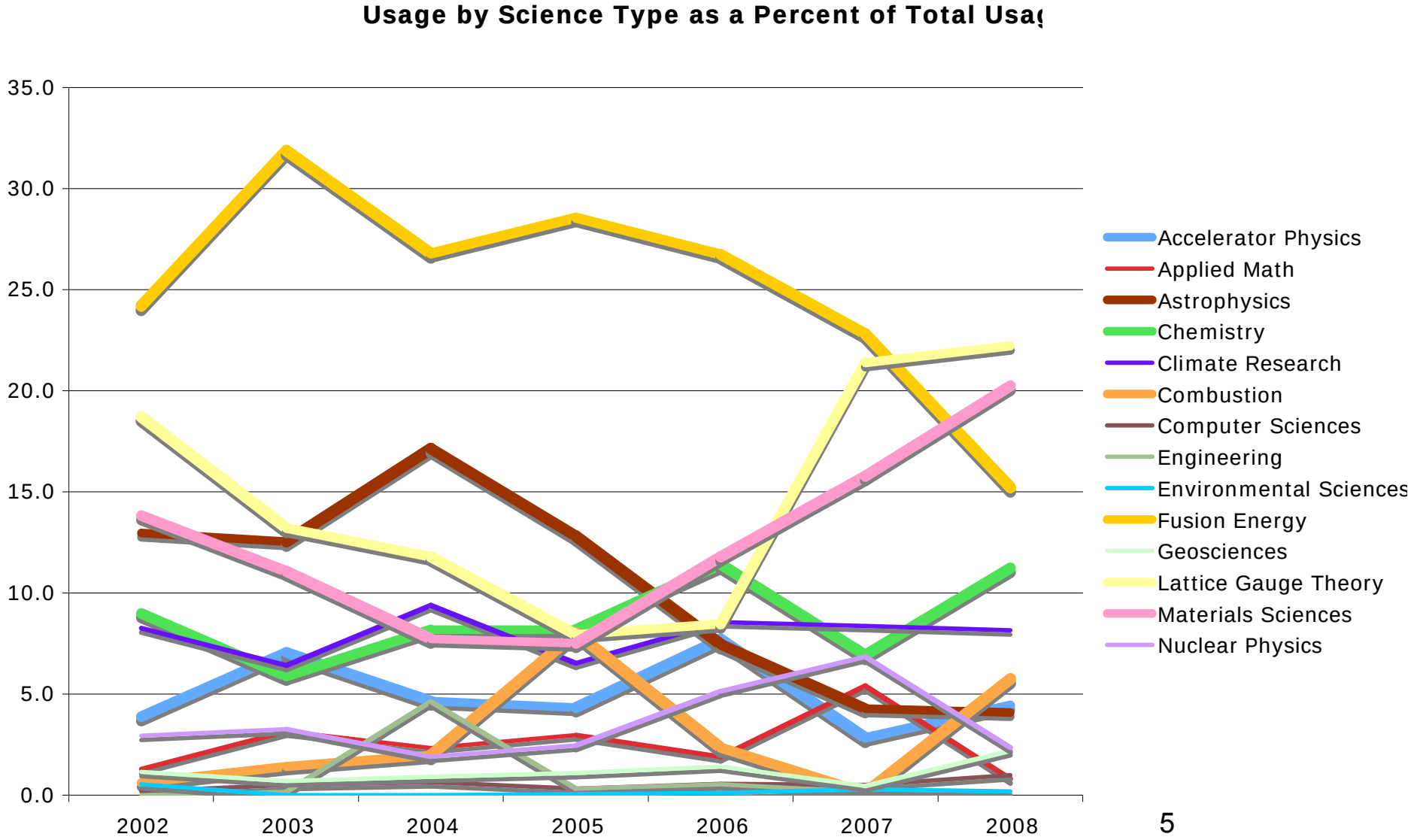
- **NERSC serves a large population of users**
  - **~3000 users, ~400 projects, ~500 codes**
- **Allocations managed by DOE**
  - **10% INCITE awards:**
    - **Created at NERSC; now used throughout SC**
    - **Open to all of science, not just DOE or DOE/SC mission**
    - **Large allocations, extra service**
  - **70% Production (ERCAP) awards:**
    - **From 10K hour (startup) to 5M hour; Only at NERSC, not LCFs**
  - **10% each NERSC and DOE/SC reserve**
- **Award mixture offers**
  - **High impact through large awards**
  - **Broad impact across science domains**

NERSC Serves DOE Mission Needs

- **DOE's SciDAC Program**
  - Brings together interdisciplinary teams
  - 55 NERSC projects tied to SciDAC
- **Focus on high end computing**
  - DOE/OMB measure of concurrency: Percent of time spent on jobs ≥ 1/8th of the machine
  - Set to 40% in 2008
- **Aside: Mid-range computing workshop being organized**



NERSC Serves Broad and Varying DOE Science Priorities



NERSC 2008 Configuration

Large-Scale Computing System

Franklin (NERSC-5): Cray XT4

- 9,740 nodes; 19,480 cores
- 13 Tflop/s sustained SSP (100 Tflops/s peak)

Upgrading to QuadCore

- ~25 Tflops/s sustained SSP (355 Tflops/s peak)

NERSC-6 planned for 2010 production

- 3-4x NERSC-5 in application performance

Clusters

Bassi (NCSb)

- IBM Power5 (888 cores)

Jacquard (NCSa)

- LNXI Opteron (712 cores)

PDSF (HEP/NP)

- Linux cluster (~1K cores)

NERSC Global Filesystem (NGF)

230 TB; 5.5 GB/s

HPSS Archival Storage

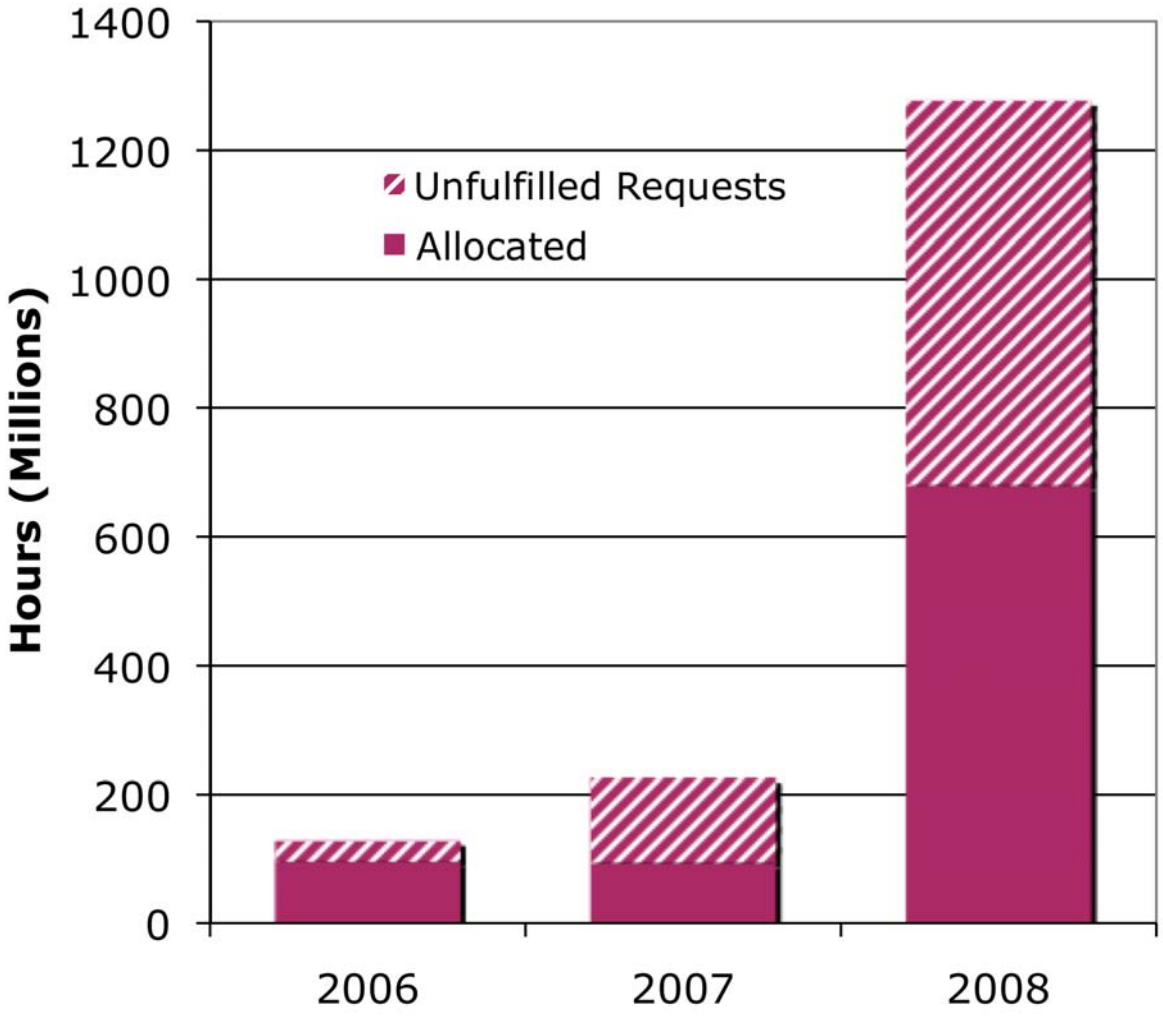
- 44 PB capacity
- 10 Sun robots
- 130 TB disk cache

Analytics / Visualization

- Davinci (SGI Altix)

DOE Demand for Computing is Growing

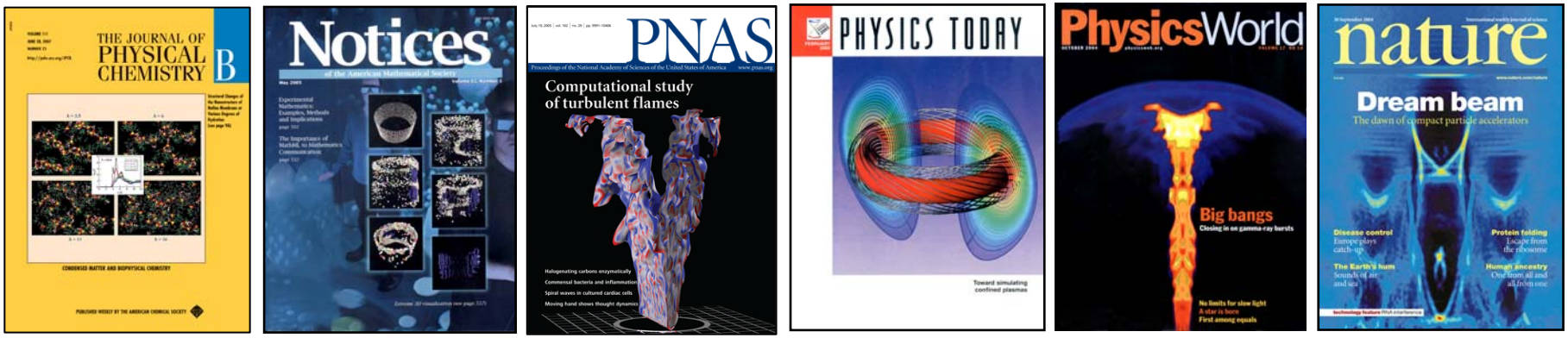
Compute Hours Requested vs Allocated



- *Each year DOE users requests 2x as many hours as can be allocated*
- *This 2x is artificially constrained by perceived availability*
- *Unfulfilled allocation requests amount to hundreds of millions of compute hours in 2008*

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Science Over the Years



NERSC is enabling new science in all disciplines, with over *1,500 refereed publications* in 2007

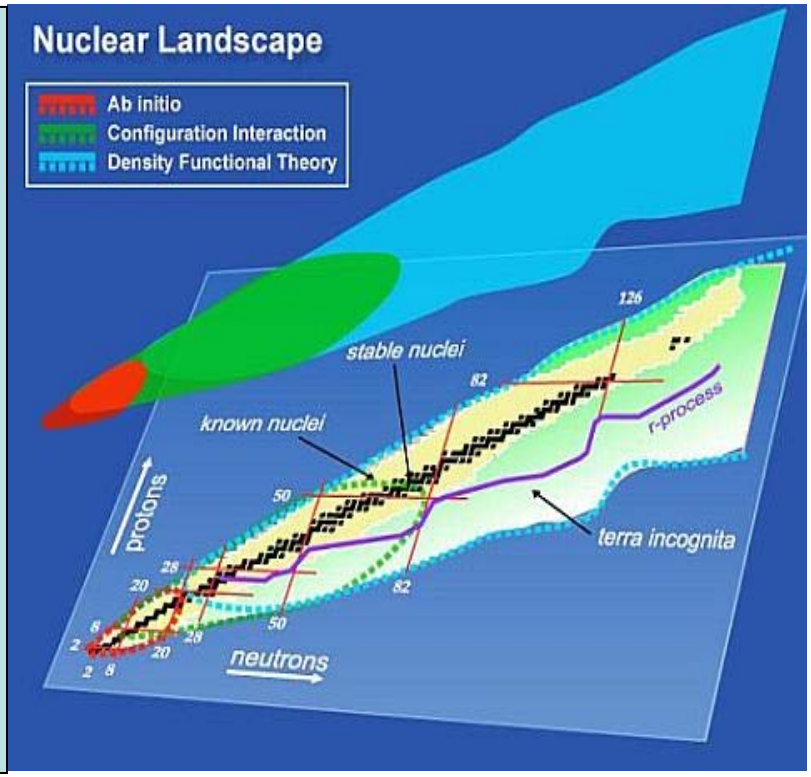
Office of Science U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Nuclear Physics

- **Calculation: High accuracy *ab initio* calculations on O^{16} using no-core shell model and no-core full configuration interaction model**
- **PI: James Vary, Iowa State**

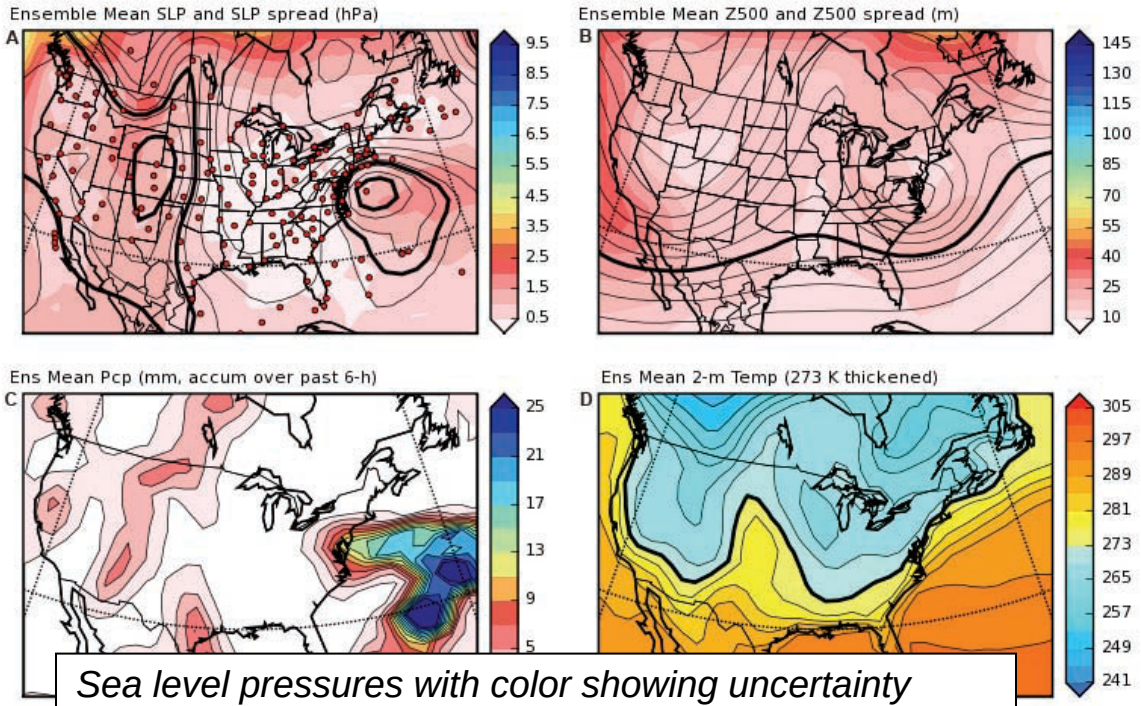
- **Science Results:**
  - **Most accurate calculations to date on this size nuclei**
  - **Can be used to parametrize new density functionals for nuclear structure simulations**
- **Scaling Results:**
  - **4M hours used; 200K allocated**
  - **12K cores; vs 2-4K before Franklin uncharged time**
  - **Diagonalize matrices of dimension up to 1 billion**



Validating Climate Models

- **INCITE Award for "20th Century Reanalysis" using an Ensemble Kalman filter to fill in missing climate data since 1892**
- **PI: G. Compo, U. Boulder**

- **Science Results:**
  - **Reproduced 1922 Knickerbocker storm**
  - **Data can be used to validate climate and weather models**
- **Scaling Results:**
  - **3.1M CPU Hours in allocation**
  - **Scales to 2.4K cores**
  - **Switched to higher resolution algorithm with Franklin access**

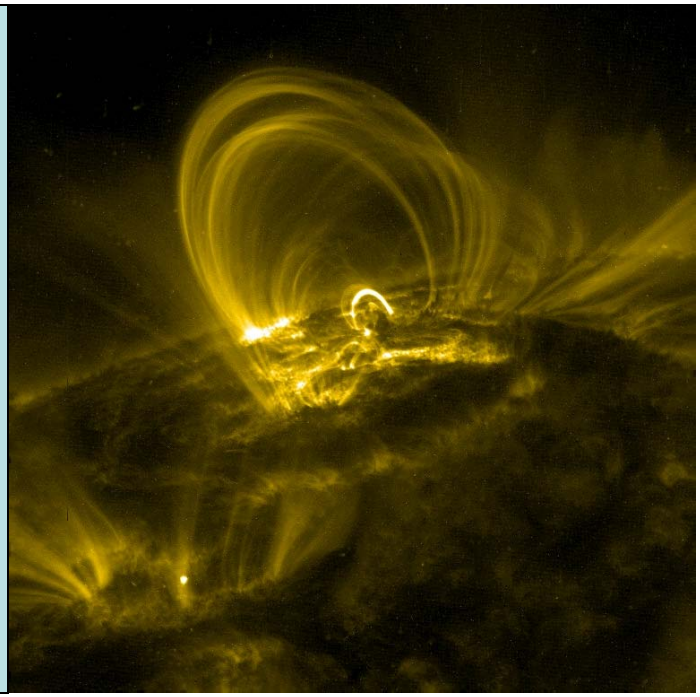


Sea level pressures with color showing uncertainty (a&b); precipitation (c); temperature (d). Dots indicate measurements locations (a).

Middle Users Capable Large-Scale Computational Science

- Calculations: AstroGK gyrokinetic code for astrophysical plasmas
- PIs: Dorland (U. of Maryland), Howes, Tatsuno

- Science Results
  - Shows how magnetic turbulence leads to particle heating
- Scaling Results
  - Runs on 16K cores
  - Combines implicit and explicit methods

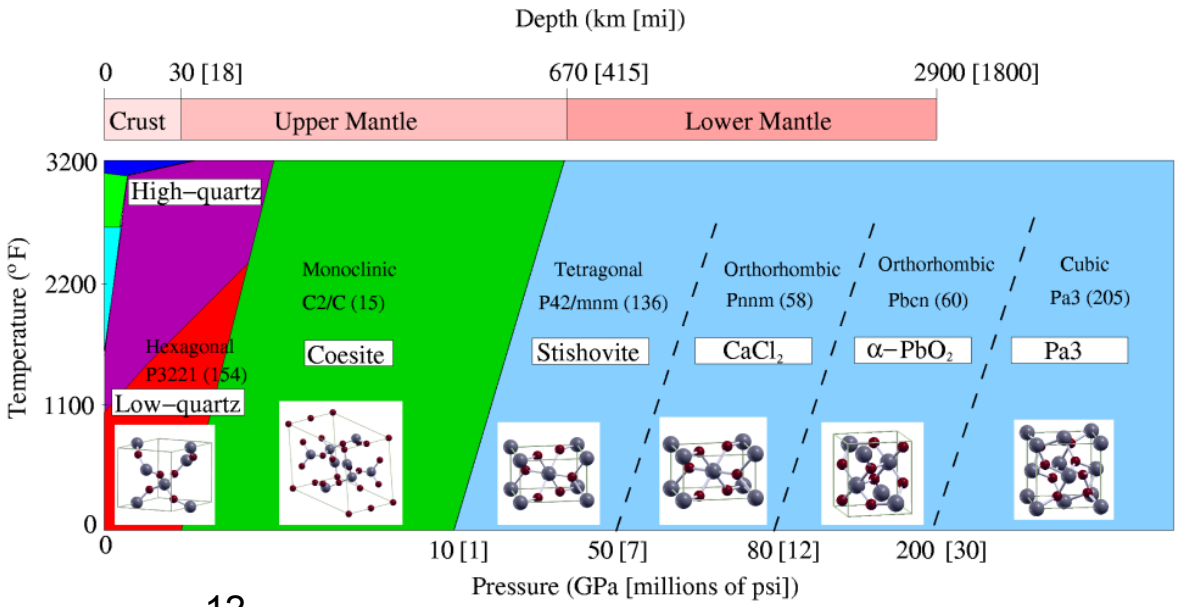


Modeling Dynamically and Spatially Complex Materials for Geoscience

- **Calculation: Simulation of seismic waves through silicates, which make up 80% of the Earth's mantle**
- **PI: John Wilkins, Ohio State University**

- **Science Result**
  - **Seismic analysis shows jumps in wave velocity due to structural changes in silicates under pressure**

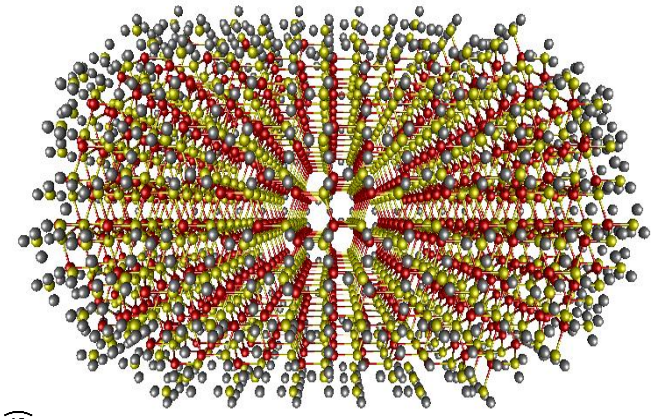
- **Scaling Result**
  - **First use of Quantum Monte Carlo (QMC) for computing elastic constants**
  - **8K core vs. 128 on allocated time**

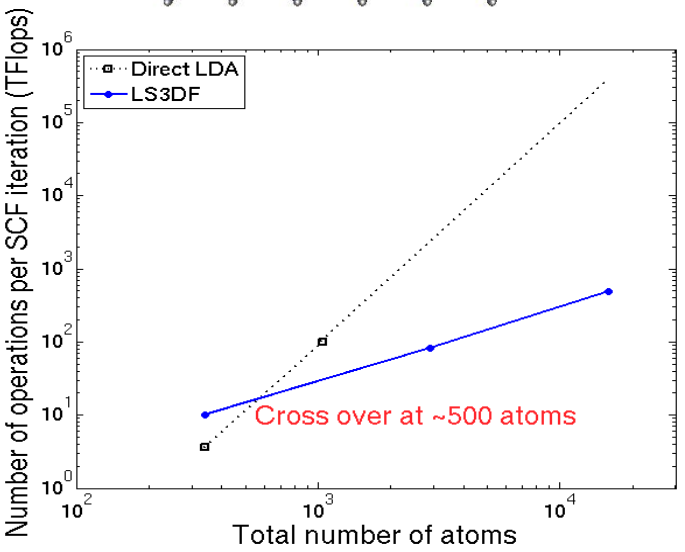


Nanoscience Calculations and Scalable Algorithms

- **Calculation: Linear Scaling 3D Fragment (LS3DF). Density Functional Theory (DFT) calculation numerically equivalent to more common algorithm, but scales with O(n) in number of atoms rather than $O(n^3)$**
- **PI: L.W. Wang, LBNL**

- **Science Results**
  - Calculated dipole moment on 2633 atom CdSe quantum rod, $Cd_{961}Se_{724}H_{948}$.
- **Scaling Results**
  - Ran on 2560 cores
  - Took 30 hours vs many months for $O(n^3)$ algorithm
  - Good parallel efficiency (80% on 1024 relative to 64 procs)





Simulation of a Low Swirl Burner Fueled with Hydrogen

- **Calculation: Numerical simulation of flame surface of an ultra-lean premixed hydrogen flame in a laboratory-scale low-swirl burner. Burner is being developed for fuel-flexible, near-zero-emission gas turbines.**
- **PI: John Bell, LBNL**

Science Result:

- **Detailed transport and chemical kinetics using an adaptive low Mach number algorithm for reacting flow.**

Scaling Results:

- **Adaptive Mesh Refinement used to save memory and time.**
- **Scales to 6K cores, typically run at 2K**
- **Used 2.2M early science hours on Franklin**

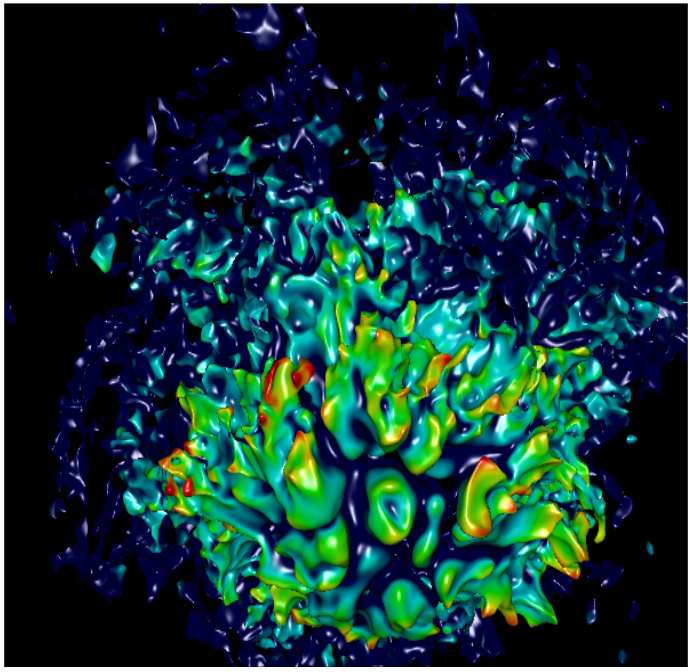
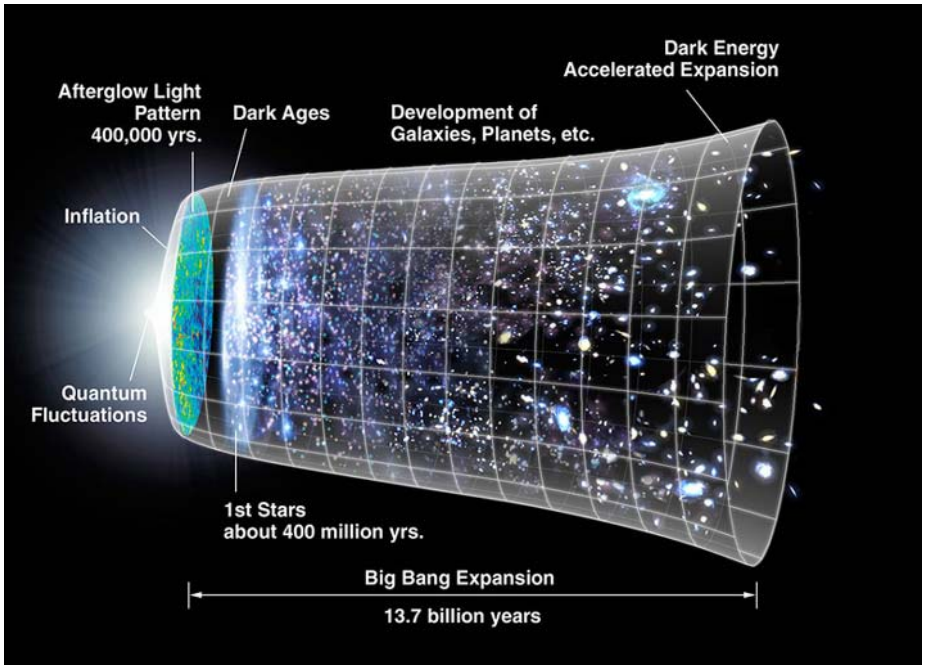


Image illustrates the cellular burning structures in hydrogen flames

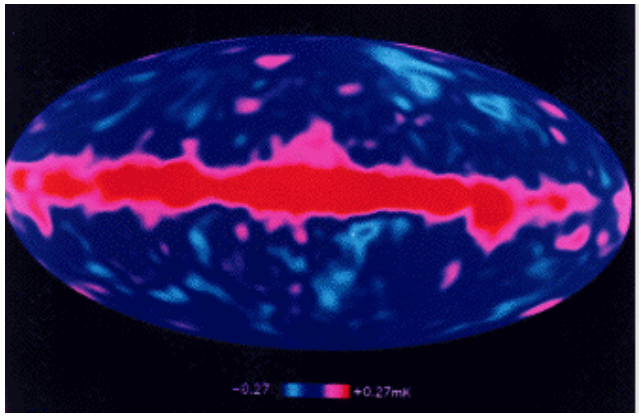
NERSC User George Smoot wins 2006 Nobel Prize in Physics



Cosmic Microwave Background Radiation (CMB): an image of the universe at 400,000 years

Mather and Smoot 1992

COBE Experiment showed anisotropy of CMB

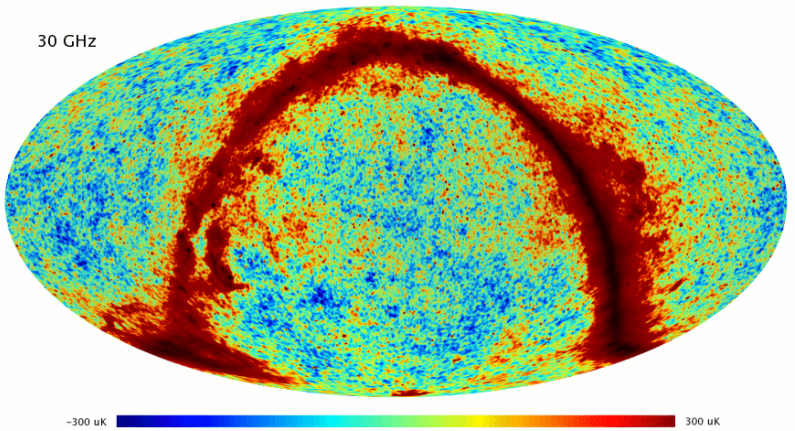


Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Impact Of High Performance Computing at NERSC

- Calculation: Planck full focal plane
  - 1 year simulation of CMB (T & P), detector noise & foregrounds
  - 74 detectors at 9 frequencies
  - 750 billion observations
  - 54,000 files, 3 TB data
- PI: J. Borrill, LBNL

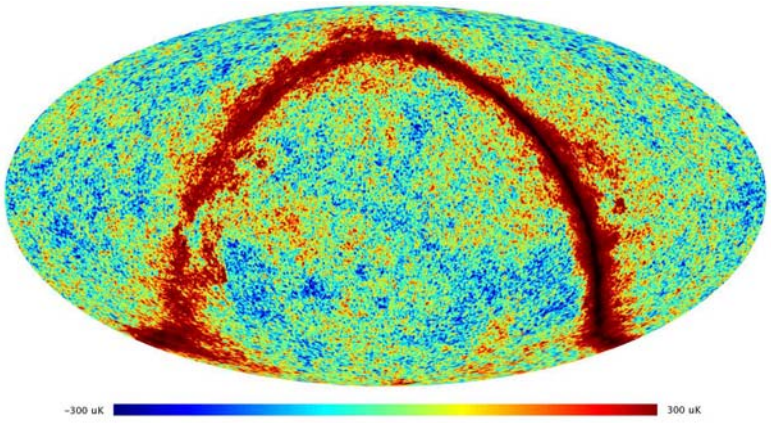


Science Result:

- **9 "routine" 1-frequency maps**
- **Unprecedented 9-frequency map with entire simulated Planck data set**

Scaling Results:

- **9-frequence problem ran for < 1 hour on 16K cores**

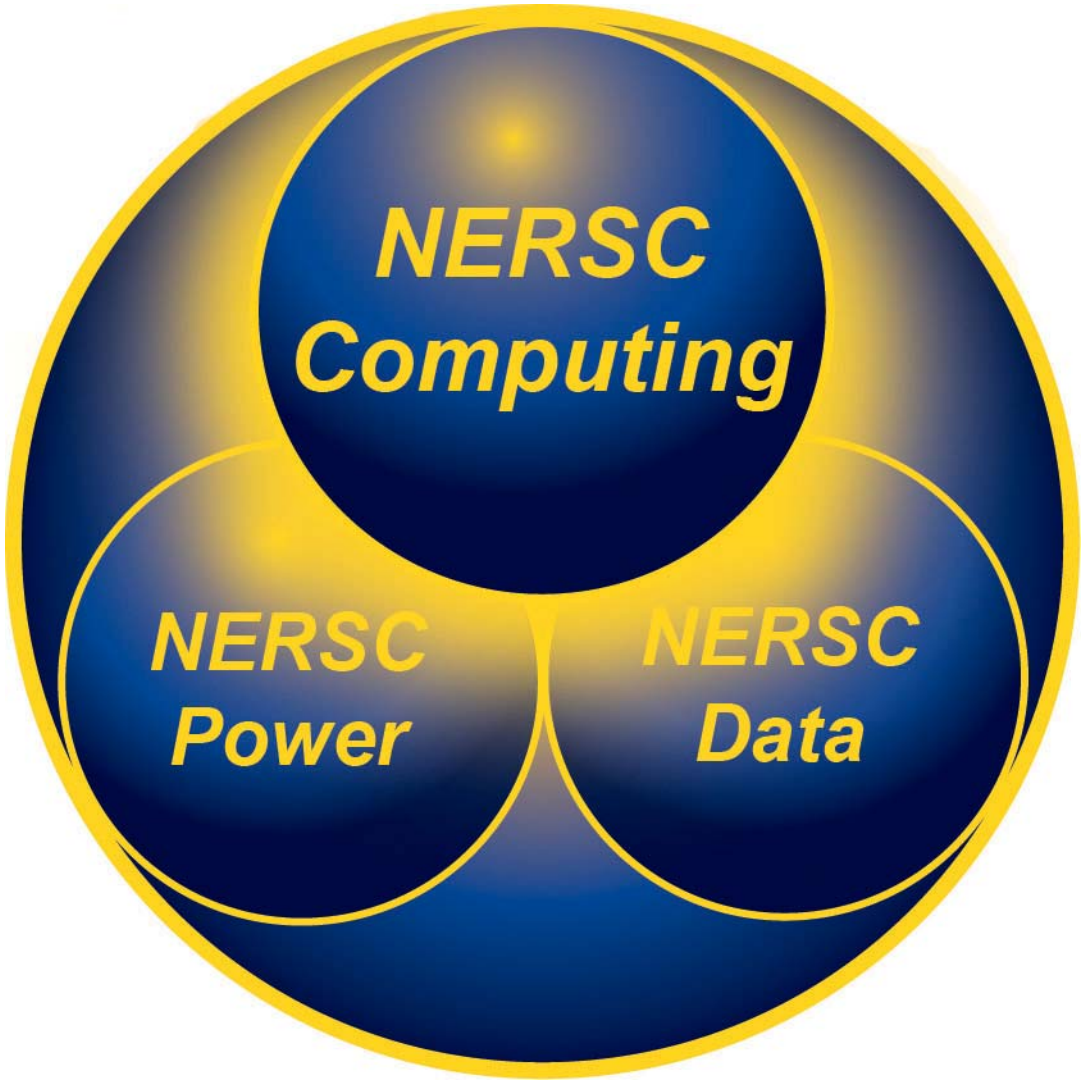


Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

NERSC Vision

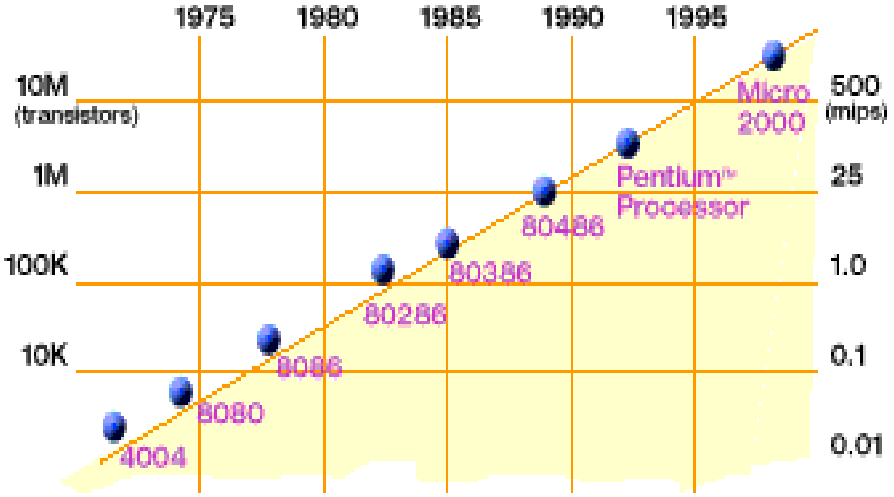
NERSC Computing



New Model for Collecting Requirements

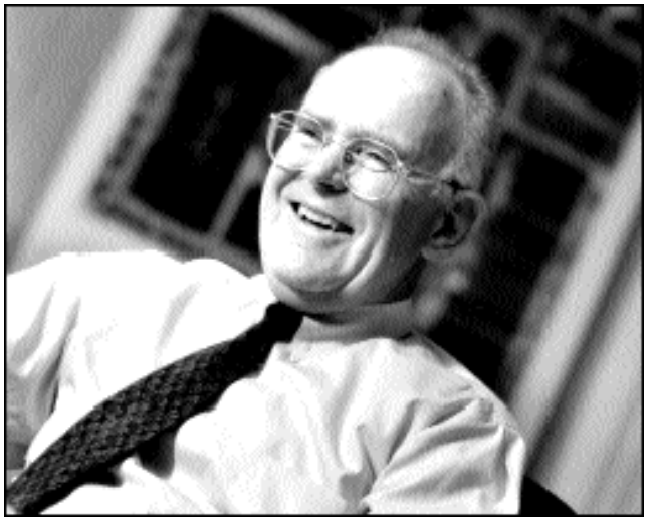
- Modeled after ESnet activity rather than Greenbook
  - Two workshops per year, starting with BER and BES
- Sources of Requirements
  - Office of Science (SC) Program Managers
  - Direct gathering through interaction with science users of the network
    - Case studies, e.g., from ESnet
      - Magnetic Fusion
      - Large Hadron Collider (LHC)
      - Climate Modeling
      - Spallation Neutron Source
  - Observation of the computing use and technology
  - Other requirements
- Requirements aggregation

Moore's Law is Alive and Well



2X transistors/Chip Every 1.5 years
Called "Moore's Law"

Microprocessors have become smaller, denser, and more powerful.

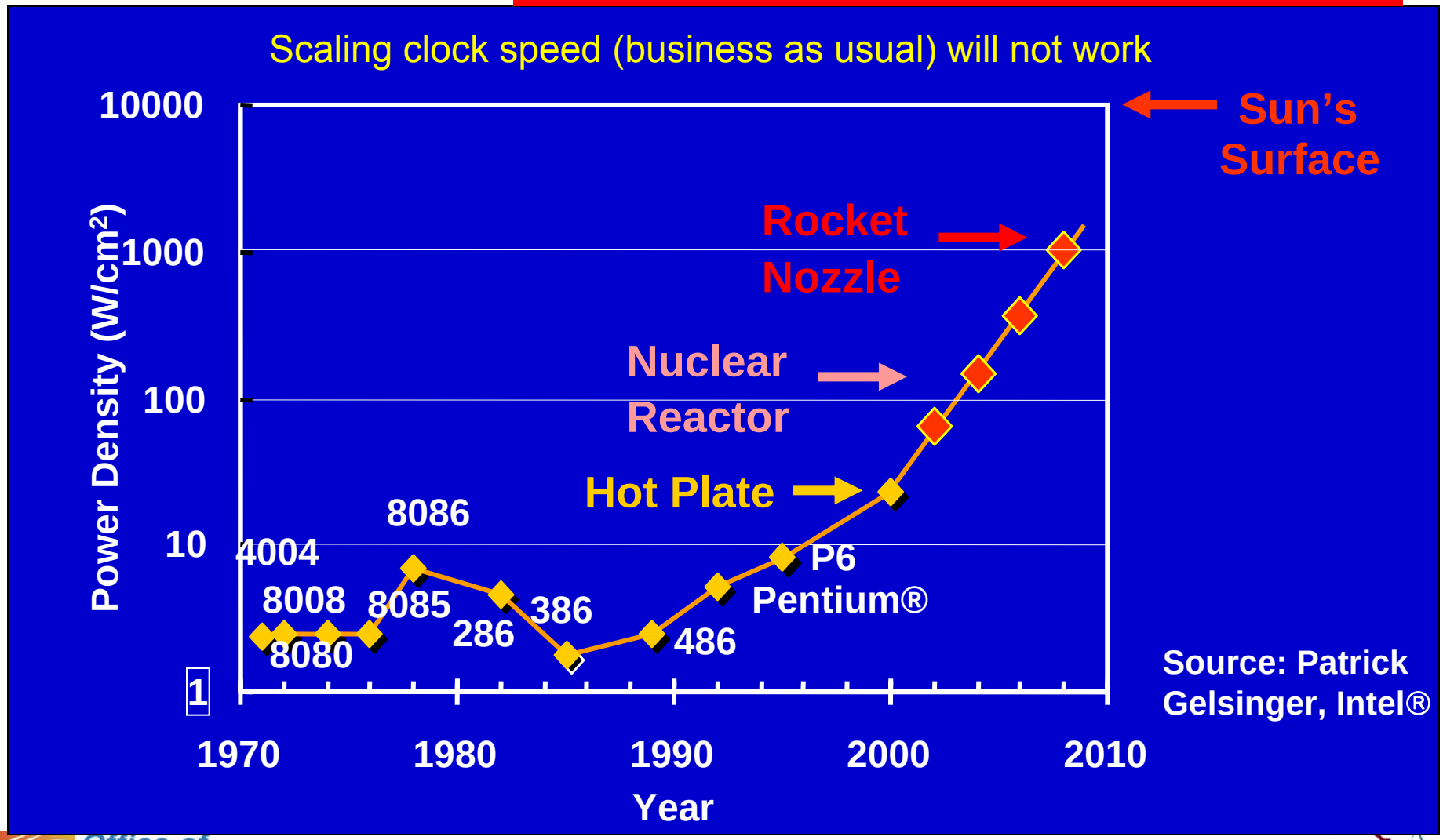


Gordon Moore (co-founder of Intel) predicted in 1965 that the transistor density of semiconductor chips would double roughly every 18 months.

Slide source: Jack Dongarra

New: Power Wall

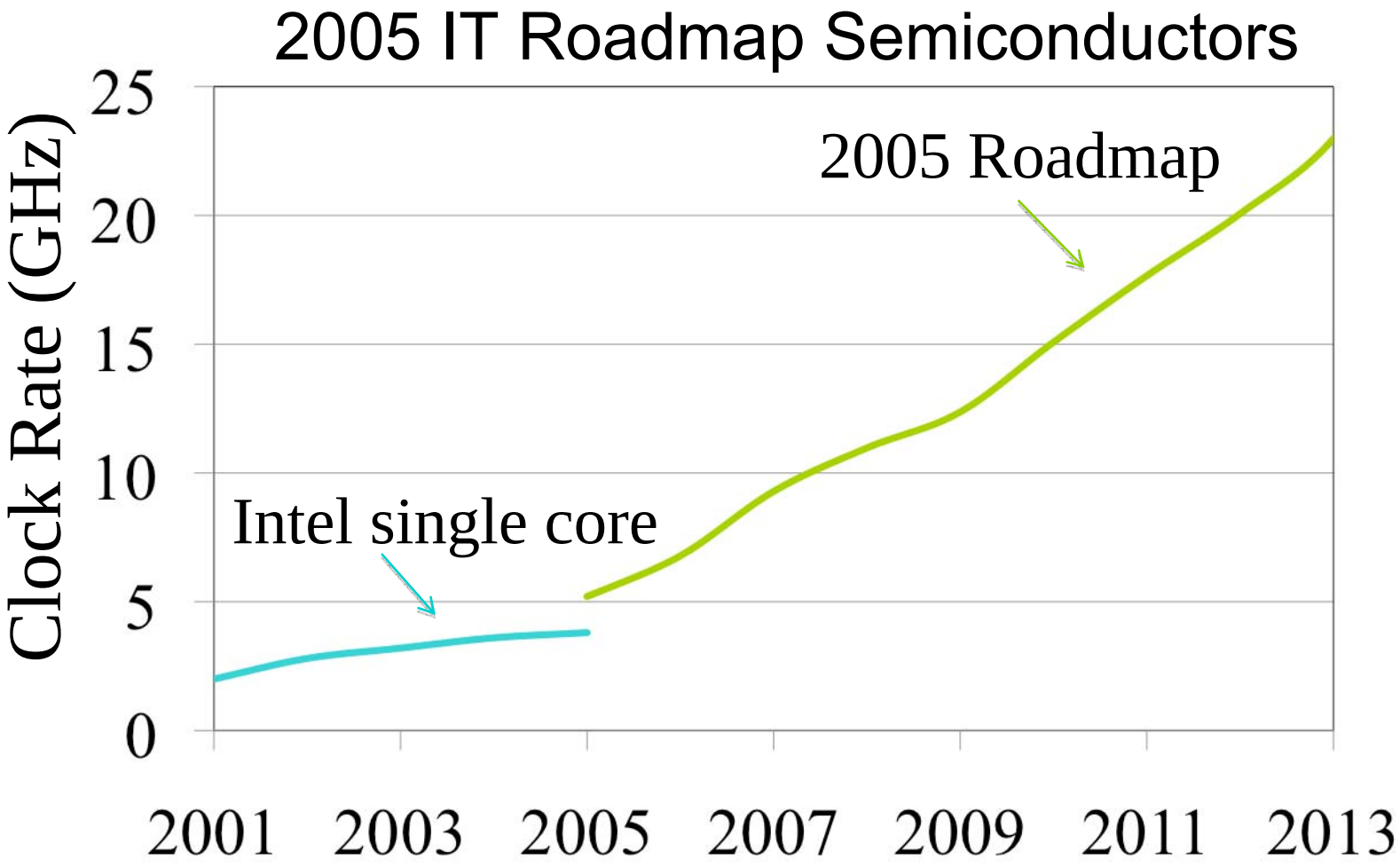
Can put more transistors on a chip than can afford to turn on



Office of Science
U.S. DEPARTMENT OF ENERGY

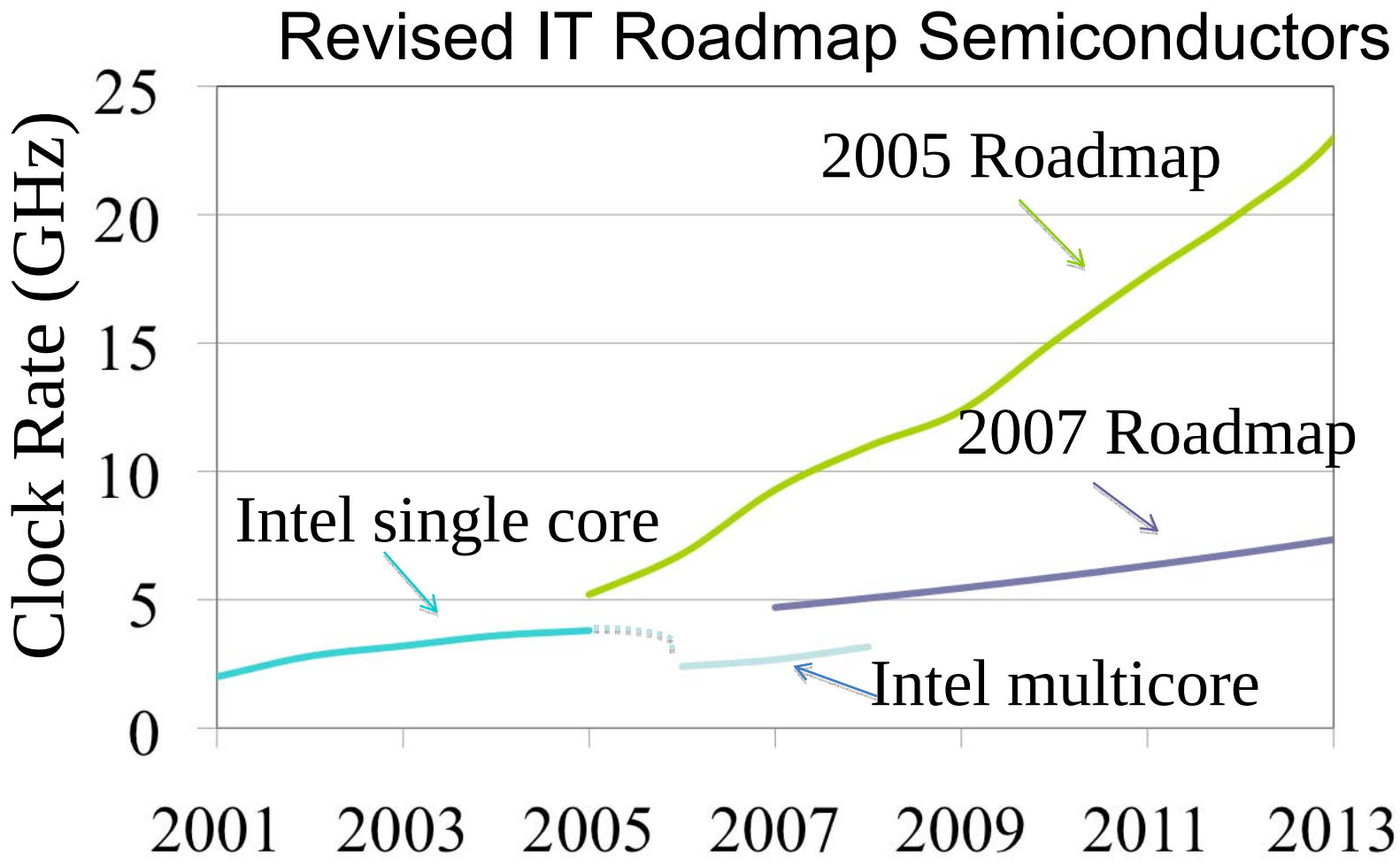
BERKELEY LAB

2005: Clock speed 2x every 2 years



23

2007: Cores/chip 2x every 2 years

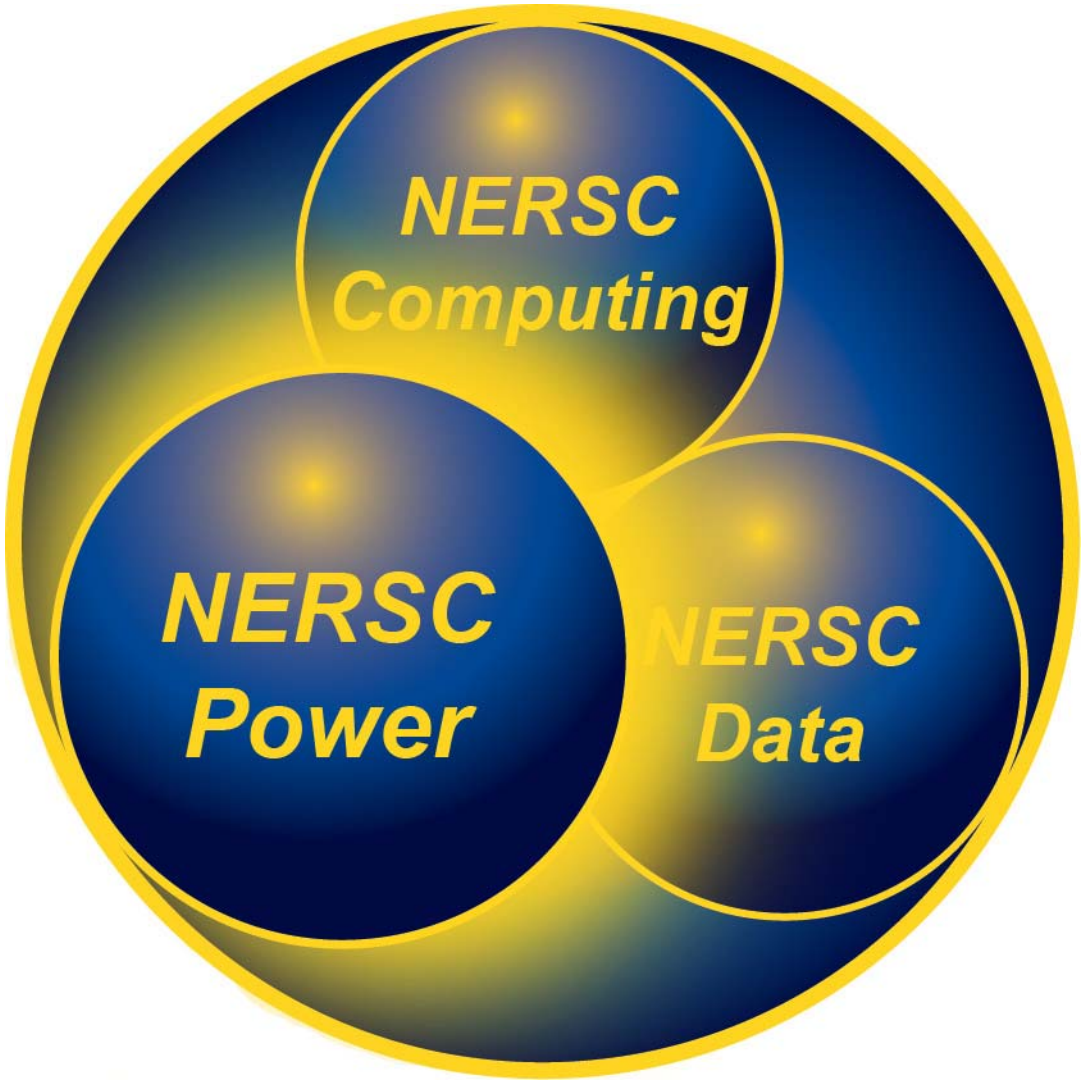


24

Parallelism is "Green"

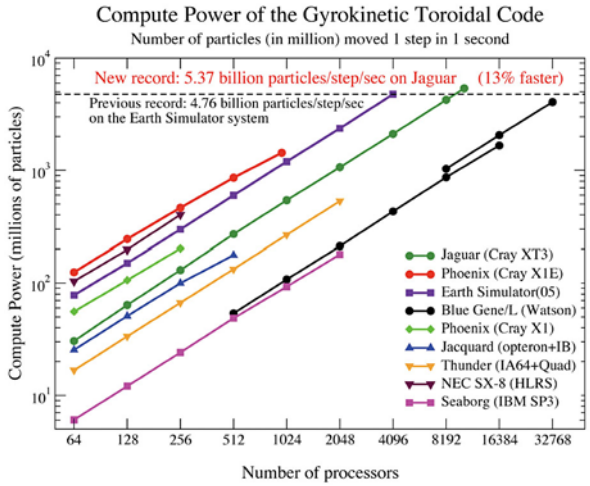
- **Highly concurrent systems are more power efficient**
  - ***Dynamic power is proportional to V^2fC***
  - ***Increasing frequency (f) also increases supply voltage (V): more than linear effect***
  - ***Increasing cores increases capacitance (C) but has only a linear effect***
- **Hidden concurrency burns power**
  - **Speculation, dynamic dependence checking, etc.**
  - **Push parallelism discovery to software (compilers and application programmers) to save power**
- **Challenge: *Can you double the concurrency in your software every 2 years?***

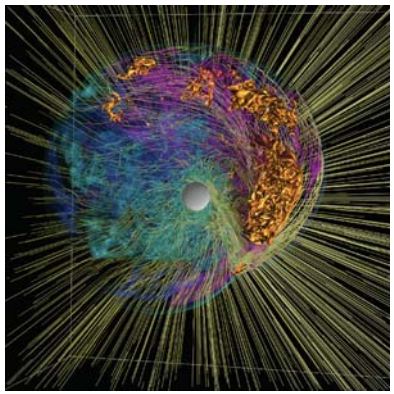
NERSC Power Efficiency

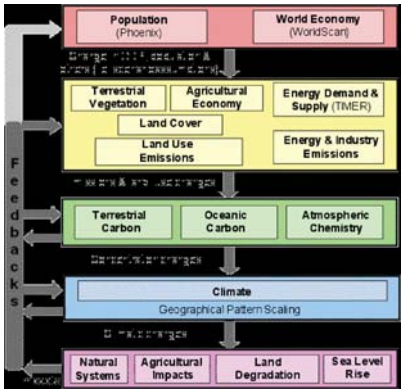


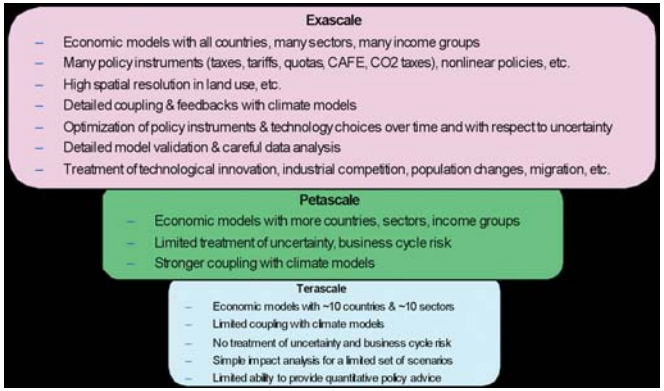
Computational Requirements of the Office of Science Are Clear

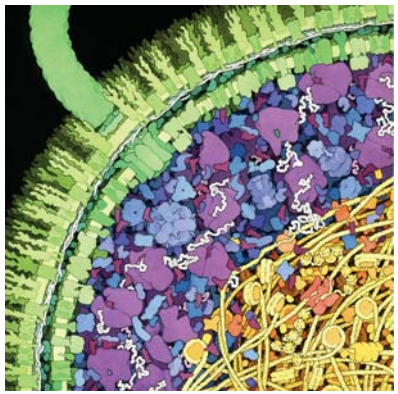
Modeling and Simulation at the Exascale for Energy and the Environment has significant requirements for Exascale

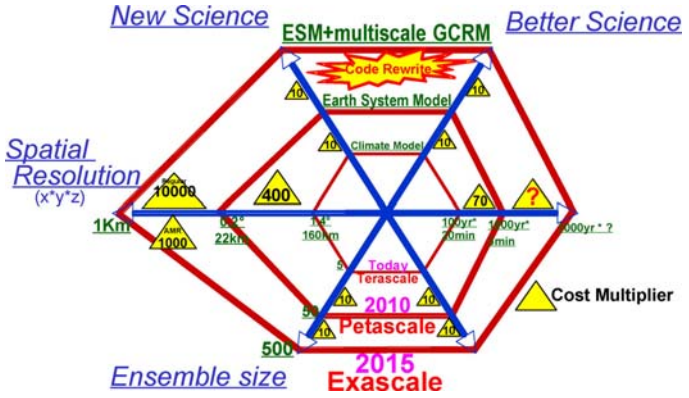










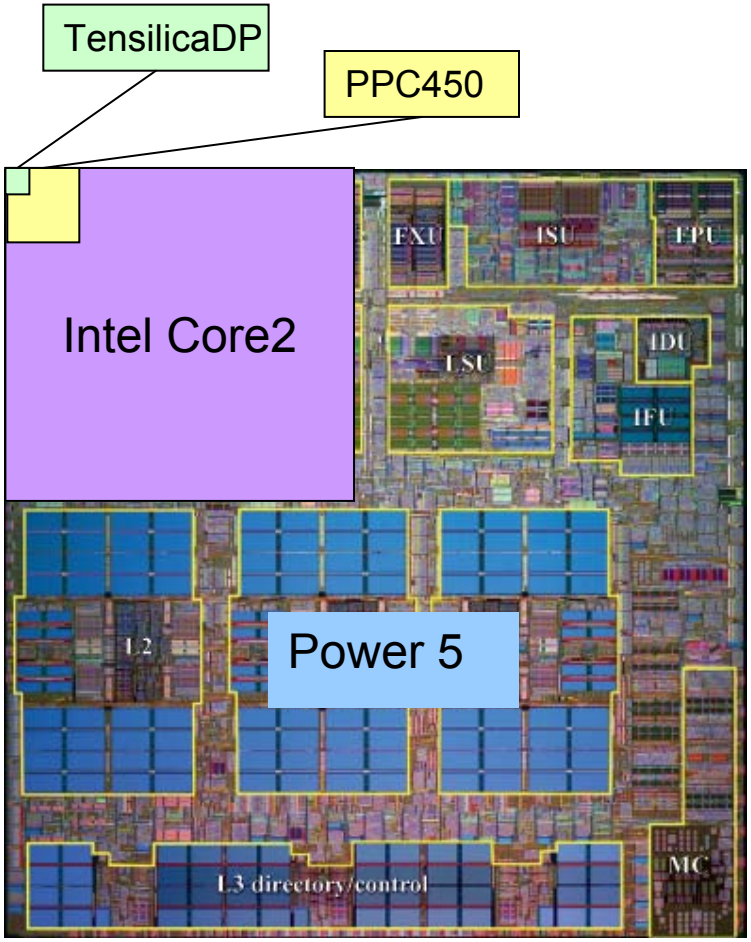


Power Demands Threaten to Limit the Future Growth of Computational Science

- LBNL Study for Climate Modeling in 2008 (Shalf, Wehner, Oliker)
  - Extrapolation of Blue Gene and AMD design trends
  - Estimate: 20 MW for BG and 179 MW for AMD
- DOE E3 Report
  - Extrapolation of existing design trends
  - Estimate: 130 MW
- DARPA Exascale Study
  - More detailed assessment of component technologies
    - Power-constrained design for 2014 technology
    - 3 TF/chip, new memory technology, optical interconnect
  - Estimate: 20 MW for memory alone, 60 MW aggregate so far
- NRC Study
  - Power and multicore challenges are not just an HPC problem

NERSC will use an innovative approach to address this challenge

Evidence of Waste

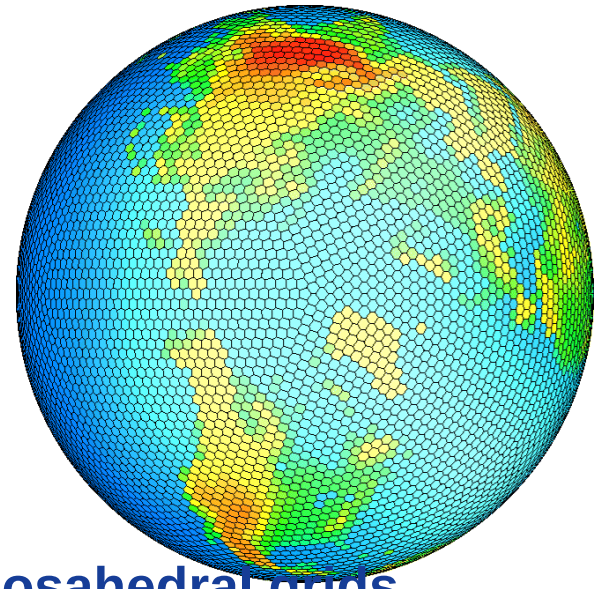


- **Power5 (Server)**
  - **389 mm²**
  - **120 W @ 1900 MHz**
- **Intel Core2 sc (Laptop)**
  - **130 mm²**
  - **15 W @ 1000 MHz**
- **PowerPC450 (BlueGene/P)**
  - **8 mm²**
  - **3 W @ 850 MHz**
- **Tensilica DP (cell phones)**
  - **0.8 mm²**
  - **0.09 W @ 650 MHz**

Each core operates at 1/3 to 1/10th efficiency of largest chip, but you can pack 100x more cores onto a chip and consume 1/20 the power!

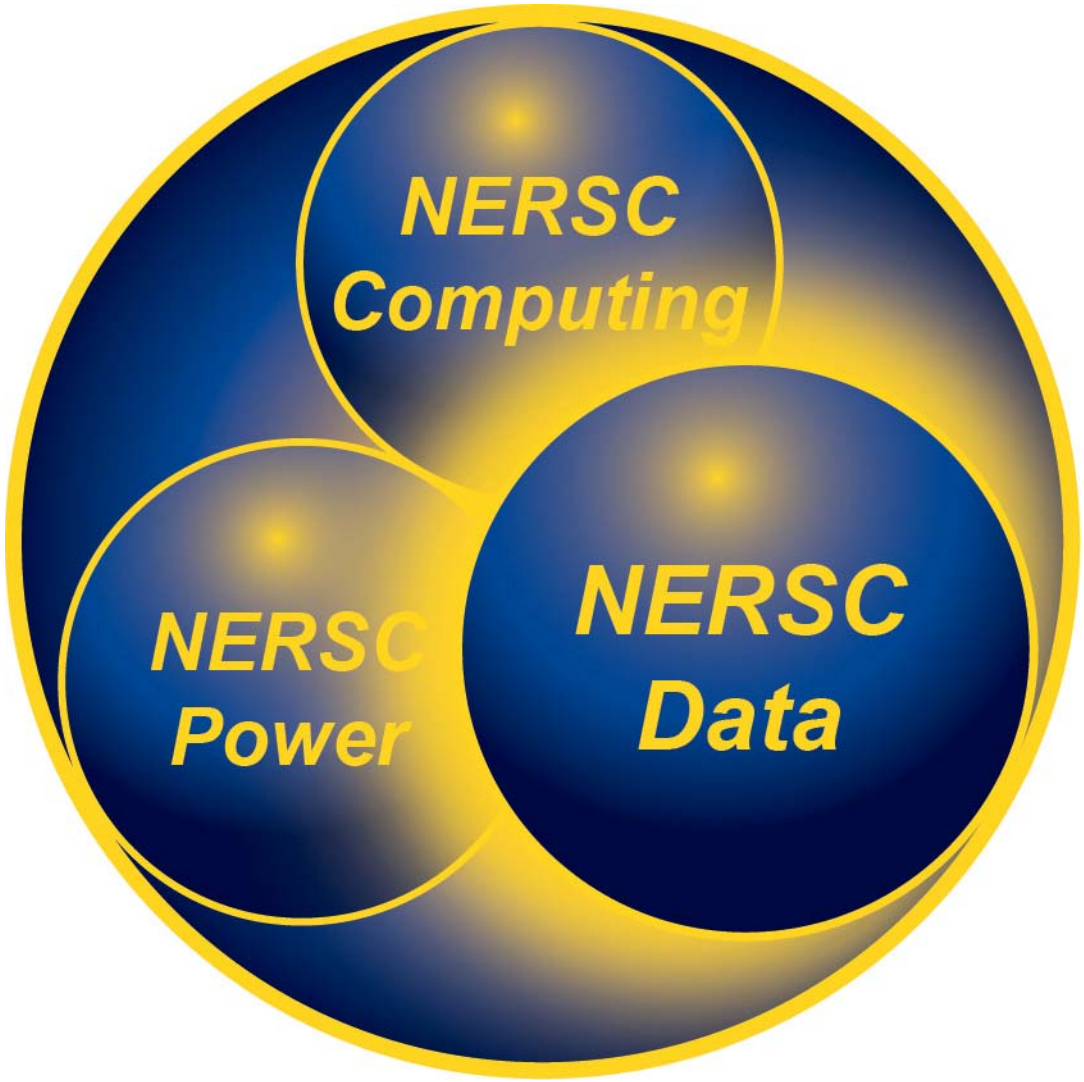
Example: Cloud Resolving Climate Simulation

- An "Exascale size" challenge is a ~1 km horizontal resolution, cloud resolving, climate simulation
- Use massive concurrency for better simulation efficiency: parallelism is power-efficient



- Requires significant algorithm work
  - E.g., Dave Randall's SciDAC work on Icosahedral grids
  - Anticipate algorithm and data structure scaling limits
- Circa 2008 estimate: 179 MW on AMD, 20 on BG/P, and 3 on Tensilica-based system tailored for Climate
- Other examples: MHD, Astro, Nano, …

NERSC Data



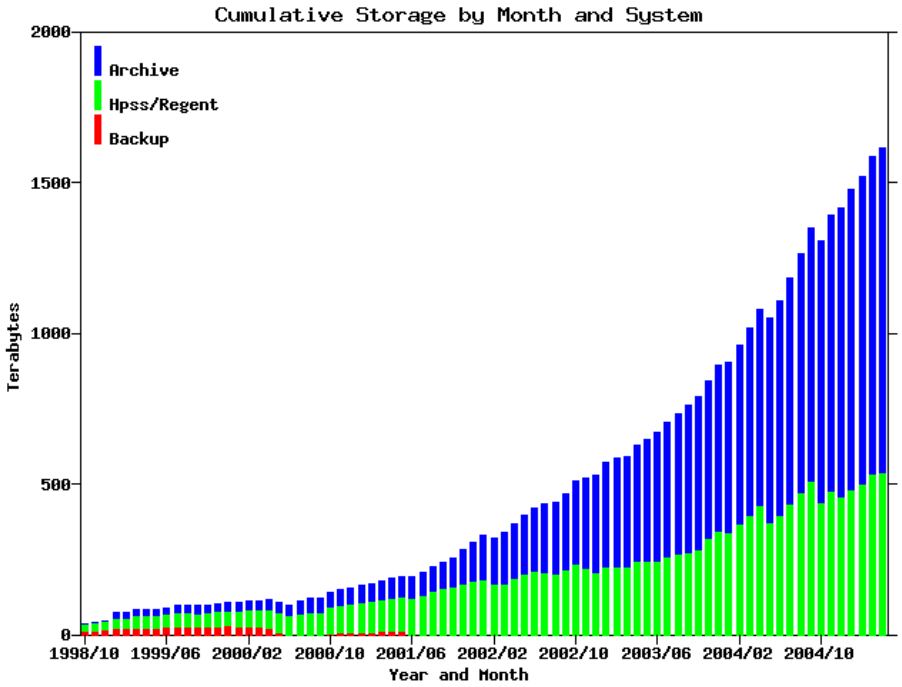
Data Tsunami

- **Soon it will no longer be sufficient for NERSC to rely solely on center balance and HPSS to address the massive volumes of data on the horizon**
- **The volume and complexity of experimental data will overshadow data from simulation**
  - **LHC**
  - **ITER**
  - **JDEM/SNAP**
  - **PLANCK**
  - **SciDAC**
  - **JGI**
  - **Earth Systems Grid**

NERSC Global Filesystem (NGF)

- After thorough evaluation and testing phase in production since early 2006
- Based on IBM GPFS
- Seamless data access from all of NERSC's computational and analysis resources
- Single unified namespace makes it easier for users to manage their data across multiple system
- First production global filesystem spanning five platforms, three architectures, and four different vendors

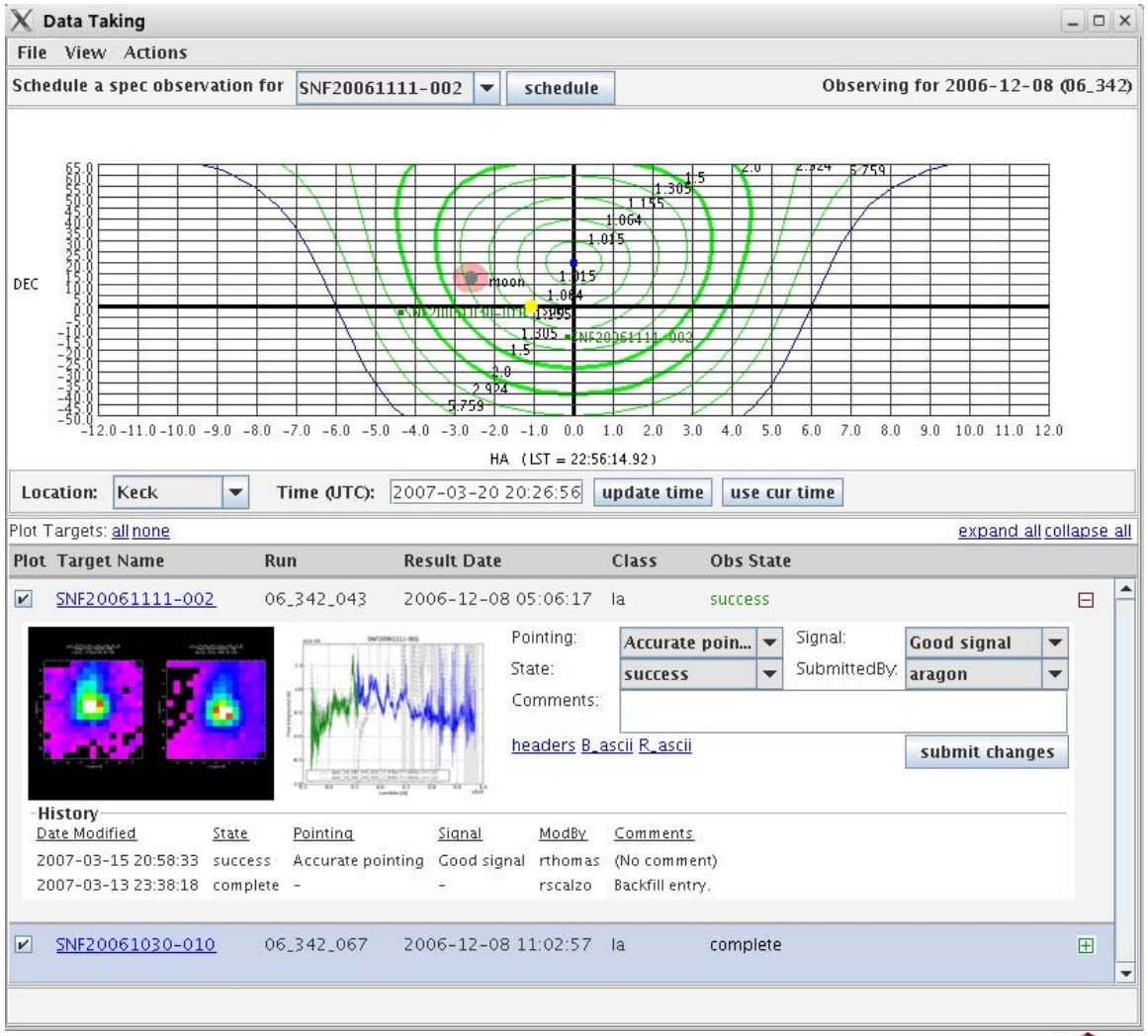
Large Storage Environment (HPSS)



61+ million files
44 PB capacity
1.7x per year data growth

Example: SUNFALL Interface

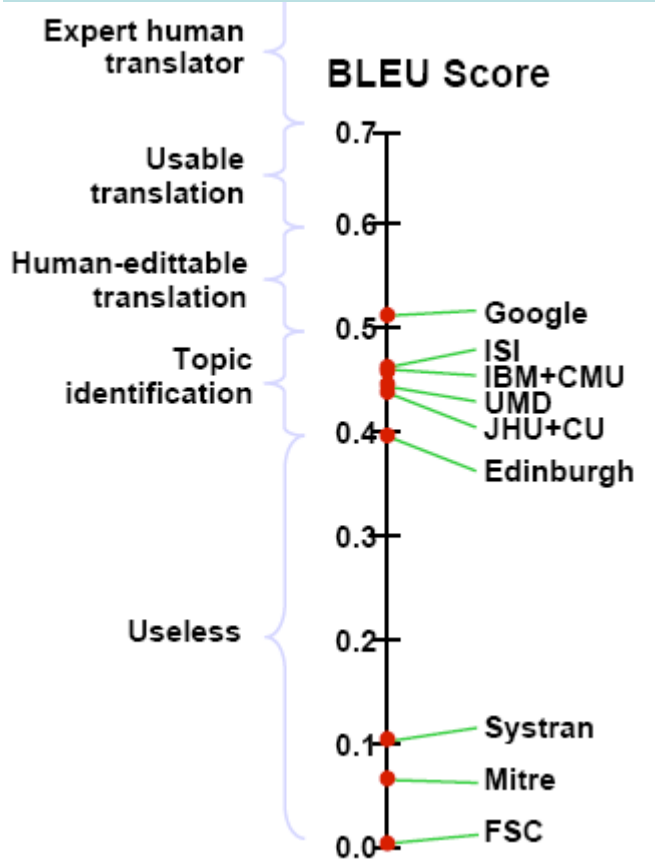
- Astrophysics data analytics
- Successful multi-disciplinary team
- Uses machine learning for discovery



"Google" for Science: Access to Data Accelerates Science

- Data helps science
  - Neanderthal genome example
  - Google has similar examples outside of science
- Google uses, MapReduce, for over 10K applications
  - Clusters, grep, machine-learning,...
  - Hides load balancing, data layout, disk failures, etc.
- NERSC will do this for science
  - Domain-specific analysis (by scientists)
  - Domain-independent infrastructure
  - Efficient use of wide area bandwidth

Arabic translation: Google with more data beats others with more specialists



National Energy Research Scientific Computing (NERSC) Division