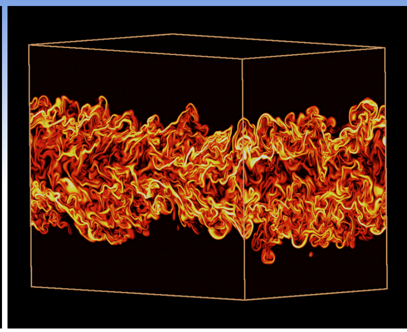
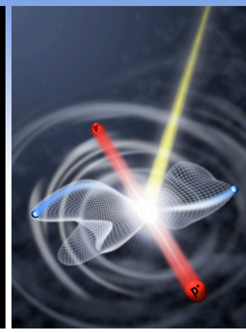
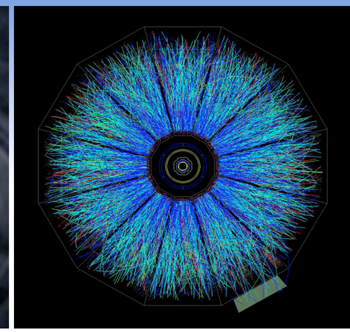
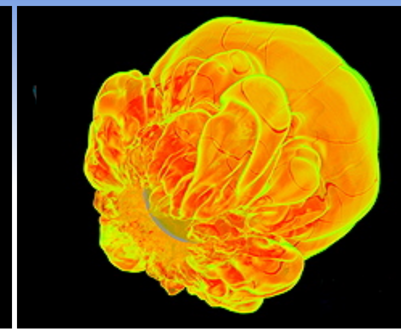






NERSC Role in High Energy Physics Research

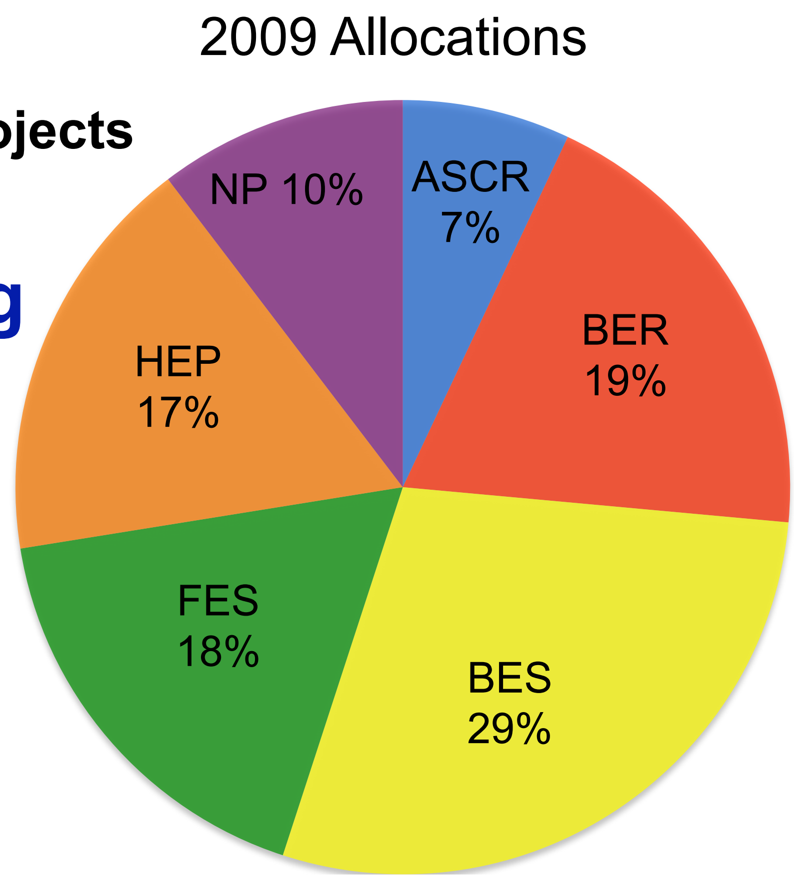
Katherine Yelick
NERSC Director

Requirements Workshop

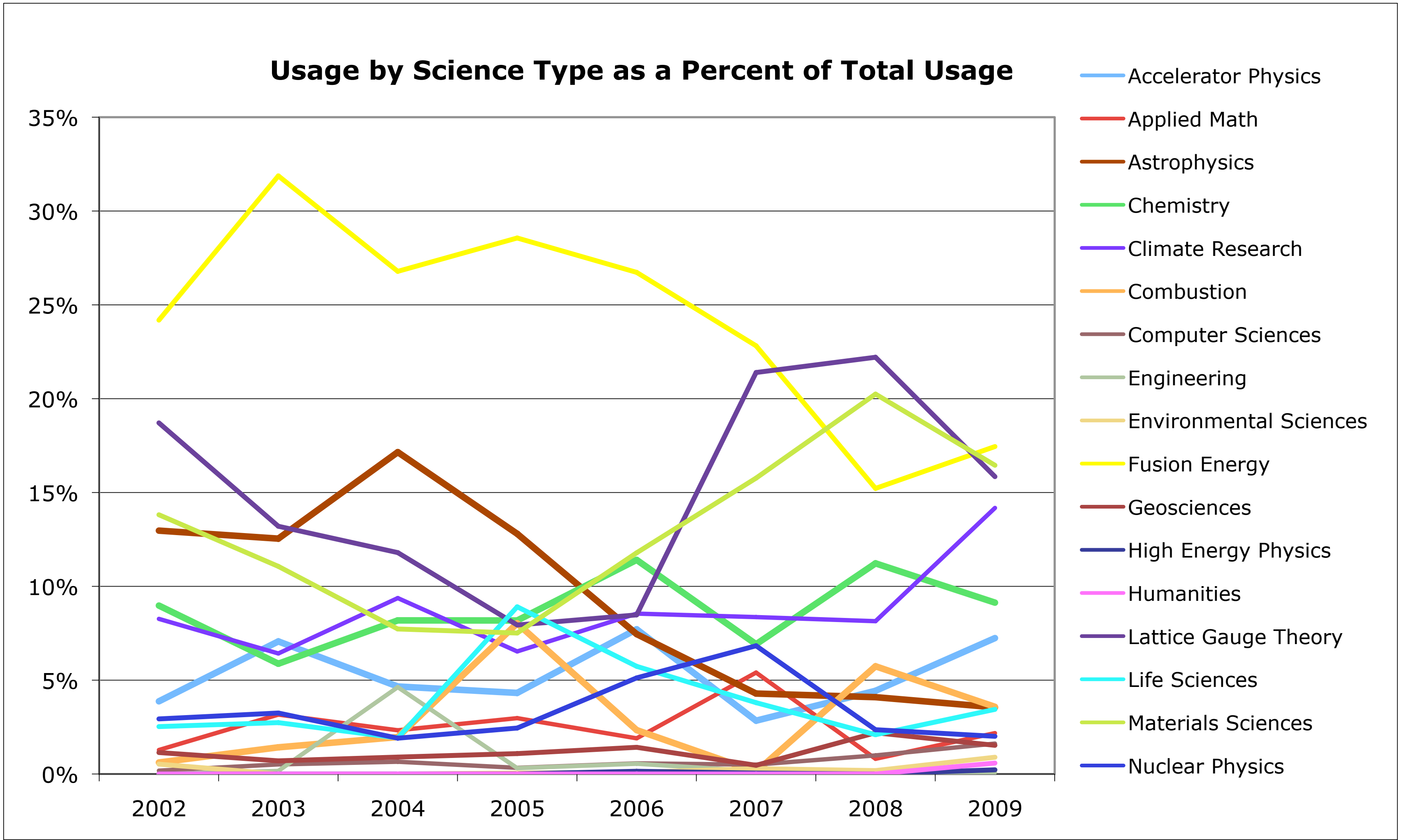
U.S. DEPARTMENT OF ENERGY | Office of Science
NERSC
BERKELEY LAB

NERSC is the Production Facility for DOE Office of Science

- **NERSC population**
  - About 3000 users in 400 distinct projects
  - About 500 code instances
- **Focus on unique computing resources for science**
  - High end computing
  - High end storage
  - Network interface
  - Expert services
- **Science-driven**
  - Machines procured competitively using application benchmarks
  - Allocations controlled by DOE based on science



DOE Priorities for NERSC Change Over Time



ASCR's Computing Facilities

NERSC at LBNL

- 1000+ users,100+ projects
- Allocations:
  - 80% DOE program manager control
  - 10% ASCR Leadership Computing Challenge*
  - 10% NERSC reserve
- Science includes all of DOE Office of Science
- Machines procured competitively

LCFs at ORNL and ANL

- 100+ users 10+ projects
- Allocations:
  - 80% ANL/ORNL managed INCITE process
  - 10% ACSR Leadership Computing Challenge*
  - 10% LCF reserve
- Science limited to largest scale; no limit to DOE/SC
- Machines procured through partnerships

NERSC 2009 Configuration

Large-Scale Computing System

Franklin (NERSC-5): Cray XT4
- 9,532 compute nodes; 38,128 cores
- ~25 Tflop/s on applications; 356 Tflop/s peak

Hopper (NERSC-6): Cray XT
- Phase 1: Cray XT5, 668 nodes, 5344 cores
- Phase 2: > 1 Pflop/s peak



Clusters


Jacquard and Bassi
- LNXI and IBM clusters
- Upgrading to Carver (NCS-c)

PDSF (HEP/NP)
- Linux cluster (~1K cores)

NERSC Global Filesystem (NGF)

Uses IBM's GPFS

440 TB; 5.5 GB/s

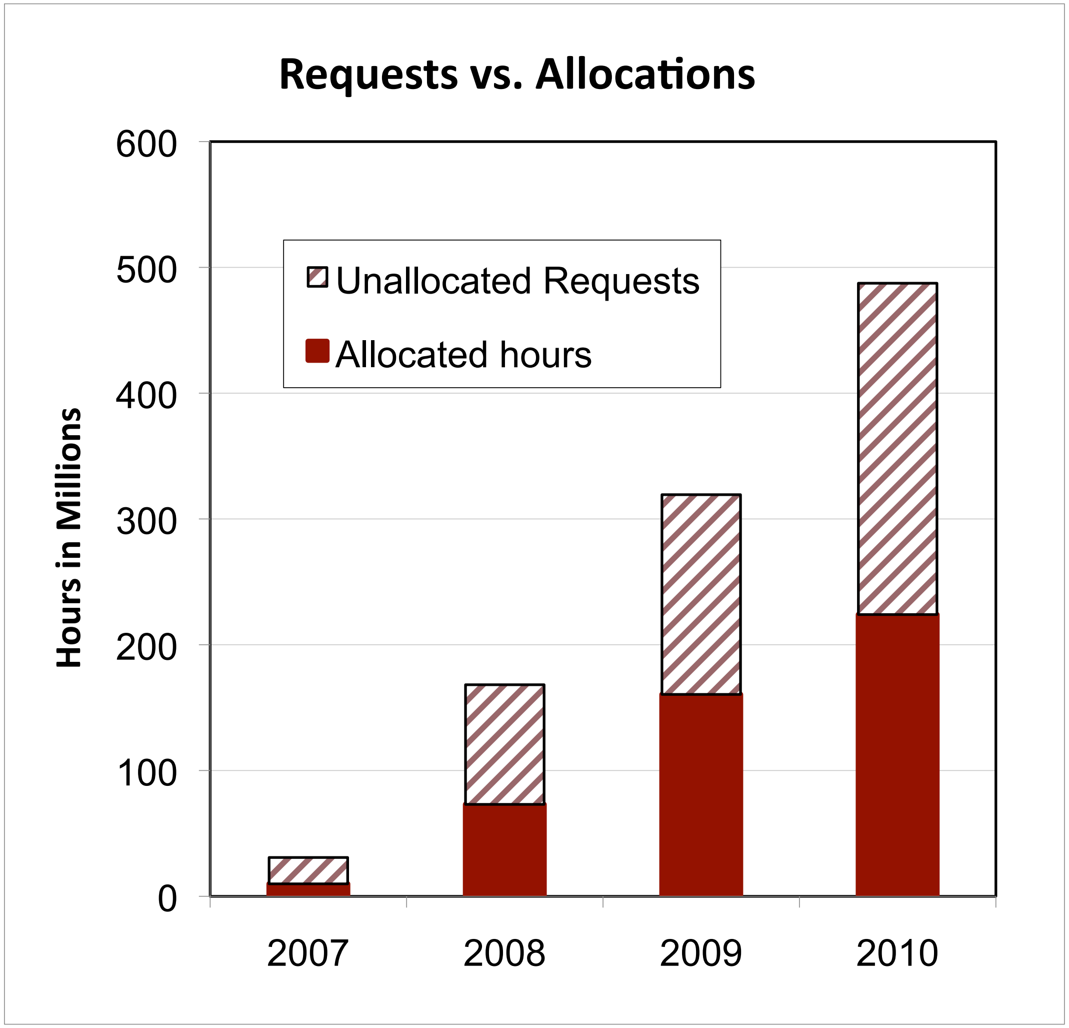
HPSS Archival Storage

- 59 PB capacity
- 11 Tape libraries
- 140 TB disk cache

Analytics / Visualization

Davinci (SGI Altix)
- Tesla testbed
- Upgrade planned

Demand for More Computing



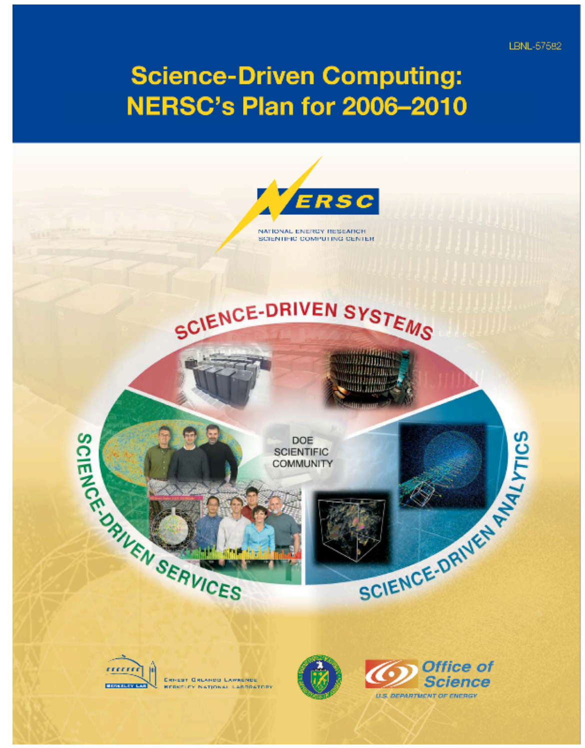
- *Each year DOE users requests ~2x as many hours as can be allocated*
- *This 2x is artificially constrained by perceived availability*
- *Unfulfilled allocation requests amount to hundreds of millions of compute hours in 2010*

How NERSC Uses Your Requirements

2005: NERSC Five-Year Plan

- **2005 Trends:**
  - Widening gap between application performance and peak
  - Emergence of multidisciplinary teams
  - Flood of scientific data
  - (Missed multicore, along with most)
- **NERSC Five-Year Plan**
  - Major system every 3 years
- **Implementation**
  - NERSC-5 (Franklin) and NERSC-6 (underway) + clusters

– **Question: What trends do you see for 2011-2015?**
  - Algorithms / application trends and other requirements



Applications Drive NERSC Procurements

Because hardware peak performance does not necessarily reflect real application performance

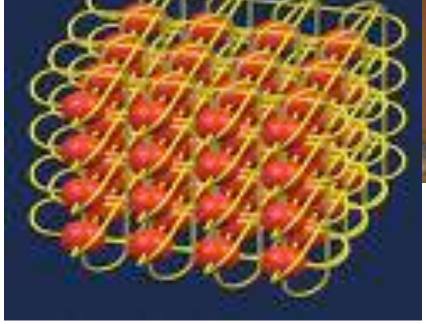
NERSC-6 “SSP” Benchmarks

CAM	*GAMESS*	*GTC*	*IMPACT-T*	*MAESTRO*	*MILC*	*PARATEC*
Climate	Quantum Chemistry	Fusion	Accelerator Physics	Astro-physics	Nuclear Physics	Material Science

- Benchmarks reflect diversity of science and algorithms
- SSP = average performance (Tflops/sec) across machine
- Used before selection, during and after installation
- Question: What applications best reflect your workload?

NERSC-5 “Franklin”

- **Largest Cray XT4**
  - 102 cabinets
  - 9,740 Quad Core nodes
  - 38,640 CPUs (cores)
  - Novel torus network for large parallel jobs
  - Direct access to parallel filesystem
- **Performance:**
  - 25 Tflop/s of sustained application performance
  - 352 Tflop/s of Peak



Benjamin Franklin,
One of America’s First Scientists, performed ground breaking work in energy efficiency, electricity, materials, climate, ocean currents, transportation, health, medicine, acoustics and heat transfer.

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

NERSC-6 “Hopper”

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Grace Murray Hopper
(1906-1992)

- Cray system selected competitively:
  - Application benchmarks from climate, chemistry, fusion, accelerator, astrophysics, QCD, and materials
  - Best application performance per dollar and per MW
  - Novel external Services for functionality and availability
  - Novel interconnect network with high bandwidth and low latency

Phase 1: Cray XT5

- 668 nodes, 5,344 cores
- 2.4 GHz AMD Opteron
- 2 PB disk, 25 GB/s
- Air cooled

Phase 2: Cray system

- > 1 Pflop/s peak
- ~ 150K cores, 12 per chip
- 2 PB disk, 80 GB/s
- Liquid cooled

3Q09	4Q09	1Q10	2Q10	3Q10	4Q10

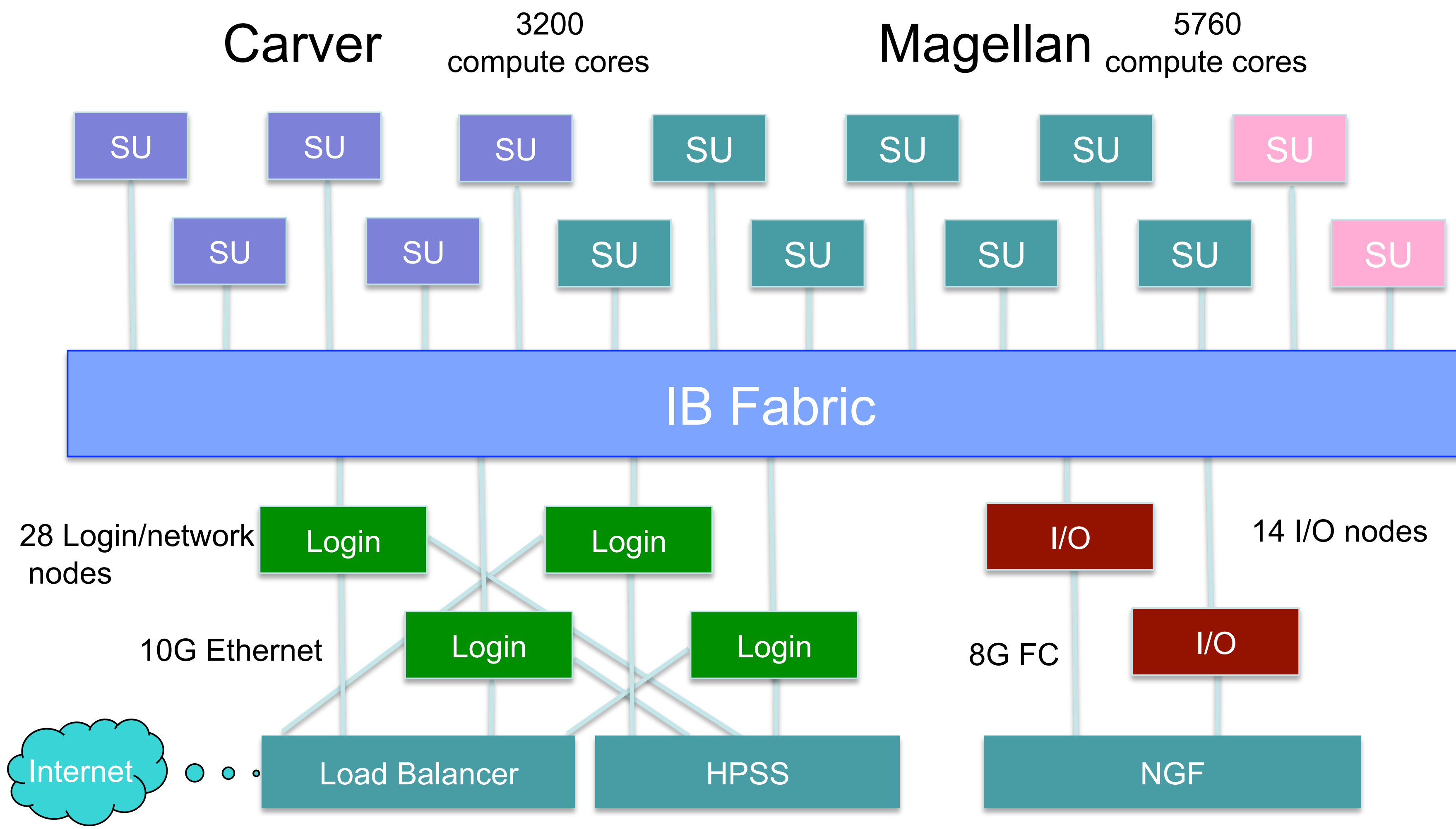
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

DOE Explores Cloud Computing

- **DOE's CS program focuses on HPC**
  - **No coordinated plan for clusters in SC**
- **DOE Magellan Cloud Testbed**
  - **$16M project at NERSC from Recovery Act**
- **Cloud questions to explore on Magellan:**
  - **Can a cloud serve DOE's mid-range computing needs?**
  - **What features (hardware and software) are needed of a "Science Cloud"? Commodity hardware?**
  - **What requirements do the jobs have (~100 cores, I/O,…)**
  - **How does this differ, if at all, from commercial clouds which serve primarily independent serial jobs**
- **Magellan not a NERSC Program machine**
  - **Not allocated in ERCAP; testbed, not production**

Cluster architecture

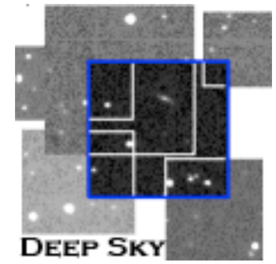


Reservations at NERSC

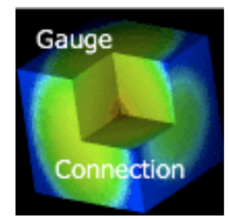
- **Reservation service being tested:**
  - Reserve a certain date, time and duration
    - Debugging at scale
    - Real-time constraints in which need to analyze data before next run, e.g., daily target selection telescopes or genome sequencing pipelin
  - At least 24 hours advanced notice
    - https://www.nersc.gov/nusers/services/reservation.php
  - Successfully used for IMG run, Madcap, IO benchmarking, etc.

Science Gateways at NERSC

- **Create scientific communities around data sets**
  - **Models for sharing vs. privacy differ across communities**
  - **Accessible by broad community for exploration, scientific discovery, and validation of results**
  - **Value of data also varies: observations may be irreplaceable**



- **A *science gateway* is a set of hardware and software that provides data/services remotely**
  - **Deep Sky – "Google-Maps" of astronomical image data**
    - **Discovered 140 supernovae in 60 nights (July-August 2009)**
    - **1 of 15 international collaborators were accessing NGF data through the SG nodes 24/7 using both the web interface and the database.**
  - **Gauge Connection – Access QCD Lattice data sets**
  - **Planck Portal – Access to Planck Data**



- **Building blocks for science on the web**
  - **Remote data analysis, databases, job submission**

Visualization Support

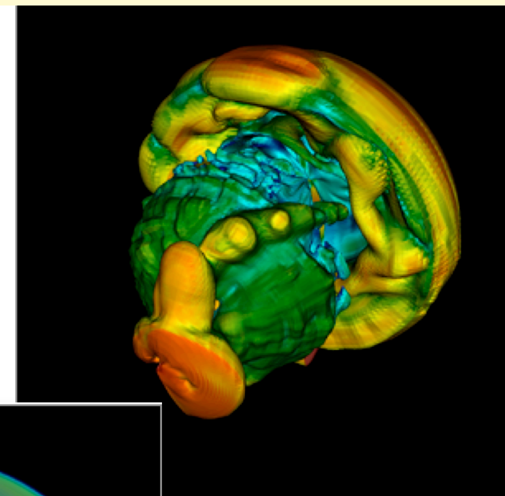
Petascale visualization: Demonstrate visualization scaling to unprecedented concurrency levels by ingesting and processing unprecedentedly large datasets.

Implications: Visualization and analysis of Petascale datasets requires the I/O, memory, compute, and interconnect speeds of Petascale systems.

Accomplishments: Ran VisIt SW on 16K and 32K cores of Franklin.

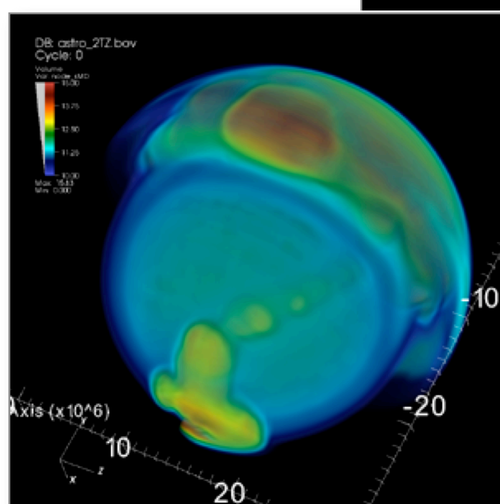
- First-ever visualization of two *trillion* zone problem (TBs per scalar); data loaded in parallel.
- Petascale visualization

Plots show 'inverse flux factor,' the ratio of neutrino intensity to neutrino flux, from an ORNL 3D supernova simulation using CHIMERA.



a

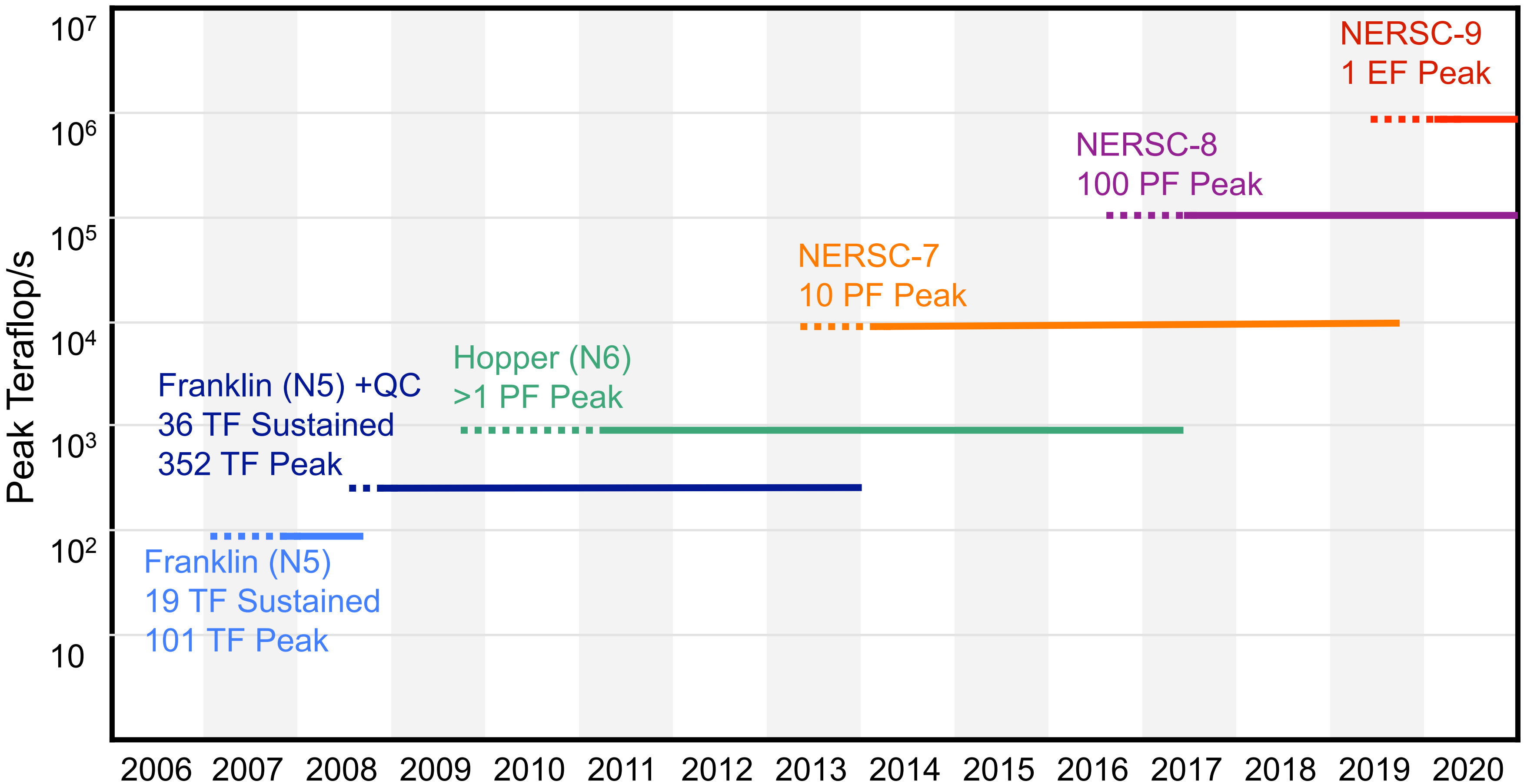
b



Isocontours (a) and volume rendering (b) of two trillion zones on 32K cores of Franklin.

Requirements Drive NERSC's Long-Term Vision

NERSC System Roadmap

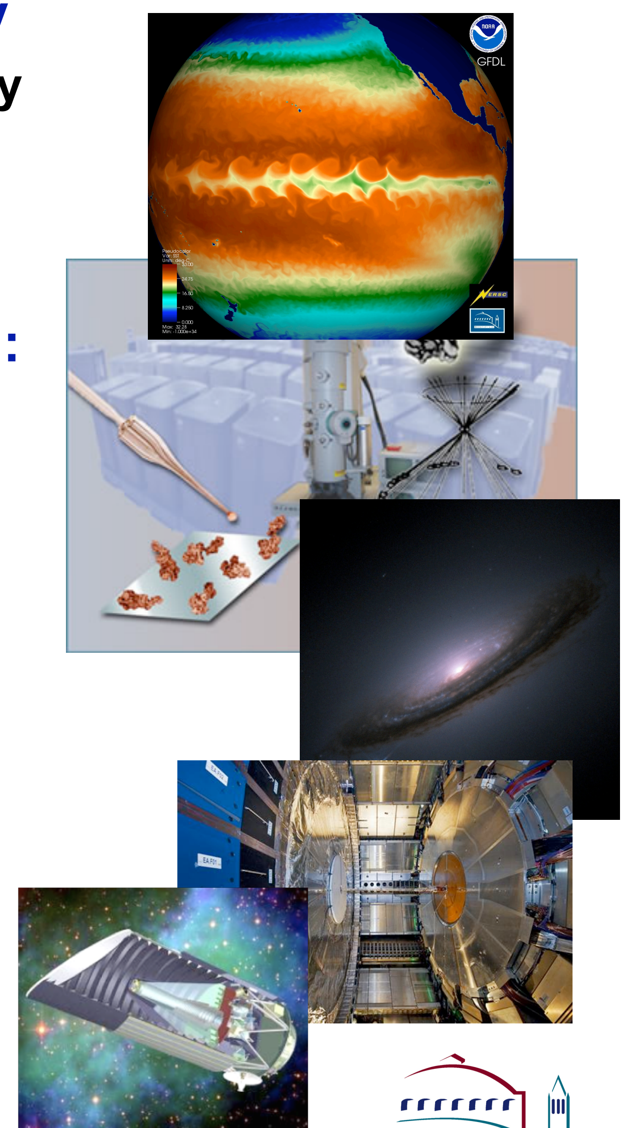


- **Goal is two systems on the floor at all times**
- **Systems procured by sustained performance**

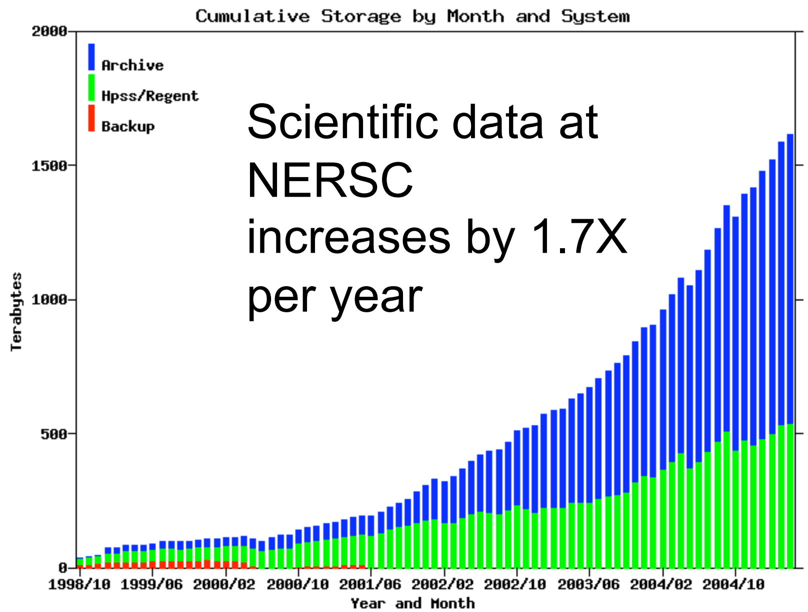
NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Data Driven Science

- **Scientific data sets are growing exponentially**
  - **Ability to generate data is exceeding our ability to store and analyze**
  - **Simulation systems and some observational devices grow in capability with Moore's Law**
- **Petabyte (PB) data sets will soon be common:**
  - ***Climate modeling:* estimates of the next IPCC data is in 10s of petabytes**
  - ***Genome:* JGI alone will have .5 petabyte of data this year and double each year**
  - ***Particle physics*: LHC is projected to produce 16 petabytes of data per year**
  - ***Astrophysics*: LSST and others will produce 5 petabytes/year**
- **Create scientific communities with "Science Gateways" to data**

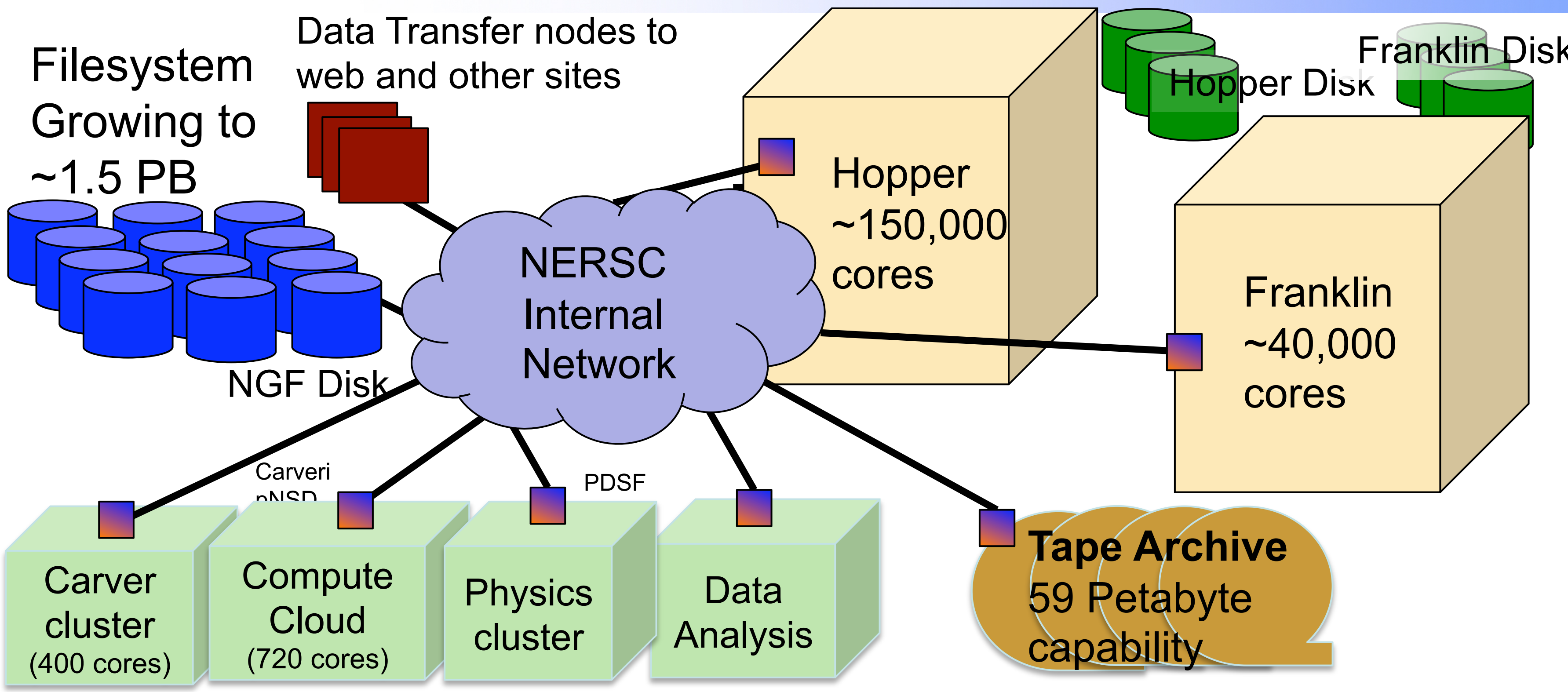


Tape Archives: Green Storage



- **Tape archives are important to efficient science**
  - **2-3 orders of magnitude less power than disk**
  - **Requires specialized staff and major capital investment**
  - **NERSC participates in development (HPSS consortium)**
- **Questions: What are your data sets sizes and growth rates?**

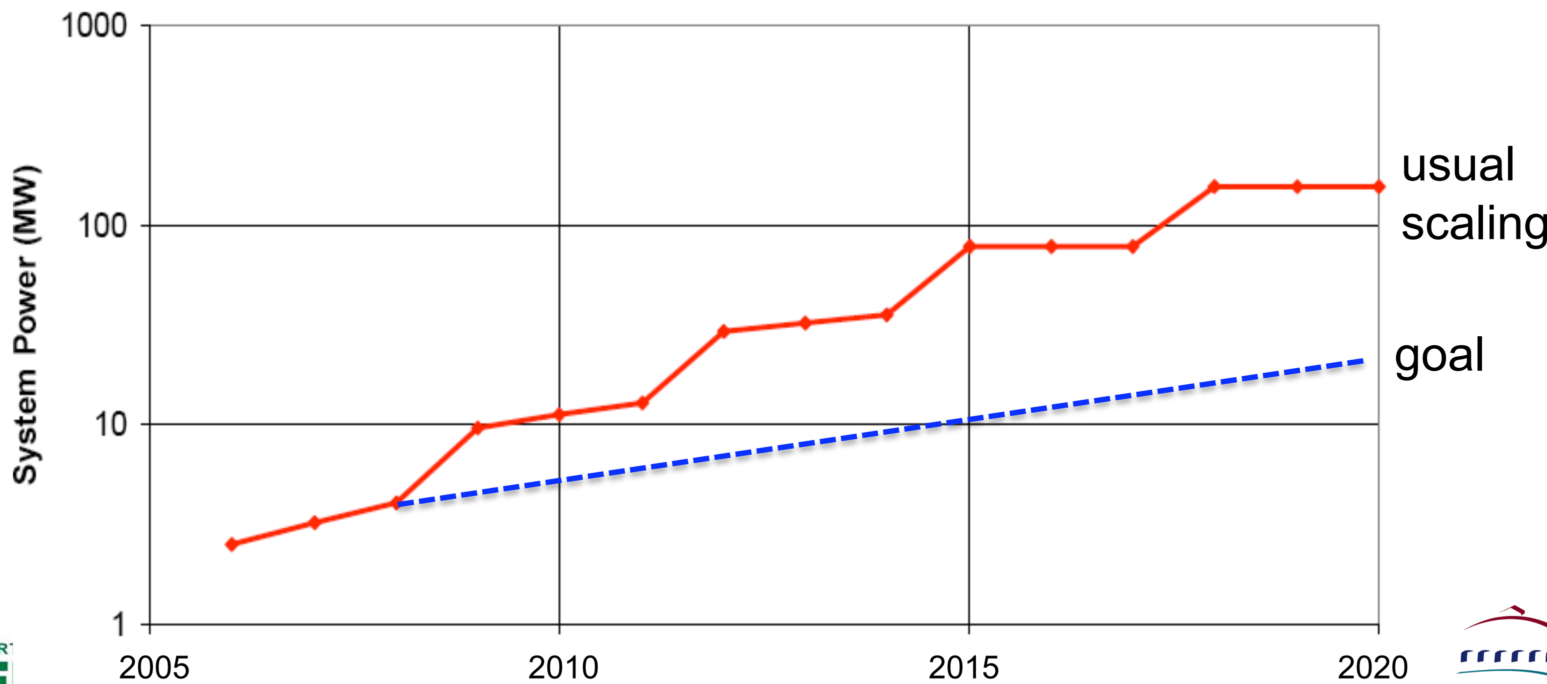
NERSC Architecture



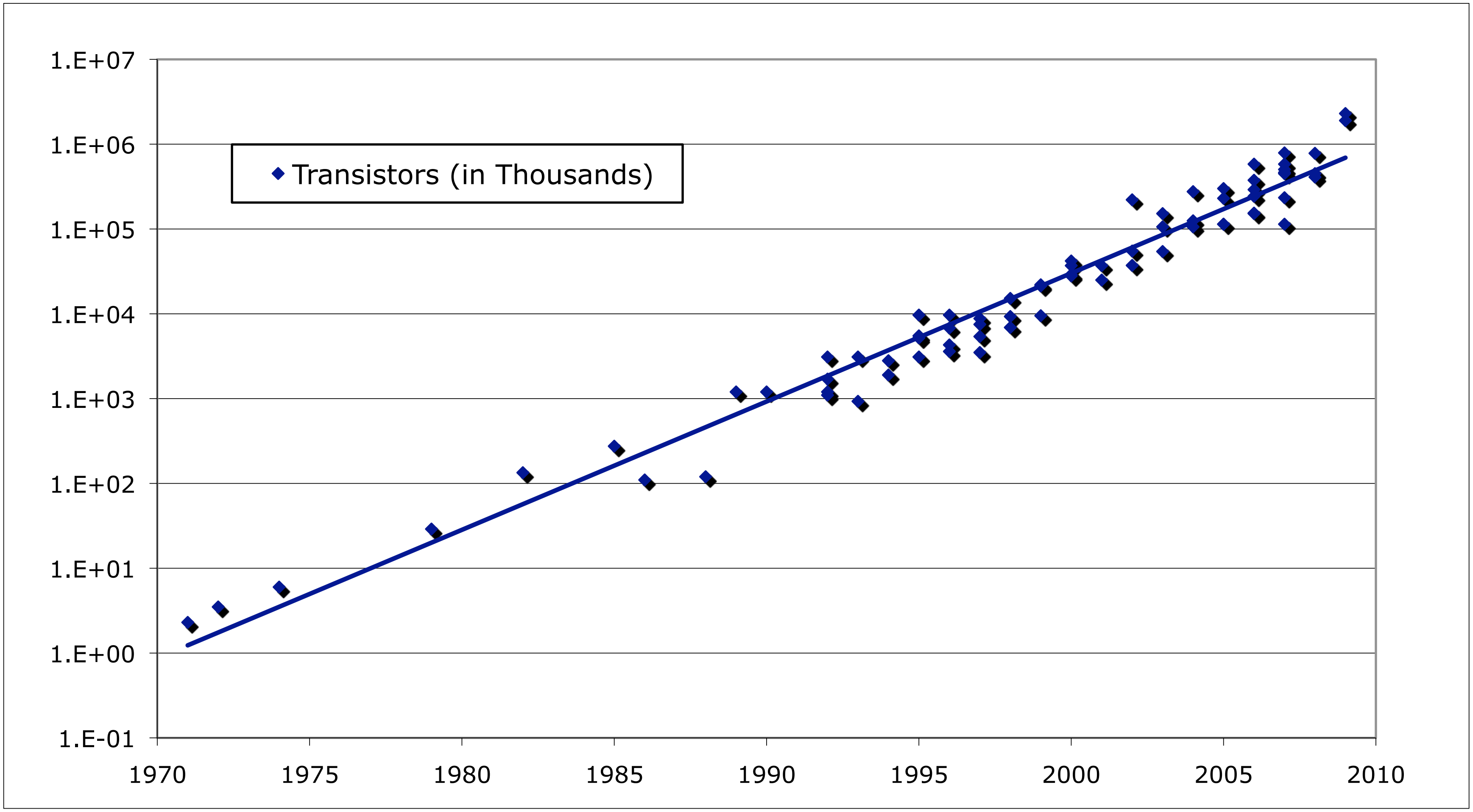
- NERSC has a modest number of “commodity systems”
- Mostly specialized science systems for compute, disk storage (parallel filesystems), and tape archives

Energy Efficiency is Necessary for Computing

- Systems have gotten about 1000x faster over each 10 year period
- 1 petaflop (10^{15} ops) in 2010 will require 3MW
  → 3 GW for 1 Exaflop (10^{18} ops/sec)
- DARPA committee suggested 200 MW with "usual" scaling
- Target for DOE is 20 MW in 2018

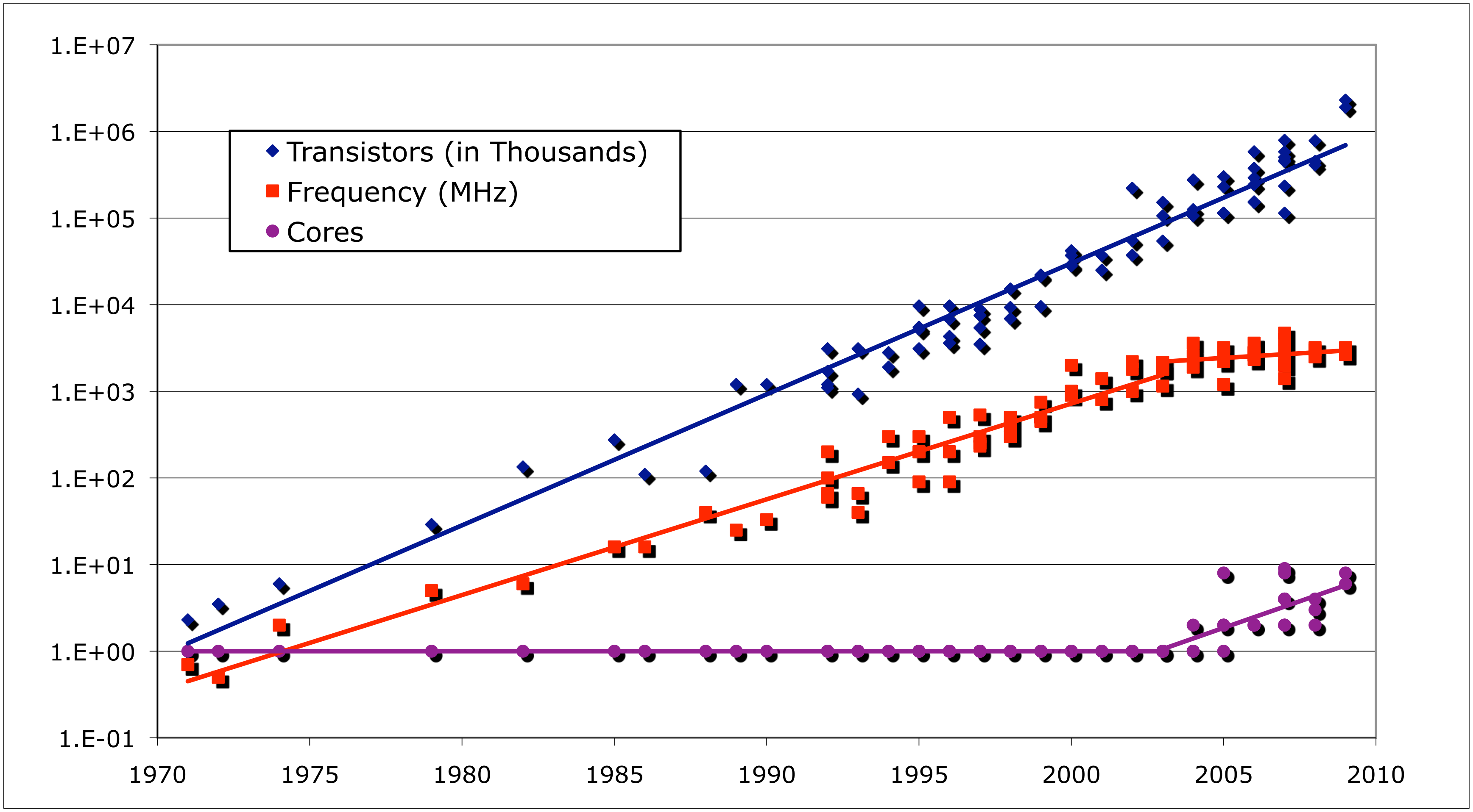


Moore's Law is Alive and Well



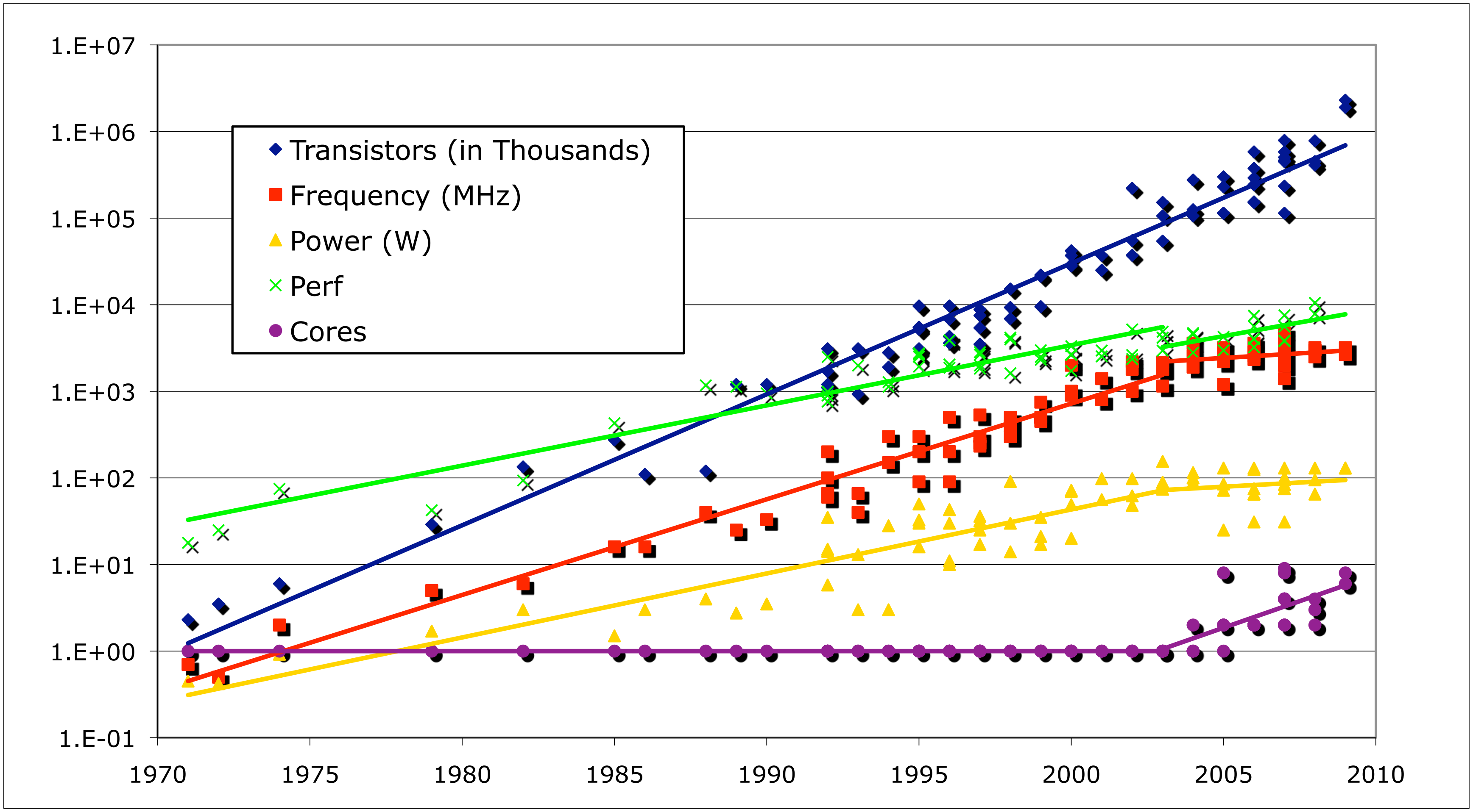
Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç

But Clock Frequency Scaling Has Been Replaced by Scaling Cores / Chip



Slide Source: Kathy Yelick. Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç

Performance Has Also Slowed, Along with Power (the Root Cause of All This)



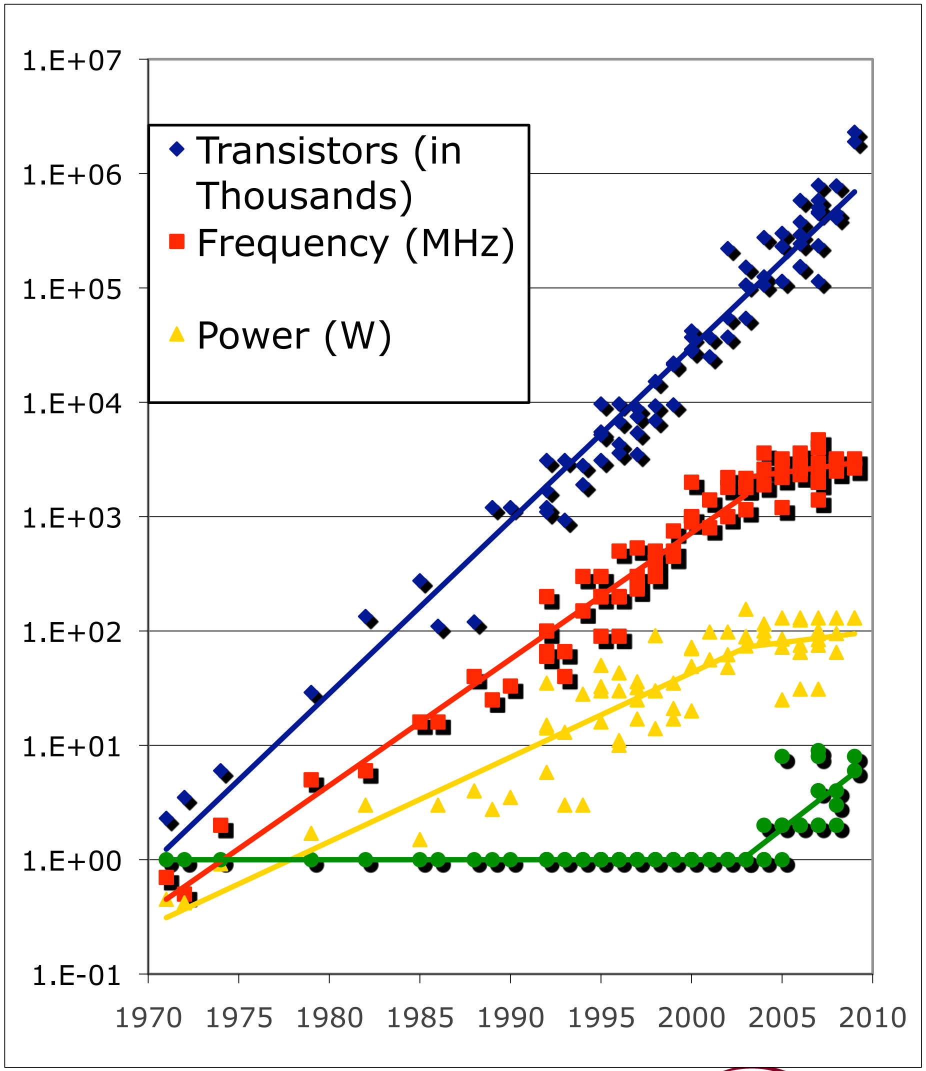
Slide Source: Kathy Yelick. Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç

NERSC Goal Usable Exascale in 2020

- **Computational scaling changed in 2004**
- **Problems also for laptops, handhelds, data centers**
- **Parallelism on-chip brings algorithms, programming into question**
- ***NERSC: Programmable, usable systems for science***

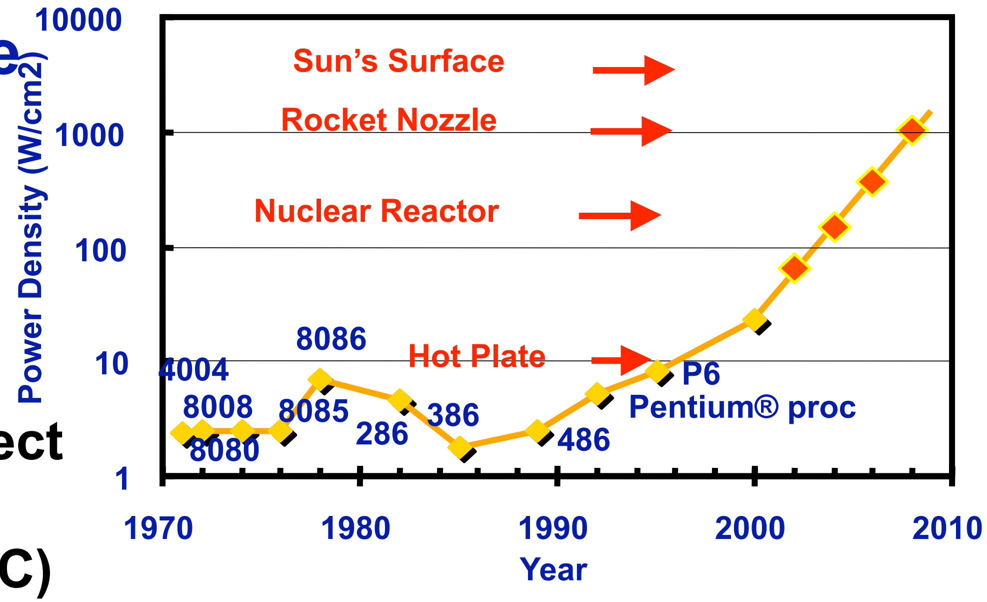
1) Energy efficient designs

2) Facilities to support scale for both high and mid scale



Parallelism is "Green"

- **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

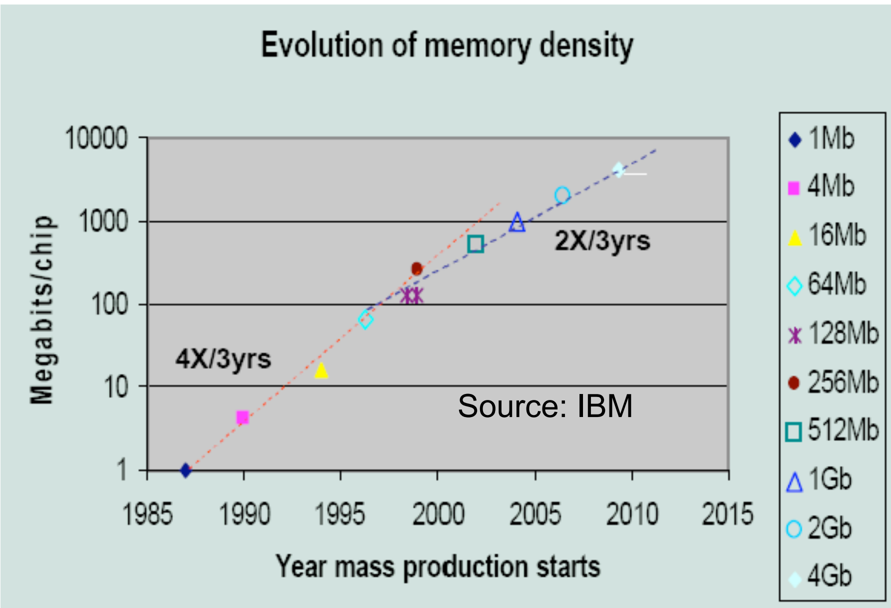


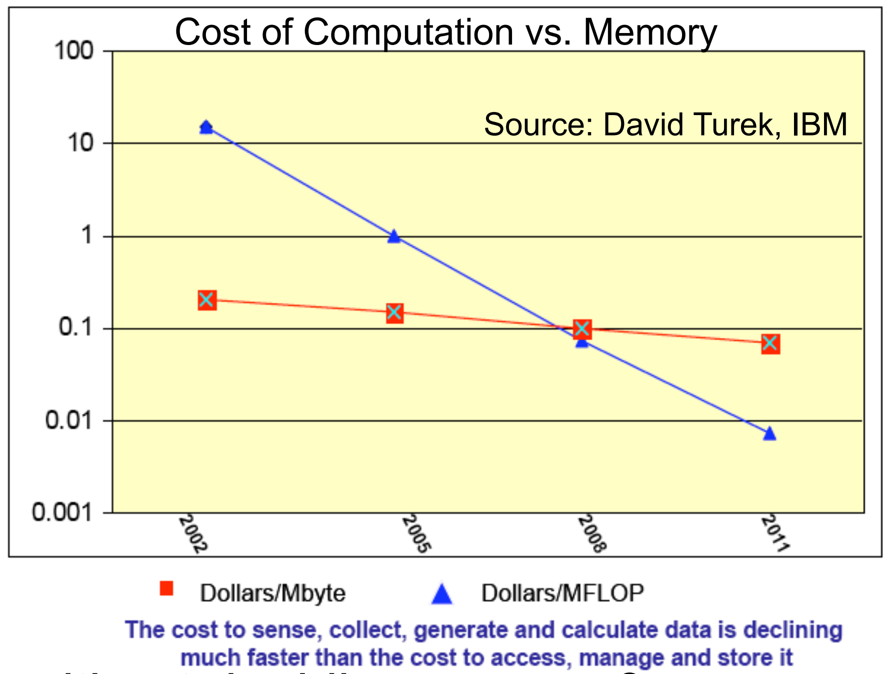
- **High performance serial processors waste power**
  - Speculation, dynamic dependence checking, etc. burn power
  - Implicit parallelism discovery
- **Question: *Can you double the concurrency in your algorithms and software every 2 years?***

Technology Challenge

Technology trends against a constant or increasing memory per core

- Memory density is doubling every three years; processor logic is every two
- Storage costs (dollars/Mbyte) are dropping gradually compared to logic costs





Question: *Can you double concurrency without doubling memory?*

Hardware and Software Trends

- **Hardware Trends**
  - Exponential growth in explicit on-chip parallelism
  - Reduced memory per core
  - Heterogeneous computing platforms (e.g., GPUs)
  - As always, this is largely driven by non HPC markets
- **Software Response**
  - Need to express more explicit parallelism
  - New programming models on chip: MPI + X
  - Increased emphasis on strong scaling
  - No more serial code scaling from hardware
- **What we want**
  - Understand your requirements and help craft a strategy for transitioning to a hardware and programming environment solution

High Energy Physics Science at NERSC

Supernova Core-Collapse

Objective: First principles understanding of supernovae of all types, including radiation transport, spectrum formation, and nucleosynthesis.

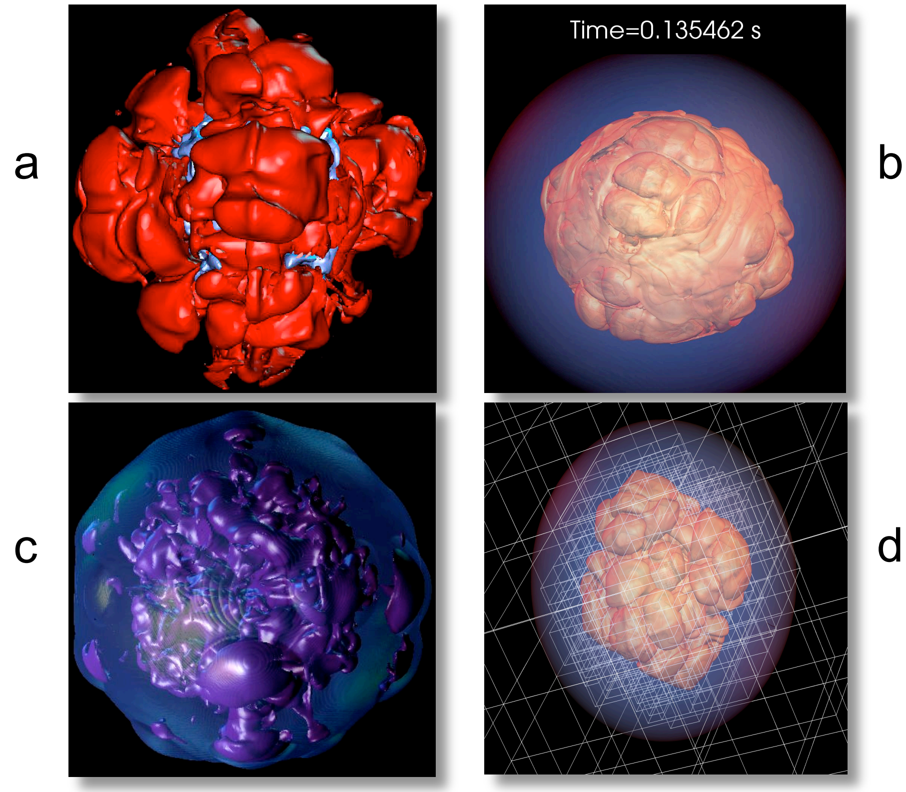
Implications: Will help confront one of the greatest mysteries in high-energy physics and astronomy -- the nature of dark energy.

Accomplishments: NERSC runs of VULCAN core collapse explain magnetically-driven explosions in rapidly-rotating cores.

- First 2.5-D, detailed-microphysics radiation-magnetohydrodynamic calculations; first time-dependent 2D rad-hydro supernova simulations with multi-group and multi-angle transport.
- CASTRO, new multi-dimensional, Eulerian AMR hydrodynamics code that includes stellar EOS, nuclear reaction networks, and self-gravity.

NERSC: 2M hours alloc in 2009; Vis support

PIs: S. Woosley (UCSB), A. Burrows (Princeton)



The exploding core of a massive star. a), b), and c) show morphology of selected isoentropy, isodensity contours during the blast; (d) AMR grid structure at coarser resolution levels."

Cosmic Microwave Background

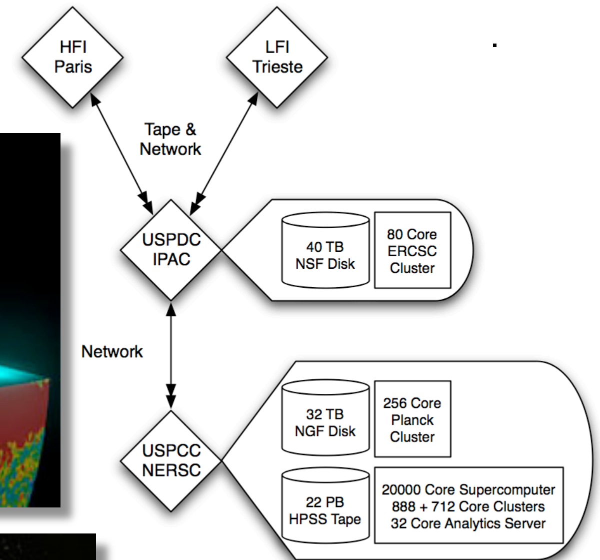
Objective: Analyze data from the Planck satellite -- definitive Cosmic Microwave Background (CMB) data set.

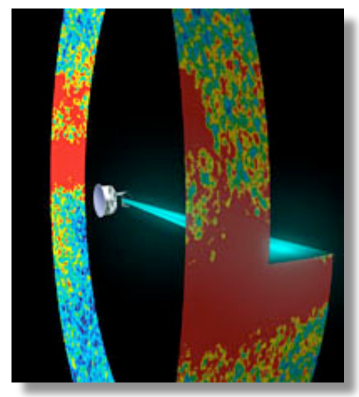
Implications: CMB: image of the universe at 400k years, relic radiation from Big Bang

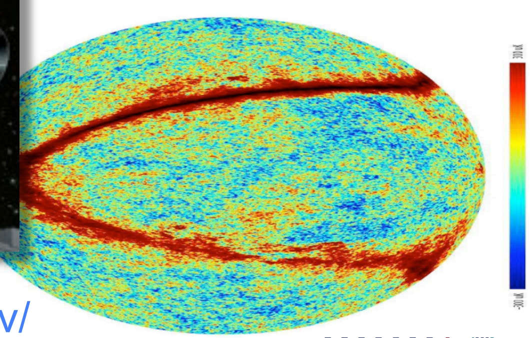
Accomplishments: NERSC provides the components of the data pipeline for noise reduction, map-making, power spectrum analysis, and parameter estimation

- 2006 Nobel Prize in Physics
- 32 TB final data set size, ~400 users
- data sets analyzed as a whole because complex data correlations; no "divide and conquer"
- Launched May09, first "light" Sept09
- Also ~10k-core XT4 MonteCarlo calibration runs, produce ~10X data
- Anticipate Moore's law growth in data set size for *15 years*

PI: J. Borrill (LBNL)







https://c3.lbl.gov/

BERKELEY LAB

HEP: Accelerator Modeling

PI: J. Qiang (LBNL)

Objective: Help design and optimize the electron beam for LBNL next-generation Free Electron Laser.

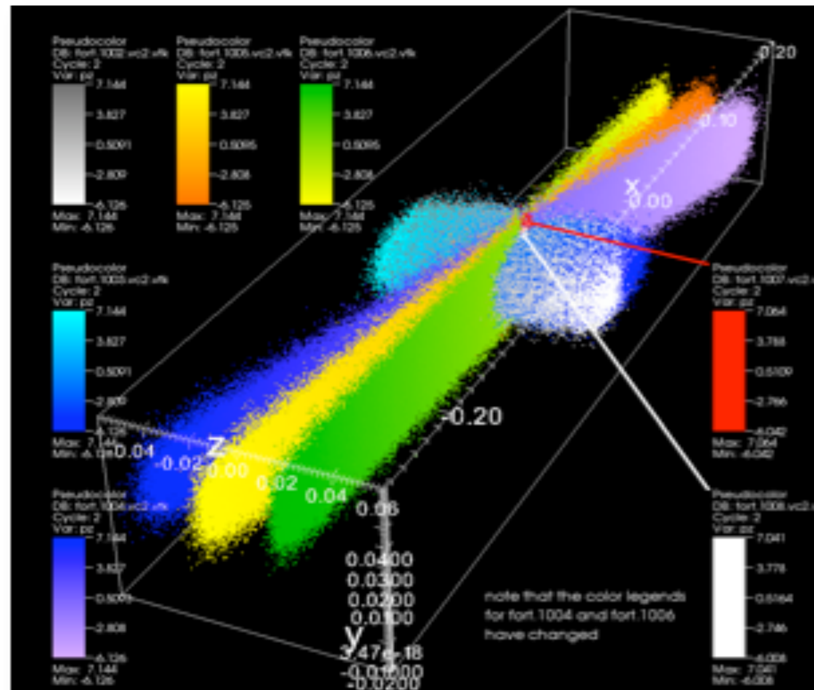
Implications: Numerically optimizing the beam lowers cost of design / operation and improves X-ray output, helping scientific discovery in physics, material science, chemistry and bioscience.

Accomplishments: Code includes self-consistent 3D space-charge effects, short-range geometry & longitudinal synchrotron radiation wakefields, and detailed RF acceleration / focusing.

- *Billion-particle simulation required*

NERSC:

- Allocated 800K hours in 2009
- IMPACT code, part of NERSC6 test suite
- NERSC provided visualization support



Visualization of an electron beam bending and changing orientation as it passes through a magnetic bunch compressor.

Proc. Linac08 Conference

Lattice QCD

Objective: Understand strong interactions that bind quarks and gluons together.

Implications: Explain new phases of matter that might form in heavy-ion collisions (in LHC, for example).

Accomplishments: Cited by DOE in 2010 Congressional Budget Request as one of 3 major accomplishments in Theoretical Nuclear Physics in 2008/9.

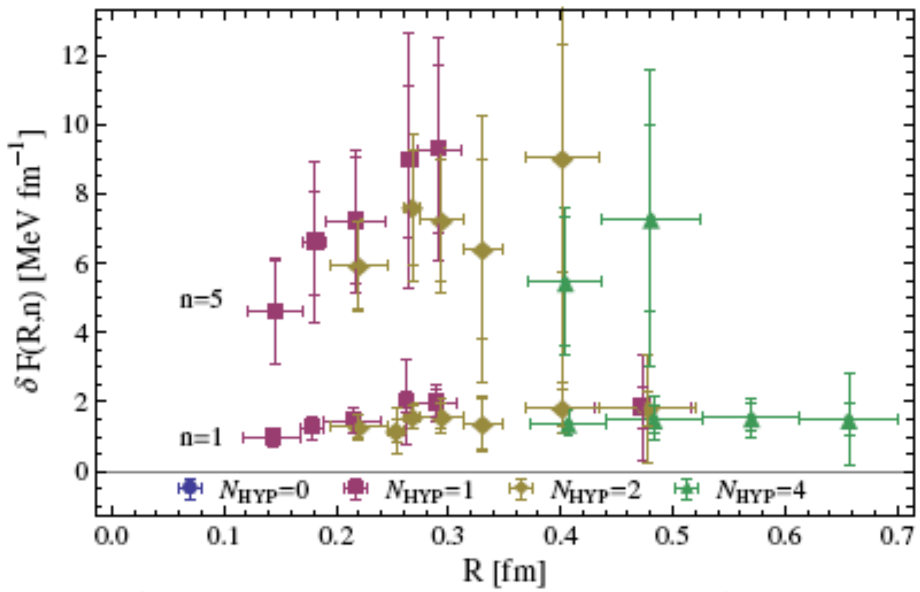
- First ever QCD calculations of:
  - Three-body force between hadrons.
  - Screening of color forces between quarks by a background of hadrons.
  - a three-baryon system.

NERSC:

- QDP++/Chroma on Franklin; 10M+ hours
- Mostly 4k cores per job

PIs: M. Savage (U. Wash.), W. Detmold (JLab, College of W&M)

Color Screening by Pions



Contribution to the radial quark-antiquark force at two pion densities. The attractive force is found to be reduced by the pion screening. This is a first step toward a more systematic exploration of hadronic effects with lattice QCD.

Phys Rev Lett, (2009)

Laser Wakefield Acceleration

PI: C. Geddes, LBNL

Objective: **Use multi-scale simulation to understand & design laser driven plasma wakefield accelerators, supporting LOASIS experiments.**

Implications: **Offers promise of accelerators orders of magnitude smaller, less costly than current machines.**

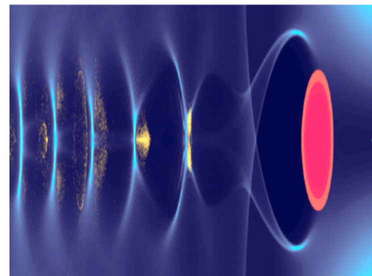
Accomplishments: **2- & 3D PIC simulations (VORPAL) reproduce electron beam charge & energy, show physical mechanisms of acceleration.**

- **Used to develop new injector technologies to improve beam quality**
- **Designing a proposed 10GeV LWFA**
- **Solutions to PIC code limitations.**
- **LOASIS and SciDAC: VACET and SDM**

NERSC:

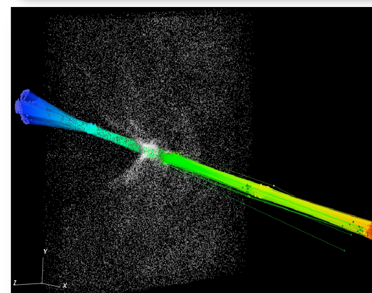
- **2.2M hours on Franklin; significant viz /analytics support; typical runs use ~10k cores**

(a)

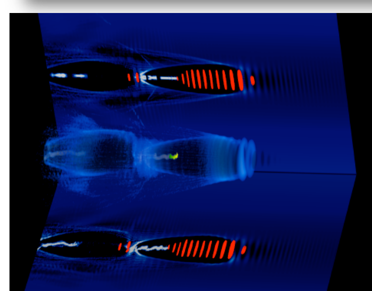


Plasma density gradient controlled injector in 2D

(b)



Particle trace of particles according to user specified criteria (momentum here; red=high)



Simulation showing 3D contours and projections of the wake (blue), laser (red), and particle bunch (yellow) in a 100 MeV LWFA

SciDAC Review Cover Story (2009)

High Energy Physics: Palomar Transient Factory

Objective: Process, analyze & make available data from Palomar Transient Sky survey (~300 GB / night) to expose rare and fleeting cosmic events.

Implications: First survey dedicated solely to finding transient events.

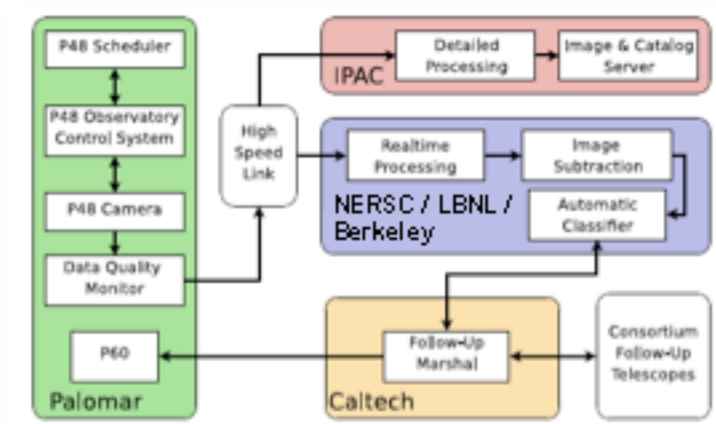
Accomplishments: Automated software for astrometric & photometric analysis and *real-time* classification of transients.

- Analysis at NERSC is fast enough to reveal transients *as data are collected.*
- Has *already uncovered* more than 40 supernovae explosions since Dec., 2008.
- Uncovering a new event about *every 12 minutes.*

NERSC:

- 40k MPP allocation + 1M HPSS in 2009; Use of NERSC NGF + gateway (next slide)

PI: P. Nugent (LBNL)



PTF project data flow

Two manuscripts submitted to Publications of the Astronomical Society of the Pacific

Deep Sky Science Gateway

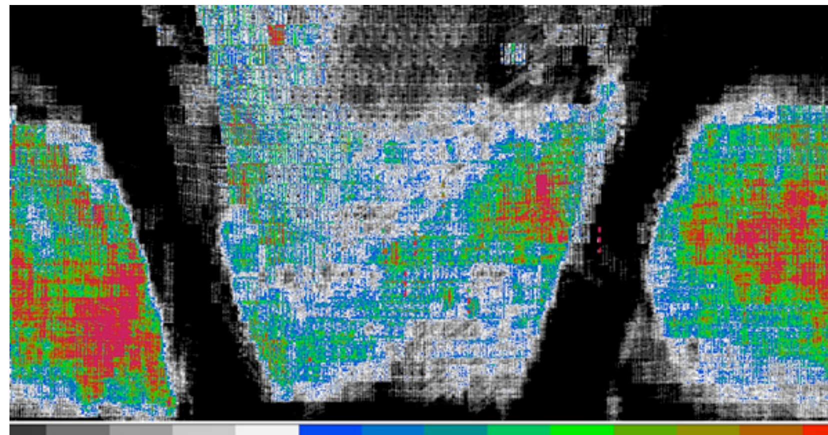
Objective: Create a richer set of compute- and data-resource interfaces for next-generation astrophysics image data, making it easier for scientists to use NERSC and creating world-wide collaborative opportunities.

Implications: Efficient, streamlined access to massive amounts of data for broad user communities.

Accomplishments: Open-source software customized to create Deep Sky database system and interface: www.deepskyproject.org

- ~ 11M 6-Mb images stored in HPSS/NGF
- DeepSky is like "Google Earth" for astronomers.
- Other NERSC gateways: GCRM (climate), Planck (Astro), Gauge Connection (QCD), VASP (chemistry/materials science).

NERSC Project



Map of the sky as viewed from Palomar Observatory; color shows the number of times an area was observed

http://www.nersc.gov/nusers/services/Grid/SG/

Other HEP Projects on PDSF Cluster

- **ATLAS, CDF, and DayaBay**
  - ATLAS is simulating the ATLAS detector at the LHC and will be doing data analysis when the LHC is online
  - CDF does data analysis and simulation of the CDF detector at the Tevatron.
  - DayaBay does simulation and analysis tools for the DayaBay Neutrino Mixing experiment

- **NERSC**
  - All three do analysis and simulation on PDSF and store data in the NGF filesystem and HPSS.
  - Requirements:
    - Need high job throughput and available bandwidth
    - ATLAS uses grid services, and thus the OSG stack support
    - DayaBay will transfer ~10MB/sec from the experiment in China for a total of ~150TB/year in 2010.

Conclusions

- **NERSC requirements**
  - **Qualitative requirements shape NERSC functionality**
  - **Quantitative requirements set the performance**

    “What gets measure gets improved”
- **Goals:**
  - **Your goal is to make scientific discoveries**
    - **Articulate specific scientific goals and implications for broader community**
  - **Our goal is to enable you to do science**
    - **Specify resources (services, computers, storage, …) that NERSC could provide with quantities and dates**

Cover Stories from NERSC Research

NERSC is enabling new science in all disciplines, with about *1,500 refereed publications* per year