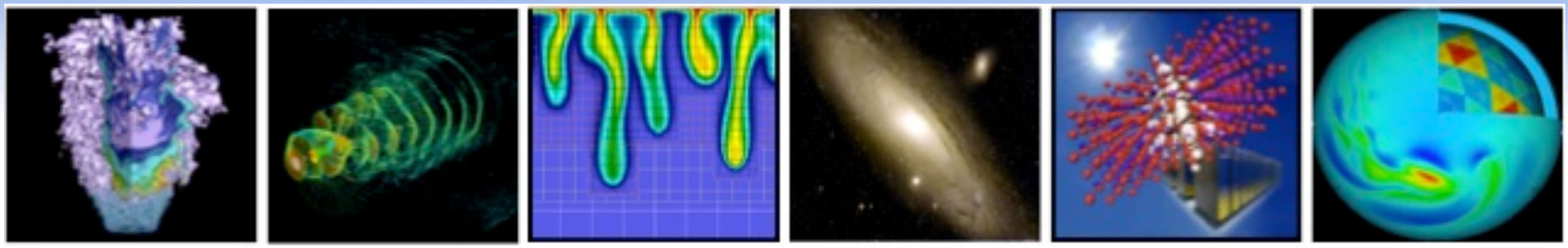


NERSC Role in Fusion Energy Science Research

Katherine Yelick
NERSC Director

Requirements Workshop

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.*

New Type of Nonlinear Plasma Instability Discovered

Objective: Study large periodic instabilities called Edge Localized Modes (ELMs) in confined toroidal plasmas using magneto-hydrodynamics code M3D.

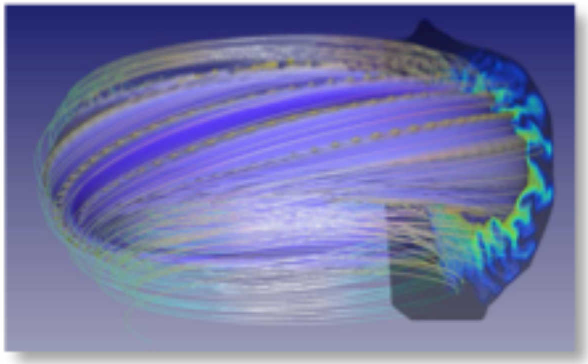
Implications: ELM properties have long resisted theoretical explanation; may be a constraint on the design of next generation fusion experiments such as ITER.

Accomplishments:

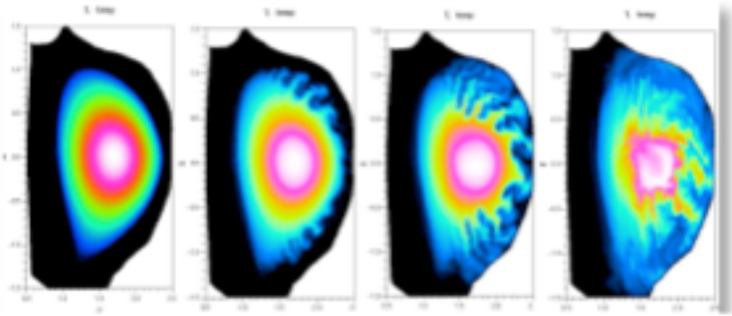
- Showed that ELMs are a new class of nonlinear plasma instability.
- The instability couples to the magnetic field, drives field perturbation deep into the plasma.
- APS invited talk + SciDAC09

NERSC: All computations, visualization done at NERSC; project used 1.2M hours in 2009.

L. Sugiyama, MIT



Temperature surface near plasma edge shows helical, field-aligned perturbation



Time evolution of an ELM.

J. Phys: Conf. Ser. 180 (2009) 012060

Simulations Explain Fast Ion Transport in Tokamak Shot

Objective: Comprehensive first-principles simulation of energetic particle turbulence and transport in ITER-scale plasmas.

Implications: Improved modeling of fusion systems is key in making fusion practical.

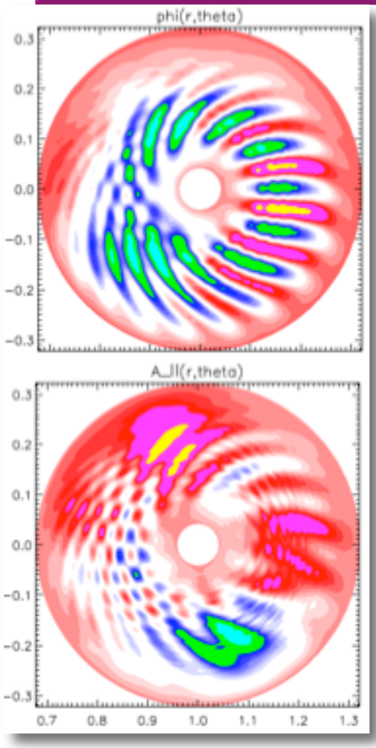
Accomplishments: GTC simulation successfully explains measurement of fast ion transport in General Atomics DIII-D tokamak shot.

- Diffusivity decreases drastically for high-energy particles due to averaging effects of large gyroradius and banana width, and fast wave-particle decorrelation.

SciDAC: Gyrokinetic Simulation of Energetic Particle Turbulence and Transport (GSEP)

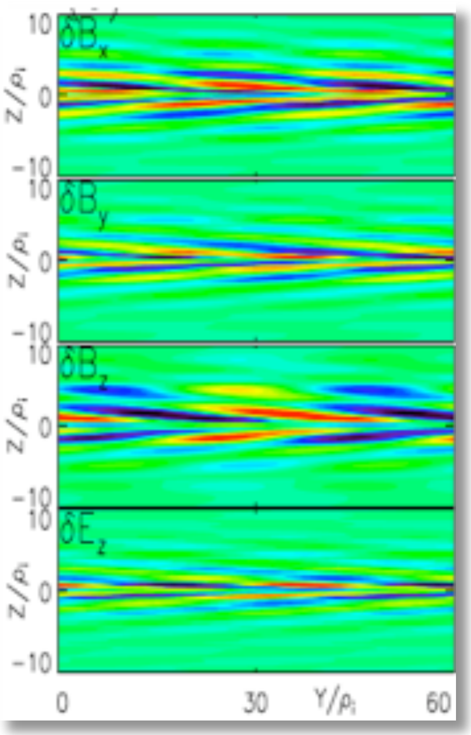
NERSC: 4M hours in 2009; GTC in NERSC-6 benchmarks and NERSC/Cray Center of Excellence study

PI: Z. Lin, UC Irvine



Gyrokinetic simulation with kinetic electrons using a hybrid model in GTC.

Comm Comp Phys (2009)



2-D Electromagnetic field fluctuations in a simulated plasma due to microinstabilities in the current.

Phys Rev Lett (2008)
Phys Plas. (2008)

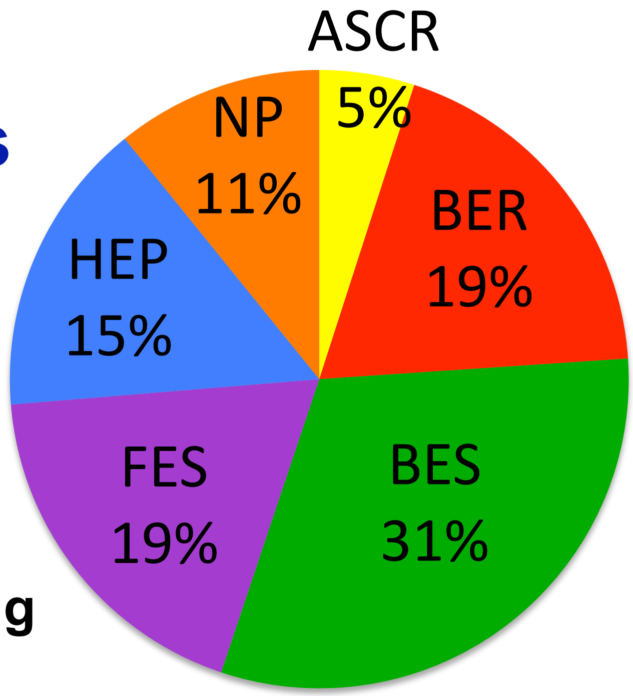
NERSC is the Production Facility for DOE Office of Science

- **NERSC serves a large population**

  Over 3000 users, 400 projects, 500 code instances

- **Focus on "unique" resources**
  - **Expert consulting and other services**
  - **High end computing systems**
  - **High end storage systems**
  - **Interface to high speed networking**
- **Science-driven**
  - **Machines procured competitively using application benchmarks from DOE/SC**
  - **Allocations controlled by DOE/SC Program Offices to couple with funding decisions**

2010 Allocations



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:
  - 60% ANL/ORNL managed INCITE process
  - 30% ACSR Leadership Computing Challenge*
  - 10% LCF reserve
- Science limited to largest scale; no limit to DOE/SC
- Machines procured through partnerships

NERSC Systems and Allocations

Large-Scale Computing Systems

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 XE6

- Phase 1: Cray XT5, 668 nodes, 5344 cores
- Phase 2: > 1 Pflop/s peak (late 2010 delivery)

Clusters

105 Tflops total

Carver

- IBM iDataplex cluster

PDSF (HEP/NP)

- Linux cluster (~1K cores)

Magellan Cloud testbed

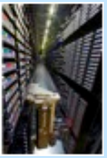
- IBM iDataplex cluster

NERSC Global Filesystem (NGF)

Uses IBM's GPFS

1.5 PB; 5.5 GB/s

HPSS Archival Storage



- 40 PB capacity
- 4 Tape libraries

Analytics

Euclid (512 GB shared memory)

Dirac GPU testbed (48 nodes)

Allocated hours will increase rough 4x once Hopper is in full production

NERSC-6 System “Hopper”

Grace Murray Hopper
(1906-1992)

- Cray system selected competitively:
  - Used application benchmarks from climate, chemistry, fusion, accelerator, astrophysics, QCD, and materials
  - Best application performance per dollar based
  - Best sustained application performance per MW
  - External Services for increased functionality and availability

Phase 1: Cray XT5

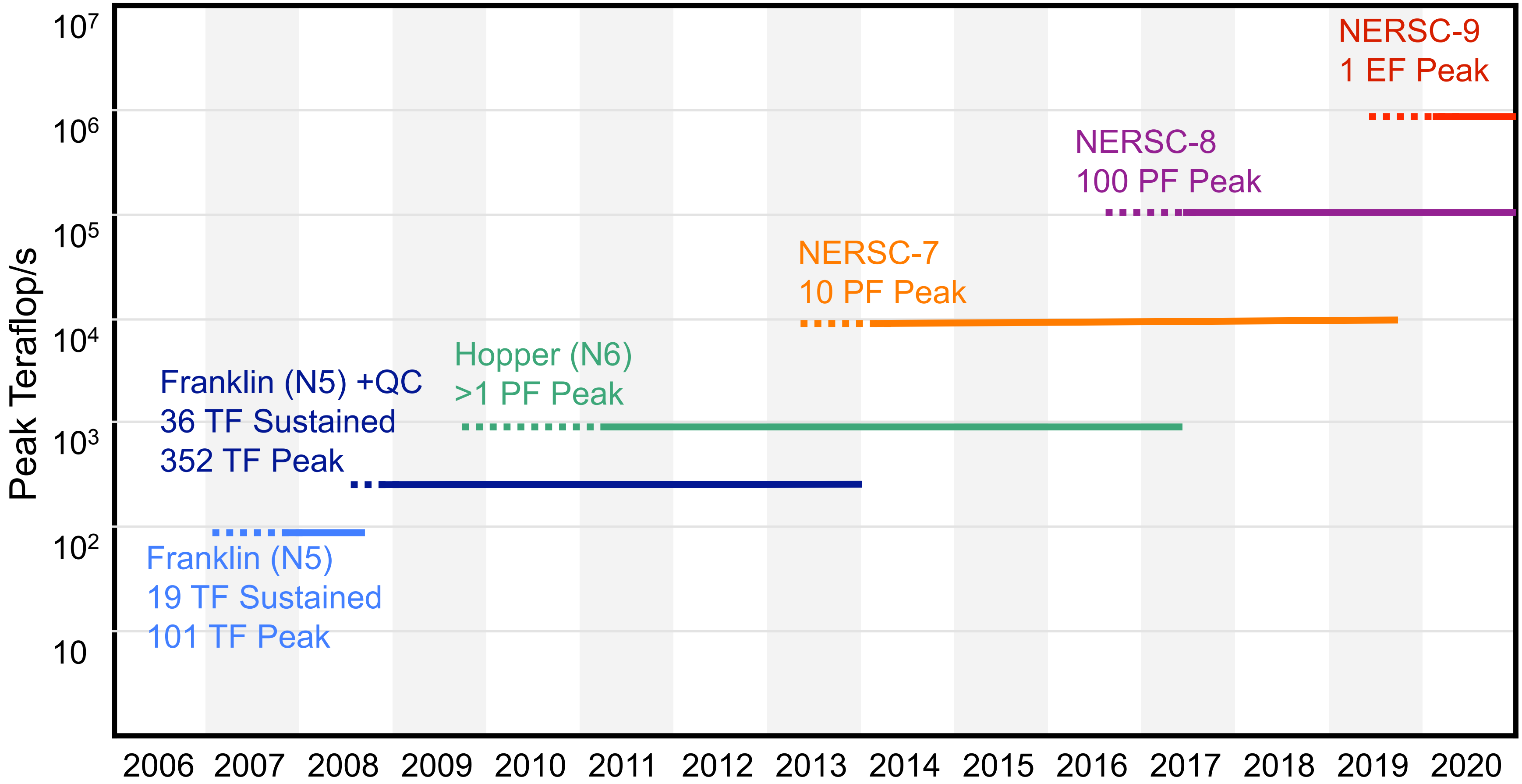
- *In production on 3/1/2010*
- 668 nodes, 5,344 cores
- 2.4 GHz AMD Opteron
- 2 PB disk, 25 GB/s
- Air cooled

Phase 2: Cray system

- > 1 Pflop/s peak
- ~ 150K cores, 6250 nodes
- 12 AMD Magny Cours chips, 2 per node (dual socket)
- Liquid cooled

4Q09	1Q10	2Q10	3Q10	4Q10	1Q11

NERSC System Roadmap



- **Goal is two systems on the floor at all times**
- **Roadmap shows technology picture with exascale investments**
- **Science requirements determine if this is necessary**

NISE: NERSC Initiative for Scientific Exploration

- NERSC Users: Open process for 10% NERSC time
- Modeled after original INCITE program from NERSC:
  - Focused computing and consulting resources
- NISE program at NERSC (started in 2009)
  - Programming techniques for multicore and scaling in general
  - Science problems near breakthrough (high risk/payoff)
  - http://www.nersc.gov/nusers/accounts/NISE.php
- ~30M hours made available to NISE projects in 2010
  - E.g., V. Izzo, GA, ITER rapid shut-down simulation
- ASCR's ALCC Program has a similar number of hours:

http://www.er.doe.gov/ascr/Facilities/ALCC.html

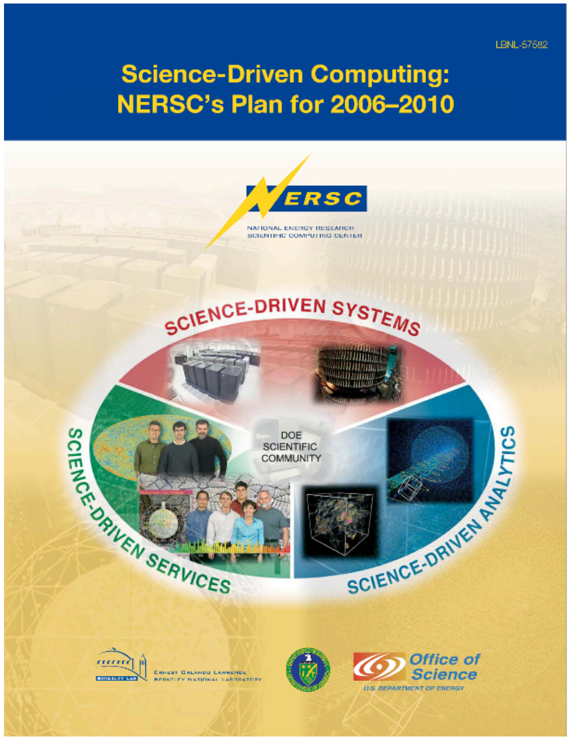
NERSC Response to Science Needs

- Installation of Hopper (NERSC-6) system
  - Phase 1 in production; Phase 2 production in 2011
- Replacement of Bassi and Jacquard by Carver
  - Saved money, space, and energy
- Replaced Davinci by Euclid for analytics
- Upgrading global filesystem (NGF) from 1.6 to 2.5 PB
  - Enabled NGF access from Franklin; available for Hopper
- Added external services to Franklin and Hopper
- Added testbeds for clouds (Magellan) and GPUs (Dirac) paid by non-program funds
- Facility upgrade from 6 MW to 9 MW

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



LBNL-57582

Science-Driven Computing:
NERSC's Plan for 2006–2010

NERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

BERKELEY LAB
ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

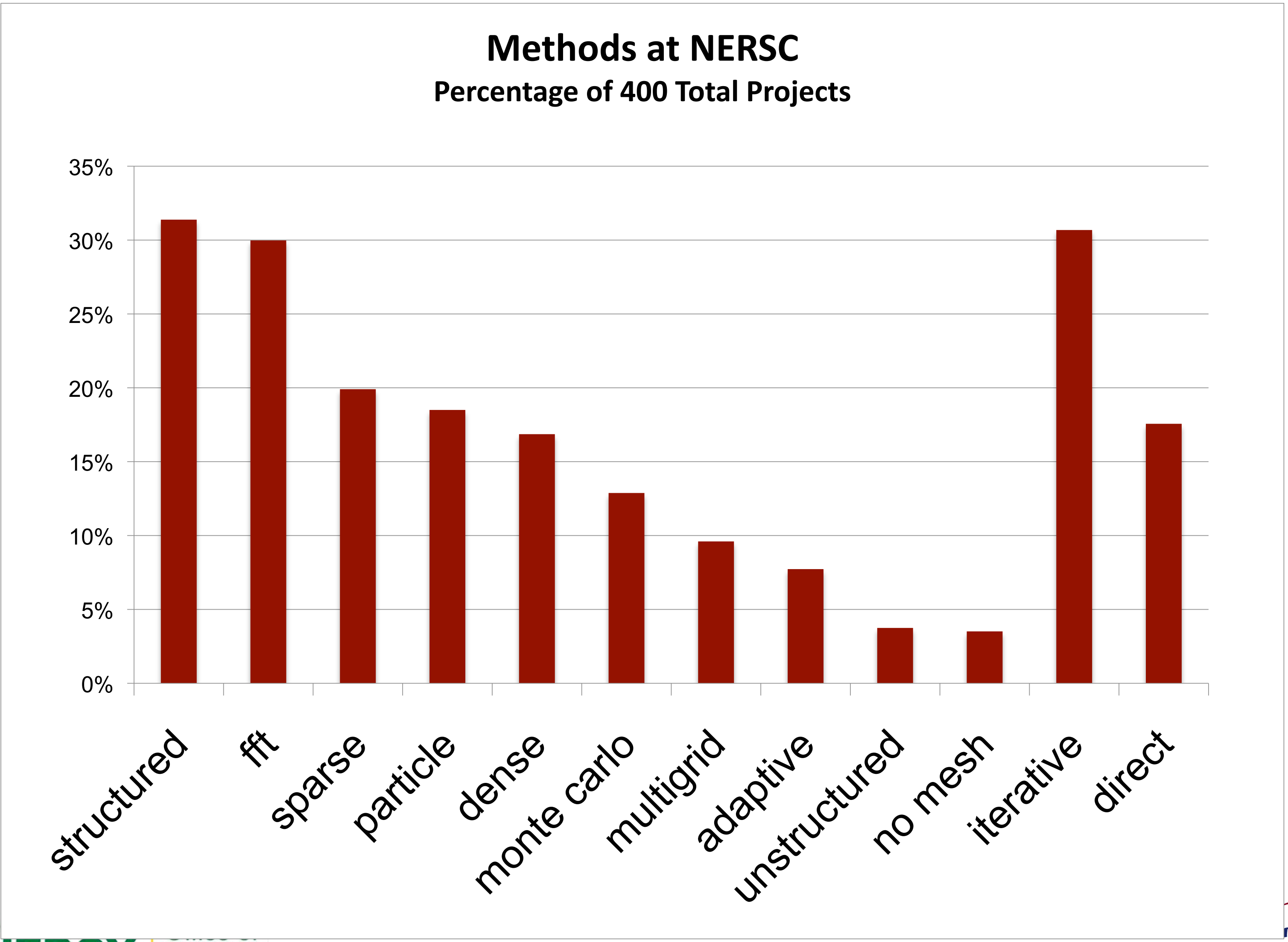
Office of Science
U.S. DEPARTMENT OF ENERGY

NERSC Strategy 2010

- **Observations**
  - **End of single processor performance gains leading to *multicore, GPU*, and other hardware innovations**
  - ***Energy* costs a growing concern for facilities and possible barrier to exascale**
  - **New capacity computing available in *Clouds***
  - ***Flood of data* is increasing from both simulations and experiments**
- **NERSC future priorities are driven by science:**
  - **Current user demand surpasses resources: Does scientific need increase by 10x every 3 years?**
  - **Balanced facility for data as well as compute needs: What are specific requirements?**

Numerical Methods at NERSC

(Caveat: survey data from ERCAP requests)



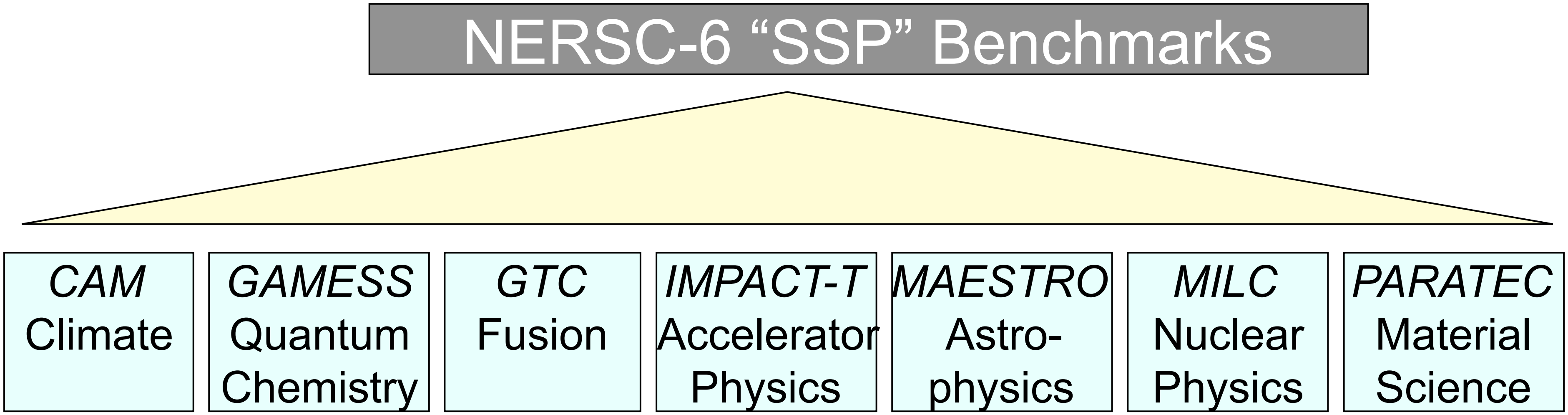
Algorithm Diversity

Science areas	*Dense linear algebra*	*Sparse linear algebra*	*Spectral Methods (FFT)s*	*Particle Methods*	*Structured Grids*	*Unstructured or AMR Grids*
Accelerator Science		X	X	X	X	X
Astrophysics	X	X	X	X	X	X
Chemistry	X	X	X	X		
Climate			X		X	X
Combustion					X	X
Fusion	X	X		X	X	X
Lattice Gauge		X	X	X	X	
Material Science	X		X	X	X	

NERSC users require a system which performs well in all areas

Applications Drive NERSC Procurements

Because hardware peak performance does not necessarily reflect real application performance



NERSC-6 “SSP” Benchmarks

CAM Climate

GAMESS Quantum Chemistry

GTC Fusion

IMPACT-T Accelerator Physics

MAESTRO Astro-physics

MILC Nuclear Physics

PARATEC Material Science

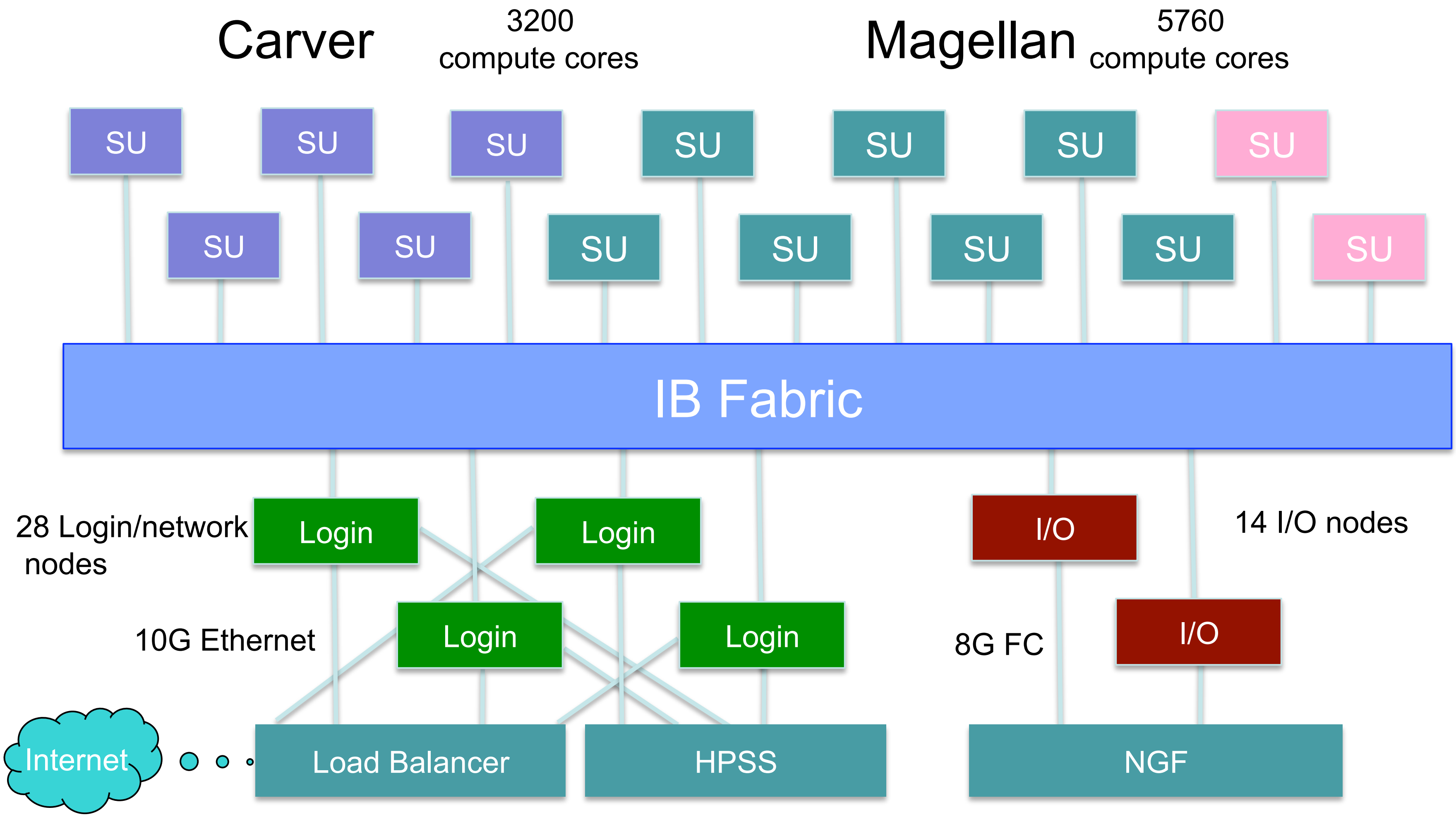
- Benchmarks reflect diversity of science and algorithms
- SSP = average performance (Tflops/sec) across machine
- Used before selection, during and after installation
- Question: What applications best reflect your workload?

DOE Explores Cloud Computing

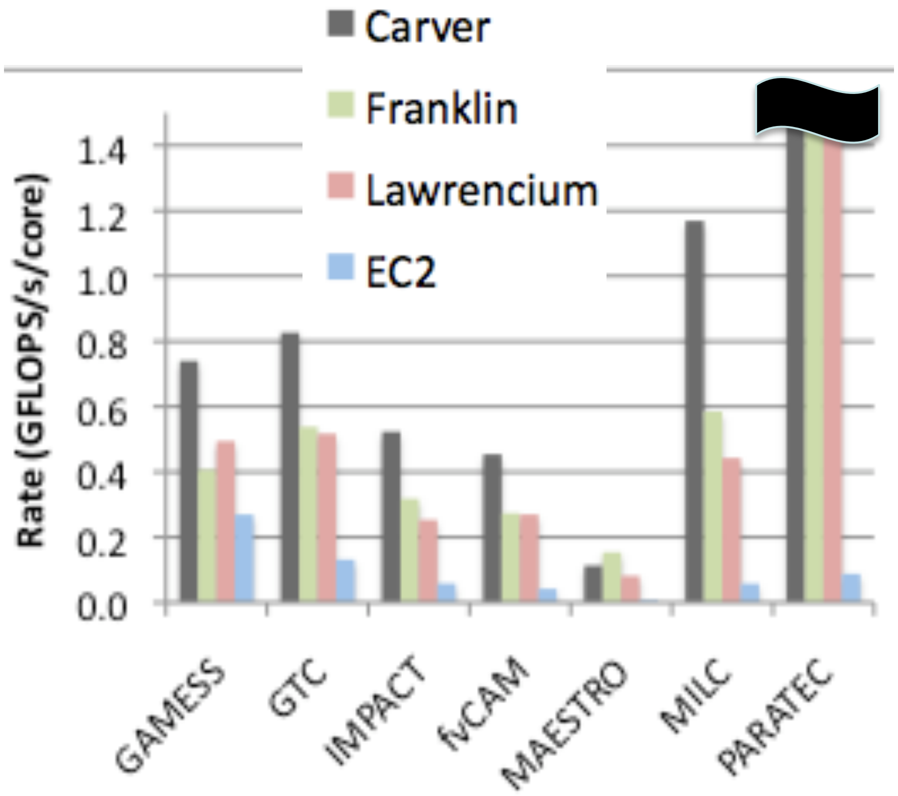
FERDINAN·MAGELLANVS·SVPERATIS
ANTARCTICI·FRETI·ANGVSTIIS·CLARISS

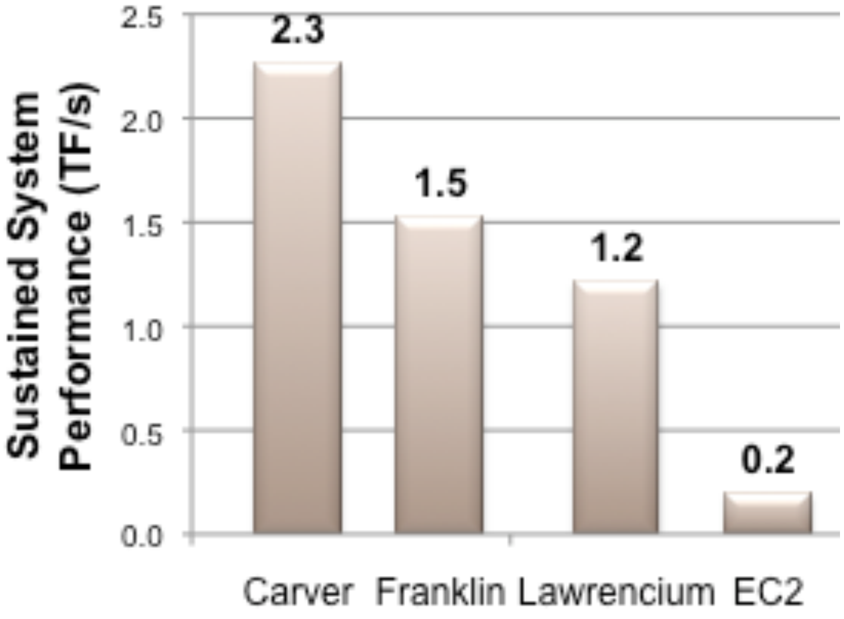
- **DOE's CS program focuses on HPC**
  - No coordinated plan for clusters in SC
- **DOE Magellan Cloud Testbed**
- **Cloud questions to explore:**
  - Can a cloud serve DOE's mid-range computing needs?
  - What features (hardware and software) are needed of a "Science Cloud"? Commodity hardware?
  - What requirements do the jobs have (~100 cores, I/O,…)
  - How does this differ, if at all, from commercial clouds?
- **What are main attractions?**
  - Elastic computing: good business model for fixed computational problems, but what about science?
  - Control over software stack (virtualization)

Cluster architecture



Application Rates and SSP





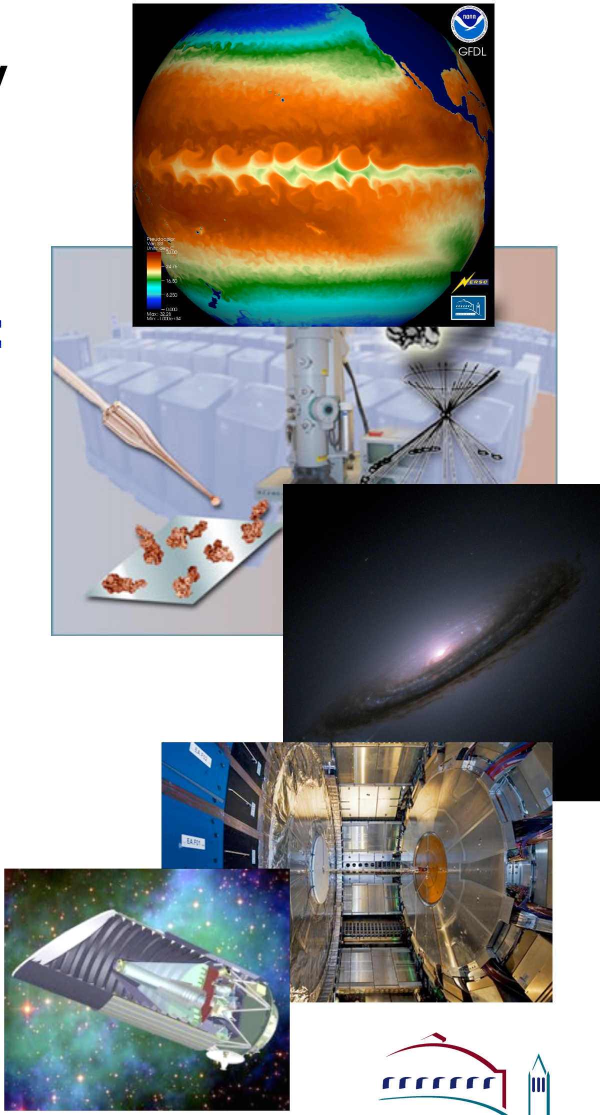
Problem sets drastically reduced for cloud benchmarking

Reservations at NERSC

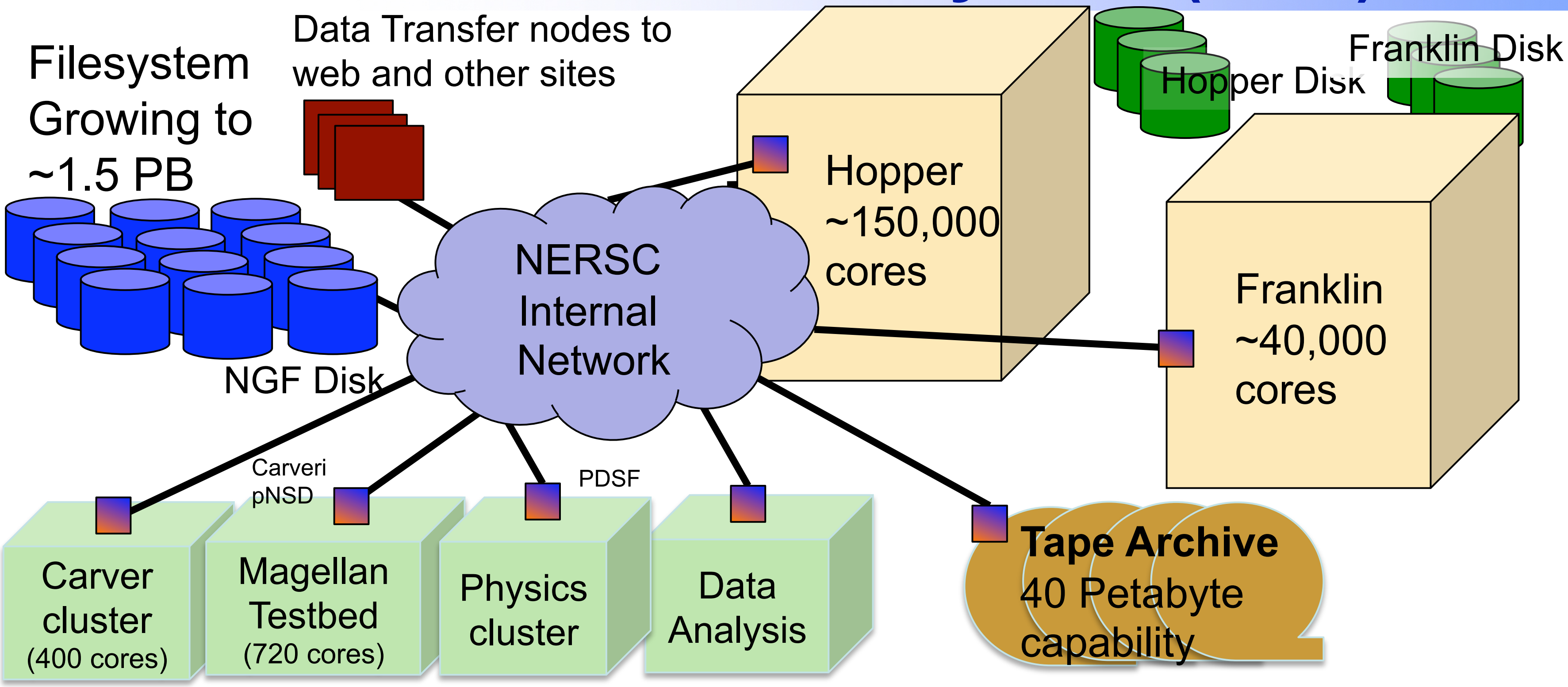
- Reservation service being tested:
  - – Reserve a certain date, time and duration
    - Debugging at scale
    - Real-time constraints in which need to analyze data before next run, e.g., daily target selection telescopes or genome sequencing pipelin
  - – At least 24 hours advanced notice
    - https://www.nersc.gov/nusers/services/reservation.php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

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**



NERSC Architecture with NERSC Global Filesystem (NGF)



Filesystem Growing to ~1.5 PB

Data Transfer nodes to web and other sites

Hopper ~150,000 cores

Hopper Disk

Franklin Disk

Franklin ~40,000 cores

NERSC Internal Network

NGF Disk

Carveri pNSD

PDSF

Carver cluster (400 cores)

Magellan Testbed (720 cores)

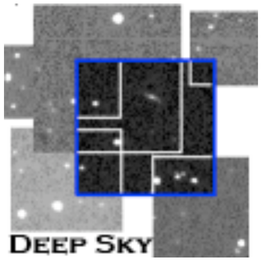
Physics cluster

Data Analysis

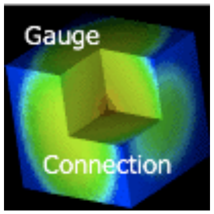
Tape Archive 40 Petabyte capability

- NERSC invests annually in storage hardware, both filesystems and the HPSS Tape archive
- Innovate to make these more convenient for users

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**



Communication Services

Since 2007, NERSC is a net data importer. In support of our users, it is important that we take on a lead role in improving intersite data transfers.

- Systems and software typically tuned only within a site
- Technical, social, and policy challenges abound:
  - High performance transfer software has too many options → hard to use.
  - Systems designed for computation can have bottlenecks in data transfers
  - Systems at different sites often often have incompatible versions of transfer software.
  - Trying to maintain security exceptions (firewall holes) for all the systems and software at each site was impossible.
- … and the list goes on.
- NERSC established Data Transfer Nodes (DTNs).
  - Reduced transfer time of 30 TB from 30+ days to 2 days
  - We formed a working group with experts at the three labs and ESNet

http://www.nersc.gov/nusers/systems/DTN

http://fasterdata.es.net

Visualization Support

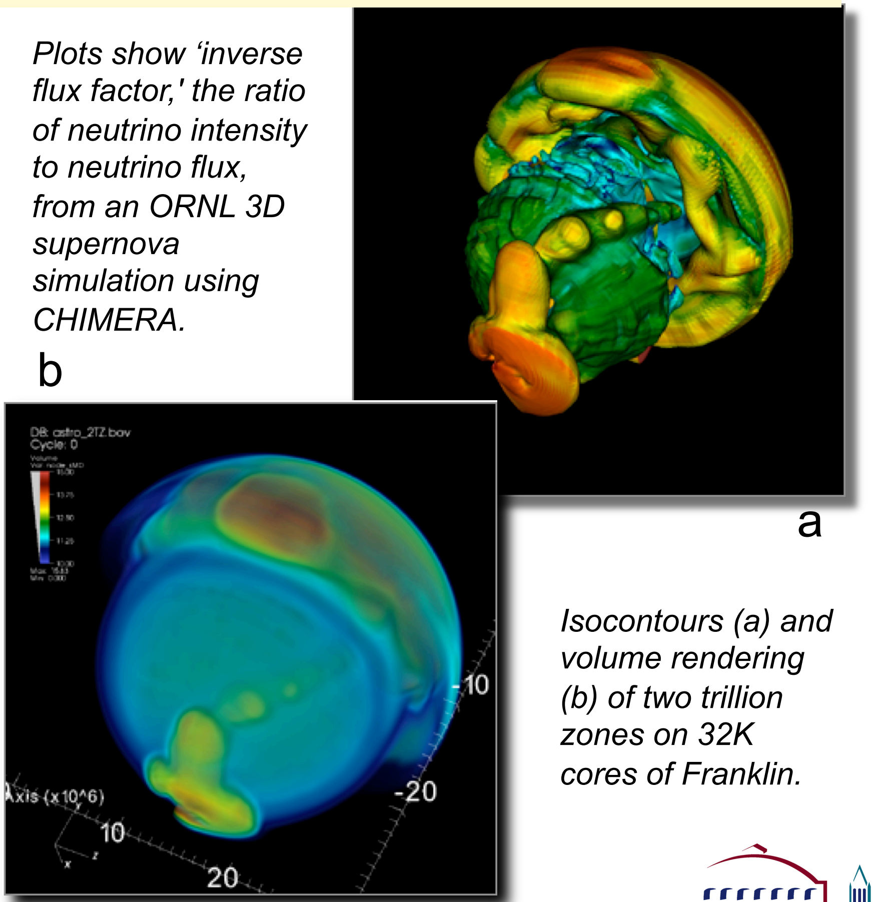
Petascale visualization: Demonstrate visualization scaling to unprecedented concurrency levels by ingesting and processing unprecedentedly large datasets.

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.

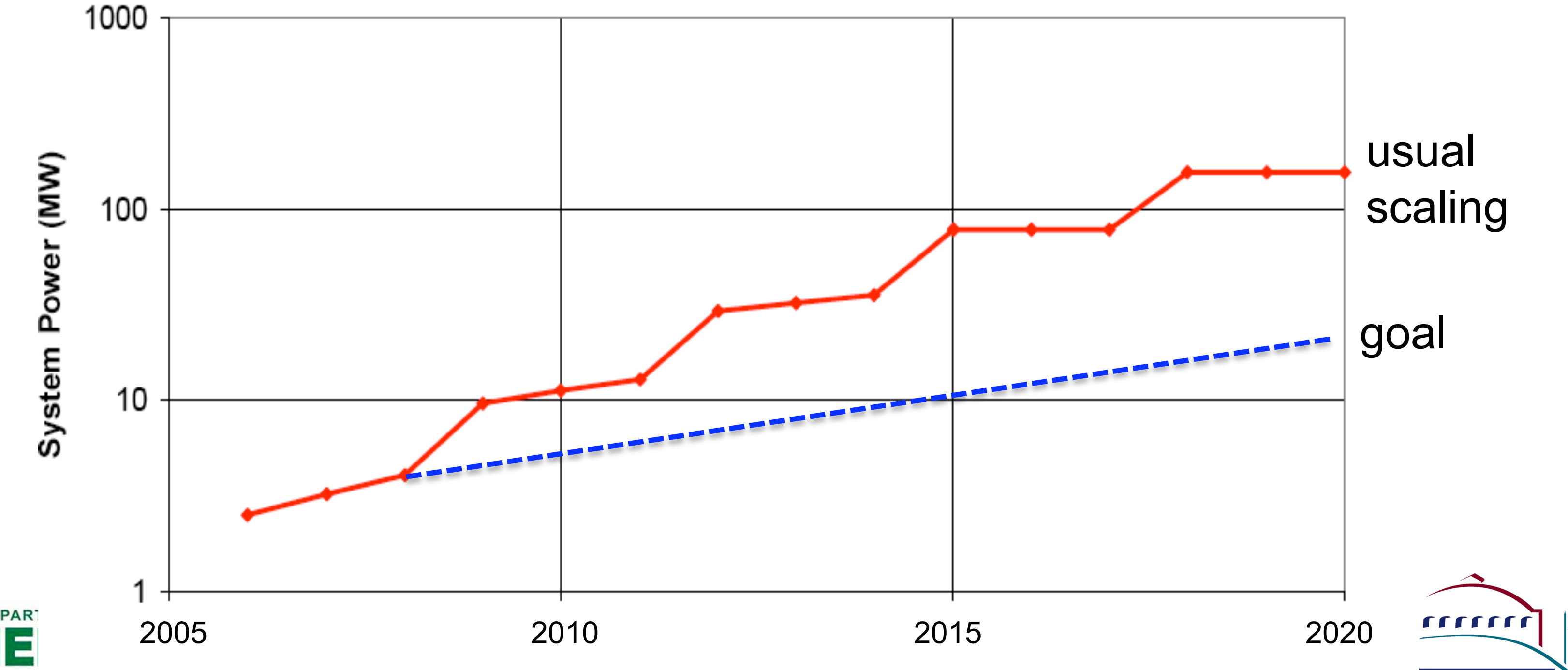


Isocontours (a) and volume rendering (b) of two trillion zones on 32K cores of Franklin.

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

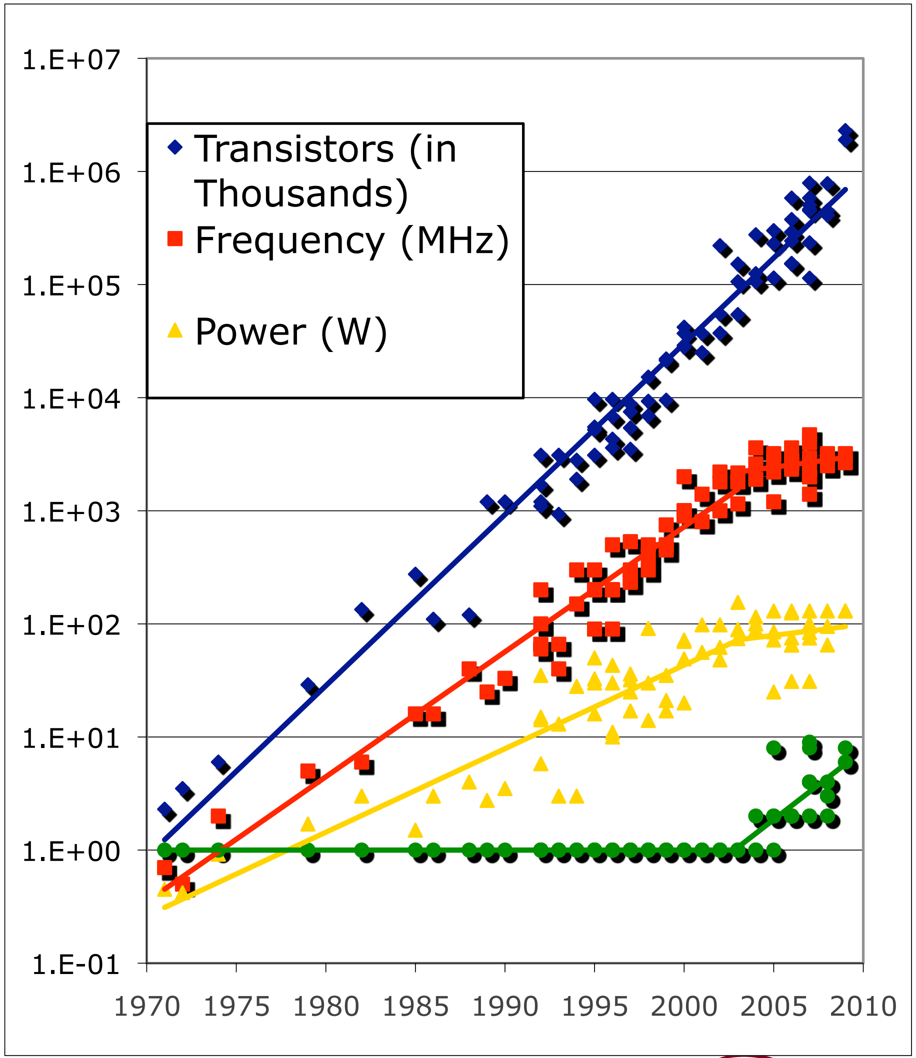


NERSC Goal Usable Exascale in 2020

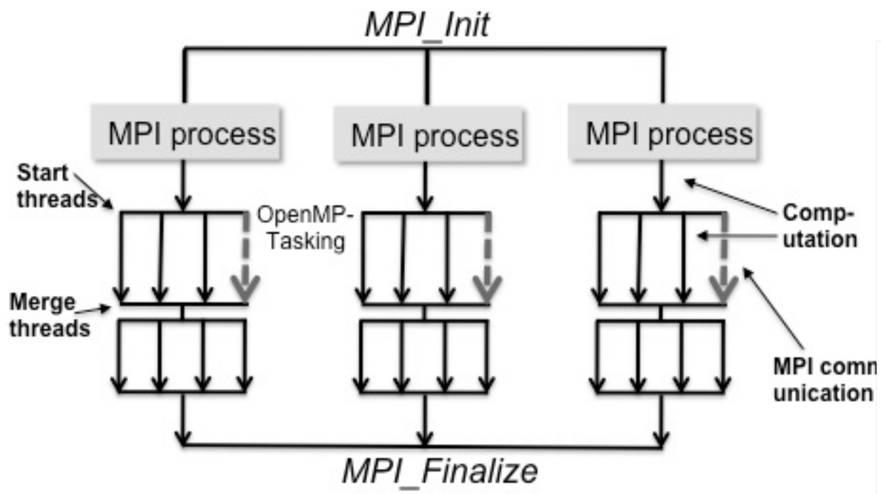
- **Computational scaling changed in 2004**
- **Problems also for laptops, handhelds, data centers**
- **Parallelism on-chip brings algorithms, programming into question**
- ***NERSC: Programmable, usable systems for science***

1) Energy efficient designs

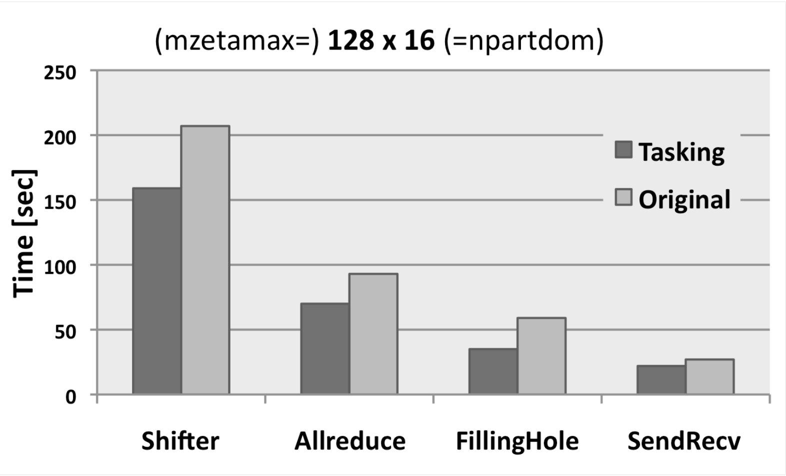
2) Facilities to support scale for both high and mid scale



Hybrid OpenMP/MPI Programming in GTS



2048 size run



- NEW OpenMP Tasking Model allows for communication / computation overlap in hybrid computation.

Work by A. Koniges, R. Preissel, N. Wichmann et al presented at Cray User Group meeting; best presentation award

Challenges to Exascale

- **System power** is the primary constraint for exascale (MW ~= $M)
- **Concurrency** (1000x today) driven by system power and density
- **Memory** bandwidth and capacity are under pressure
- **Processor** architecture is an open question, but heterogeneity is likely
- **Algorithms** need to be designed to minimize data movement, not flops
- **Programming model** memory & concurrency → new chip-level model
- **Reliability and resiliency** will be critical at this scale
- **I/O bandwidth** unlikely to keep pace with machine speed

Why are these NERSC problems?

- **All** are challenges for 100 "capacity" 100PF machines, except:
  - System wide outages and bisection bandwidth
- NERSC needs to guide the transition of its user community
- NERSC represents broad HPC market better than any other DOE center
- Hardware revolutions affect procurement strategies (Giga->Tera)

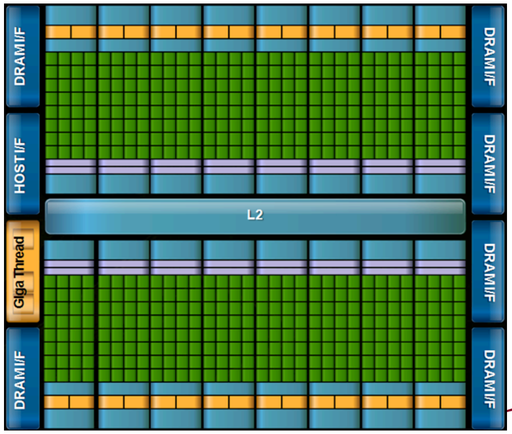
Revolutions Require a New Strategy

- **Preserve Application Performance as goal**
  - But, allow significant optimizations
  - Produced optimized versions of some codes
  - Understand performance early
- **E.g., for Fermi-based GPU system**
  - Fermi is nearly as expensive as host node
  - 48 nodes w/Fermi or 2x more nodes without
  - Minimum: Fermi must be 2x faster than host
- **Estimating value of acceleration for SSP**
  - Estimate "best possible performance"
  - Successive refinement of estimates
  - Stop if estimate < min threshold (2x host)

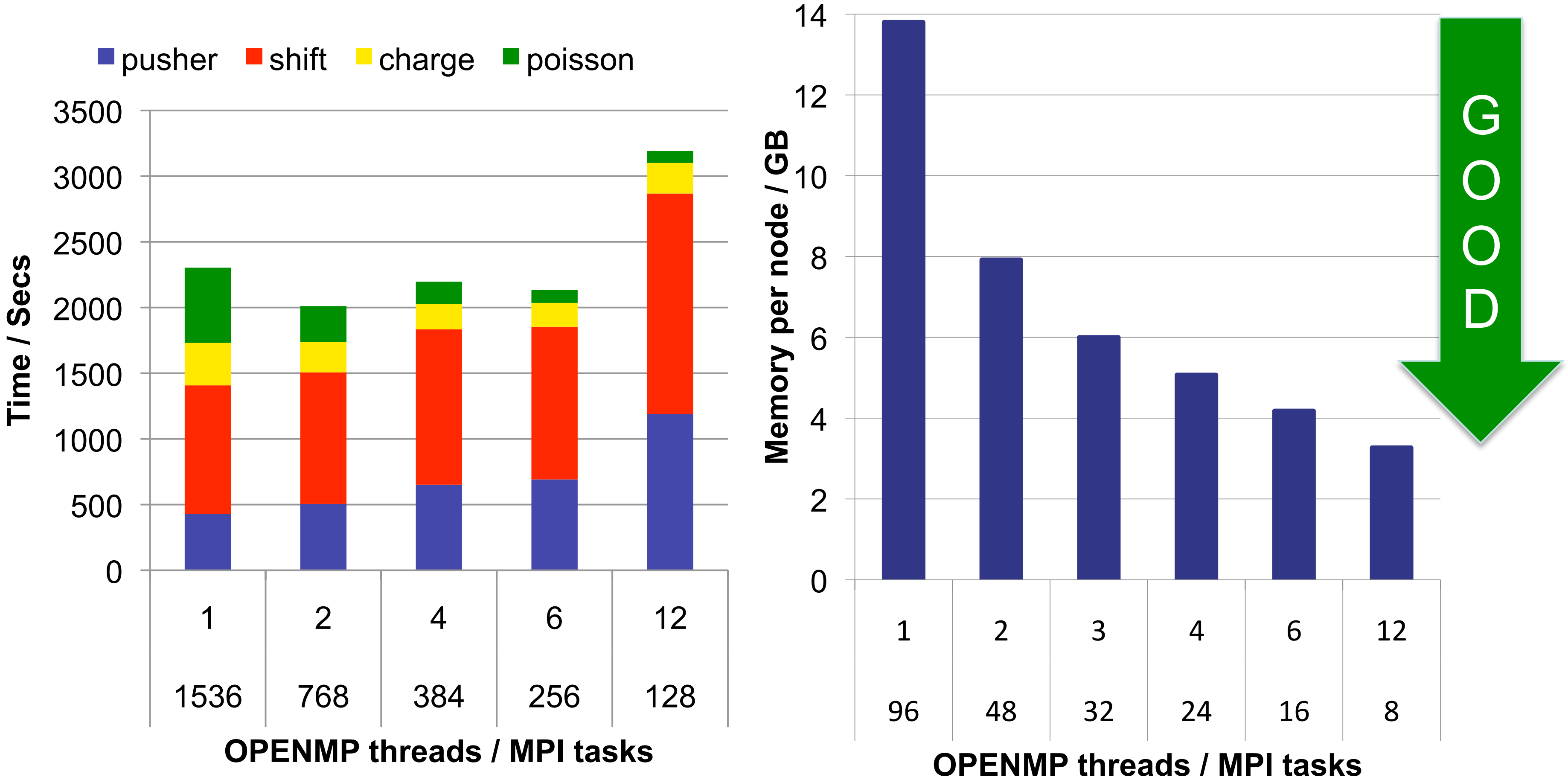
Dirac Testbed at NERSC

- Users (HEP): NERSC needs architecture / GPUs testbed
- Dirac is a 48 nodes GPU cluster
- 44 Fermi nodes
  - – 3GB memory per card
  - – ~1TF peak single precision and 500 GF/s DP
  - – @~144GB/s bandwidth + ECC
  - – 24 GB of memory per node
- 4 Tesla nodes
- QDR Infiniband network

Paul Dirac, Nobel prize-winning Theoretical Physicist



GTC MPI+OpenMP Performance



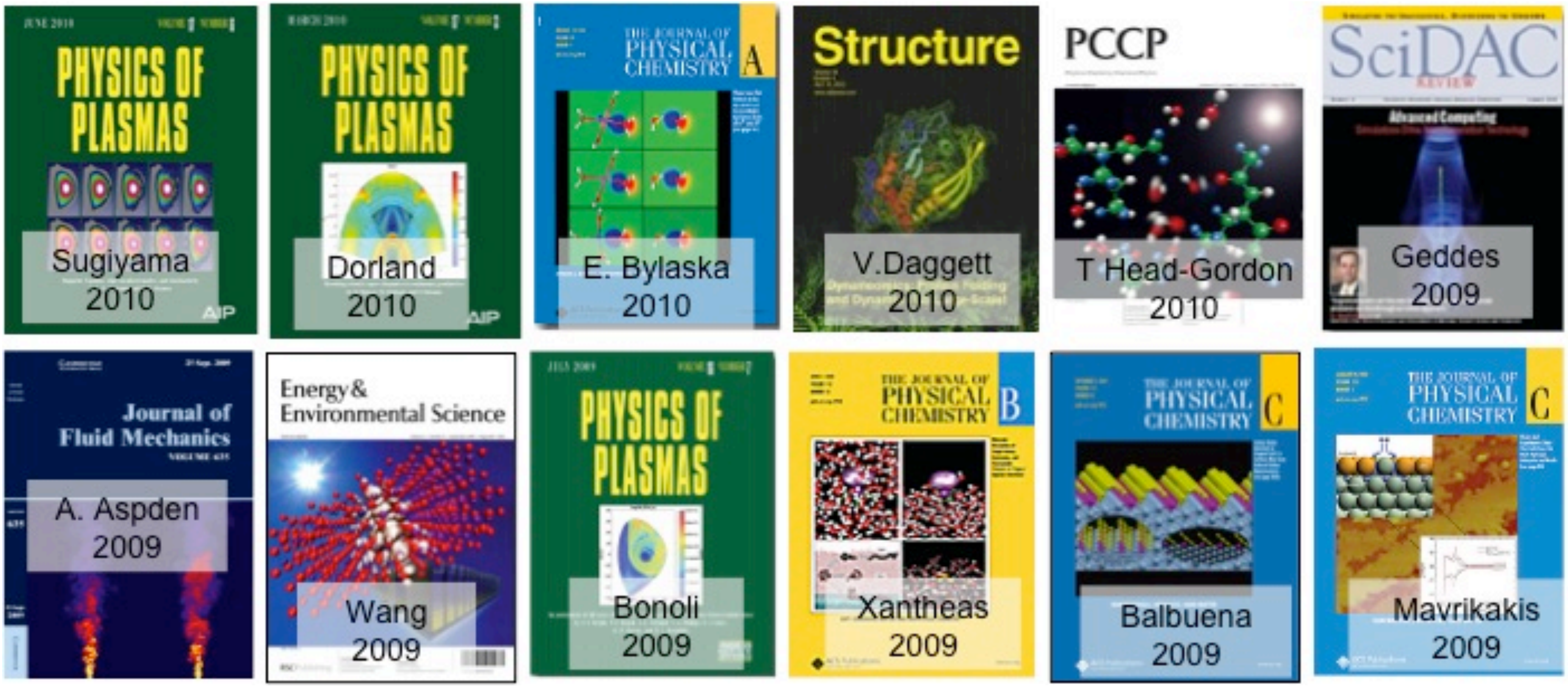
Work by N. Wright, N. Wichmann et al as part of Cray/NERSC Center of Excellence

Conclusions

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

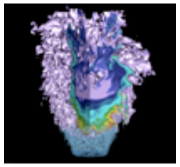
    "What gets measure gets improved"
- **Goals:**
  - Your goal is to make scientific discoveries
    - Articulate specific scientific goals and implications for broader community
  - Our goal is to enable you to do science
    - Specify resources (services, computers, storage, …) that NERSC could provide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

Recent Cover Stories from NERSC Research

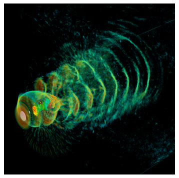


NERSC is enabling new high quality science across disciplines, with over *1,600* refereed publications last year

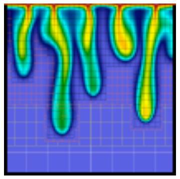
About the Cover



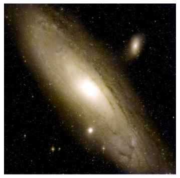
Low swirl burner combustion simulation. Image shows flame radical, OH (purple surface and cutaway) and volume rendering (gray) of vortical structures. Red indicates vigorous burning of lean hydrogen fuel; shows cellular burning characteristic of thermodiffusively unstable fuel. Simulated using an adaptive projection code. Image courtesy of John Bell, LBNL.



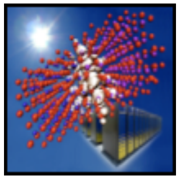
Hydrogen plasma density wake produced by an intense, right-to-left laser pulse. Volume rendering of current density and particles (colored by momentum orange - high, cyan - low) trapped in the plasma wake driven by laser pulse (marked by the white disk) radiation pressure. 3-D, 3,500 Franklin-core, 36-hour LOASIS experiment simulation using VORPAL by Cameron Geddes, LBNL. Visualization: Gunther Weber, NERSC Analytics.



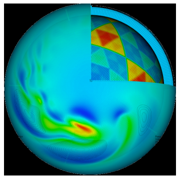
Numerical study of density driven flow for CO_2 storage in saline aquifers. Snapshot of CO_2 concentration after convection starts. Density-driven velocity field dynamics induces convective fingers that enhance the rate by which CO_2 is converted into negatively buoyant aqueous phase, thereby improving the security of CO_2 storage. Image courtesy of George Pau, LBNL



False-color image of the Andromeda Galaxy created by layering 400 individual images captured by the Palomar Transient Factory (PFT) camera in February 2009. NERSC systems analyzing the PTF data are capable of discovering cosmic transients in real time.   Image courtesy of Peter Nugent, LBNL.



The exciton wave function (the white isosurface) at the interface of a ZnS/ZnO nanorod. Simulations performed on a Cray XT4 at NERSC, also shown. Image courtesy of Lin-Wang Wang, LBNL.

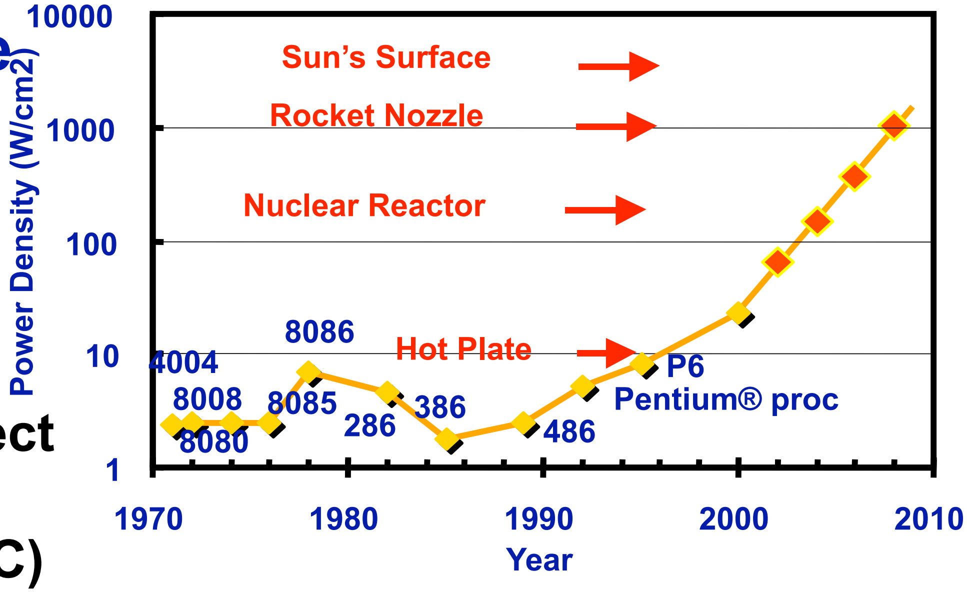


Simulation of a global cloud resolving model (GCRM). This image is a composite plot showing several variables: wind velocity (surface pseudocolor plot), pressure (b/w contour lines), and a cut-away view of the geodesic grid. Image courtesy of Professor David Randall, Colorado State University.

Backup on Technology Trends

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

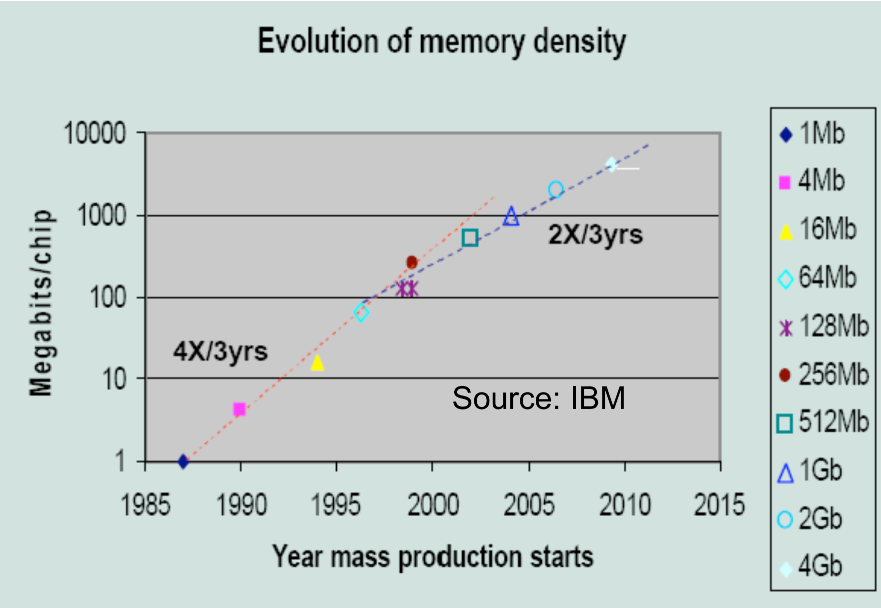


- **High performance serial processors waste power**
  - Speculation, dynamic dependence checking, etc. burn power
  - Implicit parallelism discovery
- **Question: *Can you double the concurrency in your algorithms and software every 2 years?***

Technology Challenge

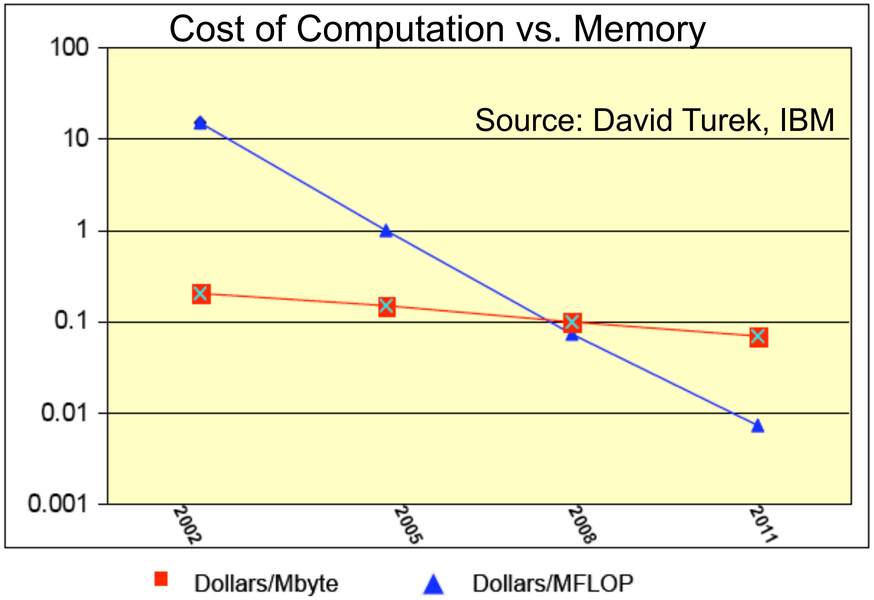
Technology trends against a constant or increasing memory per core

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



Evolution of memory density

Source: IBM



Cost of Computation vs. Memory

Source: David Turek, IBM

The cost to sense, collect, generate and calculate data is declining much faster than the cost to access, manage and store it

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