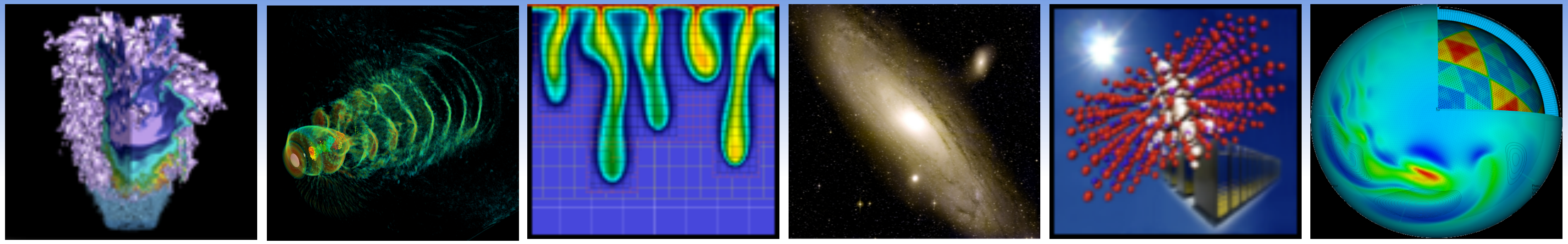


NERSC Overview

CScADS Workshop on PetaScale Applications and Performance Strategies

Zhengji Zhao

National Energy Research Scientific Computing Center
Lawrence Berkeley National Laboratory

July 19, 2010

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 to the DOE Office of Science community.

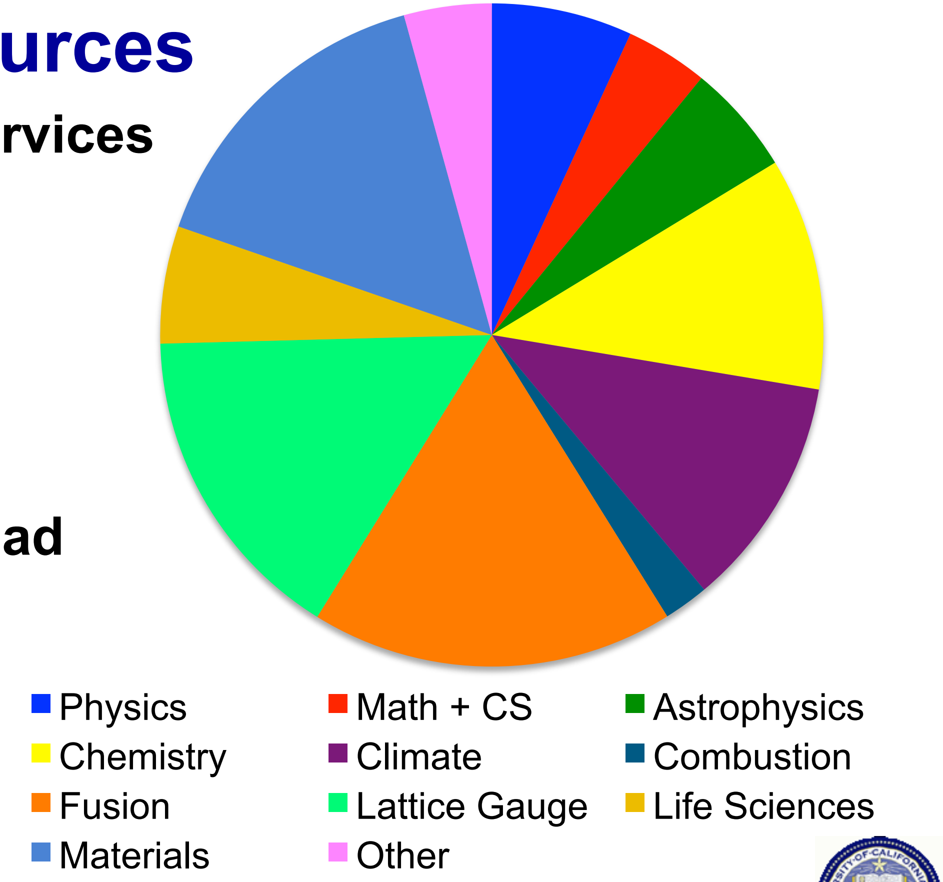
NERSC is the Production Facility for DOE Office of Science

- **NERSC serves a large population**

  Approximately 3000 users, 400 projects, 500 codes

- **Focus on "unique" resources**
  - **Expert consulting and other services**
  - **High end computing systems**
  - **High end storage systems**

- **NERSC is known for:**
  - **Outstanding services**
  - **Large and diverse user workload**



•"NERSC continues to be a gold standard of a scientific High Performance Computational Facility."
– HPCOA,Review August 2008

ASCR's Computing Facilities

NERSC

LBNL

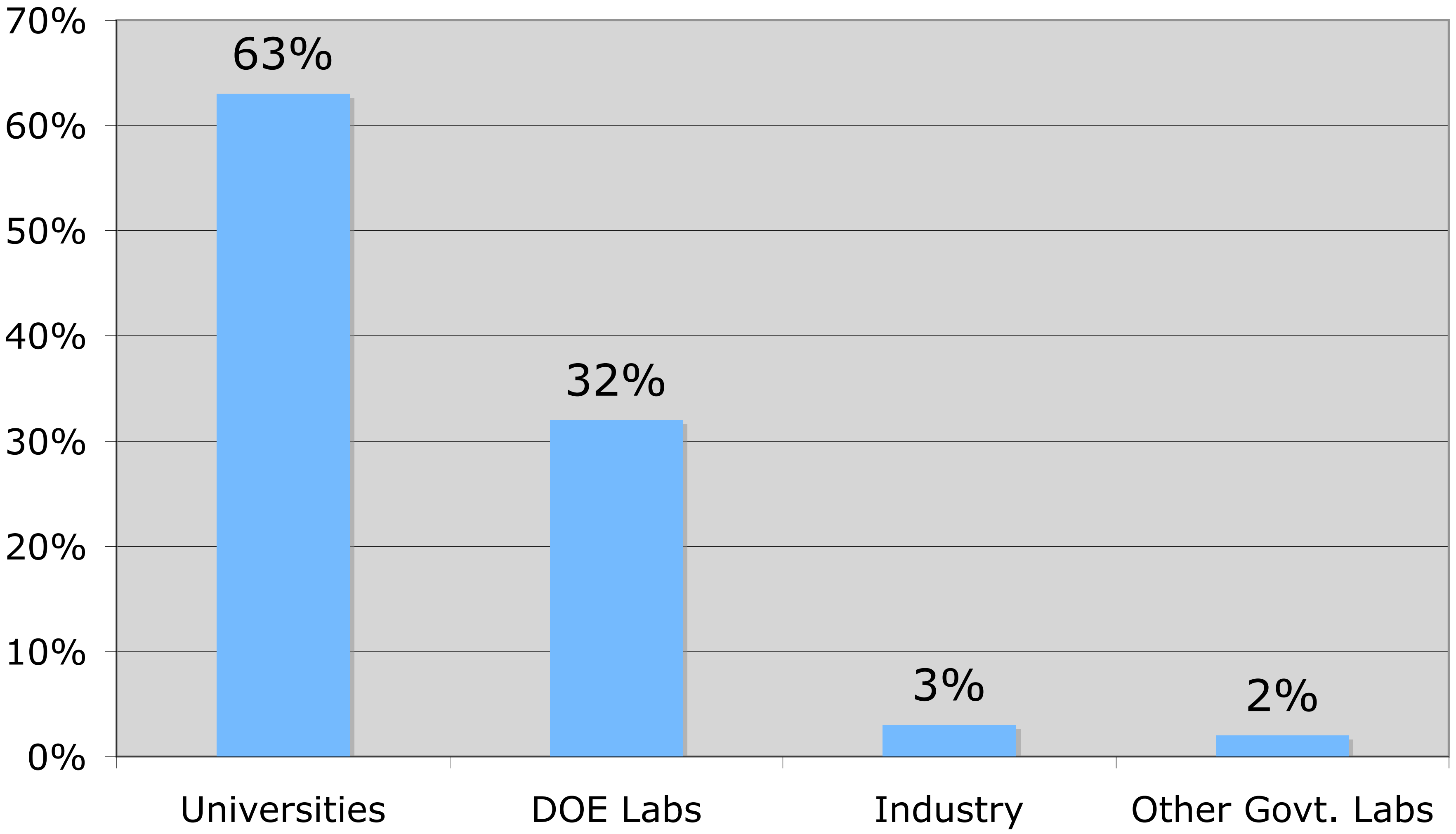
- Hundreds of projects
- 2010 allocations:
  - 70-80% SC offices control; ERCAP process
  - 10-20% ASCR (new ALCC program)
  - 10% NERSC reserve
- Science covers all of DOE/SC science

LCFs

ORNL and ANL

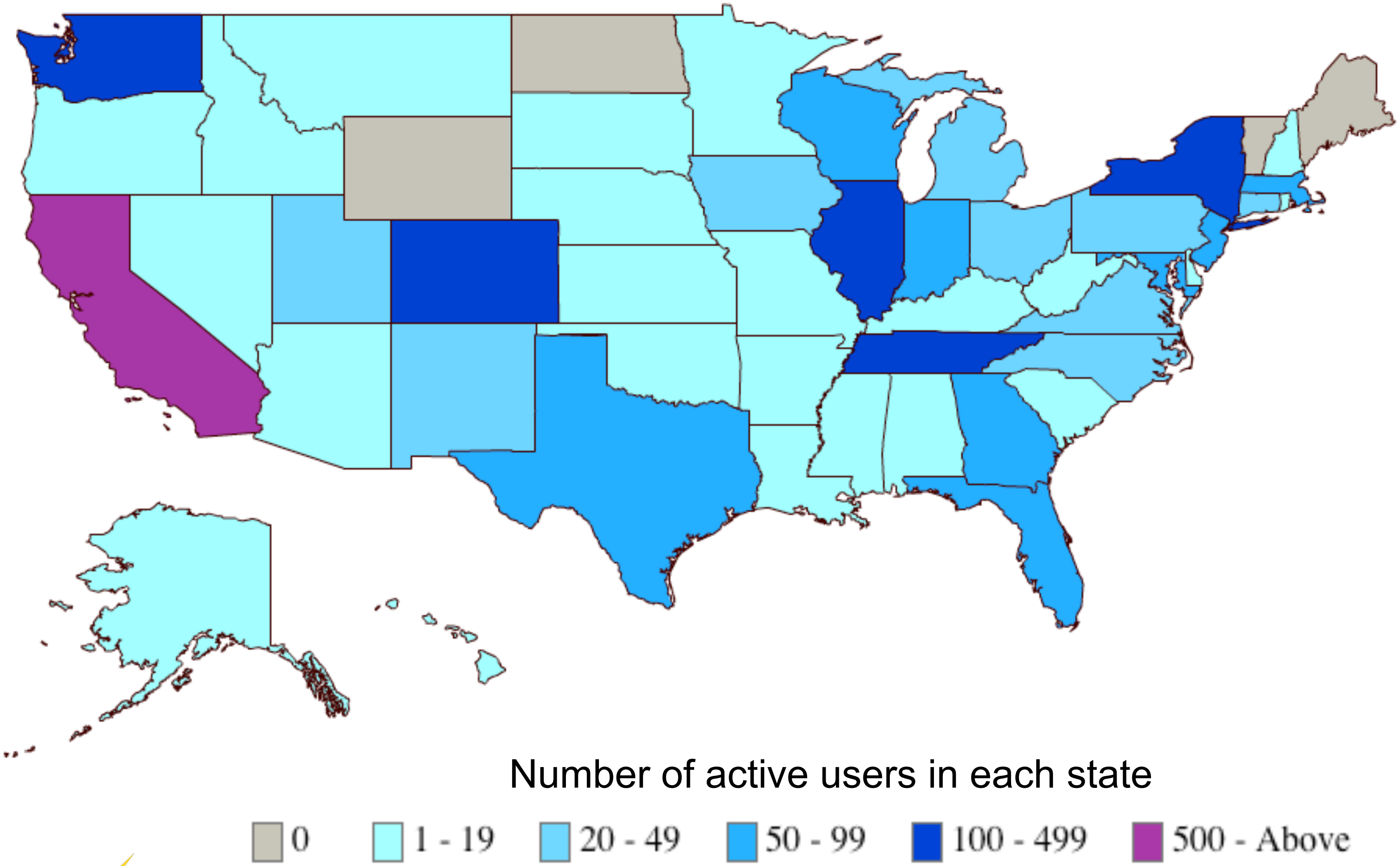
- Tens of projects
- 2010 allocations:
  - 70-80% ANL/ORNL managed; INCITE process
  - 10-20% ACSR (new ALCC program)
  - 10% LCF reserve
- Science areas limited to those at largest scale; not limited to DOE/SC

NERSC User Demographics

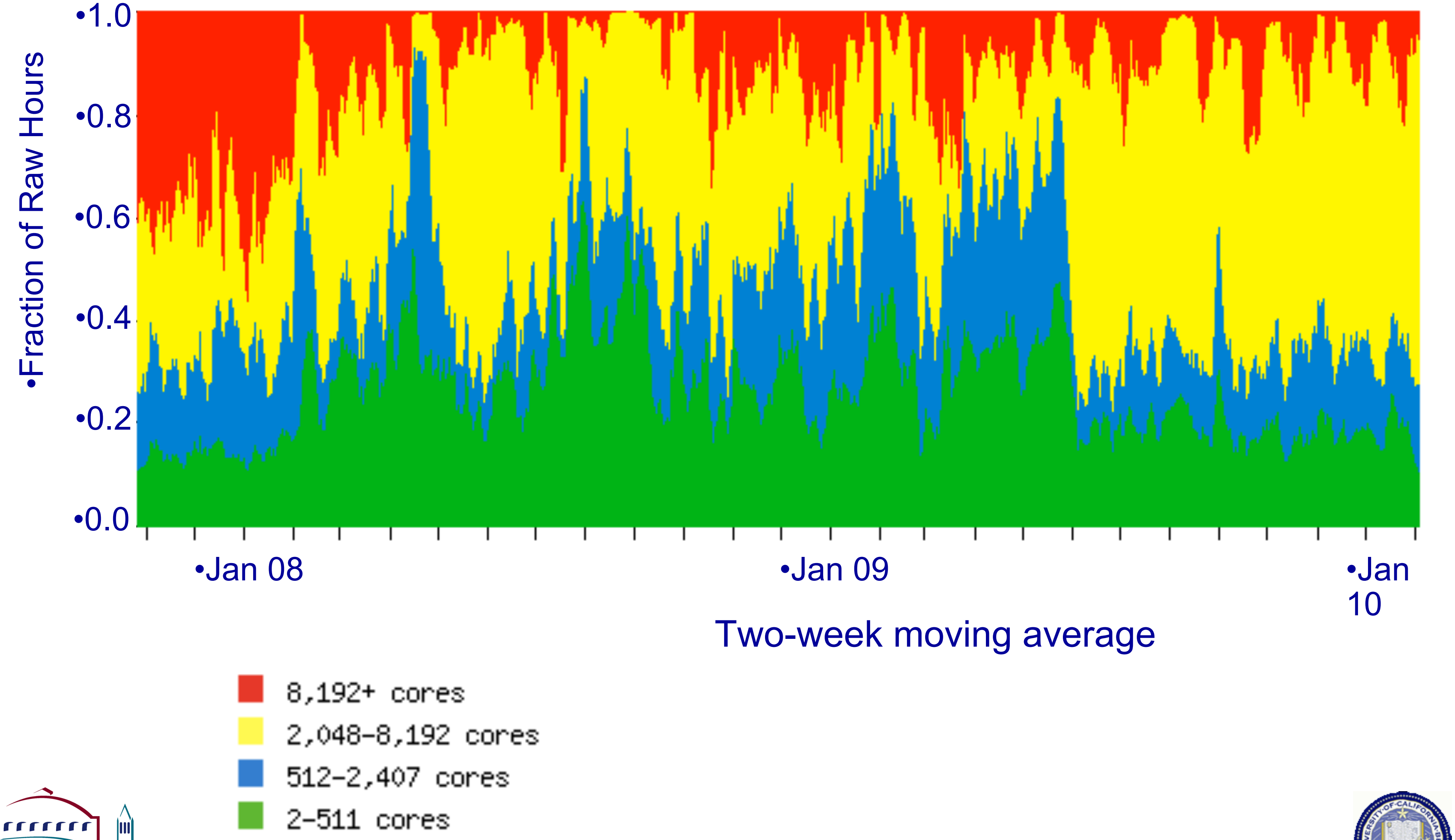


NERSC supports users throughout the country

Distribution of NERSC Users throughout the United States

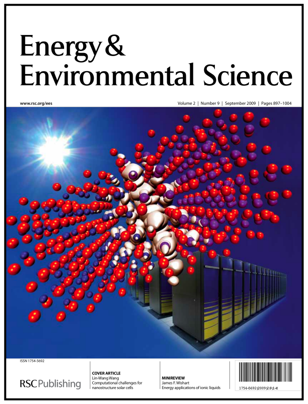


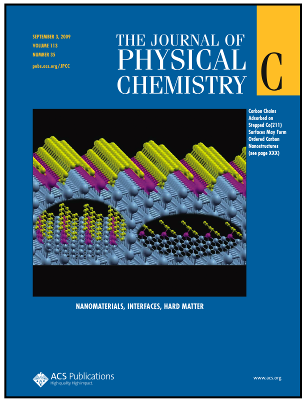
Fraction of Raw Hours by Job Parallel Concurrency

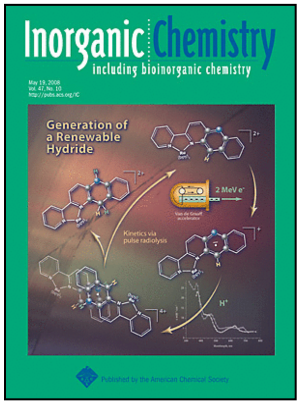


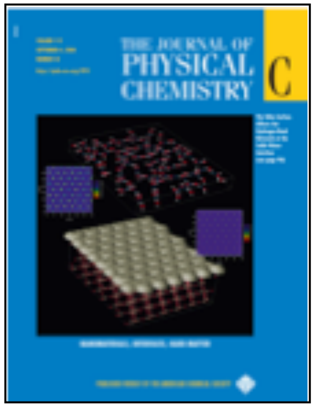
Two-week moving average

High quality science results from simulations of many different scales



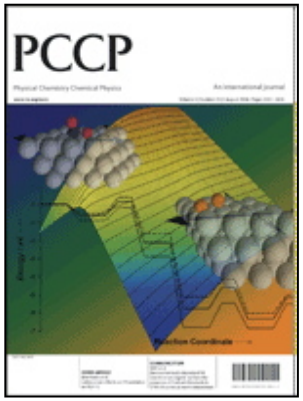

THE JOURNAL OF PHYSICAL CHEMISTRY C
NANOMATERIALS, INTERFACES, HARD MATTER
ACS Publications

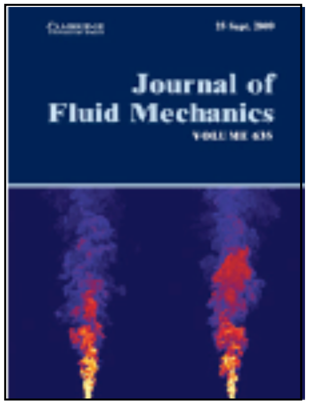

Inorganic Chemistry
including bioinorganic chemistry
Generation of a Renewable Hydride

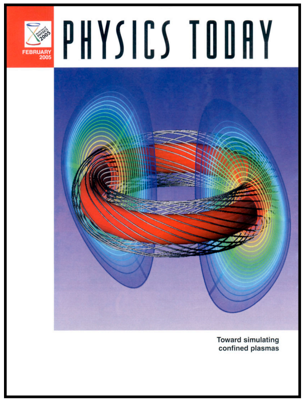


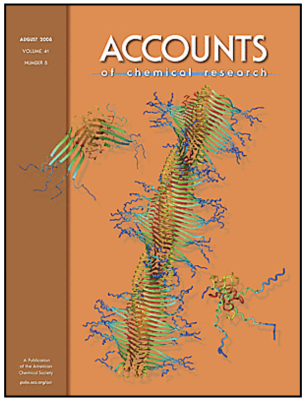
THE JOURNAL OF PHYSICAL CHEMISTRY B
CONDENSED MATTER AND BIOPHYSICAL CHEMISTRY

•64-6000 •cores	•4-128 •cores	•32 •cores	•1100-4600 •cores	•32-1024 •cores	•256 •cores





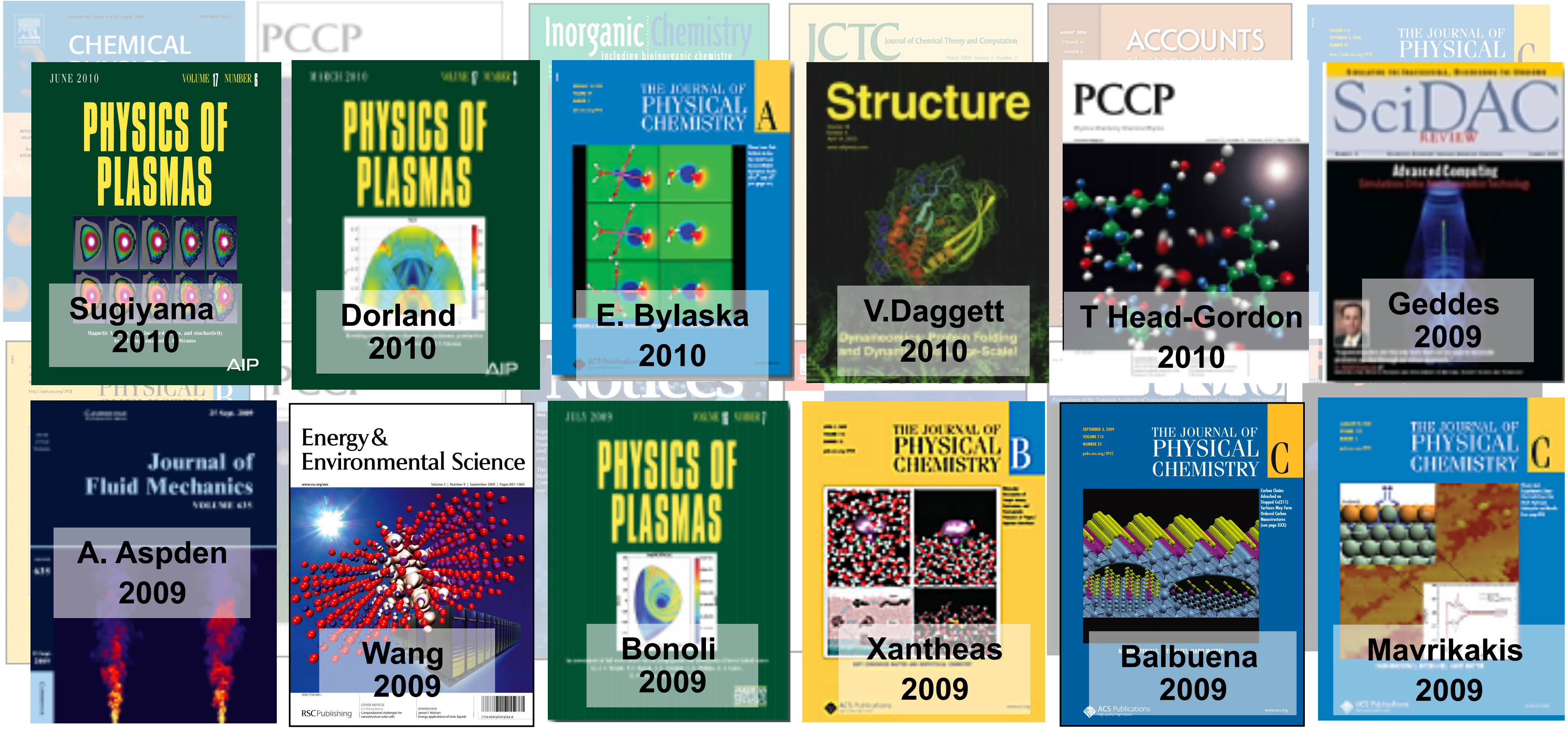




•32-256 •cores	•64-4096 •cores	•1000-4000 •cores	•80-260 •cores	•64-4700 •cores	•1100-4600 •cores

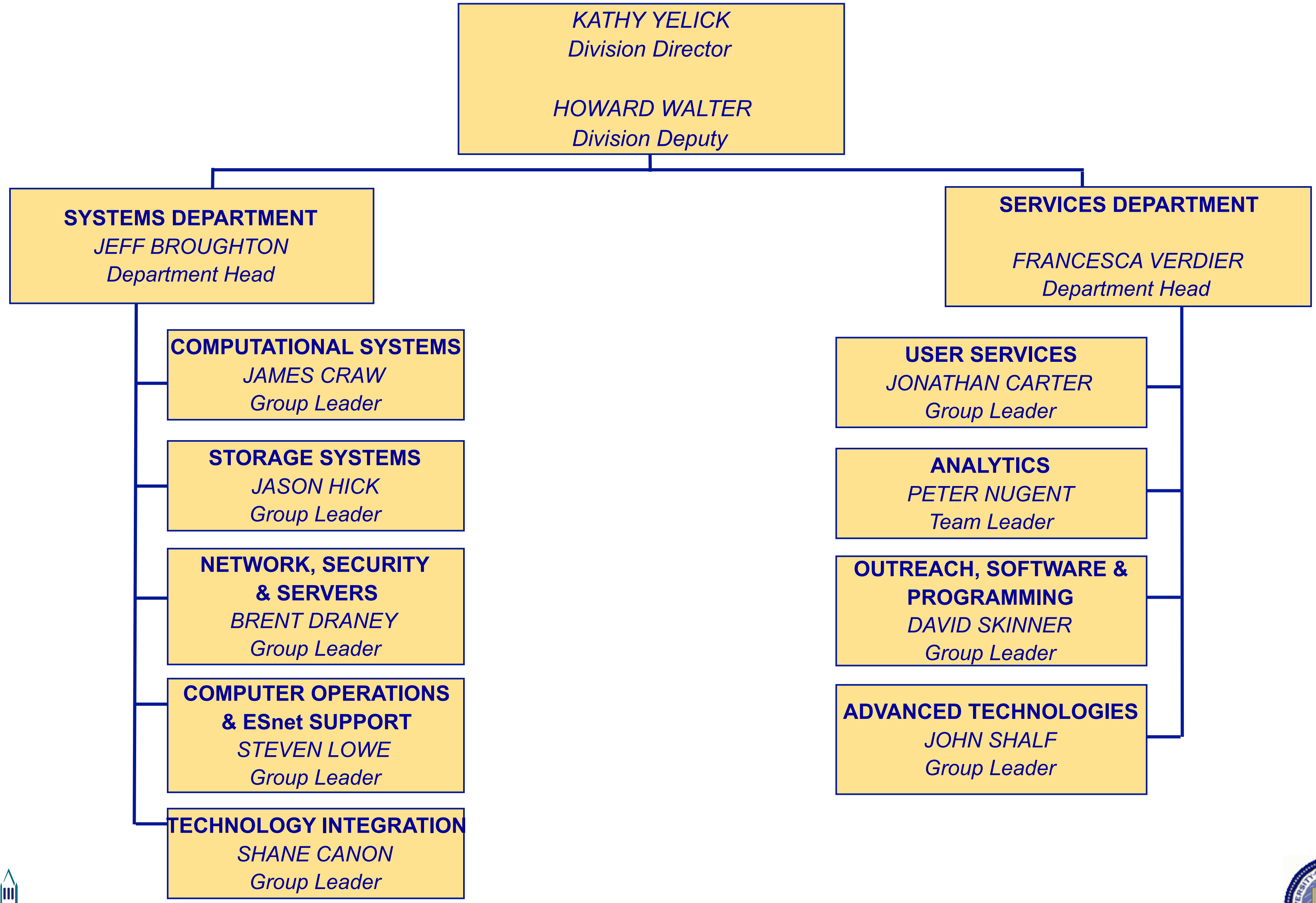
NERSC Cover Stories

Cover Stories from NERSC Research



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

NERSC Division Org Chart



KATHY YELICK
Division Director

HOWARD WALTER
Division Deputy

SYSTEMS DEPARTMENT
JEFF BROUGHTON
Department Head

COMPUTATIONAL SYSTEMS
JAMES CRAW
Group Leader

STORAGE SYSTEMS
JASON HICK
Group Leader

NETWORK, SECURITY & SERVERS
BRENT DRANEY
Group Leader

COMPUTER OPERATIONS & ESnet SUPPORT
STEVEN LOWE
Group Leader

TECHNOLOGY INTEGRATION
SHANE CANON
Group Leader

SERVICES DEPARTMENT

FRANCESCA VERDIER
Department Head

USER SERVICES
JONATHAN CARTER
Group Leader

ANALYTICS
PETER NUGENT
Team Leader

OUTREACH, SOFTWARE & PROGRAMMING
DAVID SKINNER
Group Leader

ADVANCED TECHNOLOGIES
JOHN SHALF
Group Leader

NERSC Services for Scientific Discovery: More than Hardware

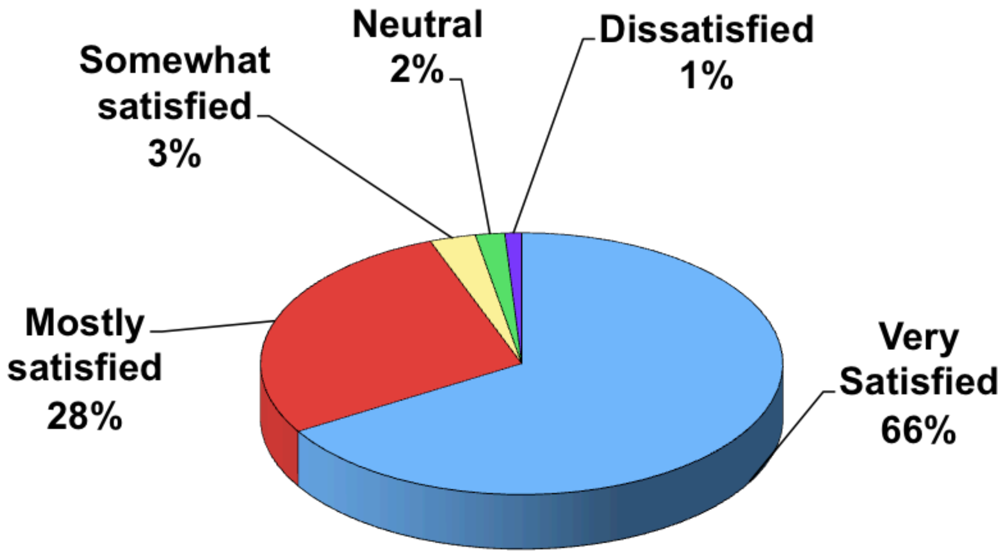
- Systems configured for productivity and usability
- Fast, high quality user services
- Easy access to data storage
- Specialized visualization and analytics services
- Highly tuned network for file transfers and connectivity
- Secure systems with minimal user interference
- Innovative and personalized web and grid
- Research on new architectures and and programming models to better serve users’ HPC needs

User Services

- 8 consultants provide broad front line support to users
- Expertise in:
  - Material Science, Chemistry, Astrophysics and Climate codes
  - Math and I/O libraries
  - Compilers
- Interact with users via:
  - Trouble ticket system/email/ phone
  - Workshops, training events
- Web documentation

•User Satisfaction with NERSC Consulting

•>350 Responses in 2009 Survey



•"The quality of the technical staff is outstanding. They are competent, professional, and they can answer questions ranging from the trivial to the complex"

•2009 User Survey

NERSC

Chemistry & Materials Applications

QUANTUMESPRESSO

CPMD consortium page CPMD

b-initio

abinit.org

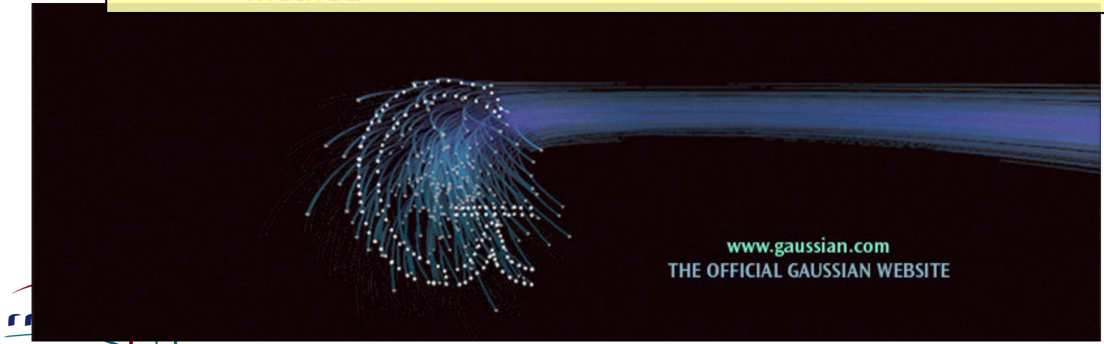
Qbox

LAMMPS

NAMD Scalable Molecular Dynamics

GAMESS

- **More than 13.5 million lines of source code Compiled, Optimized, and Tested**
- **Expert advice provided on using these applications**
- **Helped users focus on science instead of code compilations**



Q-CHEM

NWCHEM

NERSC Systems for Science

Large-Scale Computing Systems

Franklin (NERSC-5): Cray XT4

- 38,128 cores (quad core), ~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), 24-core nodes

Clusters

105 Tflops combined

Carver

- IBM iDataplex cluster

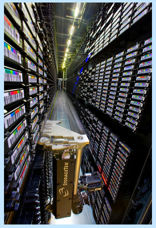
PDSF (HEP/NP)

- 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

2 Data transfer nodes

HPSS Archival Storage

- 40 PB capacity
- 4 Tape libraries
- 150 TB disk cache

Analytics

- **Euclid** (512 GB shared memory)
- **Dirac GPU testbed** (48 nodes. Fermi)

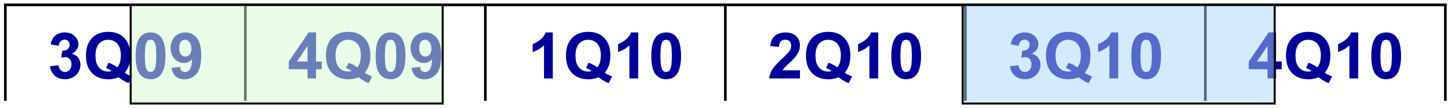
Hopper System

Phase 1 - XT5

- 668 nodes, 5,344 cores
- 2.4 GHz AMD Opteron (Shanghai, 4-core)
- 50 Tflop/s peak
- 5 Tflop/s SSP
- 11 TB DDR2 memory total
- Seastar2+ Interconnect
- 2 PB disk, 25 GB/s
- Air cooled

Phase 2 - XE6

- >6000 nodes, >150,000 cores
- AMD Opteron (Magny-Cours, 12-core)
- >1.0 Pflop/s peak
- >100 Tflop/s SSP
- >200 TB DDR3 memory total
- Gemini Interconnect
- 2 PB disk, 80 GB/s
- Liquid cooled



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

Software and Compilers

- Software will be very similar to Franklin but with shared library support
- Four different compilers
  - Portland Group
  - PathScale
  - Cray Compilers
  - GNU
- Some codes see significant performance improvements with a specific compiler
- NERSC will provide guidance and support to help users choose

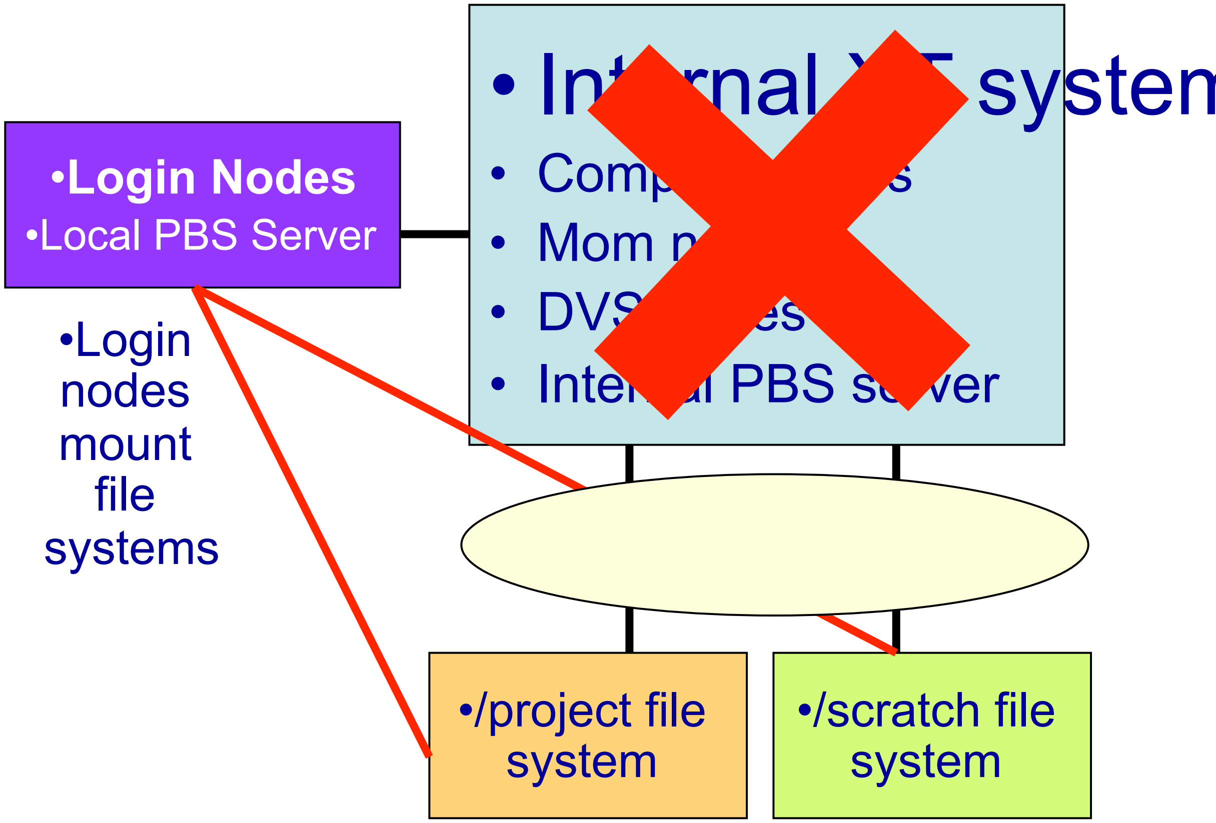
Hopper Login Nodes

- 8 login nodes external to main XT system
- Quad socket, quad-core AMD Opteron 2.4GHz
- 128 GB of memory with swap space
- Load balanced for more optimal usage
- Ability to run more intensive tools on login nodes, IDL, debuggers, etc.
- Available when XT is down

Access to data and login nodes even when XT is unavailable

- **Submit jobs when XT down**
- **Local PBS server on login nodes**
- **Holds jobs while XT is down**
- **Jobs forwarded to internal XT PBS server when XT available again**

•*Sketch of Hopper*



•**Login Nodes**
•Local PBS Server

•Login nodes mount file systems

- Int rnal X T systen
- Comp s
- Mom n
- DVS es
- Inte al PBS se ver

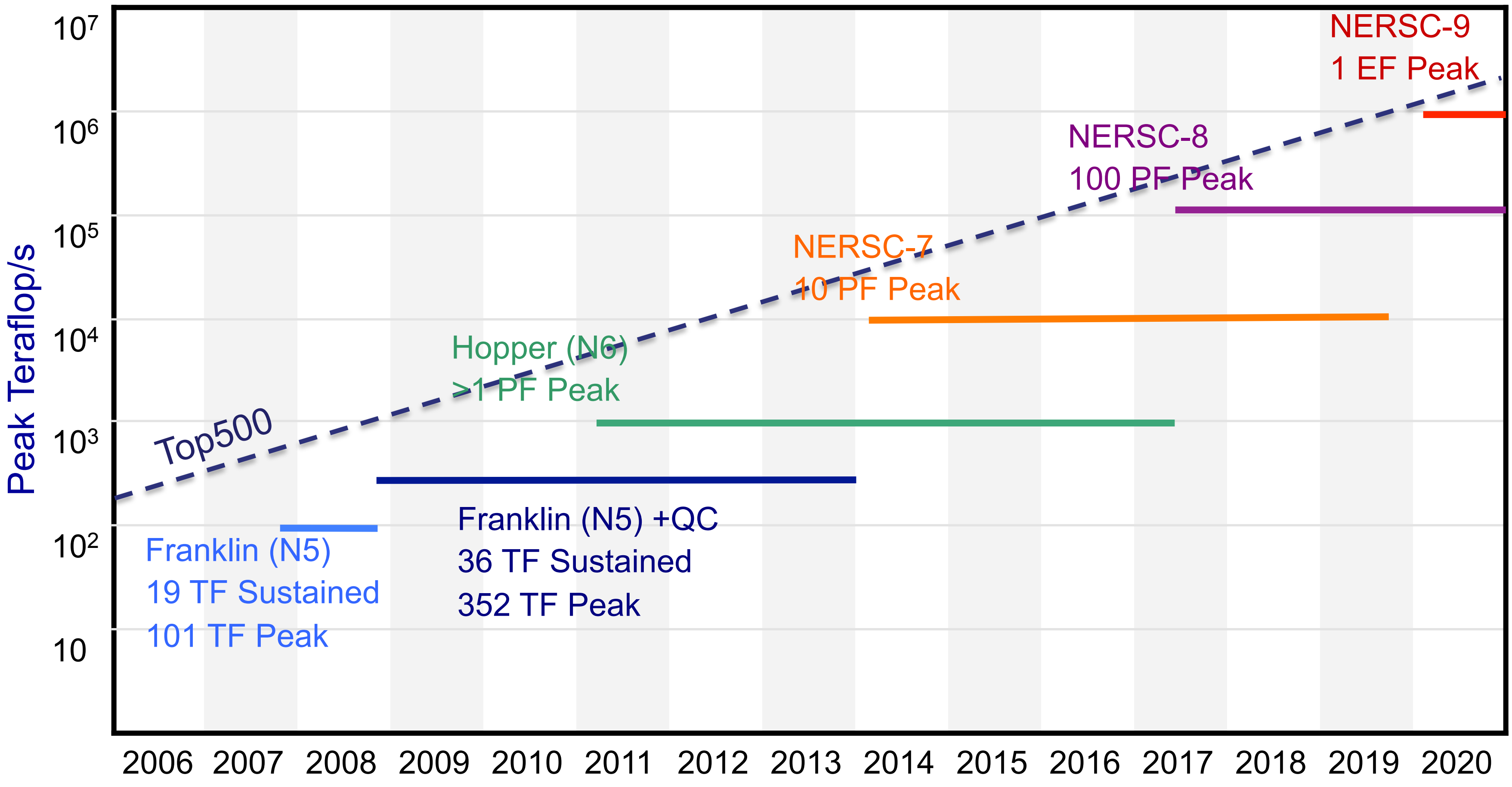
•/project file system

•/scratch file system

Dynamic and Shared Libraries

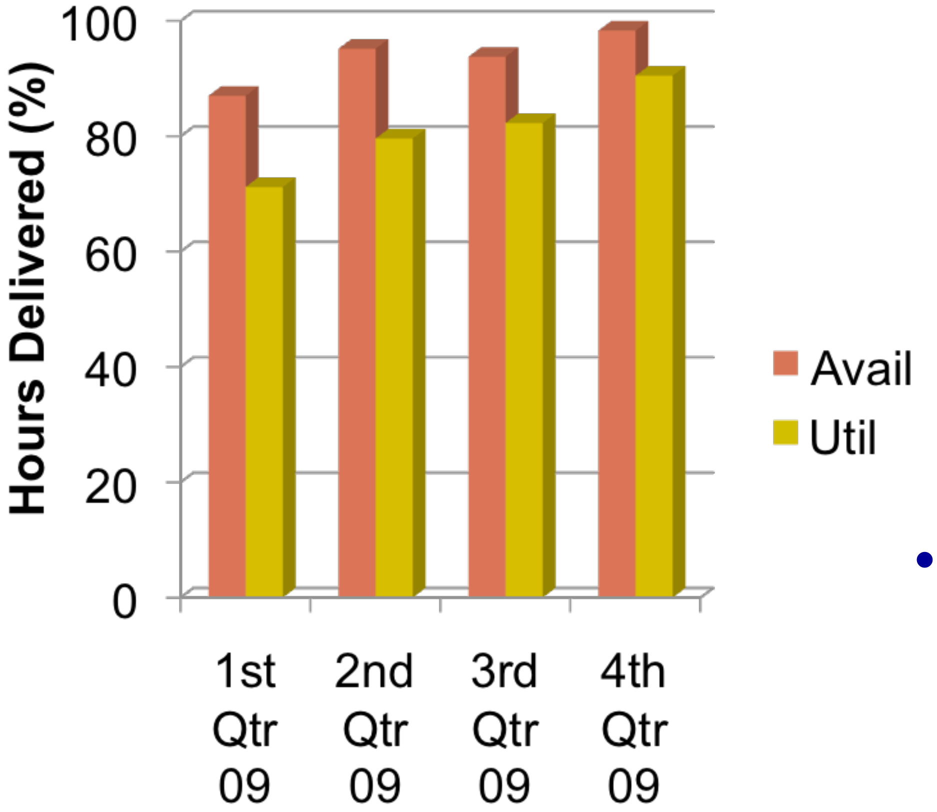
- All user software has a shared library version (mpich, acml, libsci, etc.)
- Static binaries is default environment
- Use the -dynamic compiler and linker flag
- In batch script set environment variable CRAY_ROOTFS=DSL which enables shared root file system

NERSC Roadmap



Users expect 10x improvement in capability every 3-4 years

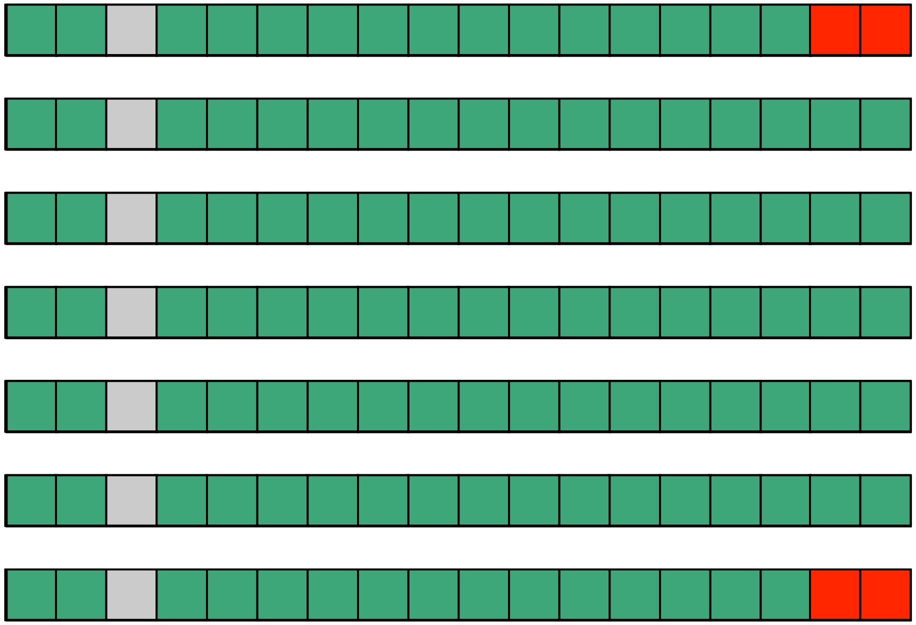
Efficient Compute Resources



- **Partnership with Cray to aggressively fix Franklin bugs**
  - More than 70% of the bugs filed by NERSC
  - Benefit our users and all sites with XT systems
  - Reduce preventative maintenances
- **Monitor queues to keep utilization high**
  - Enhance backfill opportunities
  - Reservation system for large concurrency, scaling, and debugging

Franklin Upgrade in production

- **Challenge to upgrade production system from dual- to quad-core**
- **Innovative rolling upgrade allowed Franklin to be run as two systems, allowing testing and production use simultaneously**
- **Now Cray standard operating procedure**



🟨 Dual core cabinet

⬜ Empty cabinet

🟥 Service cabinet

🟩 Quad core cabinet

Feedback from Users was crucial to Hopper Configuration

User Feedback from Franklin

Login nodes need more memory

Workflow models are limited by memory on batch ‘head nodes’

Improve Reliability and Usability

Hopper Enhancement

8 external login nodes with 128 GB of memory (with swap space)

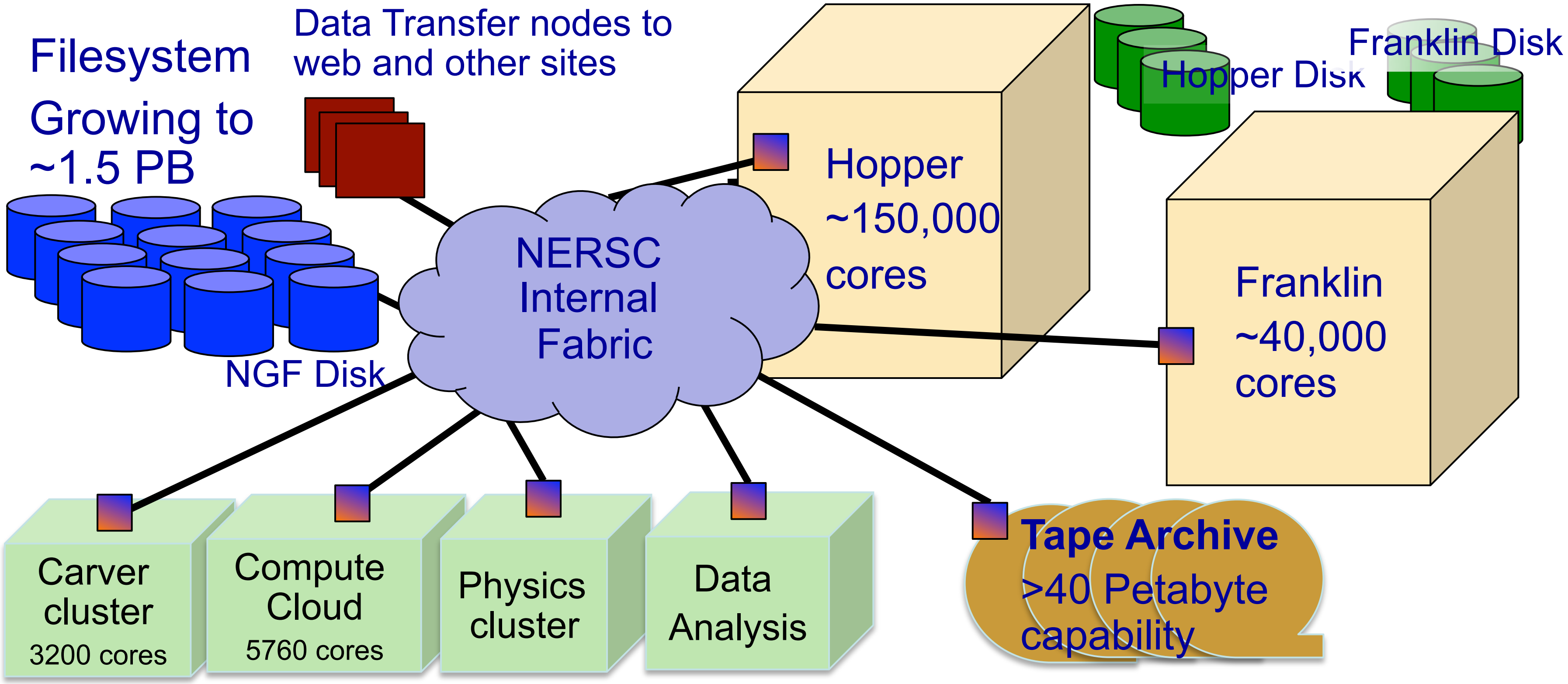
•Increased # of head-nodes per compute node ratio

•Compute nodes can be repartitioned as head-nodes

•External login nodes will allow users to login, compile and submit jobs even when computational portion of the machine is down

•External file system will allow users to access files if the compute system is unavailable

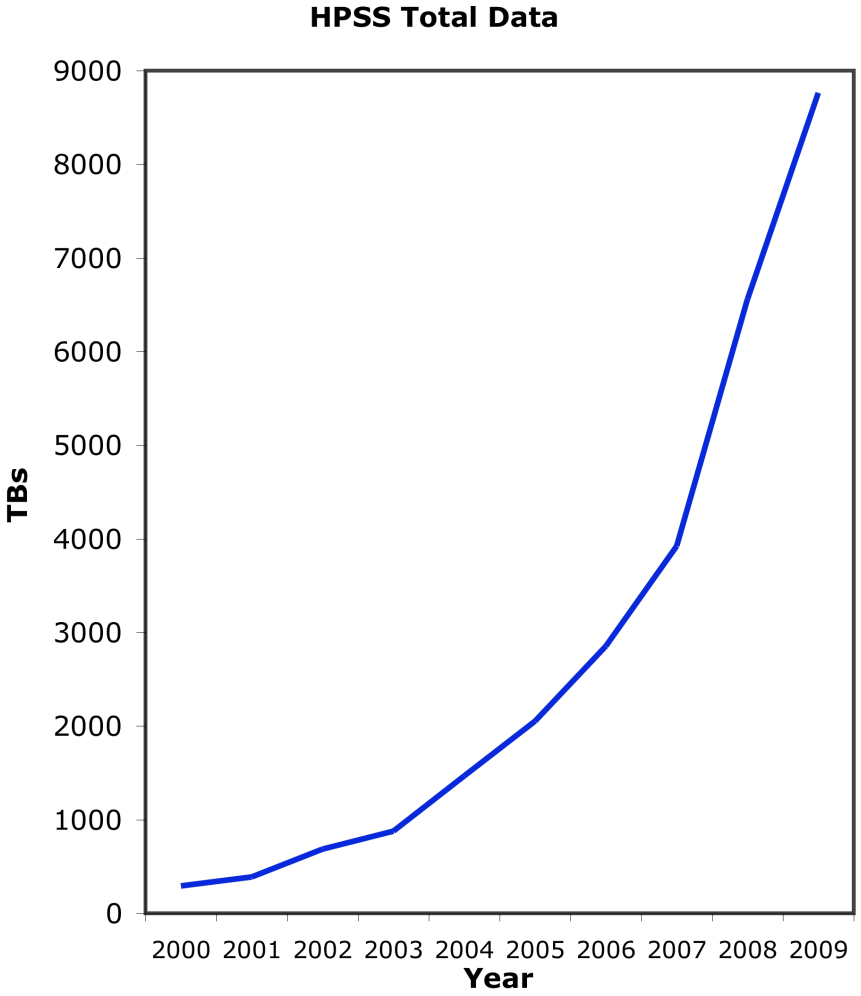
NERSC Storage Architecture



- **Share data & use most appropriate computational resource seamlessly via NERSC Global filesystem**
- **Recent work with Cray has made NGF available to parallel jobs through DVS software layer**
- **Improvements to HPSS authentication**

Tape Storage at NERSC

- As of Feb 2010, tape holds 10 PB of data with the ability to scale to over 40 PBs
- Average annual growth is 40-60%
- 30% of IO to HPSS is reading
- Largest consumers are Climate and Nuclear Physics
- Tape Strategy
  - Both a fast access and capacity tape drive
  - Early adoption of higher capacity tapes creates savings
  - Recycle older tapes by rewriting with newer drives
- Advantages
  - Energy efficient
  - Long lasting



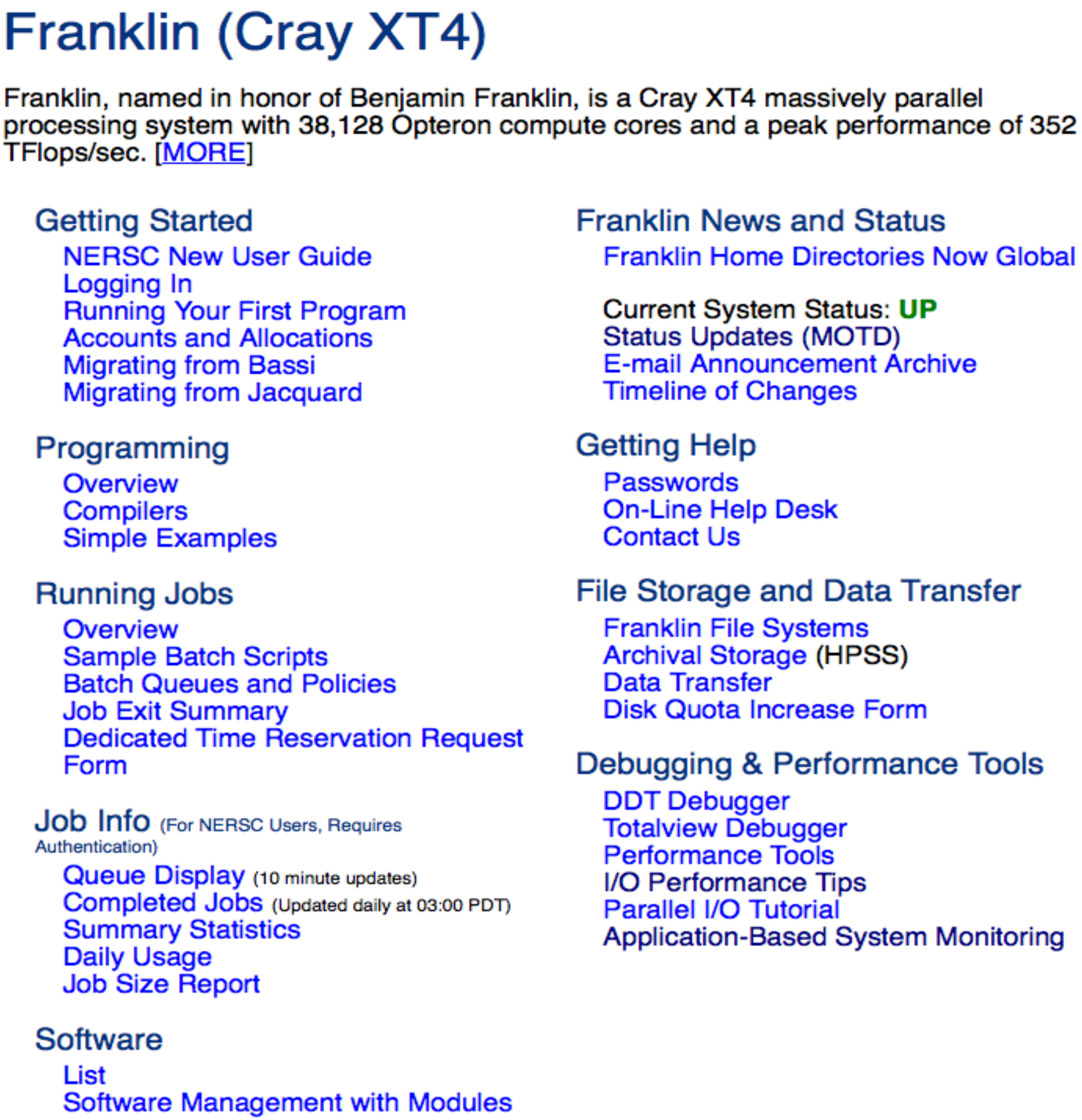
Get started at NERSC

Getting an HPC allocation at NERSC

- Not as hard as you might think
  - If you have an abstract of your research goals applying will take you 30 min or so
- Allocation types
  - Startup: A small allocation is stepping stone toward a large allocation when you need it. It helps you build a computing relationship with DOE and project reviewers. Apply anytime
  - Production: once a year
  - NISE program
- NERSC allocation web page
  - https://nim.nersc.gov/newpi.php

NERSC Training Accounts

- Training accounts available for workshop
- Access to NERSC Machines
  - “ssh train15@franklin.nersc.gov”
- Just need to sign form and I will give you password
- Queue with boosted priority already set up
  - Up to 24k cores
  - 6 hour wall clock limit
  - 20 concurrent jobs for the group
- Come talk to me at the break



Franklin (Cray XT4)

Franklin, named in honor of Benjamin Franklin, is a Cray XT4 massively parallel processing system with 38,128 Opteron compute cores and a peak performance of 352 TFlops/sec. [MORE]

Getting Started

- NERSC New User Guide
- Logging In
- Running Your First Program
- Accounts and Allocations
- Migrating from Bassi
- Migrating from Jacquard

Programming

- Overview
- Compilers
- Simple Examples

Running Jobs

- Overview
- Sample Batch Scripts
- Batch Queues and Policies
- Job Exit Summary
- Dedicated Time Reservation Request Form

Job Info (For NERSC Users, Requires Authentication)

- Queue Display (10 minute updates)
- Completed Jobs (Updated daily at 03:00 PDT)
- Summary Statistics
- Daily Usage
- Job Size Report

Software

- List
- Software Management with Modules

Franklin News and Status

- Franklin Home Directories Now Global

- Current System Status: **UP**
- Status Updates (MOTD)
- E-mail Announcement Archive
- Timeline of Changes

Getting Help

- Passwords
- On-Line Help Desk
- Contact Us

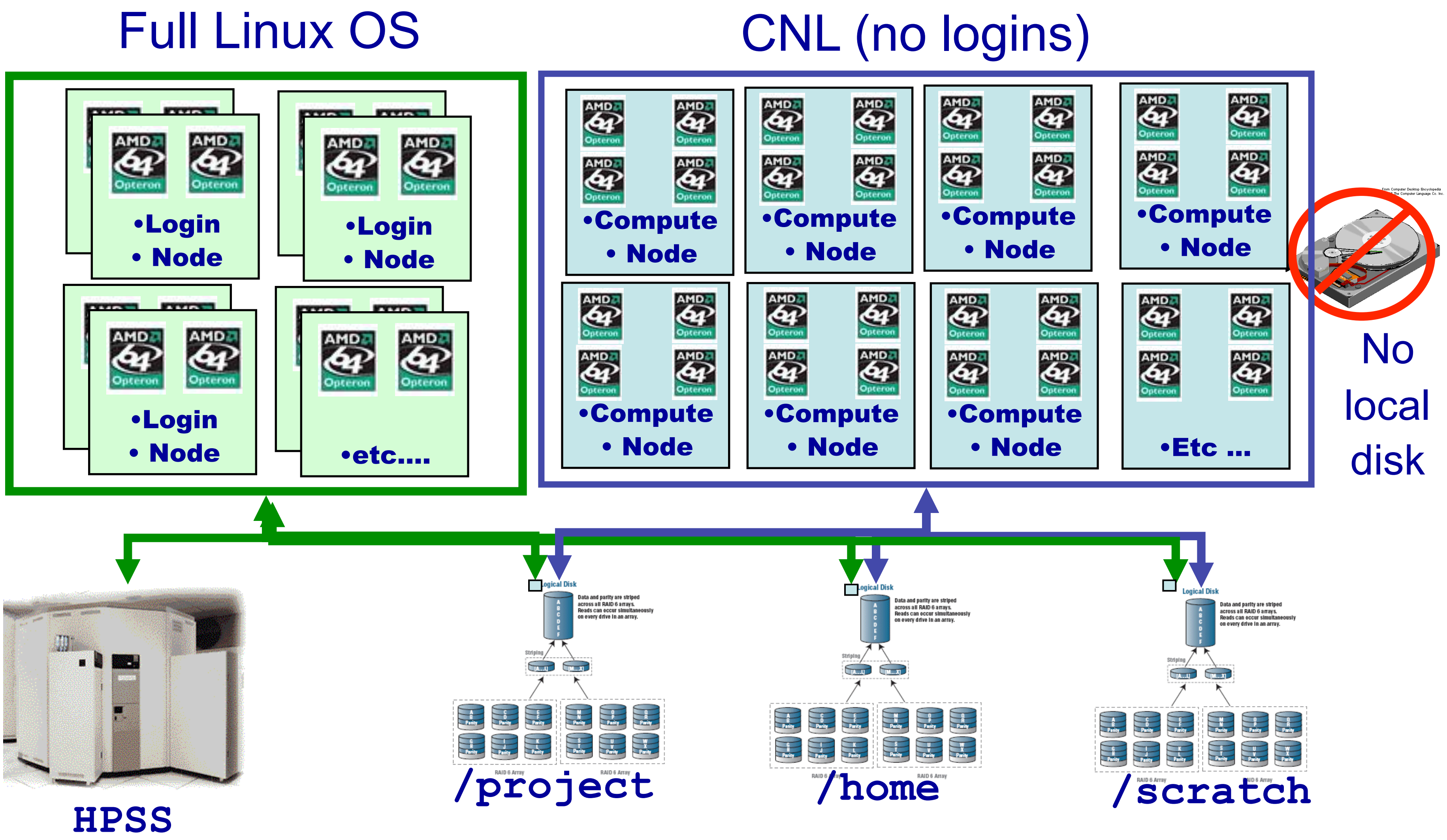
File Storage and Data Transfer

- Franklin File Systems
- Archival Storage (HPSS)
- Data Transfer
- Disk Quota Increase Form

Debugging & Performance Tools

- DDT Debugger
- Totalview Debugger
- Performance Tools
- I/O Performance Tips
- Parallel I/O Tutorial
- Application-Based System Monitoring

Franklin Overview



Full Linux OS

CNL (no logins)

- Login
- Node

- Login
- Node

- Login
- Node

- etc....

- Compute
- Node

- Compute
- Node

- Compute
- Node

- Compute
- Node

- Compute
- Node

- Compute
- Node

- Compute
- Node

- Etc ...

No local disk

HPSS

/project

/home

/scratch

AMD 64 Opteron
Logical Disk
Data and parity are striped across all RAID 6 arrays. Reads can occur simultaneously on every drive in an array.
Striping
[A...L]
[M...X]
RAID 6 Array
Parity

What kind of OS?

- Consider what kind of OS you are using
  - Limited OS
    - Depends on system but limited OS calls
    - Features which could be limited on compute nodes
      - Shared libraries
      - Scripting languages, python, perl
      - Process control (fork, exec)
      - Can’t ssh from compute node to compute node
      - Can’t call system() from Fortran parallel job
      - No Java on the compute nodes
      - No X-Windows support on compute nodes

Franklin Programming Environment

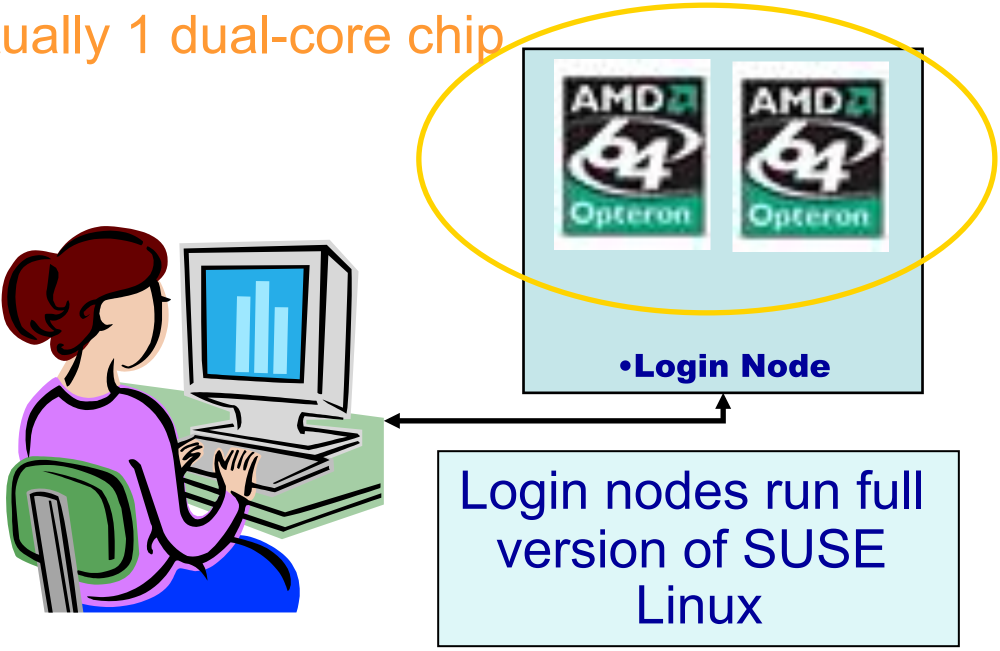
- Compilers (Fortran, C, C++)
  - PGI, PathScale, GNU, Cray
- Parallel Programming Models: Cray MPICH2 MPI, Cray SHMEM, Open MP
- AMD Core Math Library (ACML): BLAS, LAPACK, FFT, Random number generators, GNU Fortran libraries
- LibSci scientific library: ScaLAPACK, BLACS, SuperLU
- Profiling tools CrayPat, Apprentice2, IPM, TAU
- Performance API (Papi)
- Modules

Extensive 3rd Party Software

- Check to make sure your application isn’t already installed
- Use modules command to see software availability on all NERSC machines (“module avail”)
- Math - acml, aztec, dfftpack, fftw, gsl, LibSci, parmetis, parpack, petsc, pspline, superlu, sprng
- I/O - hdf5, nco, netcdf, pnetcdf
- Chemistry/Mat Sci - amber, namd, nwchem, abinit, cpmd, lammps, quantum expresso, siesta, vasp
- Visualization - idl, gnuplot, visit, ncar
- Debuggers - Allinea’s DDT,Totalview

Running a Job on Franklin

Actually 1 dual-core chip



Login nodes run full version of SUSE Linux

`www.nersc.gov/nusers/status/queues/franklin/`

NERSC Analytics server (Euclid)

On a Franklin login node:

1. Log in from your desktop using SSH
2. Compile your code or load a software module
3. Write a job script
4. Submit your script to the batch system
5. Monitor your job’s progress
6. Archive your output
7. Analyze your results

Batch Queues

- At NERSC users submit jobs to a queue and wait in line to run
- Queue policies are set to:
  - Be fair
  - Accommodate needs
    - Users
    - DOE strategic
  - Encourage high parallel concurrency
  - Maximize scientific productivity
- Special requests always given consideration
  - Reservations
  - Emergencies

Batch Queues

- debug: short, small test runs
- interactive: implicit in `qsub -I`
- regular: production runs
  - Jobs > 512 nodes given 50% discount
- premium: I need it now, 2X charge
  - Fast turn around on Franklin, not usually needed
- low: I can wait a while: 50% discount
- special: unusual jobs by prior arrangement

Memory Considerations

- Each Franklin compute node has 8GB of memory.
- Running 4 cores per node 7.38 GB of user addressable memory
  - CNL kernel, uses ~300 MB of memory.
  - Lustre uses about 17 MB of memory
  - MPI buffer size is about ~100 MB.
- Quad core MPI jobs have ~1.83 GB/task.
- Change MPI buffer sizes by setting certain MPICH environment variables.
- Hints for adjusting MPICH variables on website

Disk Quotas

- Franklin has multiple file systems
  - /home (global homes 40GB)
    - Backed up
    - Permanent
  - /scratch and /scratch2 (Default 500 GB)
    - Purged of files older than 12 weeks
    - Not backed up
    - Not permanent
  - /project (NERSC Global Filesystem)
    - Accessible from all NERSC machines
    - Currently need to request access
- Users can not submit jobs when over quota
- Projects needing larger disk quotas just need to ask

Scratch Disk Space

- Disk space is expensive and therefore limited and shared among users
- Every center must manage disk space in some way (purging, begging, quotas)
- Understand the disk usage policy at your center
- Be a courteous disk space user. We want you to run very large jobs, but then we want you to back up your files (quickly)

Performance issues

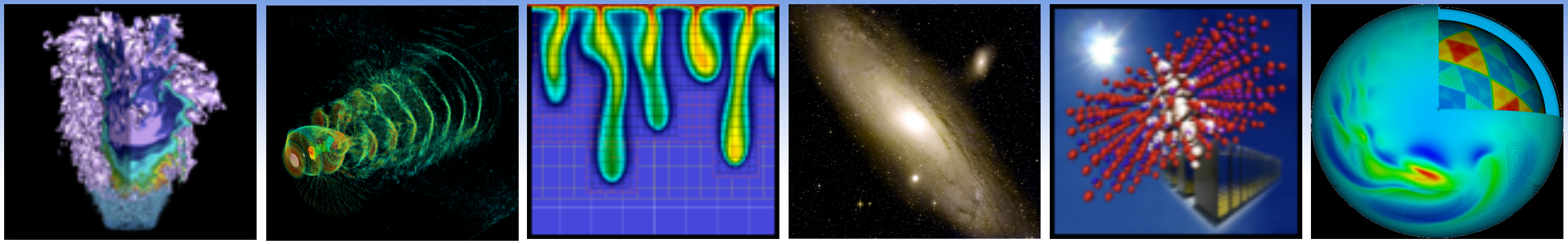
- Different compilers and compiler optimizations
- Libraries
- IO performance over the different file systems.
  - IO strategy
  - File striping (Lustre)
- Run in scale
  - Parallel scaling
  - Runtime envs

Account Support and HPC Consulting

- Account support
  - Passwords (NERSC does not use OTP keys)
  - New accounts
  - Modify accounts (add user to project)
- HPC Consulting
  - 9 Consultants to serve NERSC users
  - Aim to provide fast helpful advice from simple to complex
    - I can't submit my job
    - What library should I use?
    - My code is performing slowly
    - My code compiled on my department cluster but now …
  - Please contact the consultants!
  - We are paid to help make you more productive
  - We have often seen your problem many times before with other users

Acknowledgement

- Slides were based on the 2009 CScADS presentation of Katie Antypas, and various talks given by other NERSC staff.



Thank you!

BERKELEY LAB

NERSC

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868