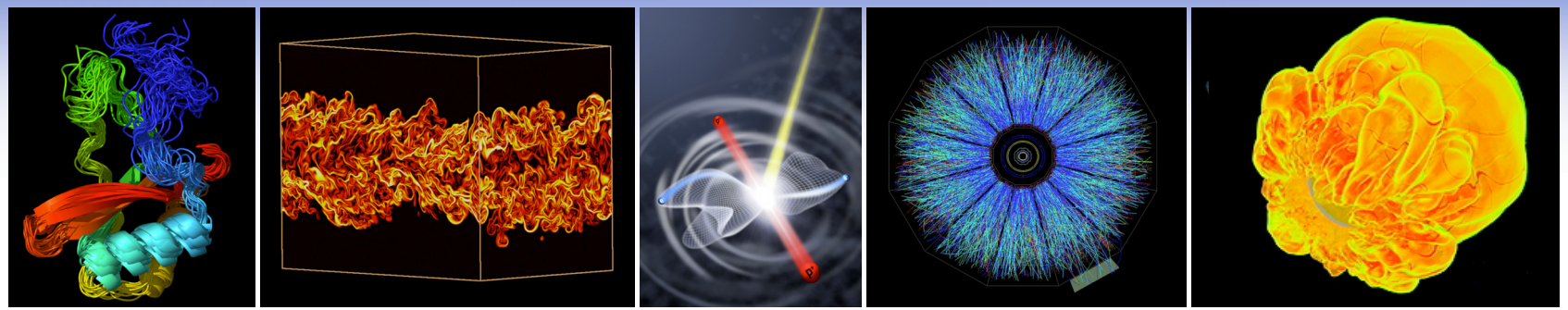


Hopper: XT5 at NERSC

XT5 Workshop Berkeley, CA

Katie Antypas
HPC Consultant

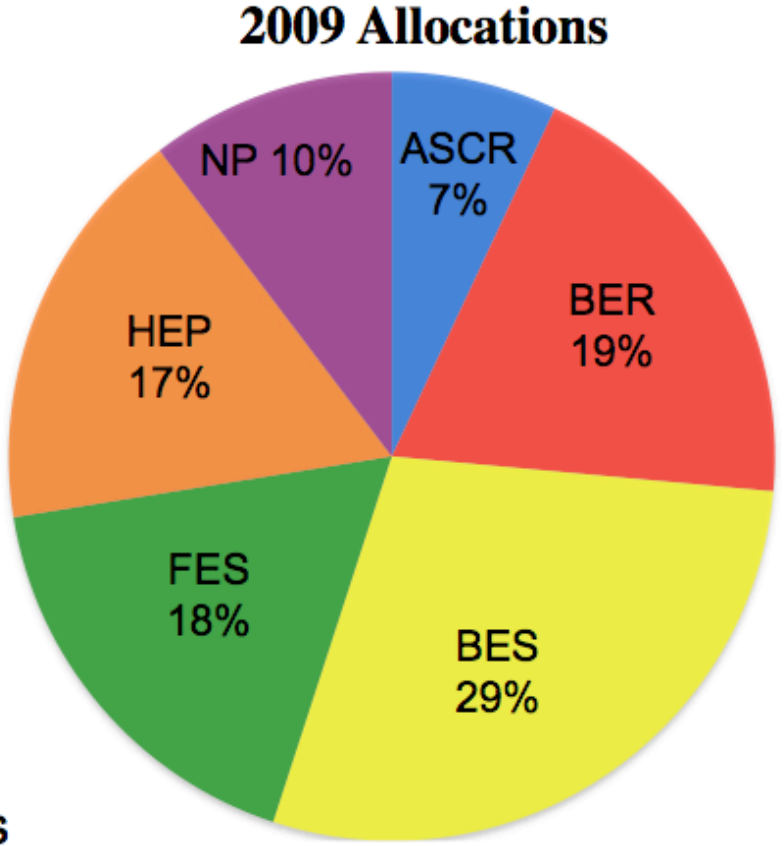
U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC

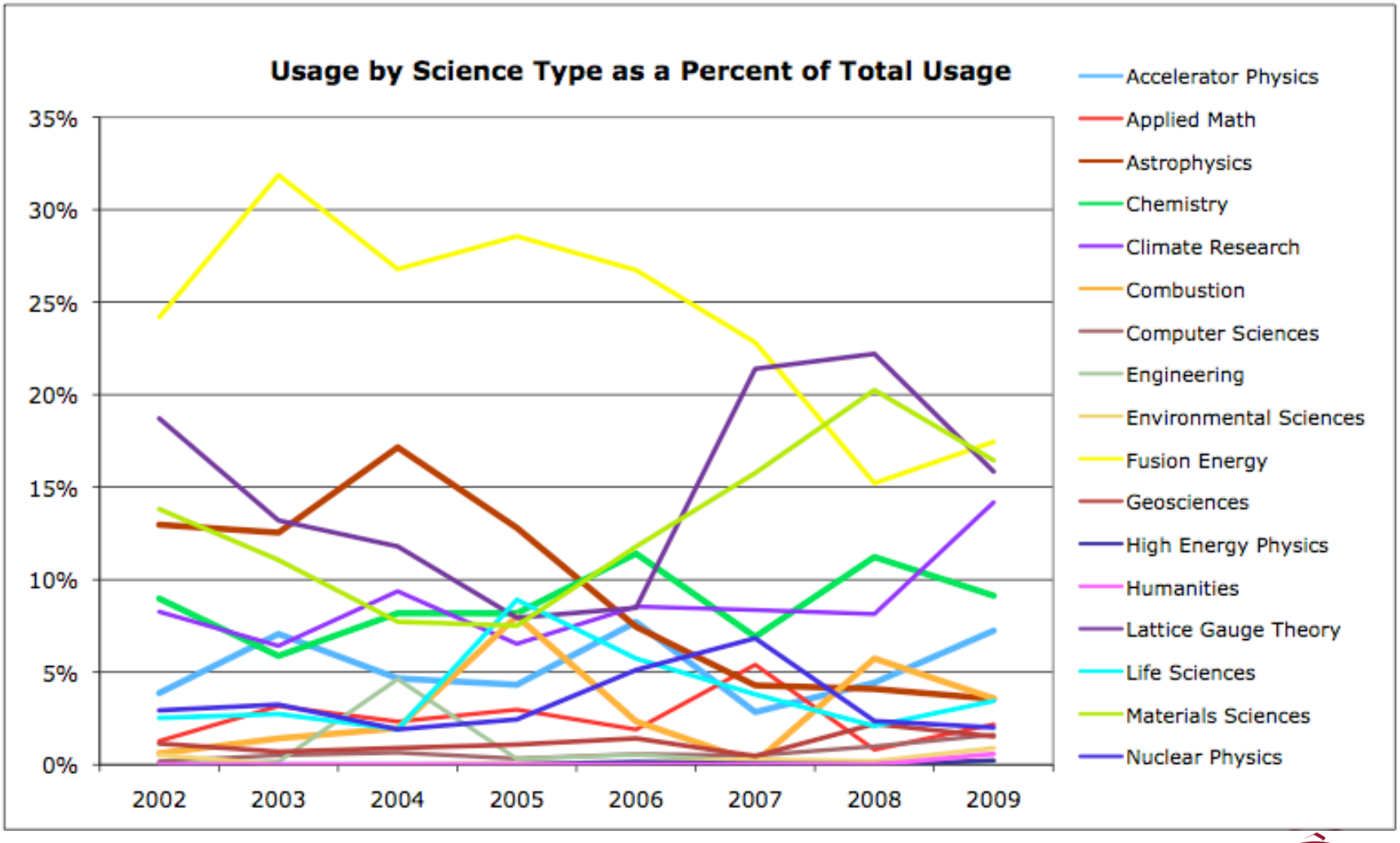
BERKELEY LAB

NERSC is the Primary Computing Facility for the Office of Science

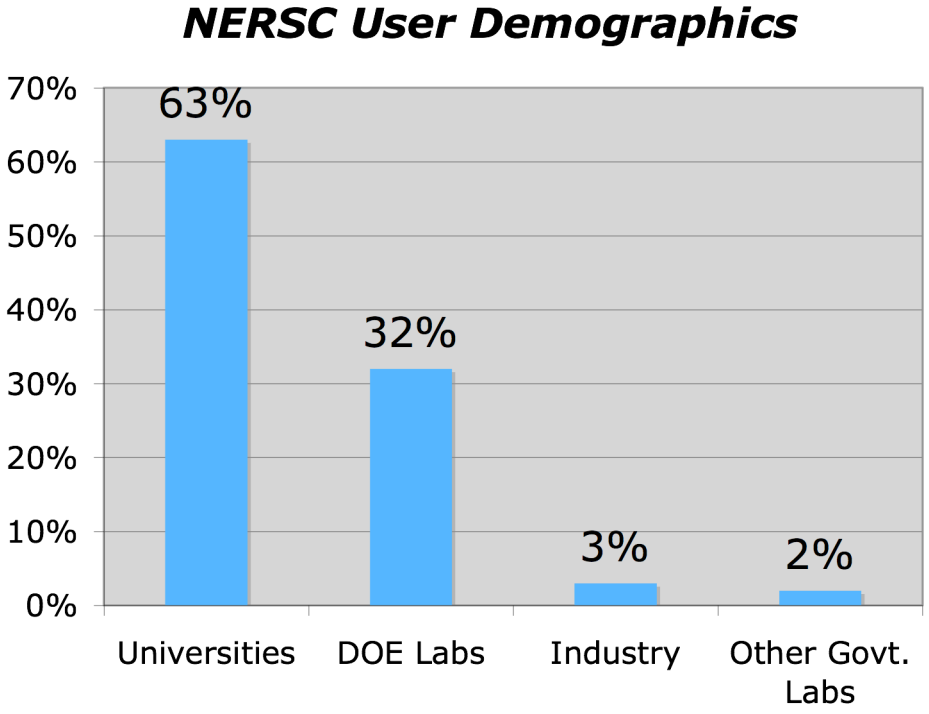
- NERSC serves a large population
  - Approximately 3000 users,
  - 400 projects, 500 code instances
- Focus on "unique" resources
  - –High end computing systems
  - –High end storage systems
    - File system and tape archive
  - –Interface to high speed networking
- Science-driven
  - –Science problems used in machine procurements and performance metrics
  - –Science services

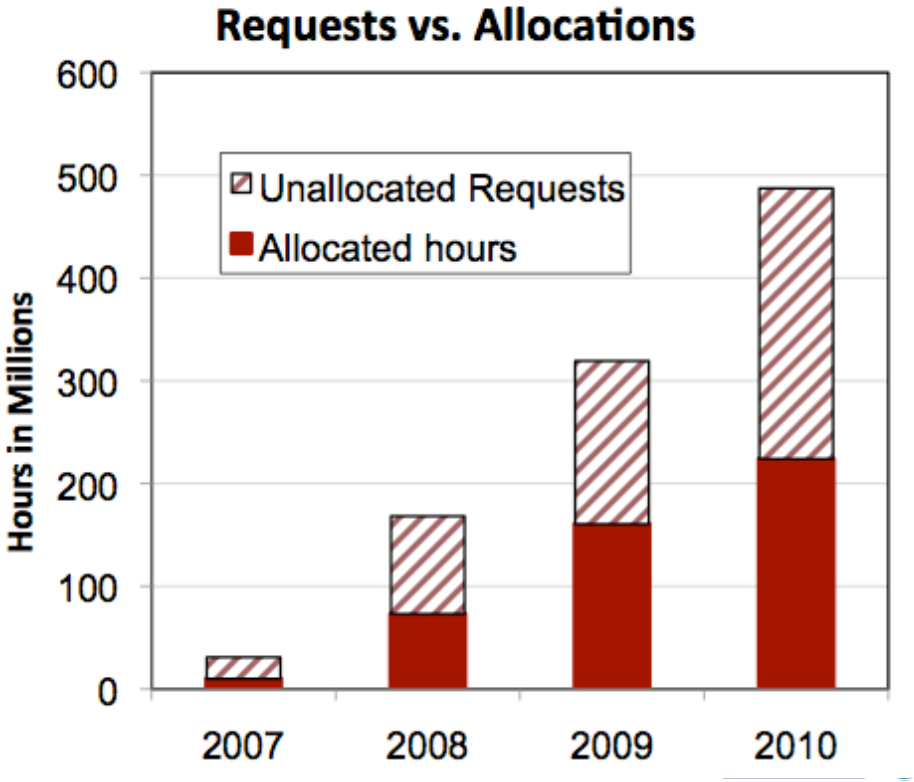


What's Changed in DOE Priorities for NERSC?



NERSC User Demographics

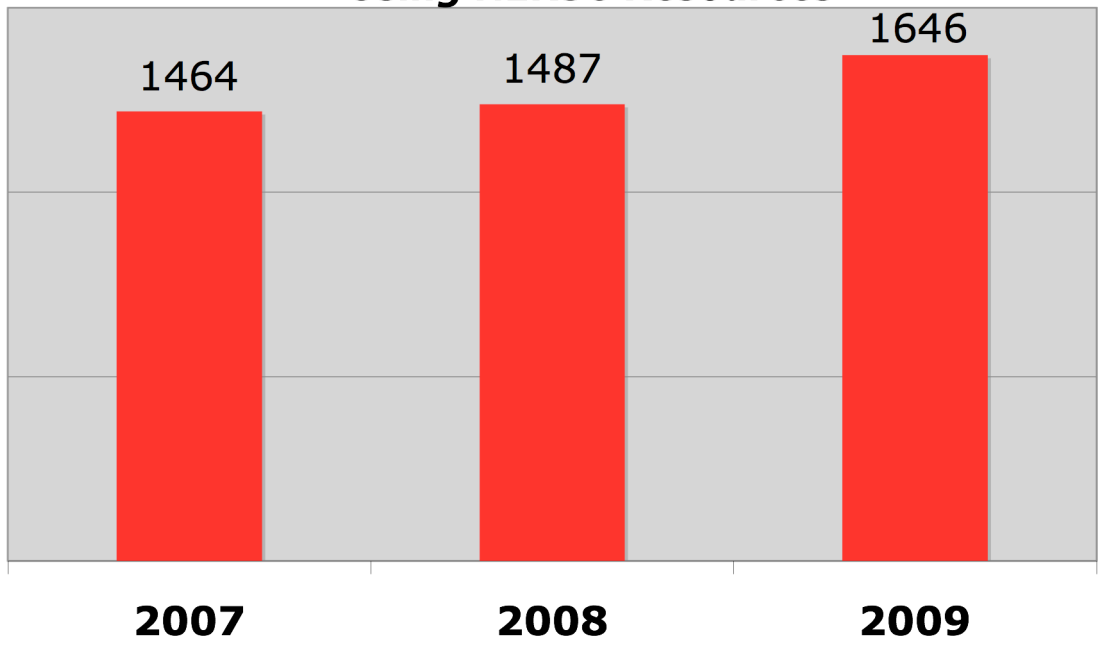


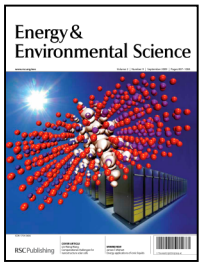


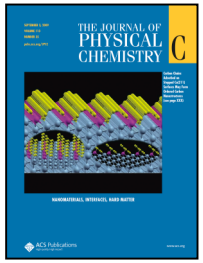
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

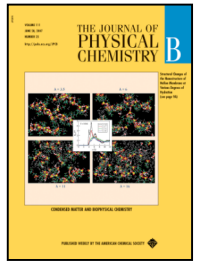
The Ultimate Gauge of NERSC's success can be Measured by Scientific output

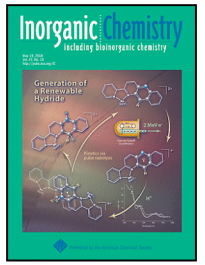
Number of Referred Publications Using NERSC Resources

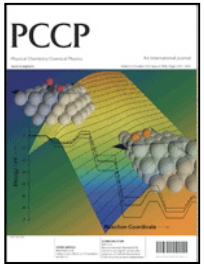


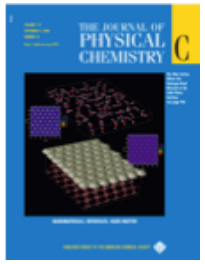


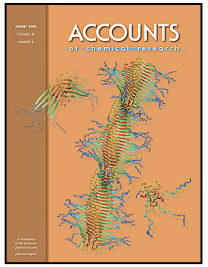












U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

NERSC Allocations

- **Allocations**
  - 80% DOE program managers control
  - 10% ASCR Leadership Computing Challenge
  - 10% NERSC Reserve
- **Start-up allocations available directly from NERSC**
  - 10,000 - 50,000 hours allocations
  - If you have an abstract of your research goals applying will take about 30 min
  - A small allocation is stepping stone toward a large allocation. It helps build a computing relationship with DOE and project reviewers.
- **http://www.nersc.gov/nusers/accounts**

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 5

- 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

Hopper System Delivered in 2 Phases

Phase I System XT5	*Phase II System X??*
• 664 Compute Nodes, 5312 cores	• Greater than 6000 nodes, over 150,000 cores
• 2.4 GHz AMD Opteron (Shanghai quad-core)	• AMD Opteron (Magny Cours 12-core)
• 50 Tflop/s peak	• > 1.0 Pflop/s peak
• 11 TB DDR2 memory	• 200 TB DDR3 memory
• Seastar2+ Interconnect	• Gemini Interconnect
• 2 PB disk, 25GB/sec	• 2 PB disk, 80 GB/sec
• Air cooled	• Liquid cooled

Feedback from NERSC Users was crucial to NERSC6 negotiations

User Feedback from Franklin	*NERSC6 Enhancement*
Login nodes need more memory	8 external login nodes with 128 GB of memory (with swap space)
Shared libraries are not supported	Shared libraries are supported. (And full Linux OS available)
Need more disk space	Includes a 7x increase in disk space over Franklin (2PB)
Increase I/O bandwidth	Includes a 3x increase in I/O bandwidth over Franklin (70 GB/sec)
Connect NERSC Global FileSystem to compute nodes	/project file system will be available to compute nodes
Workflow models are limited by memory on MOM (host) nodes	•Increased # and amount of memory on MOM nodes •Phase II compute nodes can be repartitioned as MOM nodes

Hopper Login Nodes

- 8 login nodes external to main XT system
- 128 GB of memory with swap space
- Ability to run more intensive tools on login nodes, IDL, debuggers, etc.
- Available when XT is down

Login to Hopper:

ssh username@hopper.nersc.gov

No One-Time-Password token needed

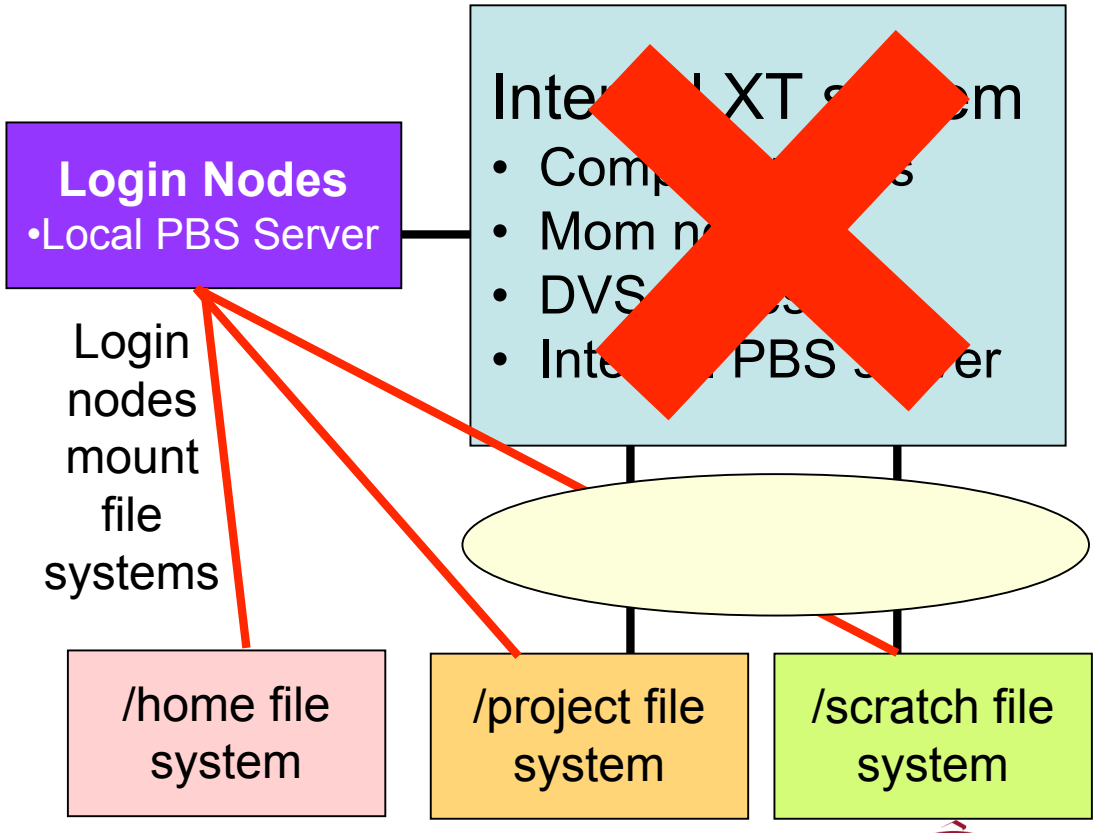
Hopper Filesystems

- $HOME
  - Store application code here, running small jobs ok
  - GPFS
  - Global file system shared by *most* NERSC machines
  - 40 GB quota
  - Peak performance ~500MB
- SCRATCH
  - Run applications from here, then move data to HPSS
  - Lustre
  - 2 file systems $SCRATCH and $SCRATCH2
  - Users can run in either, $SCRATCH2 often has less contention
  - Peak performance ~25GB/sec for each
  - Quotas and purging not yet enforced
- PROJECT
  - Primarily for groups needing to share space
  - GPFS
  - Global file system shared by all NERSC machines

Access to data and login nodes even when XT is unavailable

- Submit jobs when XT down
- Holds jobs on local PBS server while XT is down
- Jobs forwarded to internal XT PBS server when XT available again

Sketch of Hopper



Software and Compilers

- Software very similar to Franklin but with shared library support
- Four different compilers
  - Portland Group (default)
  - PathScale
  - Cray Compilers
  - GNU
- Use compiler wrappers to choose the programming environment
- Some codes see significant performance improvements with a specific compiler so we encourage users to try other compilers besides the default

- Fortran wrapper: "ftn": example: ftn myProgram.F90
- C wrapper: "cc" example: cc myProg.c
- C++ wrapper: "CC" example: CC myProg.CC

Focus on Scientific Productivity

- **Wide array of 3 party software application support**
  - **Math libraries - ACML, FFTW, gsl, LibSci, PETSc, SuperLU and more**
  - **I/O - HDF5, nco, netCDF, pNetCDF**
  - **Chemistry/Mat Sci - amber, NAMD, NWChem, abinit, cpmd, lammps, quantum expresso, VASP, and more**
  - **Visualization - IDL, gnuplot, VisIT, ncar**
  - **Debuggers - Allinea's DDT and Totalview**
- **Software environment controlled by *modules***

```
• module list
• module avail
• module load netcdf
```

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

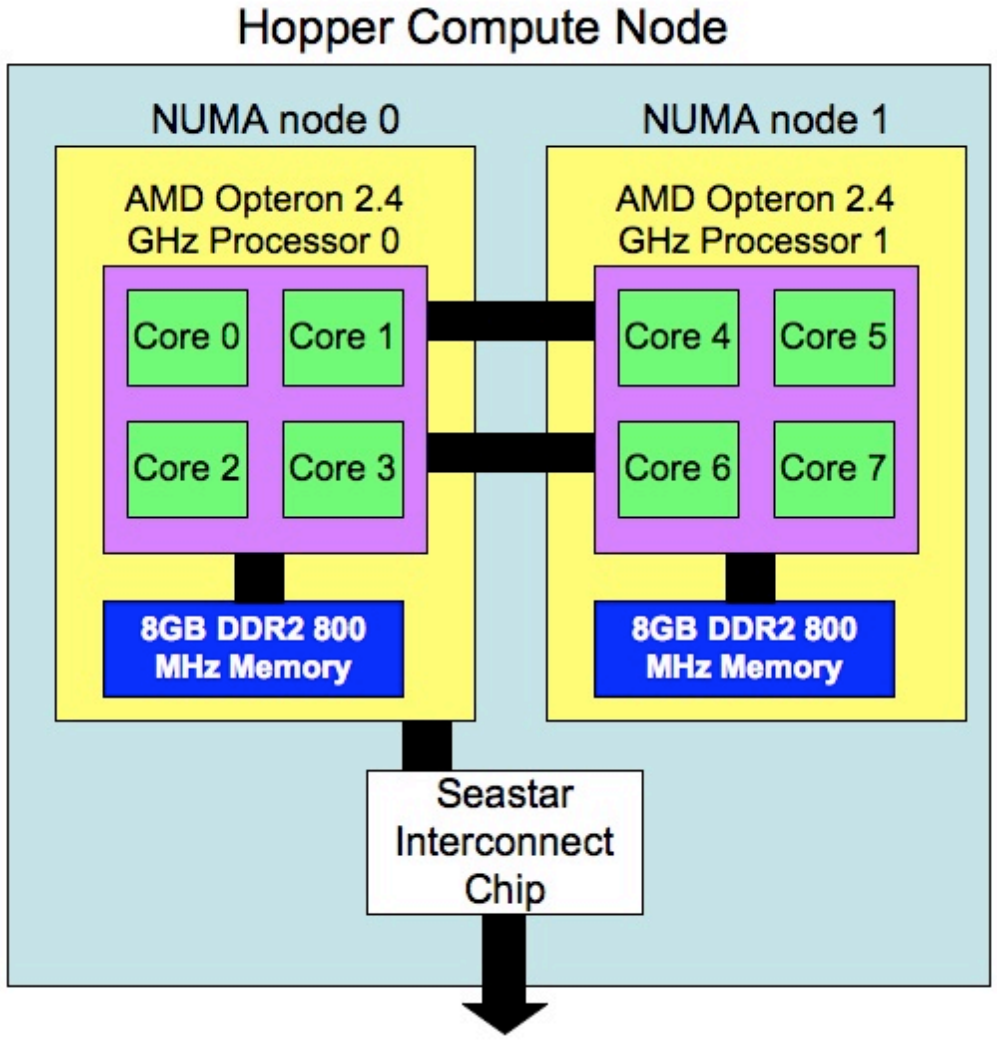
NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Running Jobs on Hopper

- Login land on a “login” node
- Parallel applications run on “compute nodes”
- MUST launch applications with “aprun” command to get them from login nodes to compute nodes
  - Batch script
  - Interactive job

aprun Options

- Hopper has 2 sockets per core, increasing the aprun options, particularly for openMP codes
- New options to specify, how many numa nodes, which numa node, cores per numa node, strict memory containment between sockets
- Afternoon talk will address these options



Account Support and HPC Consulting

- **Account support**
  - Passwords (NERSC does not use OTP keys)
  - New accounts
  - Modify accounts (add user to project)
- **HPC Consulting**
  - 8 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

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

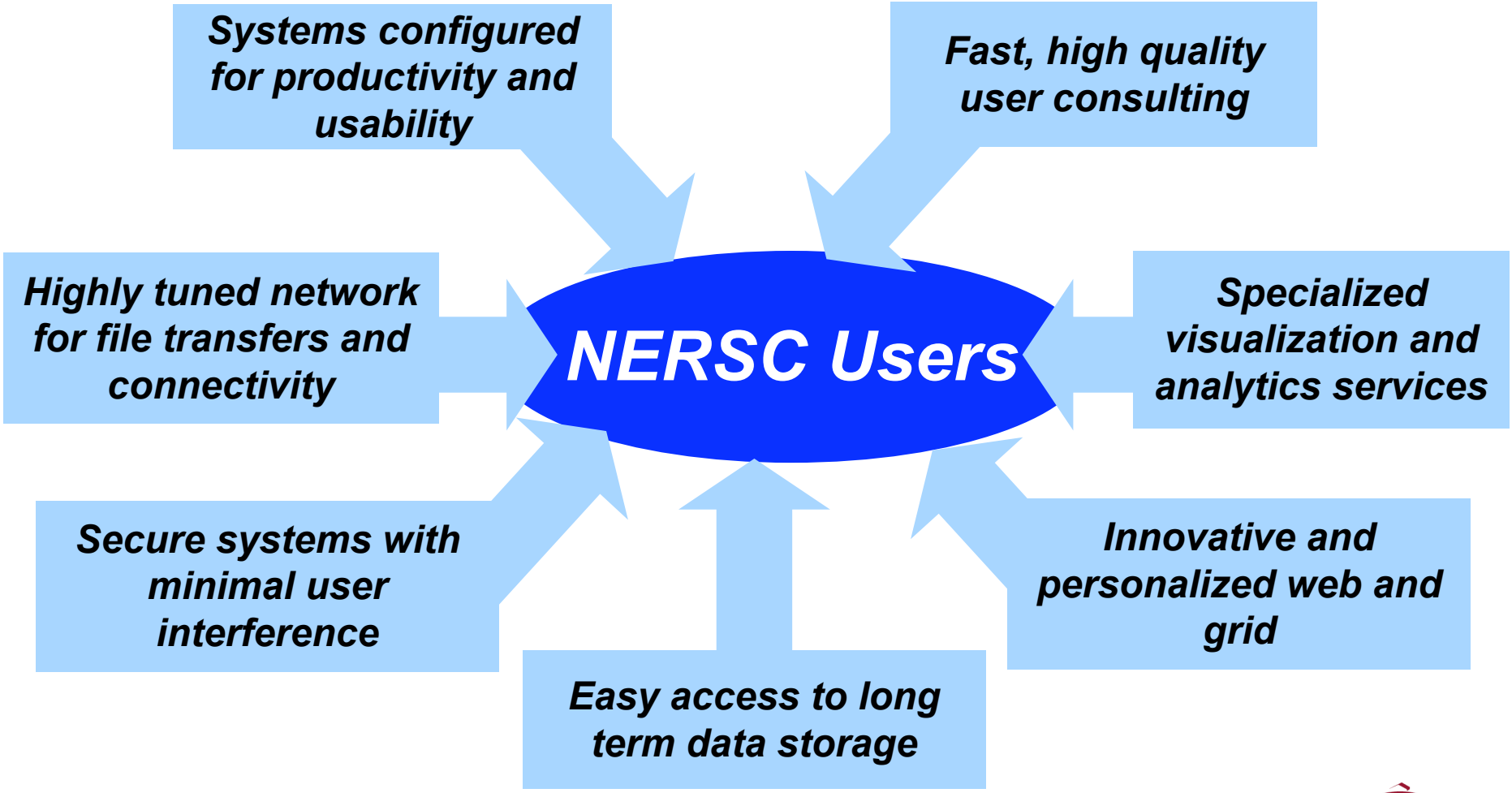
Cover Stories from NERSC Research

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

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

NERSC Services for Scientific Discovery: More than Hardware



Franklin Job Size Report

Fraction of Raw Hours by Job Parallel Concurrency

Two-week moving average