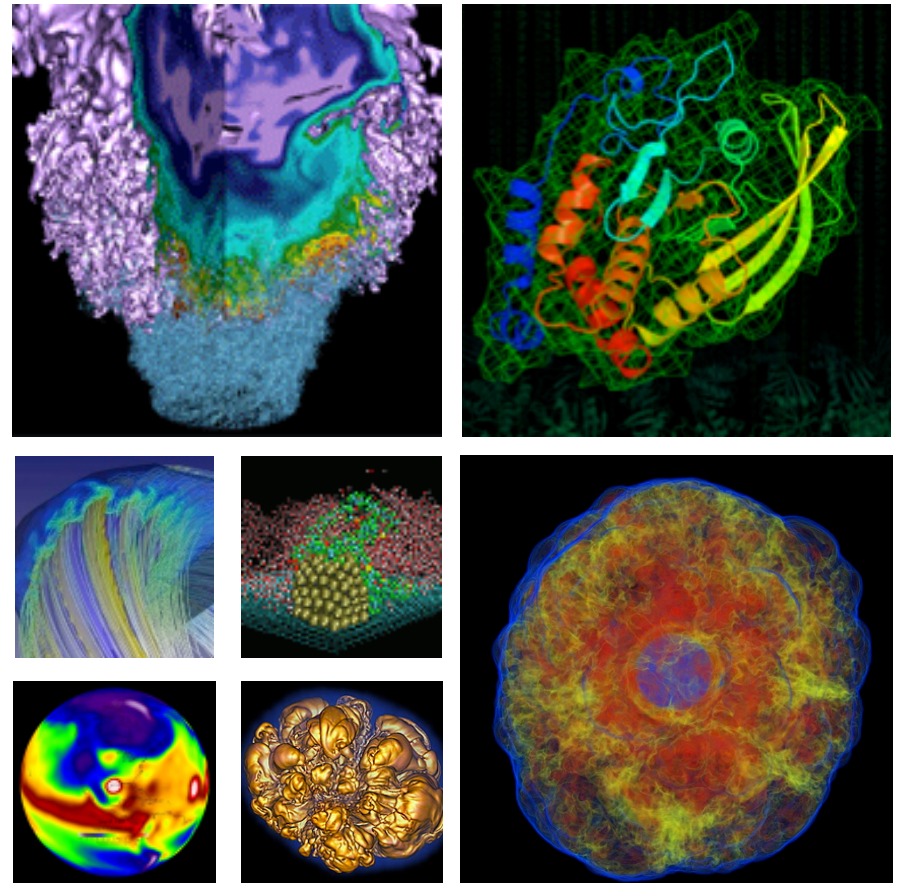


NERSC User Group Storage Update



Jason Hick
Storage Systems Group

Feb 26, 2014

The compute and storage systems 2014

NERSC

Hopper: 1.3PF, 212 TB RAM



Cray XE6, 150K Cores

2.2 PB Local
Scratch
70 GB/s

48 GB/s

Edison: 2.5PF, 357 TB RAM



Cray XC30, 130K Cores

6.4 PB Local
Scratch
140 GB/s

80 GB/s

80 GB/s

50 GB/s

5 GB/s

12 GB/s

/global/
scratch

4 PB

/project

5 PB

/home

250 TB

HPSS

**45 PB stored, 240 PB
capacity, 40 years of
community data**

Sponsored Compute Systems
Carver, PDSF, JGI, KBASE, HEP
8 x FDR IB

**Ethernet &
IB Fabric**

*Science Friendly Security
Production Monitoring
Power Efficiency*

WAN

2 x 10 Gb

1 x 100 Gb

*Science Data
Network*



Vis & Analytics, Data Transfer Nodes,
Adv. Arch., Science Gateways

Storage systems and services



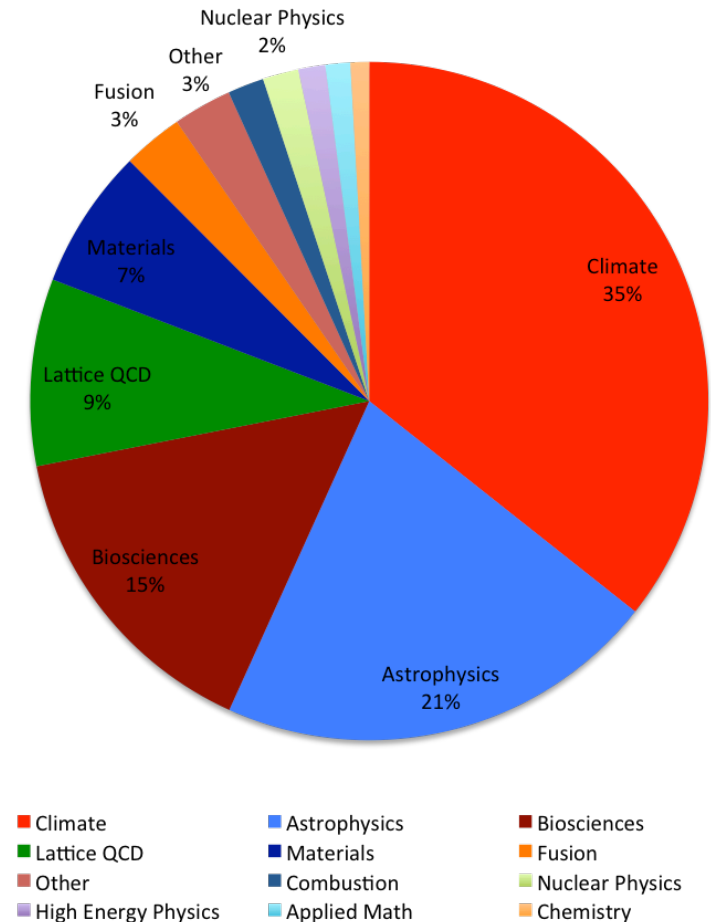
- **Parallel file systems (Lustre and GPFS) are primary storage to supercomputers**
 - Total of over 20 PBs of disk available to users
 - Some multi-PB parallel file systems backed up to HPSS (Parallel Incremental Backup System)
 - Has demonstrated it can identify backup candidates from 500M total files and process over 150TBs of backup data in a single day
- **Archival and backup systems (HPSS) are secondary storage for users**
 - 65 PBs of data stored, growing at >1PB per month
 - 30% of user IOs are read/retrieve requests from archival storage, so a very active archive
 - Focus on reliability of the system for user data by:
 - Deploying solutions to proactively monitor and maintain health of user data, and environmental parameters necessary for tape
 - Actively migrating/moving data within the system
- **Storage services highly utilized**
 - Inter-facility data transfers
 - About 50% of network bandwidth consumed by bulk data transfers
 - Science gateways for sharing data with collaborators

Storage serving the entire range of science



- Archival data (HPSS) shows a wide range of science data each month
- NGF provides 1TB project directories to every science repository at NERSC
- Users have a diverse set of storage needs that range from
 - Low latency, small file
 - Long-term data processing
 - Long-term data sharing
 - High availability web hosting
 - Supercomputing job I/O

HPSS Monthly I/O by Scientific Discipline
31 Jul 2014



Improvements to the HPSS disk cache



- **Recently deployed a new higher bandwidth and capacity disk cache for HPSS Backup**
 - Usage statistics show we are able to retain data on disk for 30 days
- **Starting to deploy onto HPSS Archive system**
 - Bandwidth will improve as well, expect 25% or more
 - 70% of all files stored to the Archive are read within 60 days

- **Next week we'll have finished upgrading all 4 DTNs**
 - dtn01 – dtn04
 - For large files, bandwidth improved by over 30% to 1.2 GB/sec
 - In 2014, DTNs moved over 4PB of data to/from HPSS
 - Support Globus Online Endpoints:
 - nersc#dtn
 - nersc#dtn_new
 - nersc#dtn_shared
 - nersc#dtn_jgi
 - nersc#hpss

Improvements to NERSC Global Filesystems



- **Global scratch replaced last year with file system capable of 80 GB/sec in aggregate**
 - Will be replaced with Cori scratch file system to be mounted across NERSC
- **Home file systems now have metadata on SSDs**
 - Over 5x improvement in backups with drastically reduced impact to users since August 2014
- **Project will get bandwidth upgrade in late 2015**
 - Upon completion of the move to CRT, and Cori scratch replacing Global scratch, project file system will have an 80GB/sec aggregate bandwidth upgrade

New Sponsored Storage option



- **Uses the /project file system**
 - Provides roughly 10GB/sec per 1PB of storage
- **Provide one-time LBL-based funding to NERSC in exchange for a large allocation in NGF**
 - Minimum allocation is 50 TBs increasing by 10 TB increments
 - Funding defers the cost of networking, storage, and server hardware
 - Supports hardware for 5 years to keep the solution running
- **Upon agreement to provide sponsored storage, a one-time recharge is conducted**
 - Generally, a NGF project directory is created with the requested allocation
 - Users may then access their data from any NERSC computational system through the /project file system
 - The /project file system is designed for scalability of capacity and bandwidth
 - High aggregate bandwidth for using your data to achieve scientific results
 - Integrated with science gateways to enable you to share your data via web portals
- **The solution and cost involves:**
 - Enterprise-class disk arrays designed for 5 years of high duty use
 - Servers and networking to provide connectivity and bandwidth to the disk from NERSC computational systems
 - Scale bandwidth and capacity about 10 GB/s per 1 PB
 - Components are redundant to avoid single point-of-failure and enable in-service maintenance activities to improve availability of data
 - /project regularly achieves 99% annual overall availability
 - Solutions are tested and evaluated before using in production
 - Service and support from vendors for a period of 5 years

- **For more information:**

<https://www.nersc.gov/users/data-and-file-systems/file-systems/sponsored-storage/>

NERSC Storage Relocation to CRT



- **Current plan is to move the global home file systems to CRT in June 2015**
 - New hardware will be installed at CRT
 - Data will be migrated over the network with file system in production
 - Expected to take several days to complete, but impact should be moderate during migration
- **Global Project file system is likely to move in Oct 2015**
 - Intend to move during Edison relocation
 - Data will be sync'ed locally at OSF, moved to CRT with file system in production
 - Requires multi-day downtime to complete final synchronization, goal is 2-day downtime, currently taking 7-days
- **Global Scratch is planned to move in Dec 2015**
 - Upon completion of the move to CRT, project file system will have new disk arrays providing aggregate bandwidth upgrade

- **Early 2015, expect improved HPSS performance from larger HPSS disk cache**
- **DTNs are refreshed with higher performing systems**
- **Global home file system metadata performance improved**
- **Project file system 80GB/sec bandwidth increase once Cori file system replaces Global Scratch**
- **Sponsored storage option helps science teams satisfy large allocation needs (> 50TB)**
- **Relocation of storage systems to CRT starting in June 2015 and planned to be complete by Dec 2015**