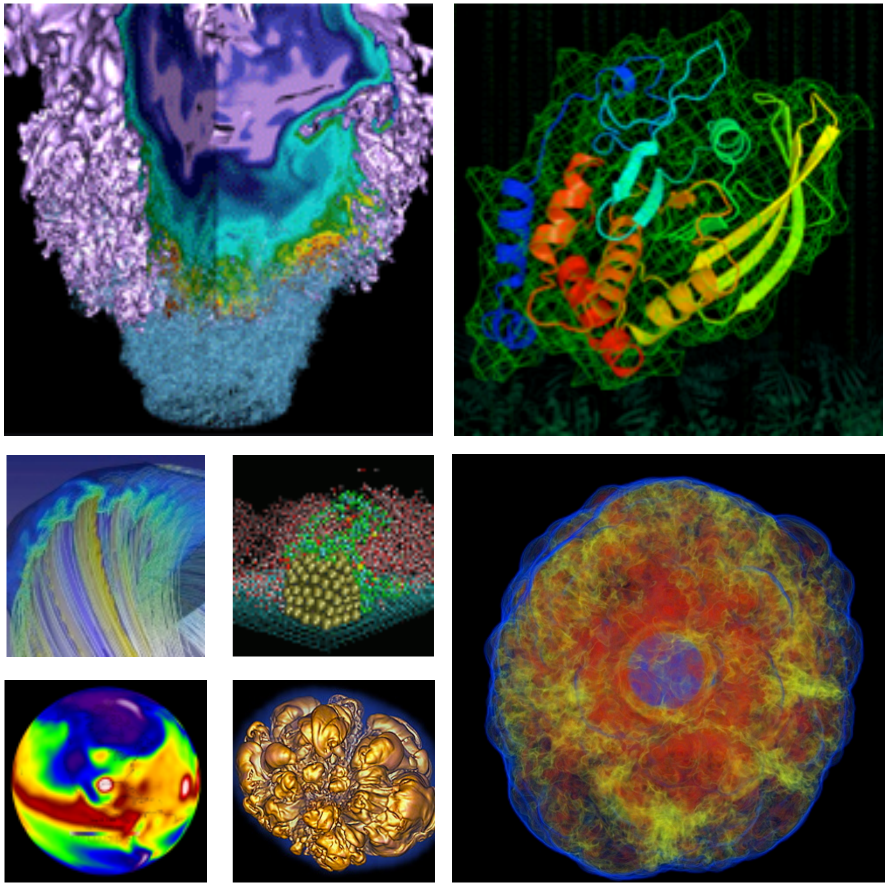
Storage Systems: 2012 and beyond



NERSC

Jason Hick

Storage Systems Group

February 12, 2013

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Science Discovery from Data Analysis

Astrophysics discover early nearby supernova

- Palomar Transient Factory runs machine learning algorithms on ~300GB/night delivered by ESnet "science network"
- Rare glimpse of a supernova within 11 hours of explosion, 20M light years away
- Telescopes world-wide redirected within 1 hour

Data systems essential to science success

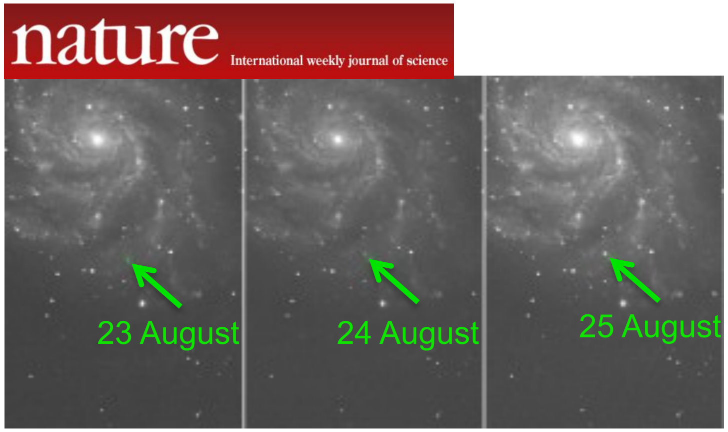
- GPFS /project file system mounted on resources centerwide, brings broad range of resources to the data
- Data Transfer Nodes and Science Gateway Nodes improve data acquisition, access and processing capabilities

PBS NEWSHOUR

ANALYSIS AIR DATE: Sept. 7, 2011

Supernova Burns Bright in a Galaxy Not So Far Away

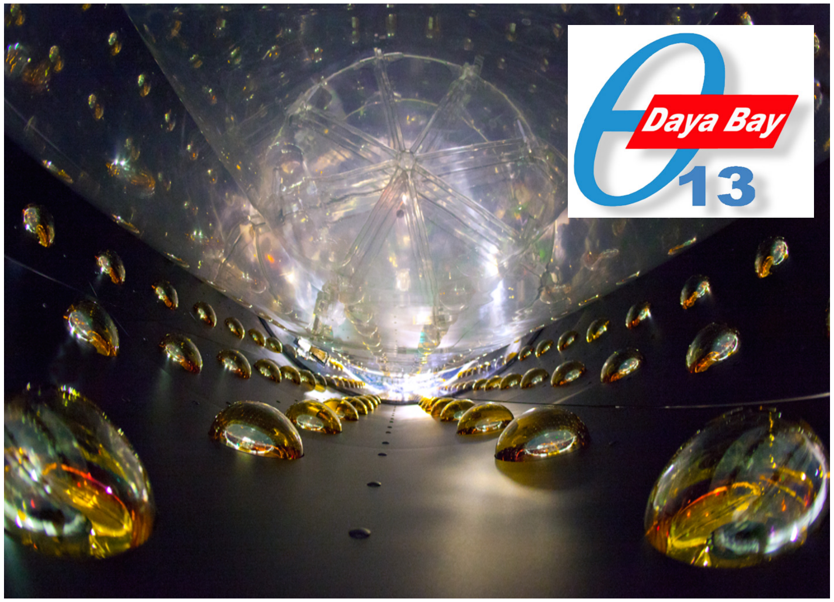
Peter Nugent
LAWRENCE BERKELEY NATIONAL LABORATORY



One of *Science* Magazine's Top 10 Breakthroughs of 2012

Discovery of θ_{13} weak mixing angle

- The last and most elusive piece of a longstanding puzzle: How can neutrinos appear to vanish as they travel?
- The answer – a new, large type of neutrino oscillation
  - Affords new understanding of fundamental physics
  - May help solve the riddle of matter-antimatter asymmetry in the universe.



Detectors count antineutrinos near the Daya Bay nuclear reactor in Japan. By calculating how many would be seen if there were no oscillation and comparing to measurements, a 6.0% rate deficit provides clear evidence of the new transformation.

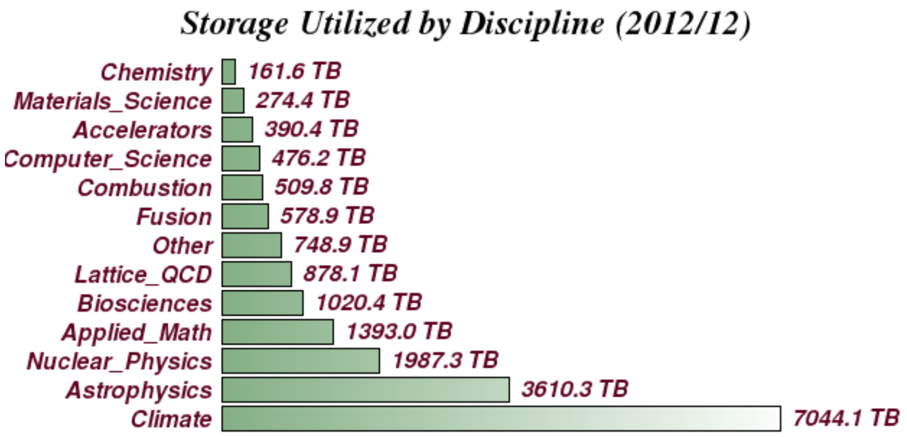
Experiment Could Not Have Been Done Without NERSC and ESNet

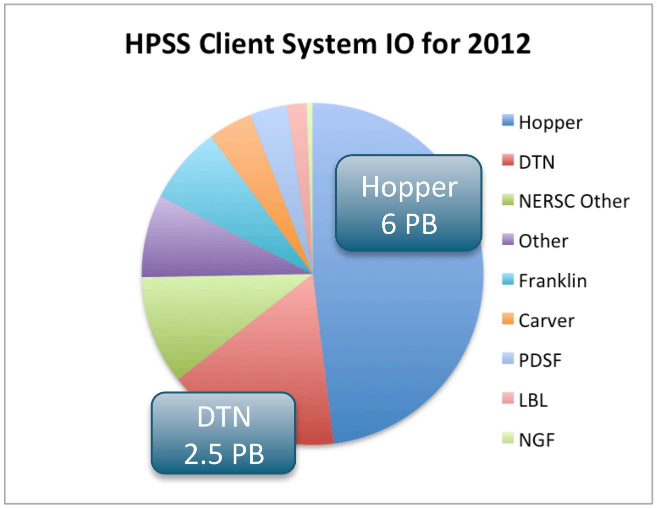
- PDSF for simulation and analysis
- HPSS for archiving and ingesting data
- ESNet for data transfer into NERSC
- NERSC Global File System & Science Gateways for distributing results

- NERSC is the *only* US site where all raw, simulated, and derived data are analyzed and archived

PI: Kam-Biu Luk (LBNL)

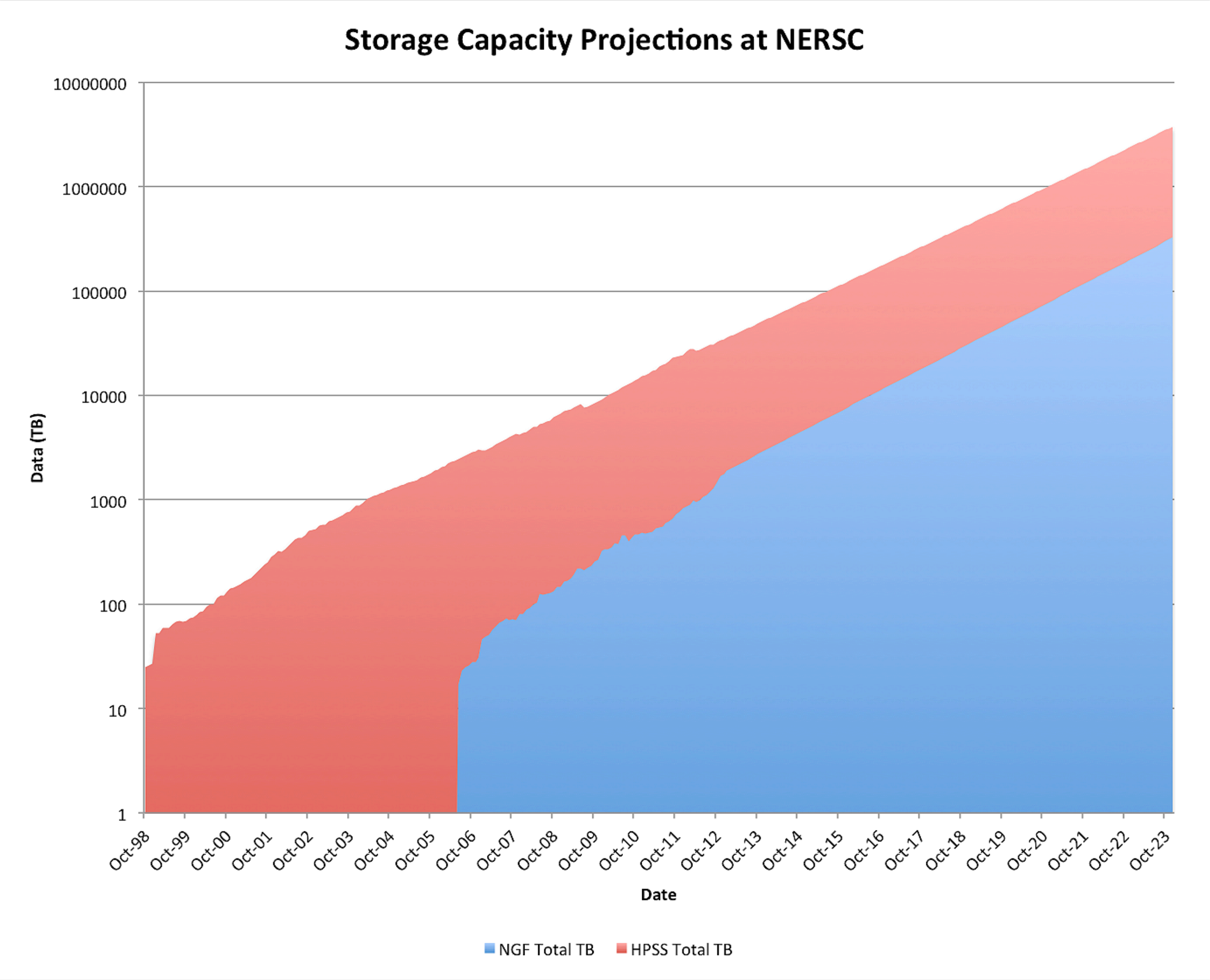
Overall statistics involving data





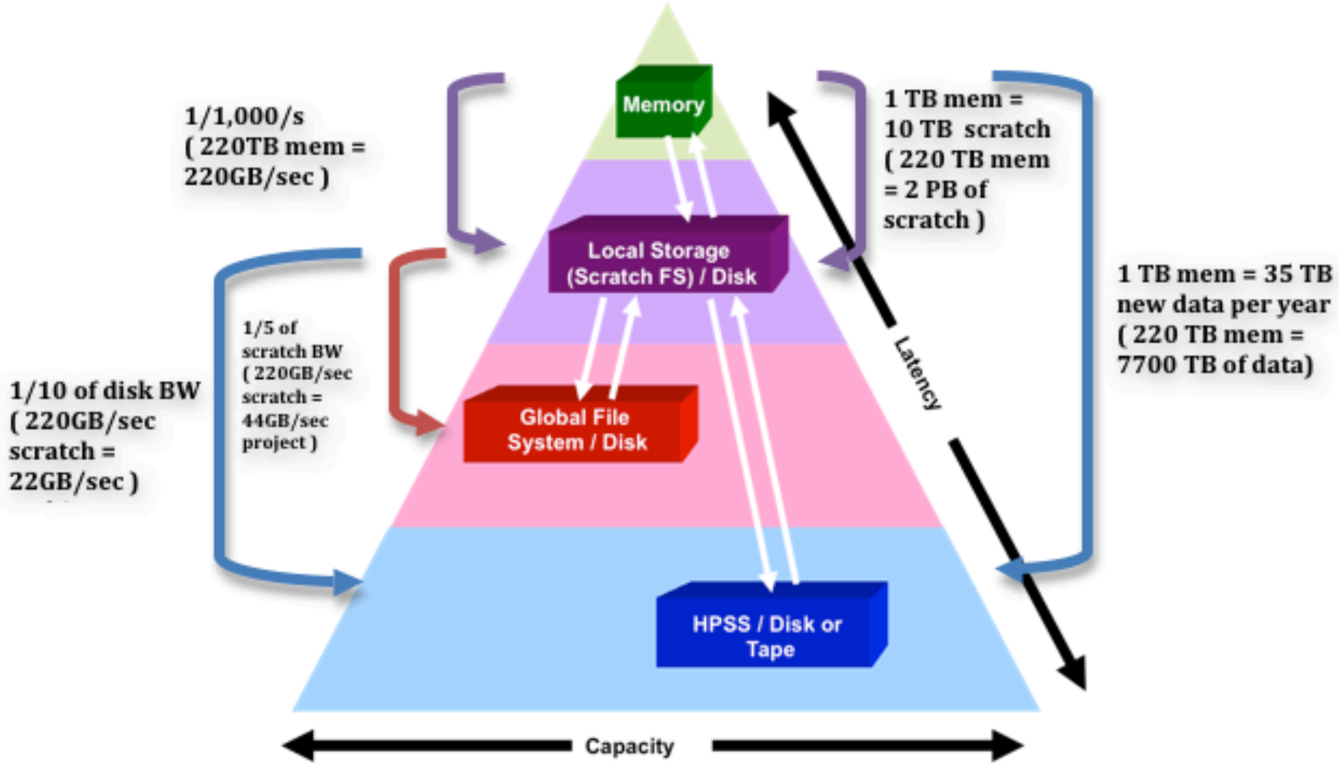
- Lately, the HPSS archive handles 2PB of I/O and grows at about 1PB each month
- Users import significantly more data to the Center than export
- Support 337 of our 700 projects with at least 4TB allocations on /project
- Incremental backups of multi-PB GPFS file systems daily (~10TB per day) and successfully transitioned backups to support users (26 user data restore operations in 2012)
- Big users of GlobusOnline (JGI, GPFS, and HPSS endpoints)

Capacity projections



- **NGF currently has 6,000 spindles and will require twice as many when it reaches 100 PB around 2021**
- **HPSS systems expected to be managing 100 PB in 2015 and will be provisioned for 1 EB capacity by 2018**

Bandwidth projections



1/1,000/s
(220TB mem = 220GB/sec)

1 TB mem = 10 TB scratch (220 TB mem = 2 PB of scratch)

1/5 of scratch BW (220GB/sec scratch = 44GB/sec project)

1/10 of disk BW (220GB/sec scratch = 22GB/sec)

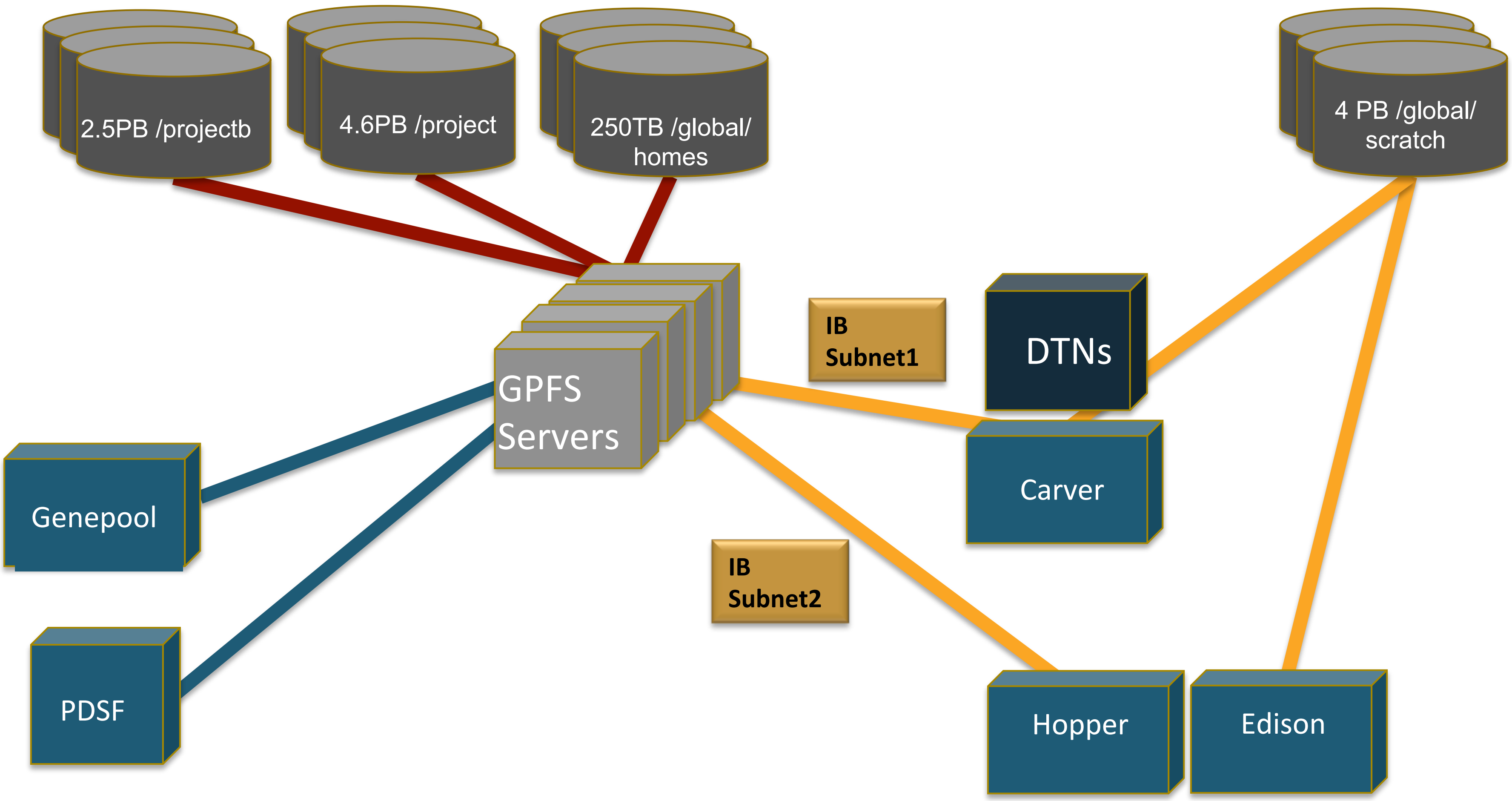
1 TB mem = 35 TB new data per year (220 TB mem = 7700 TB of data)

- The above is the NERSC aim for provisioning storage bandwidth
- With Edison, 330 TB of memory, 140 GB/sec scratch
- GPFS will add an additional 80 GB/sec global/scratch
- HPSS aggregate bandwidth will remain the same ~20 GB/s

GPFS resources

- **/project is for sharing and long-term residence of data on all NERSC computational systems.**
  - 120% growth in data stored for 2012
  - Not purged, quota enforced (4TB default per project), projects under 5TB backed up daily
  - Serves 337 projects over FC8, QDR/FDR IB, and 10Gb ethernet
  - 3.8 PB total capacity, planning to add nearly 1 PB of capacity after /global/scratch replacement
  - ~10TB average daily IO
- **/global/homes provides a common login environment for users across systems.**
  - Not purged but archived, quota enforced (40GB per user), backed up daily
  - Serves 4500 users, ~400 active per day over 10Gb Ethernet & QDR/FDR IB
  - 250TB total capacity
  - 100’s of GBs average daily IO
- **/global/common provides a common installed software environment across systems.**
  - 5TB total capacity
  - Provides software packages common across platforms
- **/global/scratch provides high bandwidth and capacity data across systems.**
  - Purged, quota enforced (20TB per user), not backed up
  - Serves 4500 users over FC8 primarily, 10Gb ethernet alternatively
  - 15GB/sec and 1PB total capacity
  - Replacing with new hardware, increasing to 80GB/sec and 4PB total by Jun 2013

GPFS Storage 2013



HPSS resources

- **User Archive System**
  - As of Feb 2012, contains 24 PB of scientific data:
    - Dating back to 1979
    - Largest file is 38 TB
  - 240 TB disk cache
  - More 5TB enterprise tape drives to improve ingest and read capability
- **Backup System**
  - Contains 14 PB of various backup data
    - ~50% is NGF/GPFS file system backups
  - 60 TB disk cache
  - 4TB enterprise tape drives to handle increase in backup/restore demand
  - Perform a user requested restore operation every other week (single file to several TBs)

Accomplishments 2012

- **Expanded HPSS bandwidth and capacity at least doubling both**
  - Production introduction of TS3500 Library with TS1140 drives (4TB tapes) and more T10KC drives (5TB tapes) enabling us to meet exponential growth needs
  - Deployed 3 new disk arrays and new HPSS p750 movers
- **Renewed GPFS contract**
  - Supporting our file system exponential growth demands through 2019
- **Testing/improvement of GlobusOnline endpoints**
  - HPSS endpoint as browser accessible using gridFTP
  - Rsync like capabilities
  - But typical new tape interface problems – no tape ordering, problems handling parallel data transfers, limited system administration tools
- **Expanded /project to enable Data Intensive Pilot awards**
  - Capacity more than doubled and we were able to deliver that capacity to 10 high demand science projects rapidly
  - Ultimately, whether it enabled science improvements or discovery will dictate whether we continue
- **User training for storage**
  - Presentations on using HPSS to various user groups
  - Best practices paper (http://www.nersc.gov/assets/pubs_presos/HSIBestPractices-Balthaser-Hazen-2011-06-09.pdf)
- **Work to improve GPFS availability**
  - Replacing problematic hardware
  - Lots of software upgrades (firmware, drivers, OS, GPFS server/clients)
  - New deployment strategy from GPFS server to clients (networking, server consolidation & specialization, direct attached storage)
  - Disk vendor engineering working on design/firmware improvements (~10 new firmware releases from 2 vendors)
  - GPFS features/bugs – FGDL, internal software fixes, new IB feature coming, problem diagnosis improvements

Goals 2013

- **User visible work**
  - NIM integration of storage services
    - /project new directories, renames, archiving, quota changes, backup notifications
  - Give and take utility
    - Enabling users to exchange copies of files
  - /global/scratch replacement
    - 80GB/s with 4PB capacity using DDN SFA12KE (embedded software on controllers)
  - /project expansion ~1PB
    - Repurposing /global/scratch hardware into /project, sunsetting older storage at same time
  - HPSS software upgrade to v7.3 and implementation of new features
    - Small file improvements
      - File aggregation for migration
      - Small file creation rate improved (~2x)
    - Checksumming for clients (HSI)
      - Validate and store checksums for HPSS files on the client
  - User storage training, advanced topics
    - Tape ordering, graphical interfaces, improving data reliability
- **System work on behalf of users**
  - DTN expansion/refresh
    - 4 DTNs today, considering expanding to 8 but need to refresh #1 & 2
  - GPFS software upgrade to v3.5
  - Continue work toward consolidated architecture of GPFS
    - Consolidation of servers
    - Move storage network from Ethernet + multiple IB to single IB storage network
  - Initiate sunsetting of 9840D and T10KB tape drives and media
    - 24x7 work from Operations to migrate data
  - HPSS 40 PB capacity increase
    - 4TB x 10,000 slot IBM tape library into full production

Goals beyond 2013

- **Plans to enable remote dual-copy option for HPSS data**
  - Guided by SRU allocation
- **GlobusOnline improvements**
  - Tape ordering
  - Enable per-transfer diagnostics/control
  - Improved retry logic
- **One-stop shopping for storage at NERSC**
  - HPSS and NGF information/services available via NIM (quota, backups, remote copies)
- **Long-term science repository of projects**
  - Today combination of /project and HPSS, consider a new solution with rich metadata/search functionality

NERSC

National Energy Research Scientific Computing Center