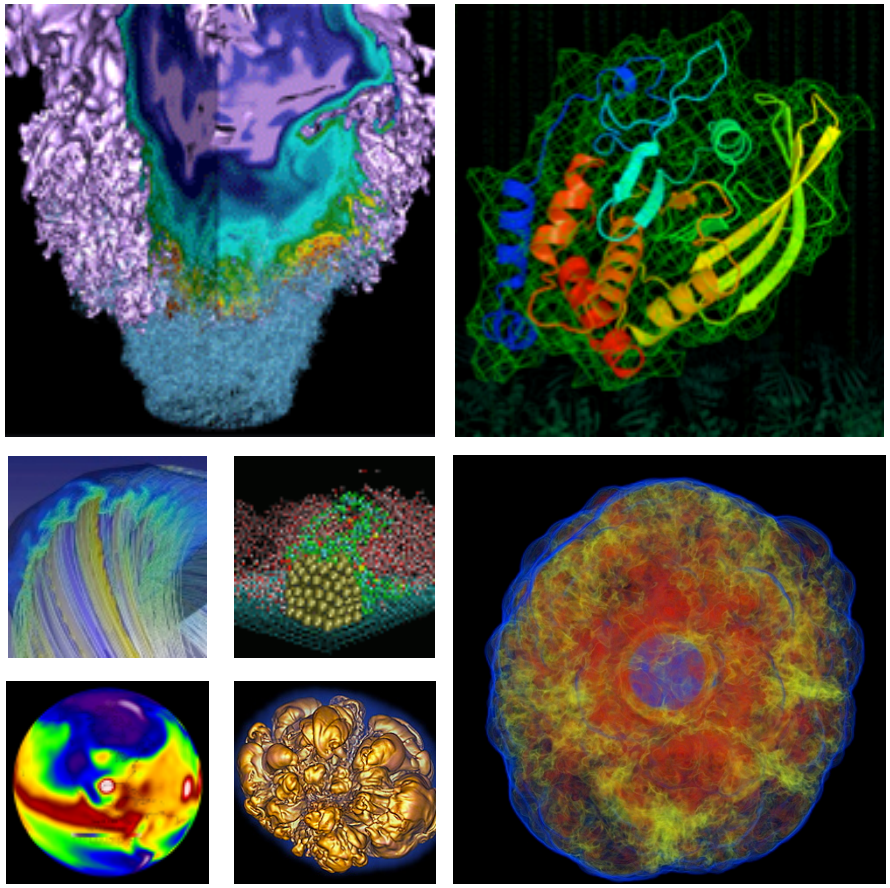
LBNL/NERSC Site Intro: HPSS in Production



Nick Balthaser

LBNL/NERSC Storage Systems Group

HUF 2013
November 5, 2013

Agenda

- HPSS at LBNL/NERSC
- Client Access
- Session Management and Fair Usage
- System Monitoring
- Tape Technology Integration
- Metrics
- Recent Challenges
- The Future of HPSS at NERSC
- Further Info

HPSS at LBNL/NERSC

- **NERSC is the production HPC division at LBNL**
  - DOE Office of Science unclassified research
  - HPSS developer site
  - ~ 5000 remote users, diverse unclassified research including climate, HEP, astrophysics, and bioscience (recent JGI merger)
    - All users are given accounts and storage space in the NERSC archive
  - 19k ft^2 data center in downtown Oakland, CA (OSF), moving to 32 ft^2 CRT facility at main Berkeley Lab site 2014 - 2015
  - HPC platforms include Cray XE6 (*hopper*), and Cray XC30 (*edison*)
- **Production HPSS systems**
  - Archive: ~30PB scientific data for users
  - Regent: ~20PB backup data for LBNL/NERSC systems
    - Limited dual-copy for select users on a case-by-case basis
  - Approximately 1PB/mo data growth over both systems (about 50%/yr)
  - Hardware/Technology:
    - Both systems are HPSS 7.3.3p9 on AIX/power servers, Oracle/STK SL8500 and IBM TS3500 libraries
  - Staff: SSG - 10 FTE: 4 NGF, 3 JGI, 1 HPSS developer, 2 HPSS systems/deployment
    - Open HPSS deployment position

Client Access

- **Standard Clients**
  - HSI/HTAR
  - PFTP
  - Standard FTP – various
- **GridFTP Clients**
  - globus-url-copy
  - uberftp
  - GlobusOnline
- **LDAP Integration**
  - In-house NerscAuth library enables token-based authentication after user authenticates with NERSC LDAP
    - Automated key-based login enabled after one-time LDAP authentication (like ssh key)
  - Unix authorization consults LDAP via AIX LAM/PAM module

Session Management and Fair Usage

- **Session Limits**
  - Users of standard clients (HSI/HTAR/PFTP) are limited to 15 concurrent logins via client mods
    - Mods allow decrease in concurrent sessions, login retry throttling, and/or user lock-out as needed
    - GridFTP/GO users can be locked out via grid-mapfile but we have no session limiting facility
- **Quotas**
  - Users are members of repositories (repos) that request HPSS Storage Resource Unit (SRU) allocations. Users exceeding their allocations can read but not write via mods to HPSS Gatekeeper (restricted status).
  - No hard limit as to how much data users can store at once as long as they do not exceed allocations
    - If they clog up migration we can limit sessions or lock out

System Monitoring

- **Many factors involved in system health assessment/monitoring:**
  - HPSS Application
    - Ops: 24x7 A&E monitoring, periodic store/retrieve checks
    - 24hr migration alerts (example at right) →
    - Db2cops, other various scripts/cron jobs
  - Libraries and Drives
    - Ops: 24x7 ACSLS and SLC monitoring
    - STA (Oracle), TSR (IBM)
    - AIX errpt– disk/tape device errors (cron)
  - Servers, Disk Arrays
    - Nagios w/in-house plug-ins
    - Cfengine – configuration, server file system space
  - Networks, Other Hardware
- **MSGStats DB**
  - Data collection/ingest:
    - ACSLS/TSR mount logs, cartridge histories
    - AIX device errors
    - HPSS configuration and errors from A&E
    - Migration/purge summary records
    - Accounting data: all successful transfers
  - Daily report/snapshot

From: **root@regent-e1.nersc.gov**
Subject: [msg] [fsg] Regent: migration candidate older than 24 hours
Date: October 8, 2013 6:05:02 AM PDT
To: Group fsg@nersc.gov

```
Date Tue Oct  8 06:05:01 PDT 2013

numrec x = 25
oldest x = Mon Oct  7 03:31:47 2013
newest x = Tue Oct  8 06:02:02 2013
  AGE x = 26:33:14

numrec 1 = 6
oldest 1 = Tue Oct  8 05:46:59 2013
newest 1 = Tue Oct  8 05:57:03 2013
  AGE 1 = 00:18:02

numrec 7 = 6
oldest 7 = Tue Oct  8 05:19:53 2013
newest 7 = Tue Oct  8 05:58:05 2013
  AGE 7 = 00:45:08

numrec 10 = 13
oldest 10 = Mon Oct  7 03:31:47 2013
newest 10 = Tue Oct  8 06:02:02 2013
  AGE 10 = 26:33:14
```

Daily Report

- **Example cartridge move report: daily cartridge movement in 4-library SL8500 complex:**

Cartridges Moved Between LSMs on SL8500 Complex (Src on Left, Dest on Top)

	1,0	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	1,10	1,11	1,12	1,13	1,14	1,15
1,0																
1,1	3				1			1	3	3	5			1	2	
1,2																
1,3																
1,4																
1,5		1						1			1					
1,6	9	12						12	1	2	5	2		2	8	
1,7		2					4		1	3	6	4		2	9	7
1,8		2					1			2		1				
1,9							3	13			2	4	3	5	9	7
1,10		1				2		10		11		3	3	5	7	6
1,11							1	1		3	7		1	5	6	2
1,12										2	1	1		2	5	
1,13							1	8		1	3	4			11	5
1,14		2					6	7		12	10	5	3	9		9
1,15								7	1	7	9	2	1	2	7	

Volume EP258000 moved 7 times, mounted 7 times

New Technology Integration

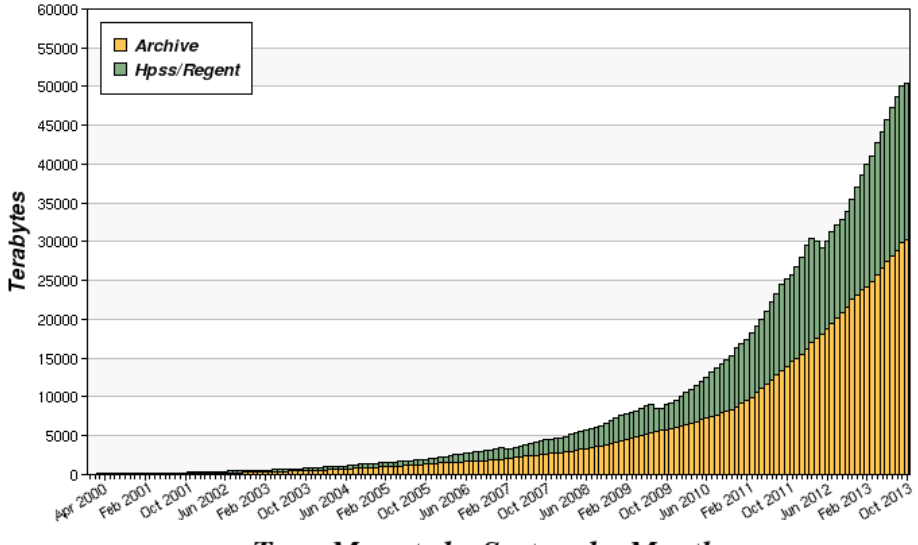
- **New media/drives are tested for performance and reliability on pre-production system. New technology is then deployed via either:**
  - Technology Insertion
    - Retire media from previous storage class
    - Re-create HPSS Storage Class with new drive/media type/label range
    - Repack files from retired media onto new media
  - Storage Class/COS Creation
    - Create new storage class for new drive/media type
    - chcos appropriately sized files from former SC to new SC
    - Repack sparse media
- **We are in a constant cycle of repack and chcos onto higher capacity media in order to preserve slot count**
- **Media is relabeled and re-used at higher capacity when possible (e.g. T10KA → B)**

Metrics

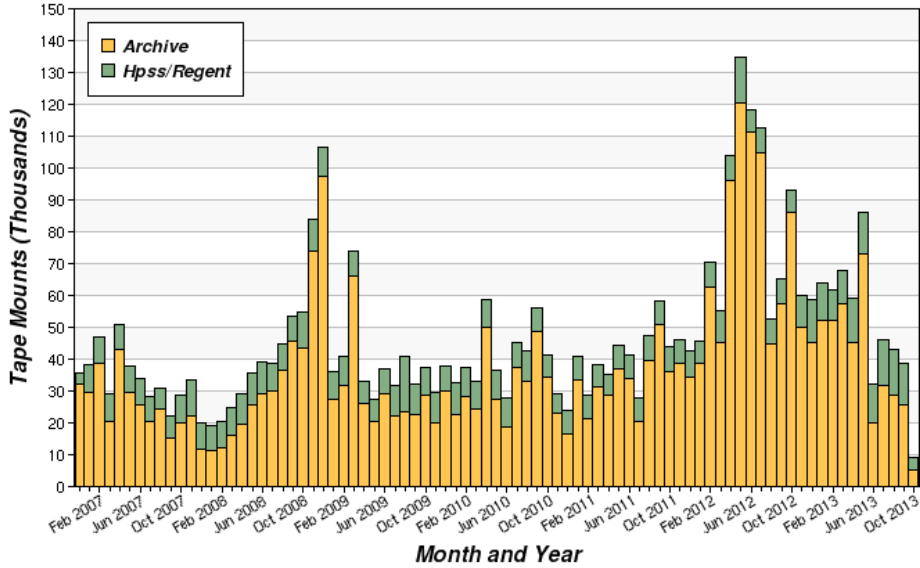
- **NERSC Web site, publicly available:**
  - Bandwidth and Transfer Activity: Daily/ Weekly/Yearly
    - Data rate vs. file size
    - Aggregate transfer bandwidth
    - Number concurrent transfers
    - Active file transfers
  - Storage Trends and Summaries
    - Cumulative bytes stored by month and system
    - Number of files stored by month and system
    - IO by month and system
    - Yearly network traffic: fetch vs. store
    - Number tape mounts
  - Storage by scientific discipline
  - System live status (up vs. down)

- **Tracked internally:**
  - Availability
  - Daily IO (R/W) – ingest vs. retrieval, by user
  - Slot counts
  - Scratch tape pool by Storage Class
  - Migration: rates and MPS stats
  - Storage clients in use

Cumulative Storage by Month and System

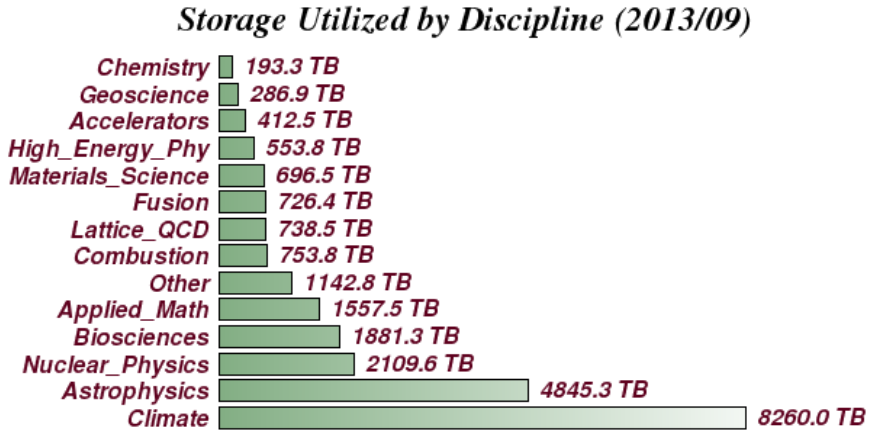


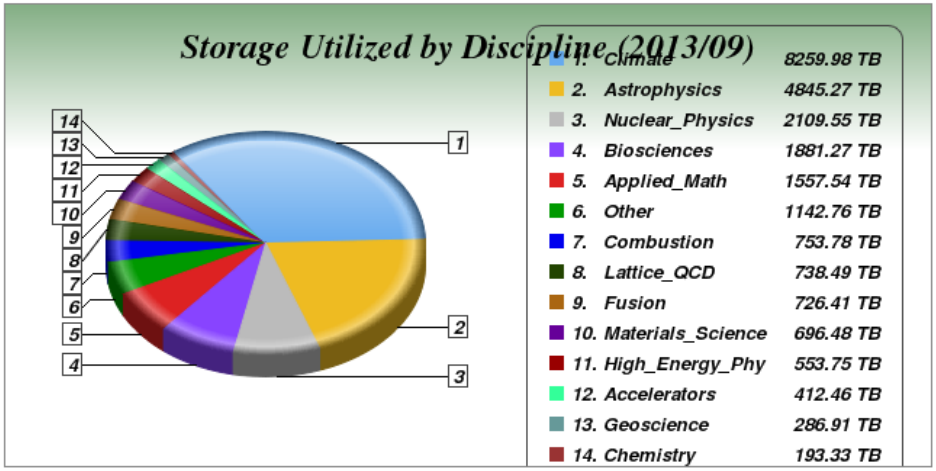
Tape Mounts by System by Month



Example Graph

- **Storage by scientific discipline:**





Recent Challenges

- **GlobusOnline/HPSS Integration**
  - Testing and deployment of NCSA GridFTP/HPSS DSI for HPSS 7.3.3
- **Tape Aggregates**
  - File retrieval: unordered vs. aggregate order performance
  - Repacking
- **IBM TS3500 Library Integration**
  - SCSI PVR idiosyncracies
    - ACSLS Drive: 1,11,1,1
    - SCSI Drive: T10:IBM|03592E07 0000078D02AC
- **Managing user behavior, issues and expectations**
  - Difficulty with standard HPSS clients
  - Optimal HPSS usage, e.g. storing small files
  - Ordering file retrievals by tape position
  - Restoring deleted user files
    - *.Trash* in HPSS?
  - Irretrievable data/damaged cartridges

The Future of HPSS at NERSC

- **2014 – 2015 CRT move**
- **Linux integration**
- **Disk cache expansion**
  - Increasing cache residency timespan (more reads from disk instead of tape)
- **Possible DR/second facility**
- **Test/enable GO/HPSS checksums**

Further Information

- **General NERSC HPSS Info:**

  http://www.nersc.gov/users/data-and-file-systems/hpss/

- **NERSC Public HPSS Stats and Metrics:**
  http://www.nersc.gov/users/data-and-file-systems/hpss/storage-statistics/

- **HPSS Admin Wiki (HPSS customer sites only):**

  https://www.mgleicher.us/hpss/admwiki/doku.php

National Energy Research Scientific Computing Center

Section Title