BNL Box

Hironori Ito
Brookhaven National Laboratory
Spring HEPiX, Budapest, Hungary
April, 2017
Concept

- All of us need the convenient method to transfer or access data in different systems
  - Users might need to copy their analysis scripts and the data between their workstations and central analysis farm separated by different network and firewalls
  - System administrators might need to transfer custom software packages to their systems for installations.
- In BNL RACF, AFS has been the storage of choice for moving small amount of data in/out of various systems.
- AFS limitation
  - Being phased out
  - Not really universally accessible.
  - Not easiest one to use in various platform.

- Commercial cloud storage seems to be popular among some of users and sys-admins.
  - Dropbox, Box, Amazon Cloud Drive, Google Drive, MS OneDrive, etc…
  - Advantages of commercial cloud storage
    - Already available for use
    - Easy to use. All of them provide https-based access.
    - Free (up to some level)
    - Available in various platforms.
  - Limitations
    - Size/Cost/Performance.
    - Archive
    - Not really meant to stream data
Target users

• All users of BNL
  • HEP/Nuclear physics communities
  • Sys-admins
  • Users from different science domains than HEP
    • NSLS-II (National user facility)
      • Massive data producers for many beamlines by many users.
    • Nano Center (National user facility)
      • Another large data producers.
  • Chemistry
  • Biology
  • Etc…
Target usage

• Transfer small data in & out of BNL between central interactive farms, workstations, laptops, and tablets/smart phones.
• Transfer large data in & out of BNL between detector data stores, central storage, remote storage of users.
• Access data to/from analysis computing farm
  • Copy to scratch
  • Stream data
• Archive data
• Owncloud Software
  • Clients are available in many popular platforms; Linux, Mac OS, MS Win, Android and IOS
  • Extremely easy to use.
    • Synchronize data automatically
    • NOTE: Requires the same amount of storage in local and remote storage.
  • Quota for each users
  • Users can share data

• Ceph Storage
  • Currently Infernalis. Targeting Kraken.
  • Reliability
  • CephFS
  • 3.8PB Raw -> 7.5PB by the end of 2017
  • Performance
    • 40Gbps for BNL Box
Mobile APP
Android / iOS

Laptop / Desktop App
MS Win / Mac / Linux

Worker Nodes

Request to restore the data from archive

Owncloud

Custom easy-to-use copy-command

Tape Restore Control

Custom XROOTd
Name-to-name
Translation to isolate a user to own data

dCache
HPSS Tape Archive

CephFS

XROOTd

70 YEARS OF DISCOVERY
A CENTURY OF SERVICE
WebDAV access and Sync

• Default sync app seems to synchronize data at the top rate of about 100MB/s per client. (100MB/s = 360GB/Hr = 8.6TB)
  • Sufficient for small data ~ less than TB.
    • Most users won’t need or physically have higher throughputs in their systems.
      • Spinning DiskI/O on desktop (~100MB/s).
      • Wifi N (max 300Mbps~40MB/s)
      • LAN (1Gbps=120MB/s)
      • Disks are not much larger (currently max at about 10TB)

• High demand users require higher throughput.
  • 10TB or more.
  • Owncloud supports standard WebDAV protocol
    • Easy to write a custom copy tool.
      • Easily achieve 150MB/s per single file transfer.
      • Concurrent multiple transfer of files will results in obtaining desired throughputs.
      • NOTE: Different SSL library seem to impact the observed throughput of WebDAV command. For an example, “curl” in RHEL 7 is compiled with NSS. This version of “curl” produces 1/5 of throughput of “curl” using OpenSSL.
Stream Access

• XROOTd and WebDAV can stream data
• Would like to separate the data-sync operations from the data-read access as much as reasonably possible.
• XROOTd can cleverly map user data in BNL Box in a very simple way.
  • Owncloud web URL maps a user data by https://host/owncloud/index.php/apps/files/MYDATA
  • This is different from how Owncloud physically stores user data in its storage as /base-directory/username/files/MYDATA
  • XROOTd can cleverly hide “username” of physical files by providing access by root://host/files/MYDATA
    • Courtesy of Andrew Hanushevsky from XROOTd
Archive data

- Some users would like to archive or store data in the tape system.
  - Will the data be read again?
- Difficulties
  - Efficiency
  - Read throughput
  - Reading small fraction in many different tapes will result in low throughput.
    - Seek is slow.
    - Mounting a tape is very slow.
- Must write in a particular way to produce the good read-IO.

- Rule
  - "/Tapes/" directory will be used to indicate data to be stored to the tape system.
  - Files smaller than certain size (1GB) will be tarred to produce a large file.
  - Tar files smaller than 1GB will be archived to tape only after a certain period.
  - Once files are transferred to the archival system, they will be removed from "/Tapes/" directory.
    - Reduce the usage of quota.
    - Create index or individual local catalog file to record the data in the archival system.
    - The above index will be synchronized by the owncloud to their local machine.
    - Also update the central catalog for archived data.
  - Restore requests will be made through a Web interface.
    - Data will be restored to different directory.
• Users only see their own directory.
• Users can share their data publicly or privately with password.
Using the provided app, users can decide what to sync automatically. For an example:

- Data and Tapes directories are not synchronized.
- Codes, Documents, Photos directories are synchronized automatically.

Desktop/Laptop apps are available in MS Win, Mac and Linux. The performance seems to be limited to the maximum of 100MB/s.
Conclusion

• Cloud storage could be potentially useful for data intensive scientific communities.

• BNL Box will provide our users with ability to store and access their data anywhere by the easy-to-use applications on various platforms.

• BNL Box allows the owners of the data to share with anyone without involvement of the system administrator.