BNL Tier1 storage

Hironori Ito
Brookhaven National Laboratory
Argonne National Laboratory
Dec 2018

BNL Tier1 Storage

- dCache storage
  - Version 3.0.x (different version on different dcache host)
    - Preparing to upgrade to 4.2.x
      - Upgrade expected in January.
  - PostgreSQL 9.6 on NVMe hot-backup for the name space.
  - Supported protocols
    - SRM, GridFTP, XROOTd, NFS, HTTPs/WebDAV, dCap
  - 17PB of the usable disk storage
    - Located behind the firewall.
    - External clients must use the proxy services on DTN nodes.

- dCache storage (continue…)
  - Very large tape read pools.
    - Much larger than it used to be or other T1 sites
    - Spit two types of tape read pools
      - Pools for newly written files.
      - Pools for staging files for HPSS.
    - To be used for the "US BNL Lake" concept.
    - Large HPSS disk cache. NOT the same as dCache read pools.
- XROOTd
  - External proxy service for BNL dCache
  - XROOTd cache for US BNL Lake.
- Globus Connect Proxy service
  - Data transfers to HPC sites.

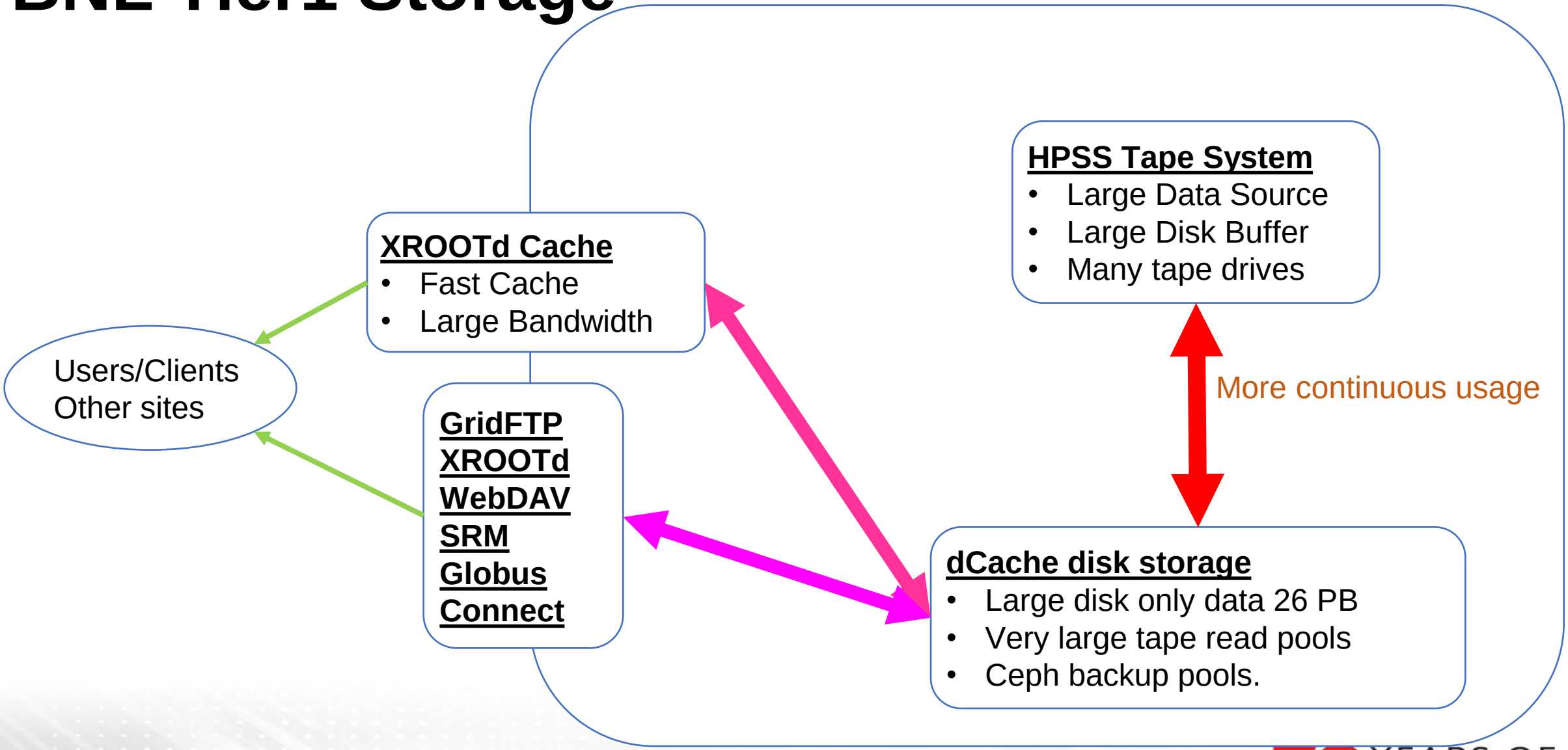
Maintenance of storage hardware

- Currently ~5 year cycle.
  - Maintaining many aging hard-drives require a lot of man power.
    - The older (more used) the disks are, the higher the rate of disk failures.
      - AFR, annualized failure rate: 1% for new disks to several % for aging disks.
    - More than 4000 disks (with resiliency).
      - A few die every week for aging storages.
        - Eg. 5%AFR + 4K disks = 200 dead disks/yr ~ 1 dead disk per 2 days.
      - Each disk failure require a few to several hours of man power; replacing disks, sending broken disks back and receiving a new disks.
    - RAID 6
      - Must fix the broken very quickly or see the dead lun.
      - Dead lun means >> 100 TB of lost data.

2019 new hardware addition

- 24 new storages.
  - For retirement of old storage and full fill new BNL share 26PB
  - >20PB of new storage (JBOD with 102 disks). No small task!! Manual installation of 2448 disks !!!
- Deployment of 10PB Ceph storage as a resilient backup using the retired storage.
- More tape read pools and tape drives for US BNL Lake.
- XROOTd cache deployment.
  - Two servers with 2 x 40 Gbps.
  - Each server with 2 x 22 TB NVMe disks

BNL Tier1 Storage



Highlight of change

- Very large data cache
  - Large pools of tape backed dCache storage. >> PB (compared with 100s TB)
  - More tape drives
  - Larger tape cache ~ PB
  - Don’t copy to disk area. **No (or reduce) TAPE to DISK copy!**
- Very fast XROOTd cache
- Better usage of man power with resilient Ceph storage.
- More APIs
  - Globus
  - 3rd Party XROOTd and HTTPs.

Big lake with large flow