Storage and data management at SDCC

Hironori Ito
Brookhaven National Laboratory

BROOKHAVEN NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY
BROOKHAVEN SCIENCE ASSOCIATES

Outline

- Nuclear and HEP experiments supported at BNL.
- Worldwide Data Distribution
- dCache Storage at BNL
- Quality of Service
- BNLBox
- Conclusion

Experiments at SDCC

- NP
  - STAR
    - XRootD
      - 10PB
    - Tier 0
  - PHENIX
    - dCache
      - 13PB
    - Tier 0

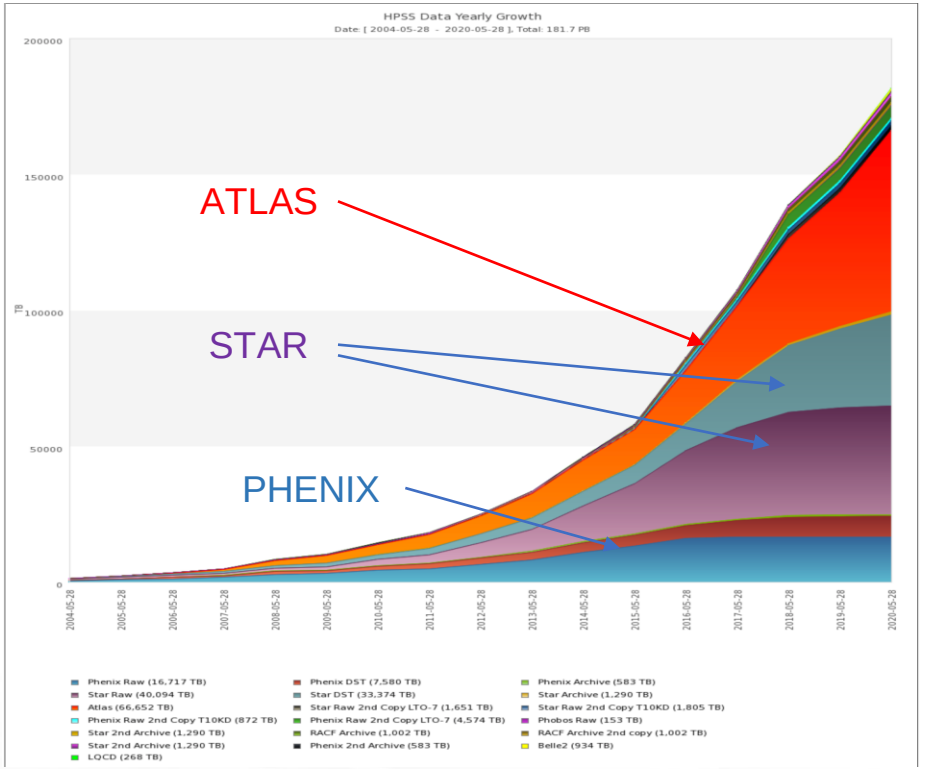
In House

- HEP
  - ATLAS
    - dCache
      - 22PB resilient
    - Tier 1
      - 23% of RAW
      - Tier0 at CERN
  - Belle II
    - dCache
      - 2.5PB
      - Raw Data Center
        - 100% RAW
        - Tier0 at KEK
      - Calibration Center

Distributed Computing

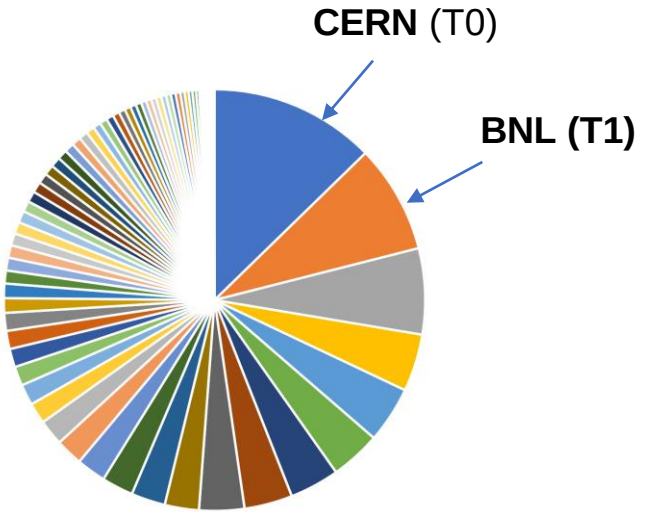
Data Volume in BNL Tape Archive

- The volume has steadily increased over the years.
- The total volume exceeds175PB.



Data distribution

- In ATLAS and Belle II, the data is distributed worldwide over the network.
  - For an example, there are over 70 sites in ATLAS with site storage.
- There are two copies of RAW data; one in T0 and the others distributed in T1s.
- The sites are connected with LHCOPN/LHCONE network.
- Data is managed by RUCIO



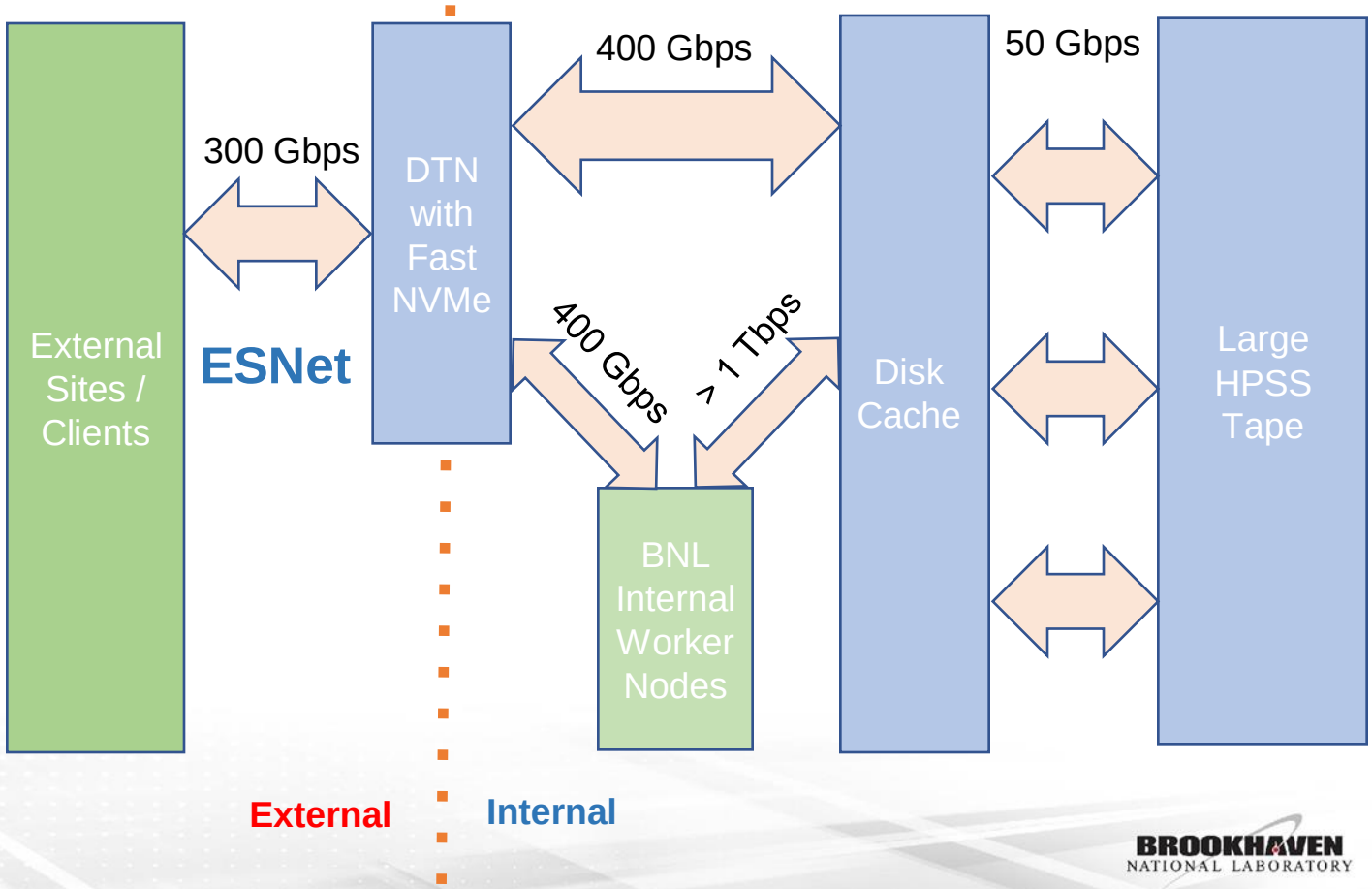
Basic diagram of ATLAS dCache at BNL

The data is meant to be distributed world wide.

DTNs are access points into the storage at BNL.

DTNs accept various protocols;
--- SRM
--- GridFTP
--- https/WebDAV
--- XRootd

Transfers are managed by FTS.



File Transfer Service, FTS

Open source software for data transfer.

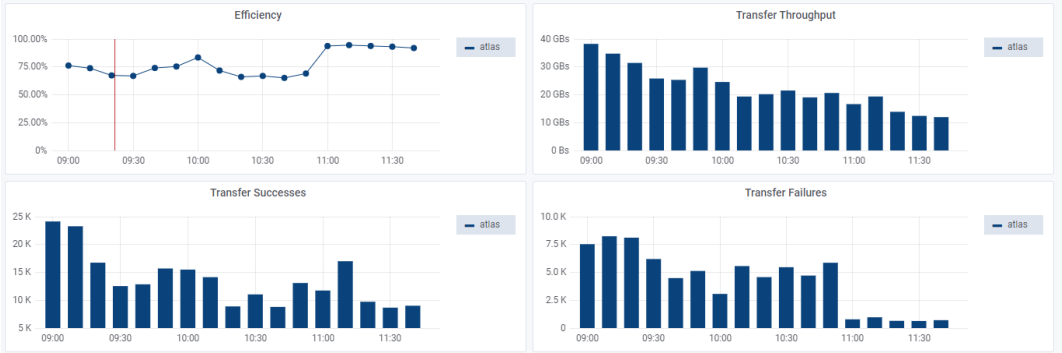
It is developed at CERN.

FTS is used by many experiments. The services are deployed at many sites including BNL (used by ATLAS and Belle II)

It acts as a scheduler for the data transfer between the various storage end points.

It can stage files for the tape backed endpoint before the transfers.

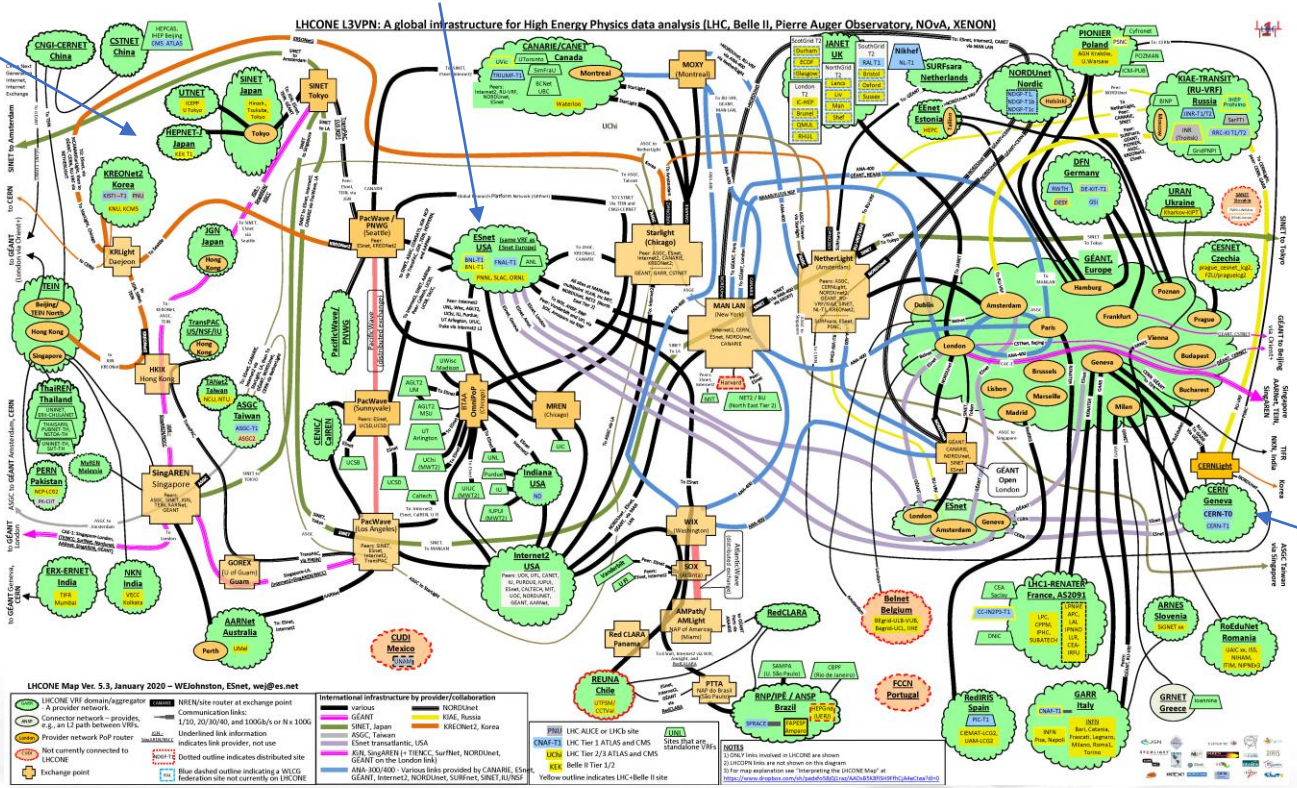
It initiates 3rd Party transfers. The transfers are between the two endpoints.



https://fts.web.cern.ch/fts

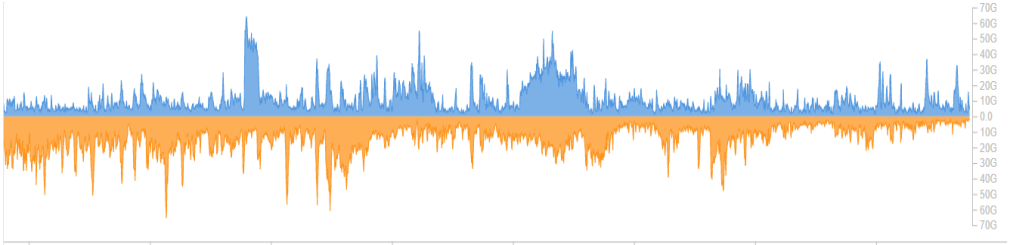
LHCONE Network

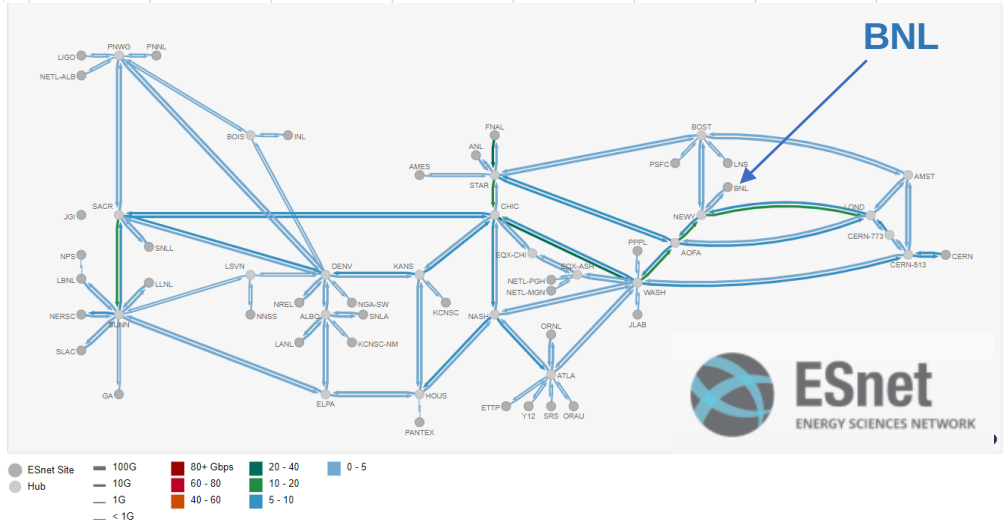
LHCONE network is being used by LHC(CERN), Belle II(KEK/Japan), NOvA(FNAL/USA), XENON(Italy), Pierre Auger Observatory (Argentina)



Network to/from BNL

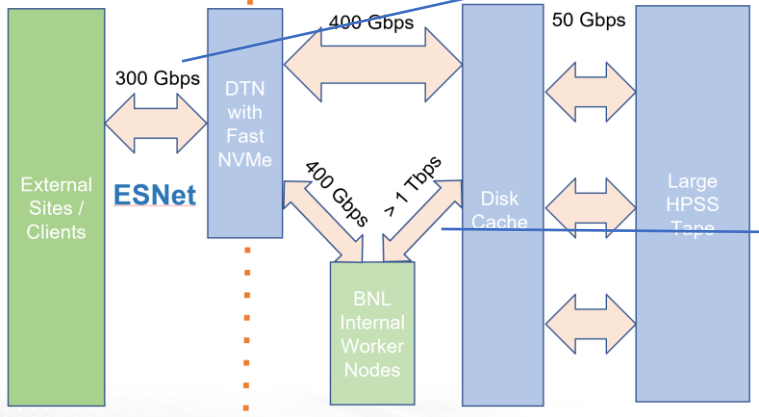
- ESNet manages all external network to/from BNL.
- BNL has 300Gbps (3x100Gbps) throughput.
- It regularly reaches the ~100Gbps.

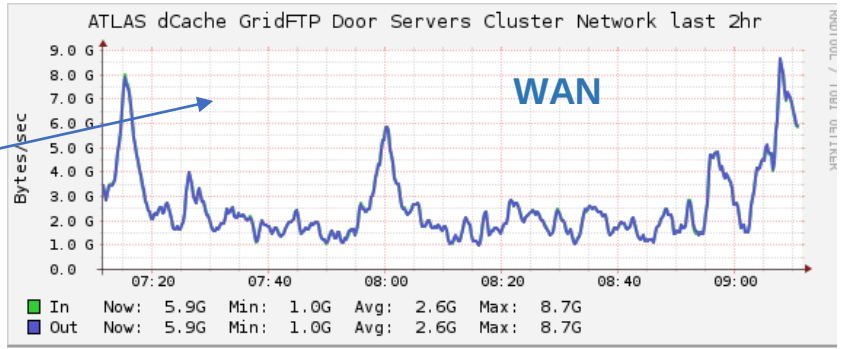


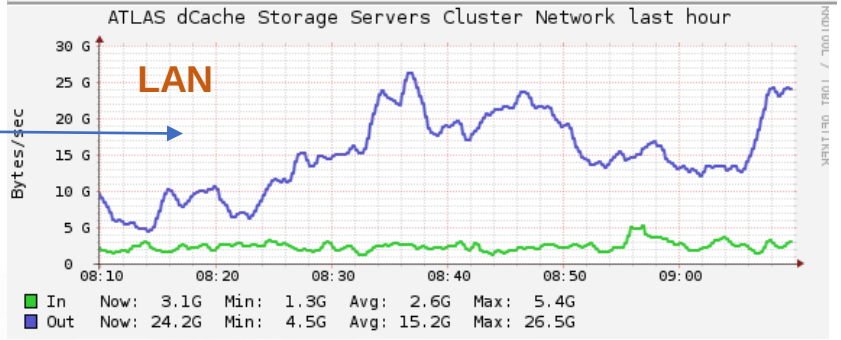


Disk Storage Throughput

- Internally, BNL network regularly deliver the data at 30GB/s.
- Externally, the rate reaches around 10GB/s often.





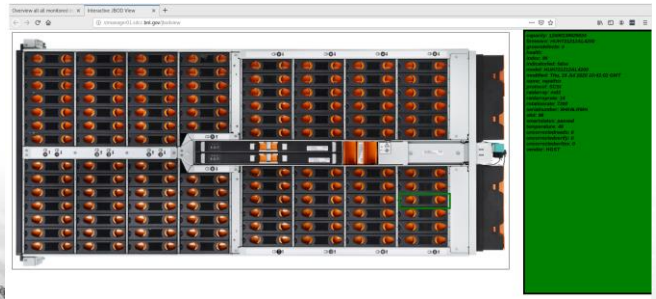


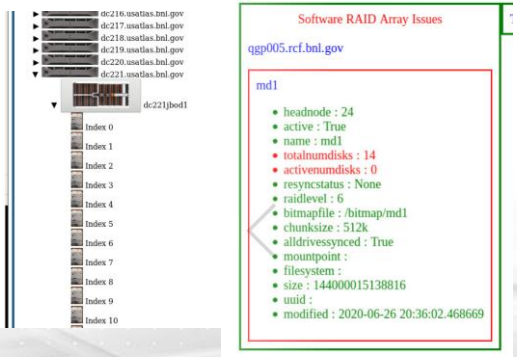
File access

- Files are typically accessed through SRM/GridFTP, Xrootd, https/WebDAV.
- SRM/GridFTP is typically used between two sites.
  - Due to the end of the support by Globus, this will be replaced by WebDAV.
  - There is a on-going effort by WLCG to migrate to WebDAV. (WLCG Data Organization, Management and Access, DOMA, project)
- XRootD is typically used by worker nodes (jobs) as well as users.
  - Xcache; Data cache service used by user analysis. (DOMA activity)
- Token Based access in stead of X509 Certificate

JBOD

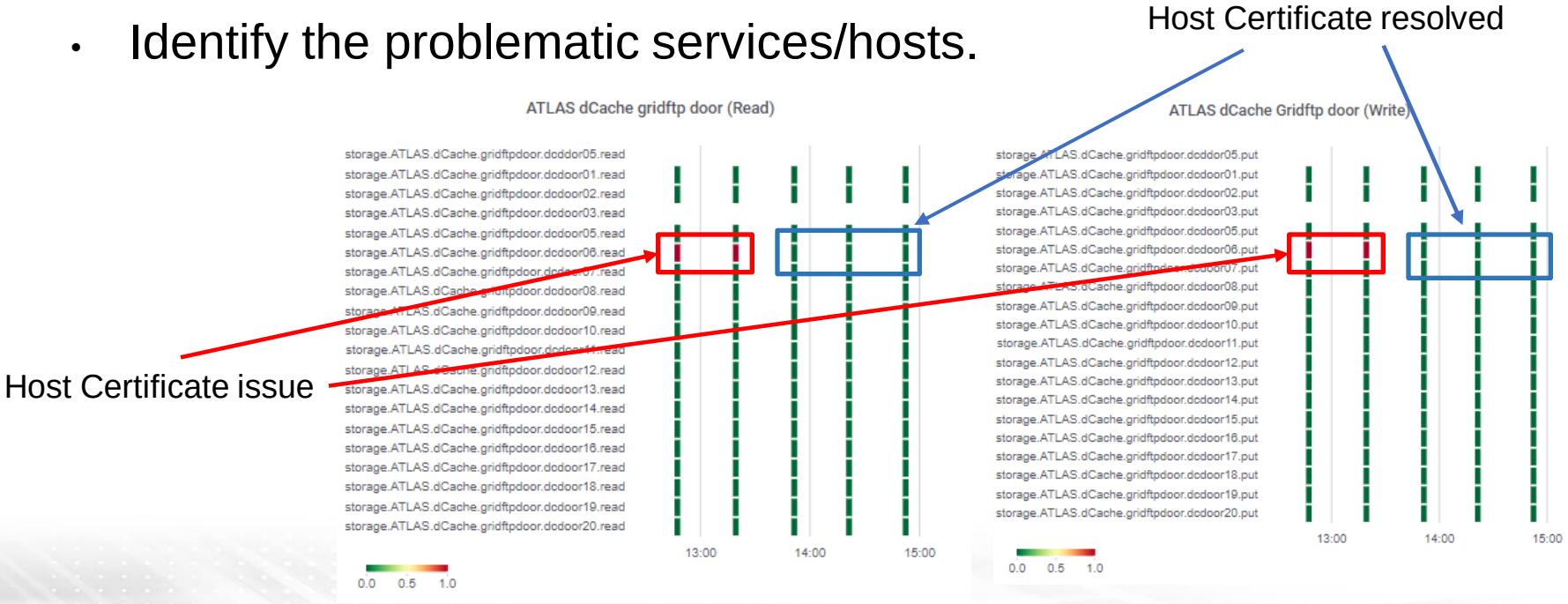
- The balk of the storage are from JBODs.
- The most of them have 102 hard drive disks with the capacity range of 10TB, 12TB or 14TB.
  - The single 4u JBOD can provide 1.2PB usable space.
- JBODs are cost effective. However, it requires more effort in the management due to the lack of management software which typically comes with the hardware-controller based storage.
- Created JBOD management software managing the total of 7000 disks.





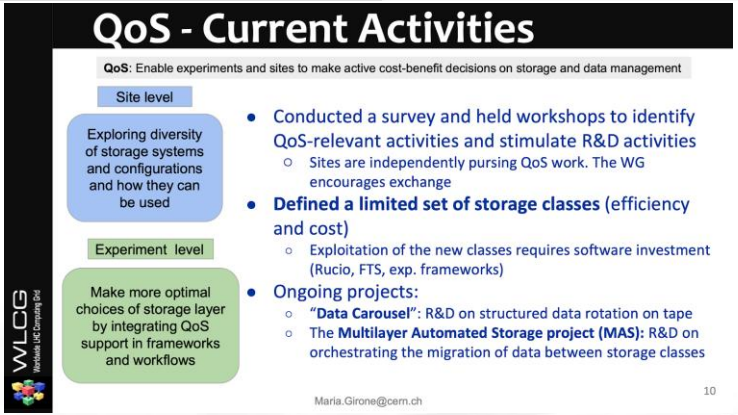
Monitoring Example

- Self-testing various storage interfaces
- Identify the problematic services/hosts.

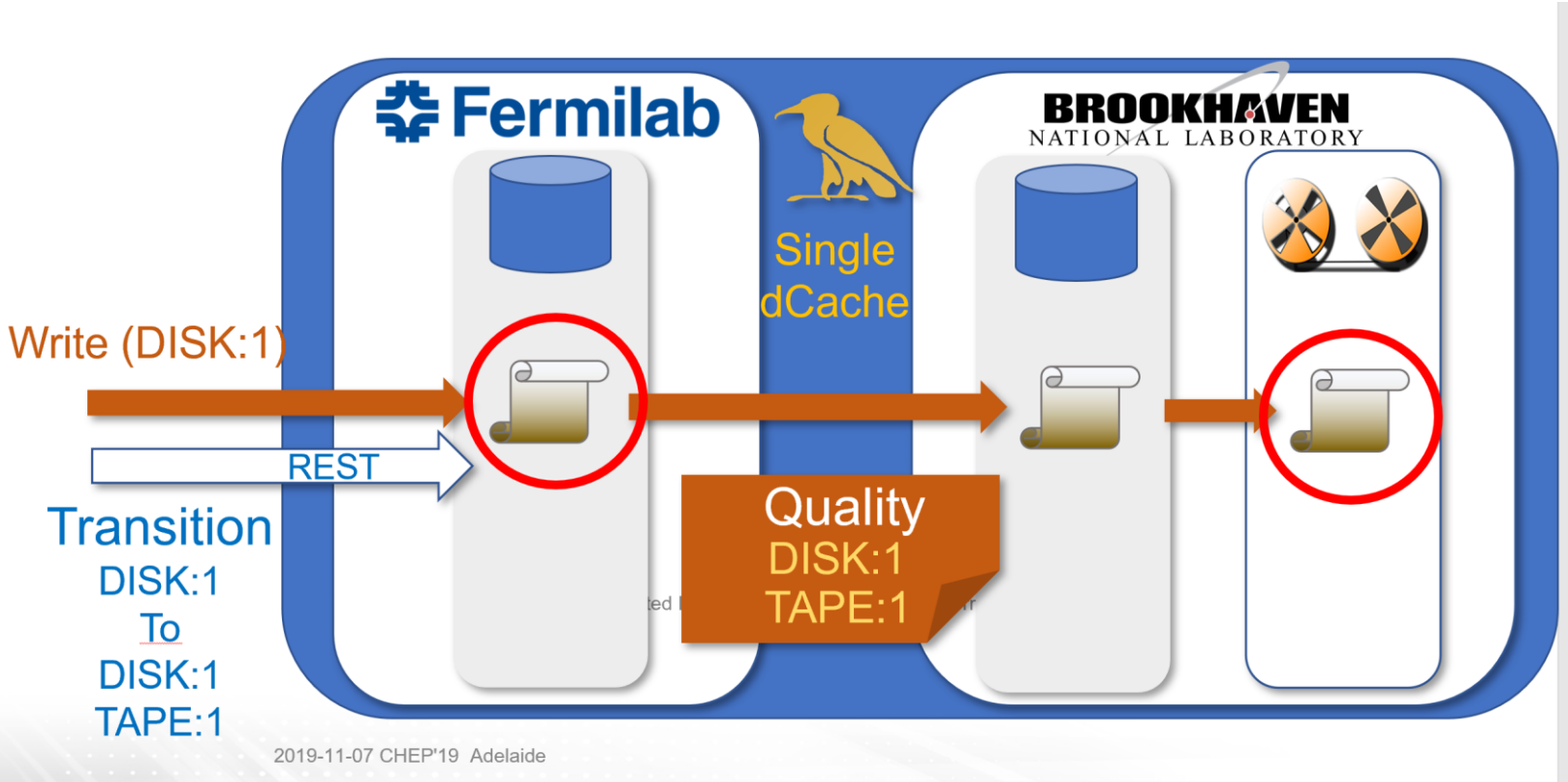


Quality of Service, QoS

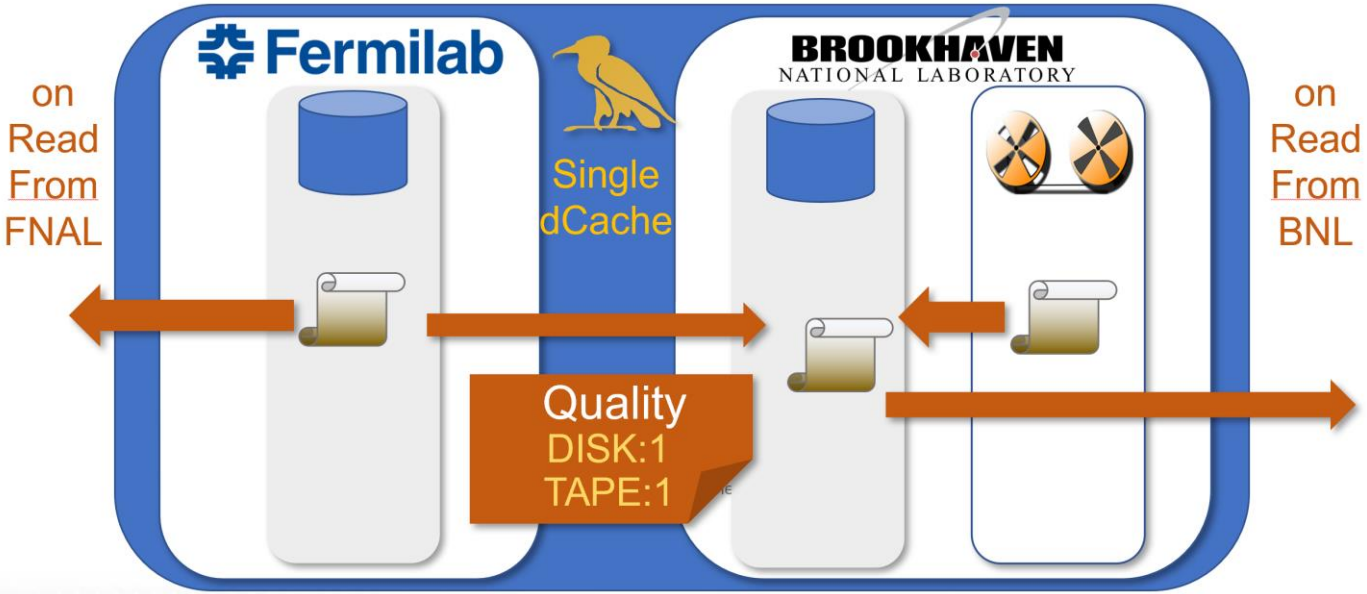
- The demand for the data volume is increasing. It is possible that it grows faster than the funding.
- Typically, the user/experiment specifies the total available disk volume without specifying the required performance.
- Typically, the large fraction of data on the disk are not being used.
  - They are often used for short period and never read again.



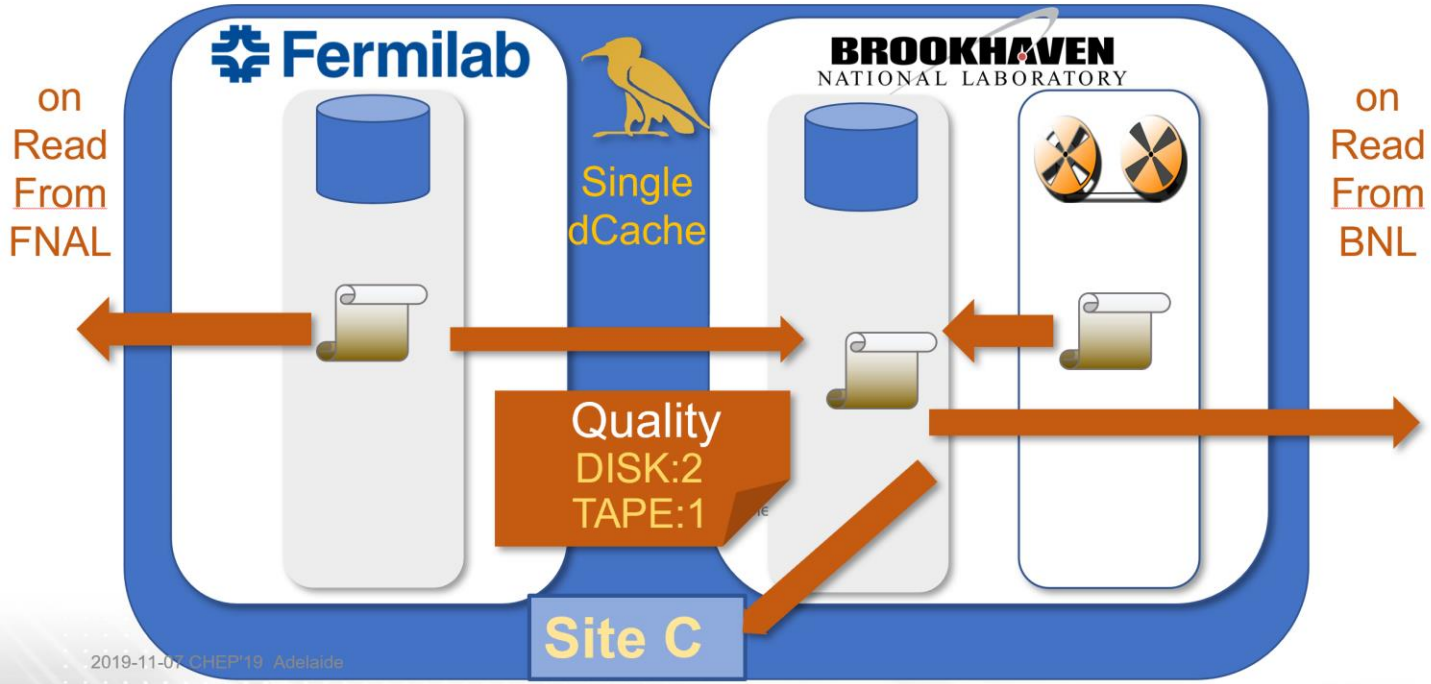
QoS with dCache at BNL and FNAL



QoS with dCache at BNL and FNAL (continue)…

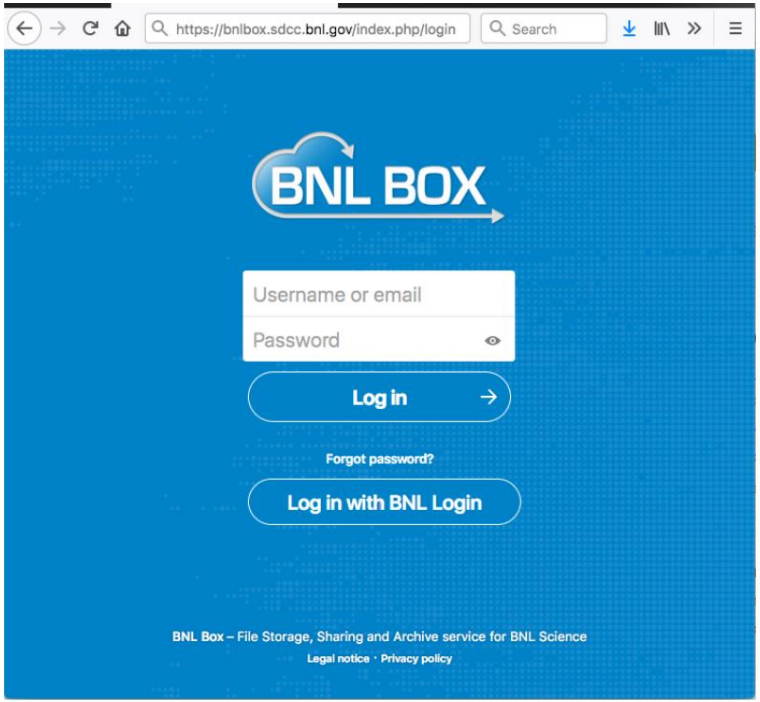


QoS with dCache at BNL and FNAL (continue)…



BNL Box

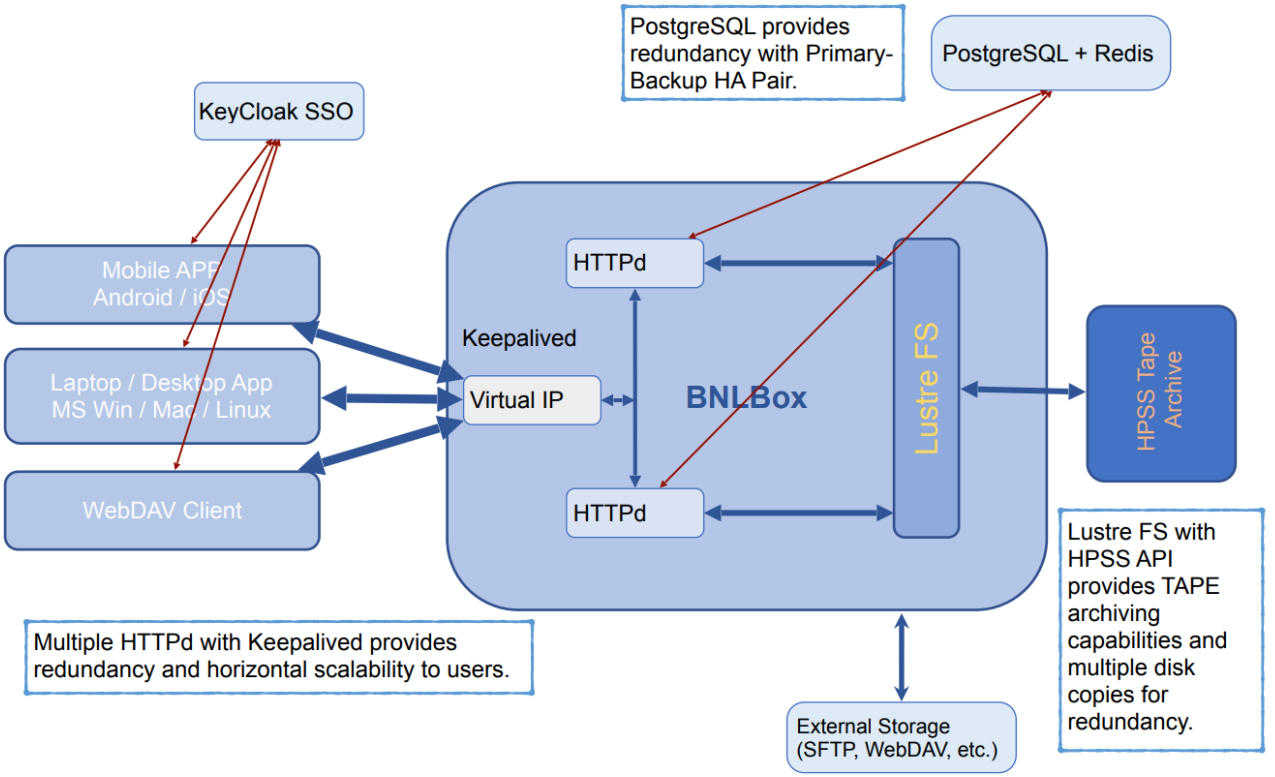
- BNL Box is an Enterprise File Sync and Share Service (EFSS) integrated into the SDCC to provide flexible, easyto-use, unified cloud storage for all BNL scientific users.
  - Based on Nextcloud version 17. 100% free open source software.
  - PostgreSQL backend with a HA Primary/Backup pair for redundancy, and fast NVMe flash storage for performance.
  - Lustre file system, with multiple copies for redundancy, an HPSS API for archival storage and TSM backup. Currently 1PB of usable storage
  - Apache server front-end pair using KeepAlived for redundancy and loadbalancing.
  - 40 Gbps WAN-to-storage network capacity.
  - Authentication via KeyCloak OIDC integrated with SDCC LDAP and BNL AD



https://bnlbox.sdcc.bnl.gov

BNLBox Tech Talk by Ofer Rind https://indico.bnl.gov/event/6868/

BNLBox Network Diagram



Conclusion

- In ATLAS and Belle II, it is designed to distribute the data across the network to many sites.
- The successful operation of the distributed data system requires the effort from the local storage administrators, the regional network engineers and operation teams of various central data management system like FTS, RUCIO, etc…
  - Access to larger volume of the storage than the single site can accommodate.
- The demand of the data might out-pace the funding.
  - QoS might be the way to use the limited disk resource more efficiently.
- New Sync-and-share type storage might have a role in the scientific computing.