dCache service for HEP at BNL

Carlos Fernando Gamboa

Scientific Computing and Data Facilities
Brookhaven National Laboratory

Workshop on Future Organization and Evolution of Storage, 12/11/2025

@BrookhavenLab

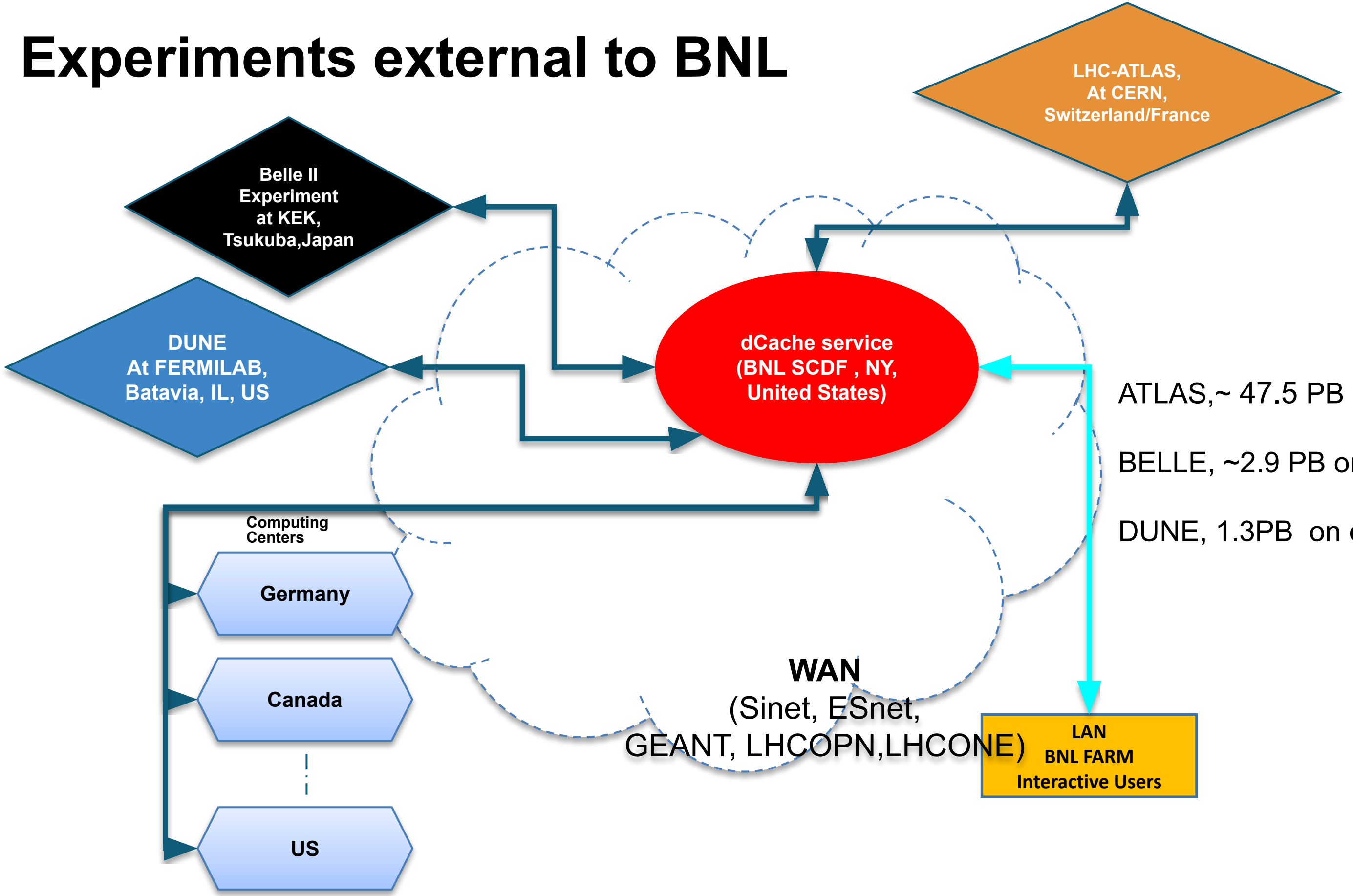
Goal

To provide a general overview of BNL's dCache service infrastructure for High Energy Physics (HEP)

Storage services at BNL Scientific Computing and Data Facilities (SCDF)

- BNL SDCF supports different storage services for a variety of Scientific Communities (SC) like NSLSII, Nuclear and High Energy Physics
- Diverse storage technologies are used to support the communities: dCache, Lustre and GPFS, please see past HEPIX BNL site report for specifics
- dCache services for LHC-ATLAS, BELLE2, DUNE store and manage ~192PB (27% DISK) of data
  - **Scientific Community data is produced outside BNL:**
    - CERN (Switzerland,France),
    - KEK (Tsukuba,Japan),
    - Fermilab(IL,US)

Experiments external to BNL



ATLAS,~ 47.5 PB on disk, 137.7 PB on tape

BELLE, ~2.9 PB on disk, 2.75 PB on tape

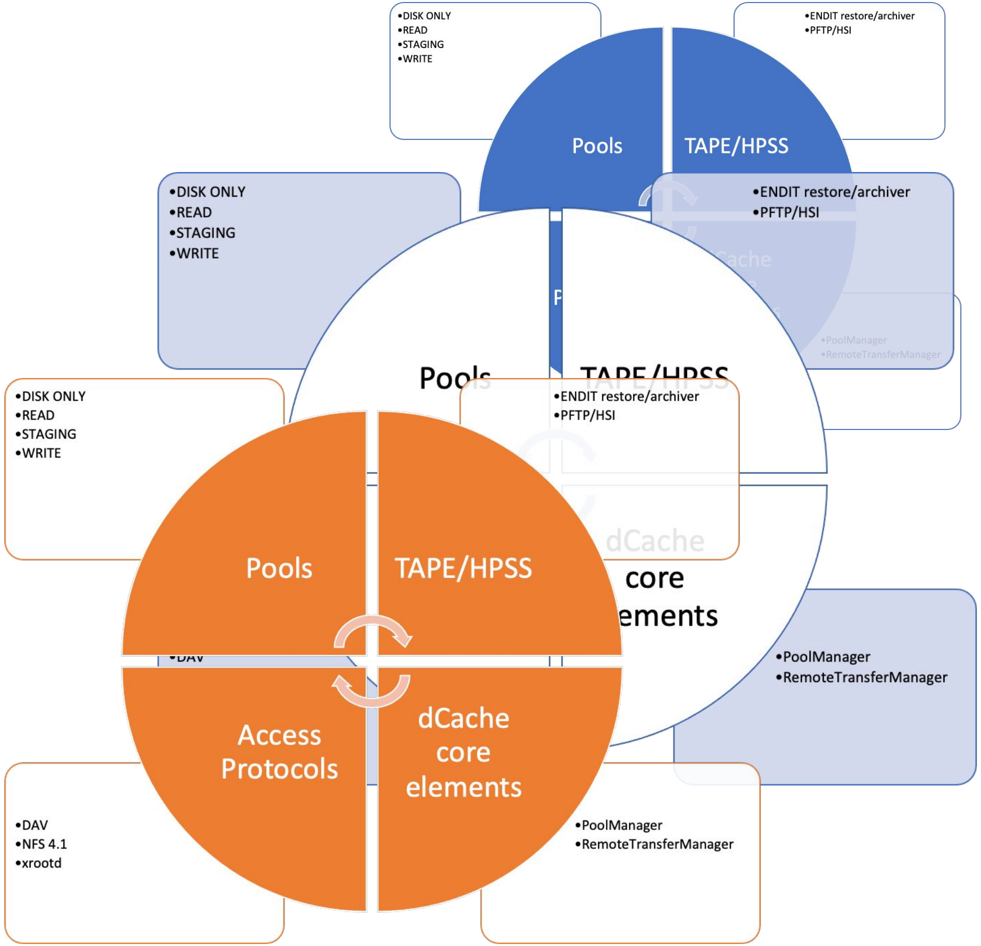
DUNE, 1.3PB on disk

dCache.org

distributed storage for scientific data

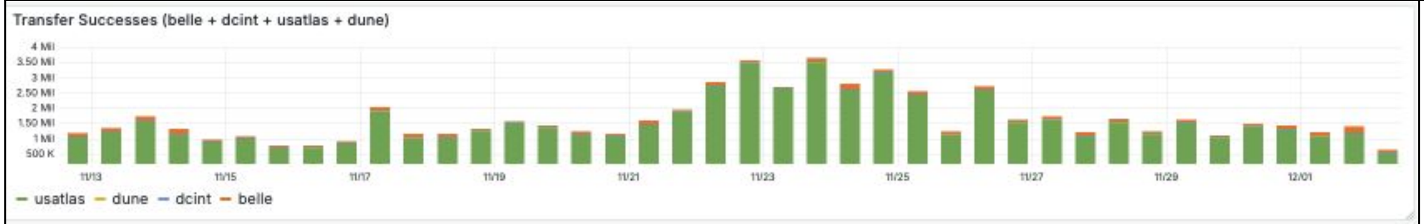
dCache instances are isolated per SC

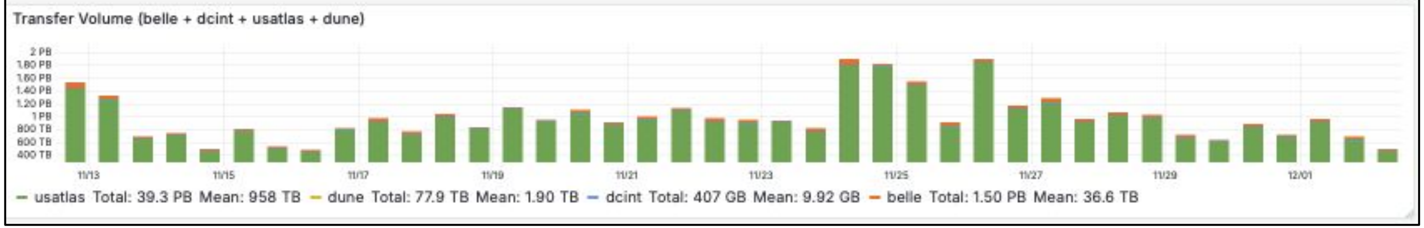
- SC diverge in their requirements
- Procurement and resource control
- Infrastructure supported on physical and virtual Machines



Storage technology flexible to support SC individual requirements

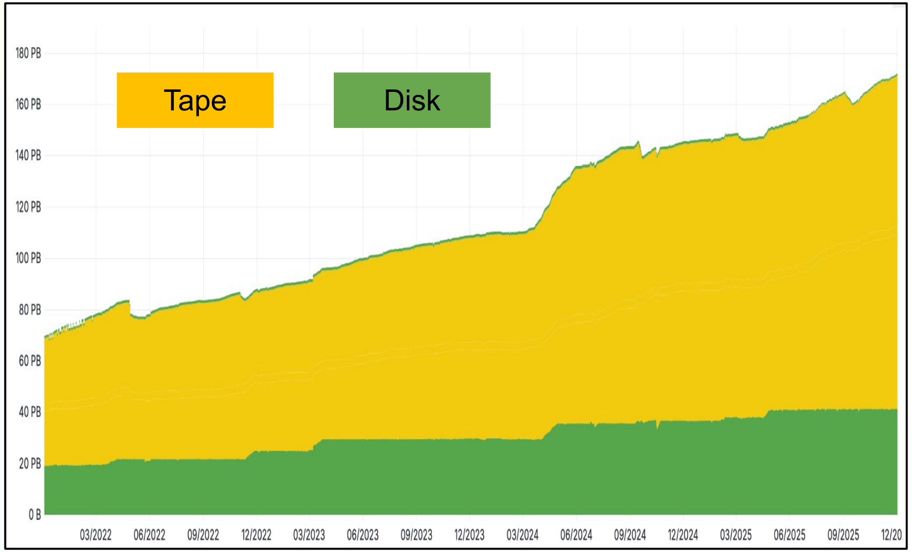
20-Day usage of dCache-based service
(source storage accounting)





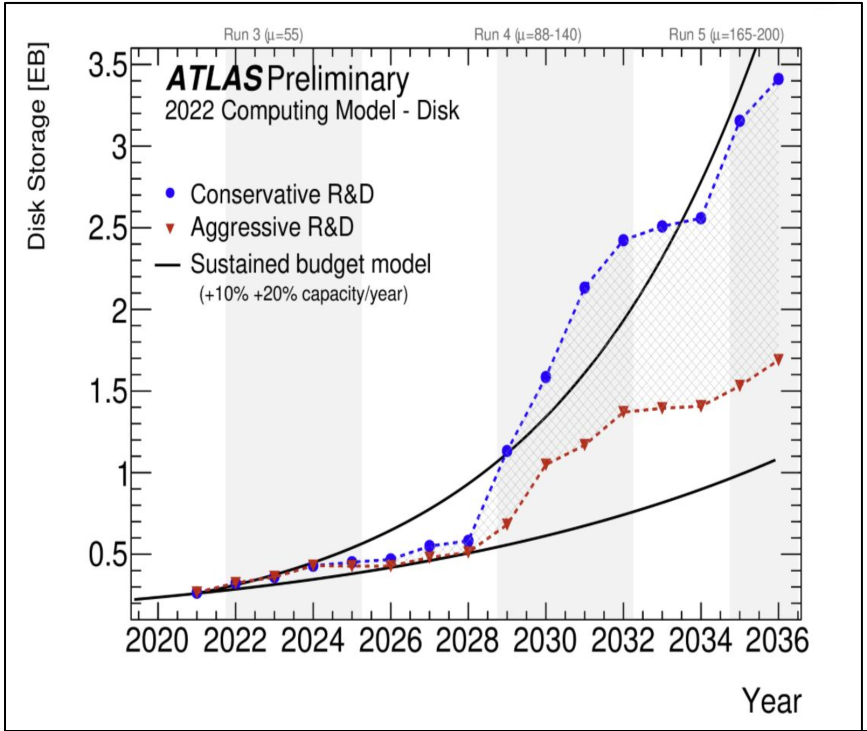
ATLAS SC community driving the storage usage compared to other HEP SC supported at BNL

Evolution of Atlas SC storage



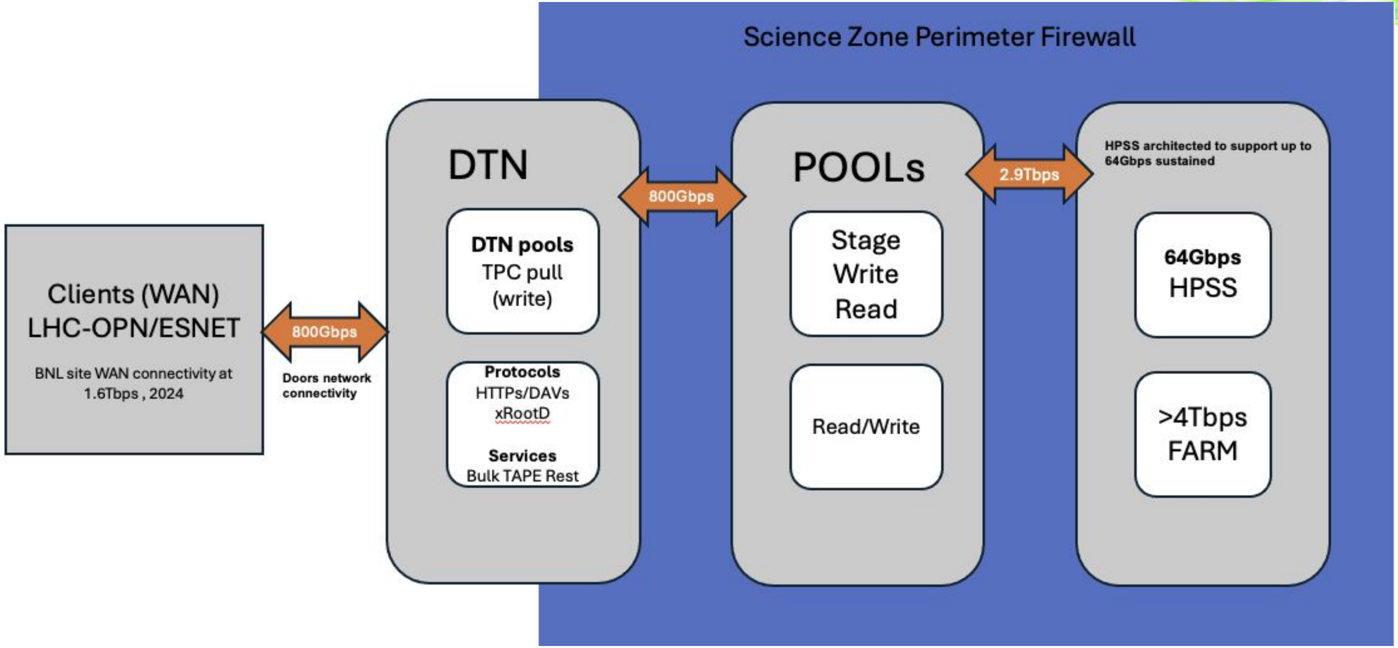
Brookhaven National Laboratory (BNL) provides over 150 PB of storage to ATLAS as part of its responsibilities as a U.S. Tier-1 center for the ATLAS collaboration

Within ATLAS distributed storage system,
BNL contributes approximately 10.35% of all ATLAS Disk capacity and 23% of all ATLAS Tape capacity



The main challenge coming is HL-LHC and with the simple model of 3 to 4 order of magnitude increase in 10 years from now

dCache general layout (ATLAS)



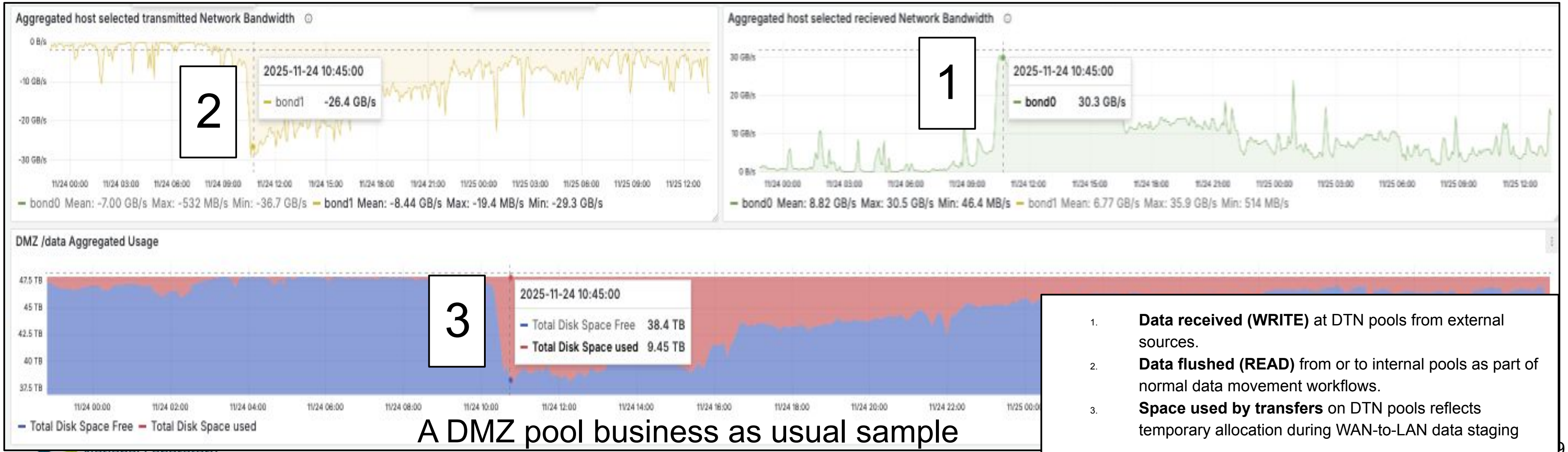
Comply with BNL cybersecurity policy disaggregation among external and internal resource accessibility

Reference deployment to be used as building block for other SC

Brookhaven National Laboratory

Improving throughput for WAN access

- WAN simple READ and WRITE request are proxied via DOORs
- DMZ pools on the DTNs are used for HTTP-TPC WRITE traffic to BNL.
  - Data hopping introduces additional WRITE/READ IOPS load on the NVMe storage, affecting overall throughput
  - Have observed NVMe max throughput of ~2.5 GB/s with mixed READ/WRITE load (implies 320 Gbps current limit in this scenario)

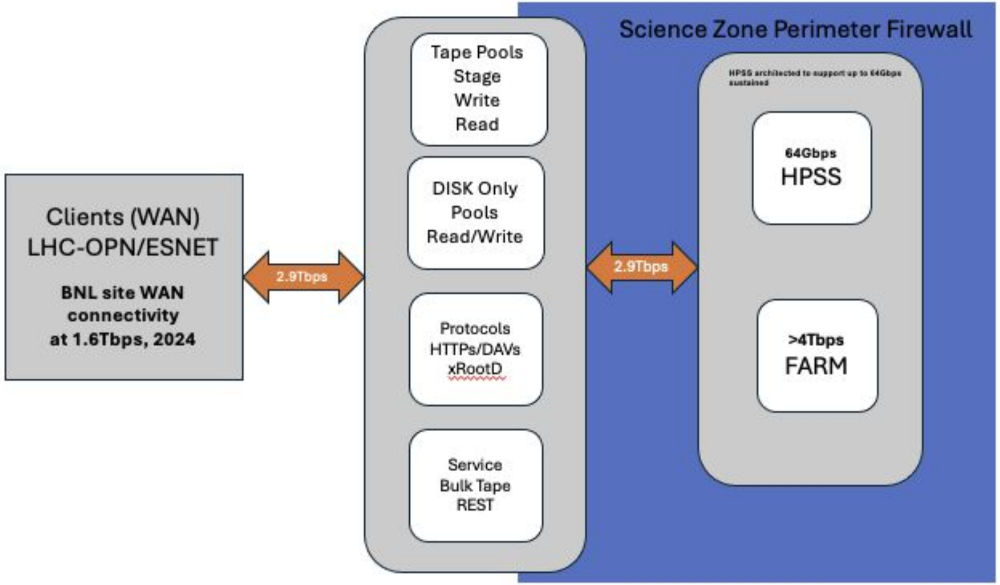


A DMZ pool business as usual sample

1. **Data received (WRITE)** at DTN pools from external sources.
2. **Data flushed (READ)** from or to internal pools as part of normal data movement workflows.
3. **Space used by transfers** on DTN pools reflects temporary allocation during WAN-to-LAN data staging

Improving throughput for WAN access (cont)

Scenarios of improvements discussed at [Workshop on ATLAS Computing and Software Activities at BNL]



Dual-Home Pool Transition

- Transitioning pools to **dual-homed WAN + LAN connectivity**
- **Phased deployment** aligned with hardware refresh cycles
- **10 / 58 pool servers** currently dual-home
- Deployment strategy will prioritized **TPC WRITE throughput** until all pools are fully dual-home.

Door proxies and DMZ pools no longer required → begin descaling door components and requirements
Once fully deployed, **dCache WAN connectivity will exceed BNL site-wide WAN capacity**

Brookhaven National Laboratory

dCache internal pools on ZFS

ZFS was adopted as the file system to host dCache data on pools.

- For ATLAS, 40 PB of data were migrated to ZFS in a rolling fashion with no downtime in 2024

ZFS layout

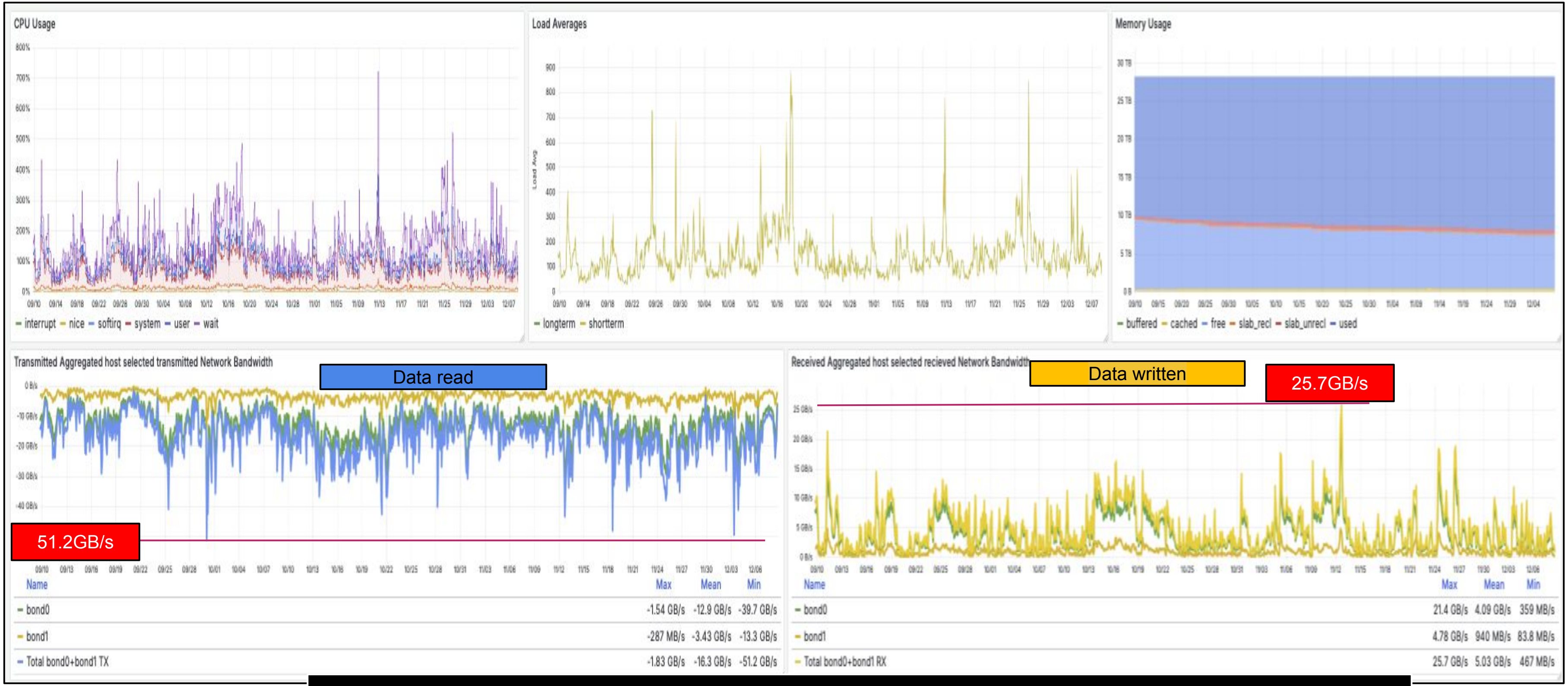
- Western Digital Ultrastar 102 disk / JBODs
- Mostly configured as draid2:8d:3s:102c (one set of 10 servers are still raidz2 7x14+4 spare)
- Recordsize=1M except on older servers that lack triple mirror SSD special device metadata where its 512k
- special_small_blocks set to 64k on pools with special device
- 5% capacity reserved for full pool COW operations

Smooth operations since transition to ZFS

ZFS dRAID storage backend, normalized across the systems in 2024 migrating for MDRAID

Brookhaven National Laboratory

Pools: a business as usual sample



Brookhaven National Laboratory

Smooth operations since transition to ZFS

Infrastructure Supporting dCache Services

Core Cells: redundant deployment to ensure high availability.

Core databases: in a primary standby replication mode using postgresql database

Doors (16): Equipped with 2×25 Gbps internal links and 2×25 Gbps external (WAN) links, providing dual-path connectivity for both LAN and WAN data flows.

DMZ Pools (NVMe) (16): Handle TPC WRITE operations from external sites. Upon completion, data is automatically flushed into the internal pool infrastructure.

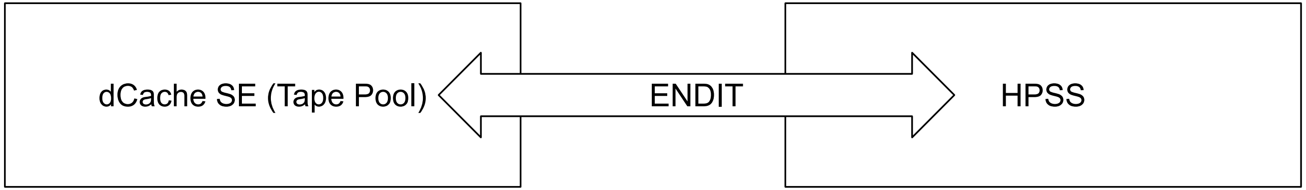
Internal Pools (48): Connected through 2×25 Gbps internal network links, supporting high-throughput data access and bulk internal workflows.

Dual-Home Pools (10): Configured with 2×25 Gbps internal links and 2×25 Gbps external (WAN) links, enabling direct WAN and LAN data movement as part of the dual-home deployment model.

dCache instance	Number of VMs+Physical Hardware(PH)	dCache Version	Notes
ATLAS	74(99%PH)	9.2.35	40 PB of data were migrated in 2024 during the transition from MDRAID to ZFS. 1 file replica
BELLE2	13(92%PH)	9.2.35	Upgrades are subject to a yearly schedule, primarily taking advantage of detector downtime. 1 file replica
DUNE	12(42%PH)	9.2.35	Legacy hardware in a resilient configuration 2 copy/file
Pre-production/Test (AKA dcint)	12(8%PH)	9.2.35	WLCG REST API test endpoint Integrated with ATLAS DDM test infrastructure Dual pool home studies EPiC tests

dCache and HPSS tape interaction via ENDIT

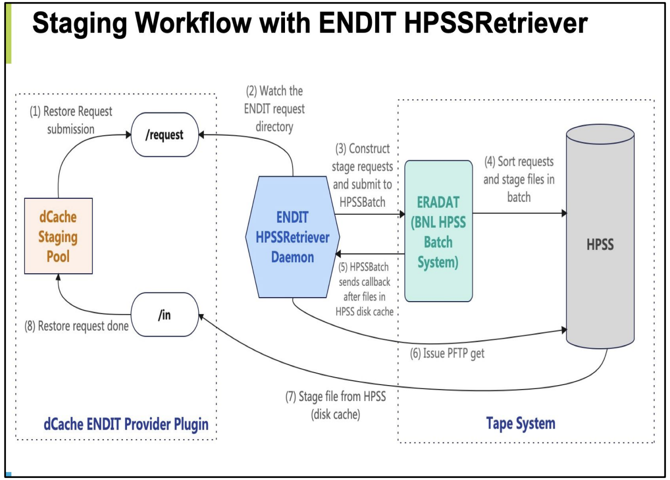
dCache and HPSS systems configured to support 200k+ simultaneous staging requests



Successful adoption of ENDIT retriever permitted the extension of usability for writing interactions to HPSS

- Allowed consolidate legacy software/code for writing to HPSS

Extended overview covered on this talk



Client storage usage

- **Authorization/Authentication:** PKI, token-based authentication, and support for NFS UID/GID
- **Door Proxies:** Used for non-TPC writes requiring WAN access
  - LAN Data Flow: Data is redirected to/from pools; this remains the default operational behavior
- **NFSv4 Access:** NFSv4 doors are accessible within the BNL LAN
  - Used by Belle II calibration farm
  - Interactive Nodes: Also configured to access dCache via user accounts

Monitoring and observability enhancements

- **Observability Migration:** Transitioning from a Grafana/Graphite stack to a Grafana /Victoria Metrics architecture

- **Unified Metrics Stream:** All dCache SEs report normalized storage-metric events via Kafka, enabling consistent and scalable ingestion of operational data

- **Search & Analytics:** Events are indexed and analyzed using OpenSearch, providing powerful search capabilities and long-term retention of operational metrics

Brookhaven National Laboratory

Future work

dCache Software Roadmap

- 11.2 Golden Release targeted for Summer 2026
  - JAVA 11 -> JAVA 17
- Initial release expected January 2026
- Planned enhancements include:
  - Improved staging workflow performance and reliability
  - Enhanced pool space-management mechanisms
  - Advancements in quota-management capabilities
- User Access Policy
  - Currently under review, including the potential transition to a quota-based management model

Infrastructure & Hardware Evolution

- Continued pool hardware refresh cycle
- Expanded deployment of dual-home pool configurations to optimize WAN/LAN throughput and streamline data workflows
- Improved OS patching deployment processes
  - KPATCH currently under evaluation for live-kernel-patching capabilities

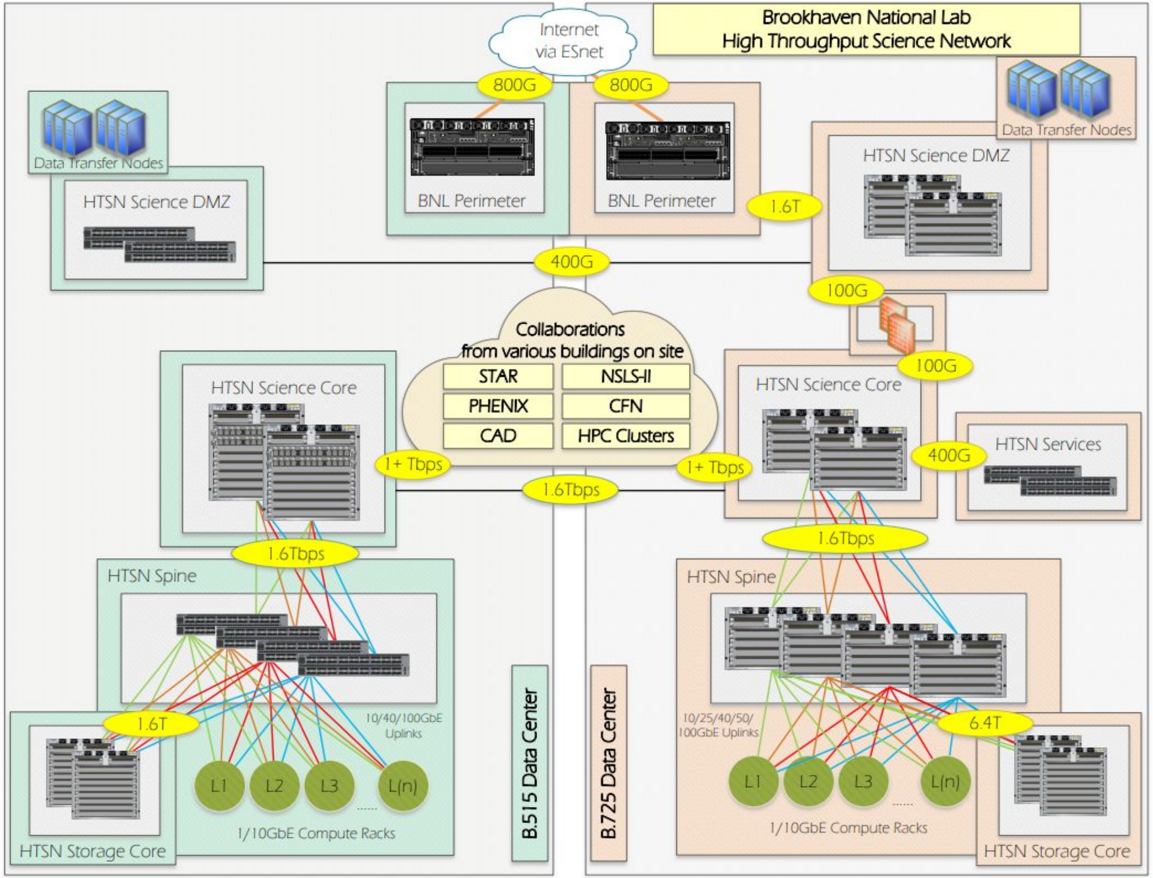
In Summary

An overview of dCache-based storage services was provided, primarily to contextualize the technology and its deployment at BNL

dCache is effectively supporting a variety of SC at BNL

Backup slides

Network



- WAN
  - 2x400 Gbps LHCOPN
  - 2x400 Gbps LHCONE
- LAN
  - 1.6+ Tbps
- Firewall isolating Science Core from DMZ and WAN