# US ATLAS T1 Storage Report : Status & Plan

*Vincent Garonne, Carlos Gamboa, Qiulan Huang, Doug Benjamin, Eric Lancon, Zhenping Liu, Shigeki Misawa, etc.*

Scientific Data and Computing Center (SDCC)
Brookhaven National Laboratory

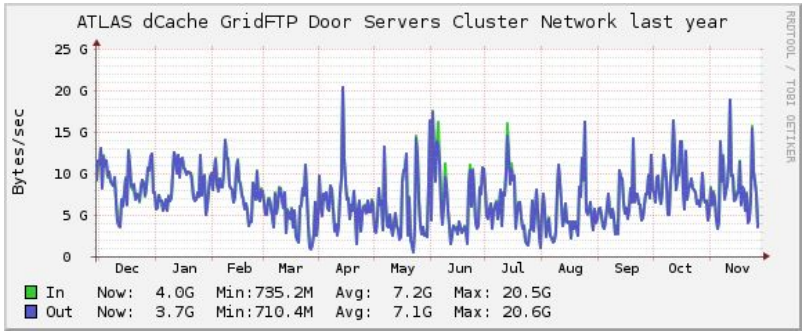
@BrookhavenLab

US ATLAS Computing Facilities Meeting 11/30/2022

# US ATLAS T1 Storage @ BNL

- 150 PB,140M files
  - DISK: 47.1 PB, 100M files
  - TAPE:  90 PB, 40M files
- 9 PB/month to the WAN
- > 99% WLCG availability & reliability report
- Large and complex
  - ~1,000 pools, 55 hosts
  - ~20 doors

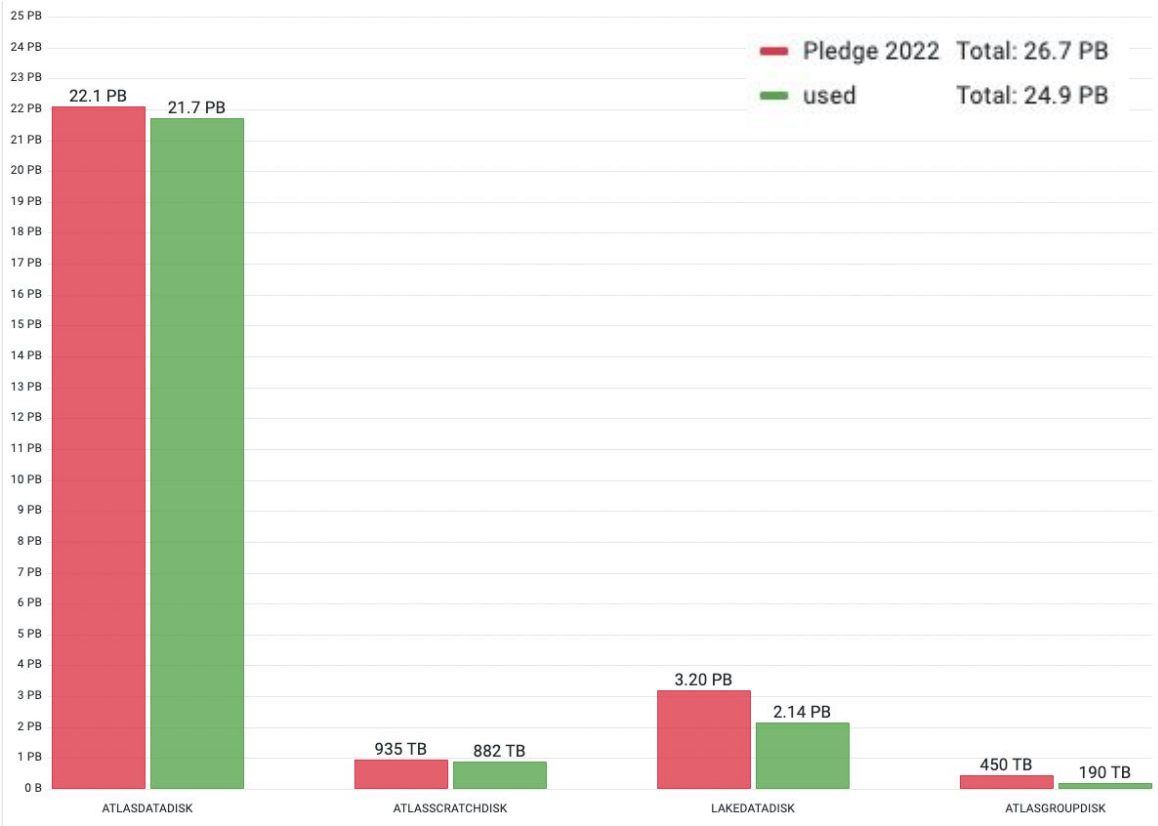
→ Constant increase and spiky load




Brookhaven National Laboratory

# Pledge 2022: DISK

- BNL must provide 23% of of Tier-1 resources to ATLAS
- 26.7 PB deployed on 04-01-2022




Brookhaven National Laboratory

# Pledge 2022 → 2023

**Major Pledge 2023 milestone: 23% of 136 PB, 31.3 PB, on 04-01-2023**

Activities to cover 2023 Pledge:

- ✓ Data deduplication of DATADISK — 2022Q4
  - Internal reconfiguration ongoing
  - No plan and funding for duplication for FY23 and beyond
- ✓ Decommissioning of LAKEDATADISK — 2022Q4
- ❑ HW Retirement (~14PB) — 2023Q1
  - ~9PB can be kept alive to help ATLAS due to EU T1 risk
- ❑ FY23 HW purchase commissioning

Data deduplication gain: + 4.6 PB

| | Pledge 2022 | Now | Pledge 2023 |
|---|---|---|---|
| ATLASDATADISK | 22.1 PB * | 25.3 PB ** | 30 PB ** |
| LAKEDATADISK | 3.20 PB | | |
| ATLASSCRATCHDISK | 935 TB | 935 TB | 935 TB |
| ATLASGROUPDISK | 450 TB | 450 TB | 450 TB |
| **Total** | **26.7 PB** | **26.7 PB** | **31.3 PB** |

* Duplicated

** Non duplicated

Brookhaven National Laboratory

# Other US ATLAS T1s Internal Milestones

2023Q1: Storage central services HW refreshment deployment and old Hardware retirement

2023Q2: Storage Service consolidation (failover, replication)

2023Q2: New TAPE HPSS (8.3.20) integration with storage software

2023Q3: (More) Storage analytics for BNL storage. C.f. LDRD Qiulan's talk

2023Q1/Q2/Q3/Q4: dCache upgrade to latest version of the dCache golden release serie (8.*,9.*, 10.*)

# dCache Upgrade 7*. → 8.*

USATLAS upgrade is scheduled for Dec 19. Testing and pre-productions instance upgraded to 8.2.4 to identify issues:

- ENDIT archiver/retriever https://github.com/dCache/dcache/issues/6868 [Fixed]
- Cleaner cell https://github.com/dCache/dcache/issues/6879 [Blocker]
- IPV6/IPV4 issue: XROOT on proxy mode (Prefers IPV6 to proxy transfers to pools) https://github.com/dCache/dcache/issues/687 [Blocker for Xrootd proxy mode]

Xrootd on proxy mode successfully tested for pools with DUAL stack

- ATLAS production some pools are only supporting IPV4
  - Possibility to consolidate pools to dual stack prior migration

HSI client testing (hpss83_u20)

- Tested on 7.2.16 (production) and 8.2.4 dCache versions
- Functional tests done includes file copy, retrieval and deletion
- Integration with HPSS testbed in preparation of HPSS upgrade

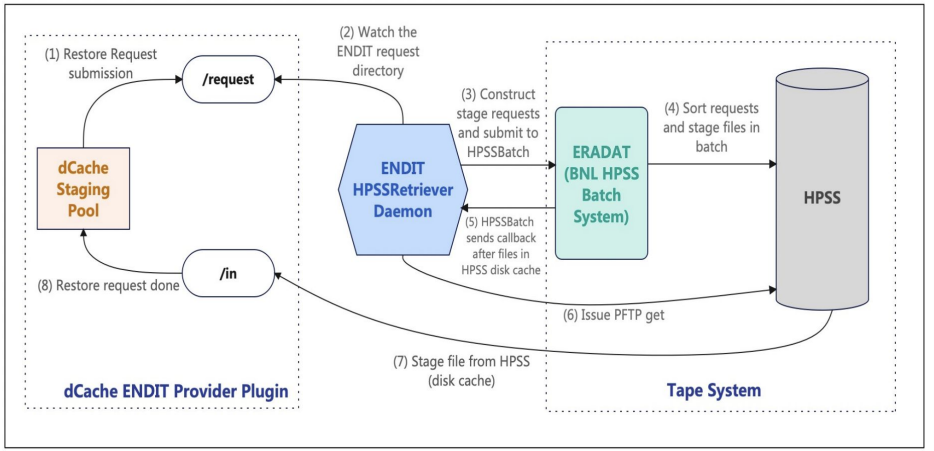
→ Regular interactions with dCache team and other T1s, e.g., Weekly dCache & T1s meeting

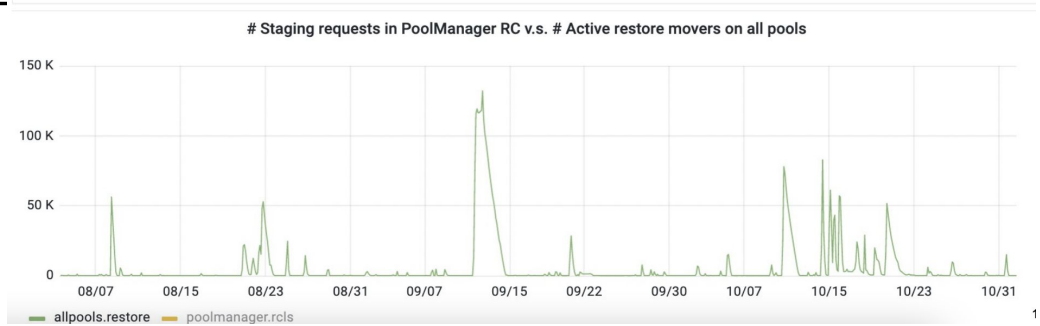
# A new efficient interface between dCache and HPSS @ BNL

Performance issues with previous dCache tape / HSM interface

New Interface is now fully in production for read and write:

- dCache ENDIT (initially developed by NDGF) with changes to meet BNL's needs
- New software developed by BNL for interacting between ENDIT and HPSS Batch System



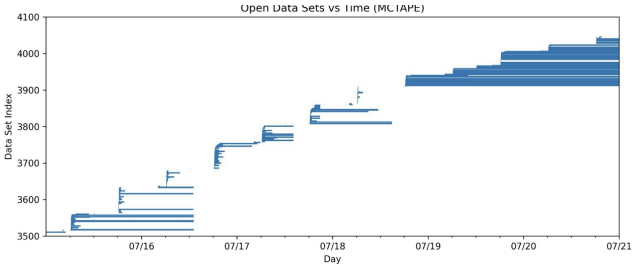




Cf. [Jane's talk at HEPIX]

Brookhaven National Laboratory

# Future TAPE

**MCTAPE Dataset Transfer Concurrency**

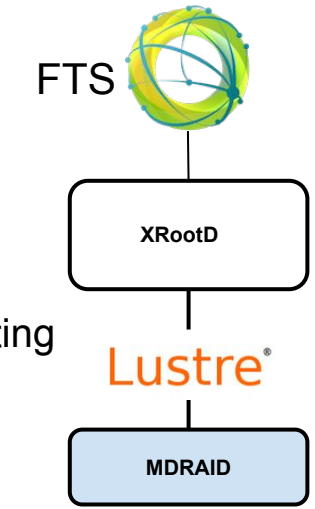



- Tape usage analysis at BNL presented by Shigeki Misawa, at the ATLAS DDM meeting https://indico.cern.ch/event/1220522/
- White paper on optimizing the use of tape resources through the expanded use of metadata for ATLAS data is being prepared for the ATLAS Technology Interchange Meeting (TIM) next week
  - https://indico.cern.ch/event/1212249/
- New Tape REST API — required dCache 8.*
  - Joining the TAPE REST API testbed common effort with WLCG / DOMA-BDT — 2023Q1 (“Leading-edge T1s will have production instances in 2023 available for tests”)

# Future and progresses on storage for ATLAS

- We are gaining expertise and operational experience on alternate/complementary storage "classes", e.g., Object stores, CEPH, etc.
- Synergy with WLCG, CERN, FNAL, etc
  - i.e., Cross FNAL/BNL storage team meetings

- Lustre instances@SDCC:
  - Production services (NSLS II, SPHENIX, BNLBOX, etc.)
  - ATLAS:
    - RSE BNLHPC_DATADISK: Xrootd Standalone server + Lustre
    - dCache and Lustre (5-3 PB) testbed for functional, performance testing and comparison wrt ATLAS use cases
      - Cf. Previous talk



# Performance testing : XROOTD+Lustre and dCache w/localdisk

DISCLAIMER
Not final numbers !

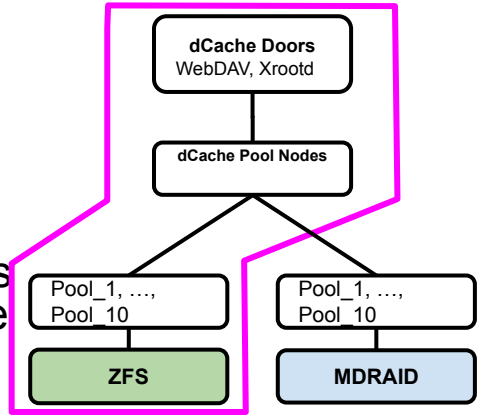
| Davs TPC | XRootd w/ Lustre (No affection via FTS tuning) | dCache w/local disk (Before FTS tuning) | dCache w/ local disk (After FTS tuning) |
|---|---|---|---|
| traffic per door | Effective traffic is ~3.2GB/s per door | 1.6 GB/s per door | +2.0GB/s per door |
| CPU Usage | <10% per door | ~50% per door | ~50% per door |
| Checksum | Compute checksum needs reading back, which causes extra traffic | Checksum computed on the fly | Checksum computed on the fly |
| Success rate | >98.5% | >97.4% | >97.4% |

**Initially a factor 2**

Tuning FTS parameters, i.e, number of active requests, helped

- When the limit value>150, the performance of dcache w/local disk does not increase but decrease, 25% gained by tuning(1.6GB/s→+2.0GB/s, results under active discussion with dCache people:
  - Some possible limiting factor for Davs TPC (remote transfer manager),seems not there for direct Davs access
  - The dcache team has reviewed the configuration and suggested more tests with the latest dCache
- XRootd w/Lustre behaves better to support more active FTS requests. The performance will not decrease until the limit value >600

# OpenZFS




ZFS has efficient data compression, snapshots, copy-on-write, clones and automatic self healing Cf. Previous talk. However, it comes at the cost of higher overhead:

- 1.21x (mdraid raid6) vs 1.27x (ZFS 14 X 7) capacity overhead

New pools (FY21/22 HW purchase) are configured with ZFS

Converting old MDRAID pools to ZFS is a time-consuming and expensive procedure

→ We will start gradually with few hosts to assess the gain - 2023Q1

Expected gains:

- Better data integrity
- Reduced manual intervention: No manual intervention on reboot required and less complicated / manual steps for disk replacement
- We do keep track of incidents and interventions (e.g., offline pools, HW actions, MTBF, etc)

# Other Noteworthy Events

GRIDFTP@BNL closed on 2022/8

- Several issues found with gsiftp dependant components
- Error tracking, debugging and reporting issue to FTS team, e.g., GGUS-158503

Issues with Standalone Xrootd server — Xrootd WAN access@BNL

- IPV4/IPV6 dual stack & zone issues
- “Timeout issues from PanDa/pilot jobs when accessing data”
  - Intermittent stalled gfal clients behaviour observed (gfal, rucio, xrootd, dCache):
  - Report to xrood team, contribution to Rucio 5989

1-2% of T0 export failures from CERN to certain T1s (e.g., TRIUMF, BNL, ..)

  - Collaborate with CERN- IT EOS ops team

# Recent talks / Contributions

- A Scalable and Efficient Staging System between Dcache and HPSS at BNL, Zhenping Liu et al. HEPIX Autumn 2022, https://indico.cern.ch/event/1200682/
- Tape usage analysis at BNL, Shigeki Misawa, ATLAS DDM meeting https://indico.cern.ch/event/1220522/
- Multi-experiment Storage service at BNL, Carlos Fernando Gamboa et al. HEPIX Autumn 2022, https://indico.cern.ch/event/1200682/contributions/5094113/
- Exploring Future Storage Options for ATLAS at the BNL/SDCC facility, Qiulan Huang, Vincent Garonne et al. CHEP 2023, https://docs.google.com/document/d/1Zn-OpI7k49_mrL_kFuOdqsM1uRQwiw6s9rMwEiqYJ9I/edit

# Job Opportunities at Brookhaven National Lab / SDCC

Please come and discuss with us!