Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

US ATLAS Tier 1

BNL Site Report

Presented by Shigeki Misawa

Pre-GDB meeting
November 5, 2023

@BrookhavenLab

Tape Storage at the US ATLAS Tier 1

- Based on HPSS v8.3*
  - ATLAS shares the HPSS system with other programs
    - Hardware sharing limited to the HPSS core server
- Accessed through dedicated dCache instance for ATLAS
  - Three Rucio Storage Elements (RSE) associated with ATLAS tape storage
  - Single HPSS Class of Service backend for the three RSEs
  - dCache ENDIT Provider plugin used to interface dCache to HPSS
    - Locally written ENDIT to HPSS script connects plugin to HPSS
  - ERADAT (“HPSS batch”) bulk recall optimizer used to enhance file staging performance

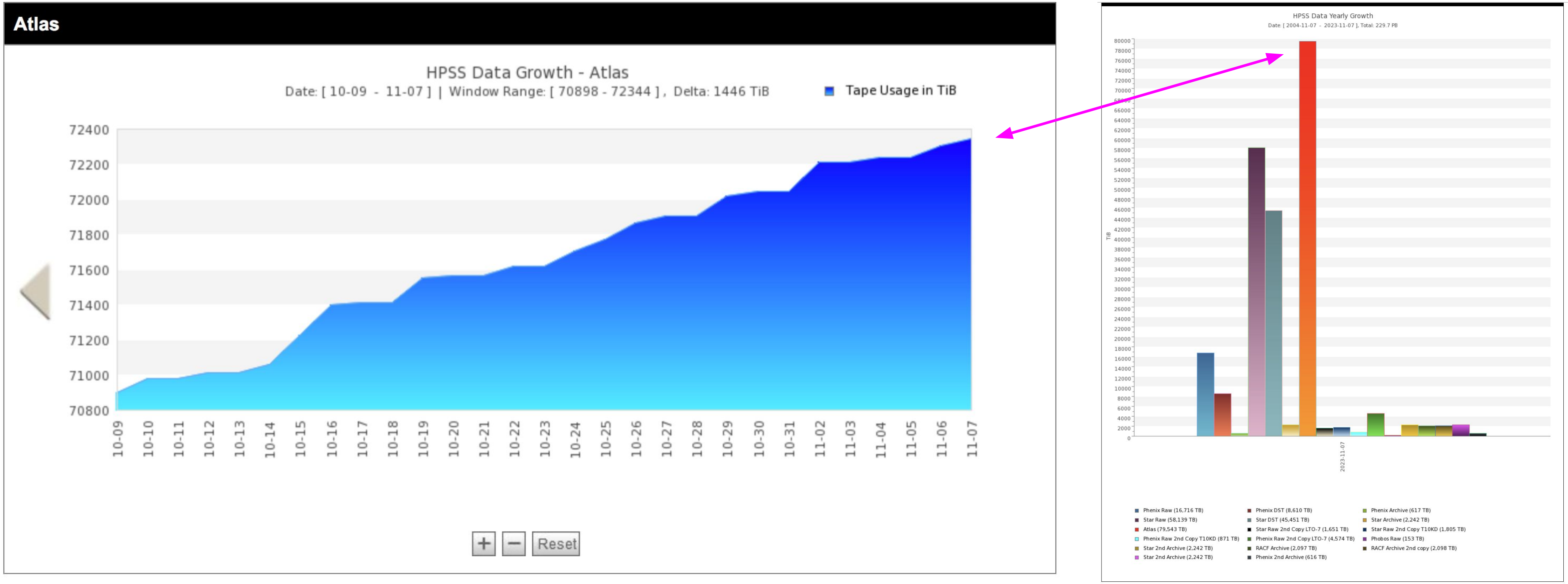
* High Performance Storage System - hpss-collaboration.org

Hardware Configuration

- Designed for 8 GB/sec sustained writes to tape
  - Supports 16 GB/sec burst writes into HPSS
- Tape resources dedicate to ATLAS for new incoming data
  - Four disk movers
  - 1 PB (petabyte) disk cache (22 LUNs)
  - Two IBM TS-4500 libraries
  - 64 LTO-8 tape drives
  - LTO-8 media
  - 2 x 25 GbE mover connectivity
- Future additions
  - Third library in 2025 if necessary
  - Move to LTO-10 when available

* High Performance Storage System - hpss-collaboration.org

ATLAS Data in HPSS



* High Performance Storage System - hpss-collaboration.org

Data Center Migration

- Tape operations split between data centers
- Bldg 515 - Original “legacy” data center
  - Hosts data primarily from before run 3
  - 3 ATLAS Oracle SL-8500 libraries
  - ~11K LTO-7, 6K LTO-6 tapes with ATLAS data
- Bldg 725 - New, energy efficient and highly available data center
  - Hosts data from Run 3
  - HPSS core server
  - ATLAS HPSS disk cache
  - ATLAS IBM TS-4500 libraries
  - LTO-8 tapes containing new data
  - ATLAS LTO-8 tape drives

* High Performance Storage System - hpss-collaboration.org

HPSS in the new Data Center

* High Performance Storage System - hpss-collaboration.org

Tape Optimization

- “Migration by Directory”
  - Lowest level directory in HPSS namespace = ATLAS dataset
  - Files in lowest level directory written to tape as a group
  - Expected to enhance dataset recall performance by increasing data locality on tape
- Longer time between migration of data on disk to tape
  - Increases # file written to tape together from a dataset
- ATLAS now reading data on tape written with migration by directory enabled
  - Anecdotal evidence suggests higher effective bandwidths reading data off these tapes

* High Performance Storage System - hpss-collaboration.org

Metadata & Tape Optimization

- Information about data that might be used to “optimize” tape access
- Information may not be relevant or actionable at all tape sites
- Metadata examples
    - dataset is quantum of data retrieval (This is true for virtually all ATLAS access)
    - dataset is never read (“cold”), dataset may be read (“tepid”), dataset will be read (“warm”)
    - If dataset A is read, datasets B and C will also be read (“correlated” datasets)
    - Size of dataset

* High Performance Storage System - hpss-collaboration.org

Work in Progress

- Changes to bulk file retrieval (ERADAT/ENDIT)
  - Rewriting engineering ERADAT
  - Testing quaid/lori “native” bulk retrieval mechanism in HPSS v10.2 as the backend engine
- Motivation
  - Ability to handle more queued file stage requests
  - Utilize Recommended Access Order (RAO) technology in newer tape drives
  - Reduce dependency on locally developed code
  - Improved fair share scheduling

* High Performance Storage System - hpss-collaboration.org

In the Queue

- Possible finer grained use of "file families"
  - Requires major re-write to code connecting dCache to HPSS for writes
  - Algorithm(s) used to map files to file families - **TBD**
  - Cost/benefit analysis
    - Can potentially increase dataset staging performance
    - But can negatively impact write performance and staging performance if the mapping is poorly chosen
    - Also increases the complexity of dCache to HPSS glue code
    - Algorithms that change frequently over time or are too numerous are not sustainable

* High Performance Storage System - hpss-collaboration.org

In the Queue

- File/dataset metadata for tape optimization
  - Mechanism for changing tape behavior based metadata needed
  - Different metadata may suggest different strategies
  - Most likely coupled to re-write of dCache to HPSS glue code
  - Possible utilization of additional dCache features
- Searchable file metadata
  - Possible investigation into off line metadata store for fast, no-impact metadata searches
  - Utlizes features in HPSS version 10
- Evaluation of new HPSS features, including updated batch code, in preparation for next software upgrade cycle

* High Performance Storage System - hpss-collaboration.org