News, status, and future of the HSM/Tape storage interface at BNL

Zhenping (Jane) Liu, Vincent Garonne, Douglas Benjamin, Tim Chou, Carlos Gamboa, Christopher Hollowell, Qiulan Huang, Shigeki Misawa, Jason Smith, Yingzi (Iris) Wu

Scientific Data and Computing Center, Brookhaven National Laboratory

Abstract:

To address the staging performance bottlenecks discovered during the WLCG tape challenges, BNL took inspiration from ENDIT (Efficient Northern dCache Interface to TSM), which was initially created by the Nordic Data Grid Facility. BNL then designed and developed its own site-specific ENDIT system customized for its needs, enabling efficient dCache interfacing with HPSS tape storage.

The ENDIT HPSSRetriever component offers a scalable staging solution and significantly improves overall tape staging performance by eliminating high load issues on dCache and increasing the number of simultaneous staging requests to BNL's HPSS Batch (ERADAT) system. The ENDIT HPSSArchiver component provides a more flexible and controllable solution for flushing requests due to its asynchronous nature. A further benefit of ENDIT is the adjustable control of the simultaneous reads and writes to HPSS, preventing overburdening a pool host and stressing the HPSS gateway. In addition, the changes allow for the addition of new features such as monitoring and analytics capabilities, which we have already begun implementing and will continue to do so in the future, including metadata support and smart writing.

The ENDIT HPSSRetriever component has been running stably for over a year and demonstrated noticeable improvements in performance since being deployed to ATLAS dCache production. It has alleviated heavy loads on stage hosts, and the system has handled 140K staging requests without any issues. It reached up to 7 GBytes per second during the 2022 WLCG Tape challenge. In the future, we intend to pursue more aggressive staging testing through the WLCG Tape challenge. The BNL ENDIT HPSSArchiver component has been successfully deployed to ATLAS dCache for approximately a year and is operating efficiently.

We plan to extend this development as a standard solution to other experiments like BELLE-II in the future. Additionally, ENDIT will serve as a centerpiece for upcoming changes and improvements in future tape usage of the experiments, such as smart writing, user-defined metadata propagation to HPSS for different writing/reading strategies, and more.

Problem

dCache at BNL has been in production for almost two decades. For years, dCache relied on the default driver included with the dCache software to interface with HPSS tape storage systems. However, the synchronous nature of this approach and the high resource demands resulting from periodic script invocations significantly limited scalability. During the WLCG tape challenges, bottlenecks in dCache staging were identified. Performance issues on staging servers arose due to high load and out-of-memory problems, causing staging servers to become nonfunctional under heavy restore request levels of 120K or more.

Components of BNL ENDIT System

- ENDIT provider (plugin developed by NDGF)
  - Customized minor changes for BNL
- ENDIT Daemons
  - HPSSRetriever daemon:
    - Submits stage requests to HPSS Batch
    - Retrieves files from HPSS using PFTP when Batch is complete
  - HPSSArchiver daemon:
    - Flushes precious files from dCache tape write pools into HPSS
- Two Cron jobs on HPSSBATCH
  - Checks the list of new ENDIT staging requests; submits only unique and non-existing requests to BATCH
  - Sends callbacks to the requesting pool(s) when the file processing in batch is complete

dCache ENDIT Provider plugin

- Adapted from NDGF, with changes to accommodate BNL's requirements
- Mechanisms:
  - Use predefined file actions in a set of specified directories
    - File actions include creating, modifying, and deleting specific files in designated directories
  - WatchService: Monitors file events triggered by these file actions
- Java's and Google Guava's concurrent frameworks
  - Handles file events asynchronously
- The ENDIT directory must reside on the same file system as the pool's data directory
- Several directories under the pool directory are recognized by the ENDIT provider:
  - ./request: ENDIT provider sends stage/migration requests here
  - ./in: ENDIT provider checks for completed staged files in this directory
  - ./out: ENDIT provider places files to be written to tape here

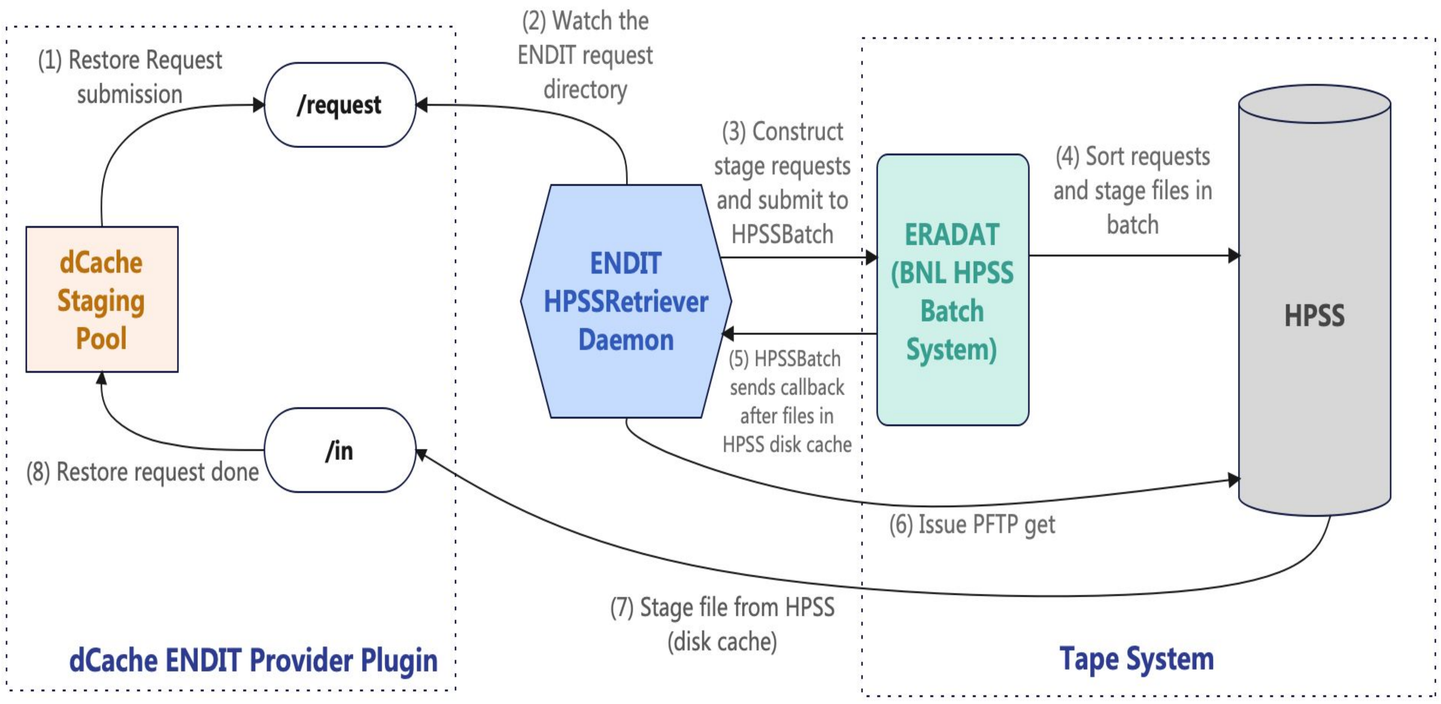
ENDIT HPSSArchiver Daemon

HPSSArchiver is a daemon to flush dCache tape area files into HPSS.

HPSSArchiver Daemon – Workflow

1. New flush requests are created under the ENDIT request directory by the ENDIT Provider.
2. The daemon monitors the ENDIT request directory and detects incoming new flush requests.
3. The daemon invokes PFTP to flush files to HPSS.
4. A file is flushed to HPSS successfully.
5. ENDIT Provider detects the completion of a flushing process and marks the end of the flush request.

Staging Workflow with ENDIT HPSSRetriever



ENDIT HPSSRetriever Daemon

HPSSRetriever is a daemon to stage files from HPSS to dCache pools

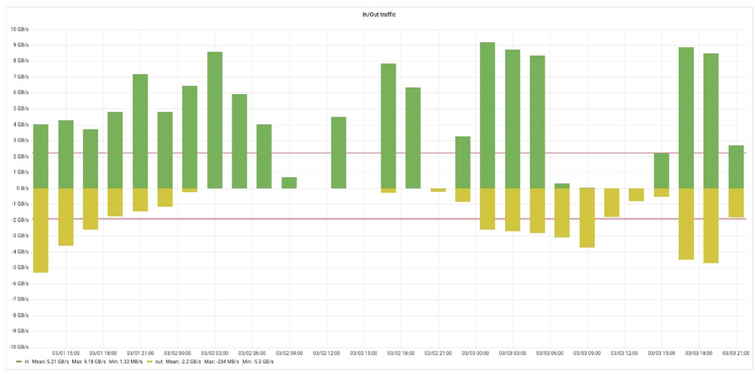
HPSSRetriever Daemon – Workflow

1. New stage requests are created under the ENDIT request directory by the ENDIT Provider.
2. The daemon monitors the ENDIT request directory and detects incoming new stage requests.
3. The daemon constructs stage requests and submits them to the ERADAT (HPSSBatch) request queue.
4. ERADAT (HPSSBatch) sorts requests and stages files in batches.
5. ERADAT (HPSSBatch) sends a callback after a file is in the HPSS disk cache.
6. The daemon checks the callback content. If it's good, the daemon invokes PFTP to retrieve the file from HPSS.
7. The file is staged from HPSS (disk cache) to the ENDIT ./in directory.
8. ENDIT Provider detects the data file, moves it from ./in to the pool data directory, and marks the end of the stage request.

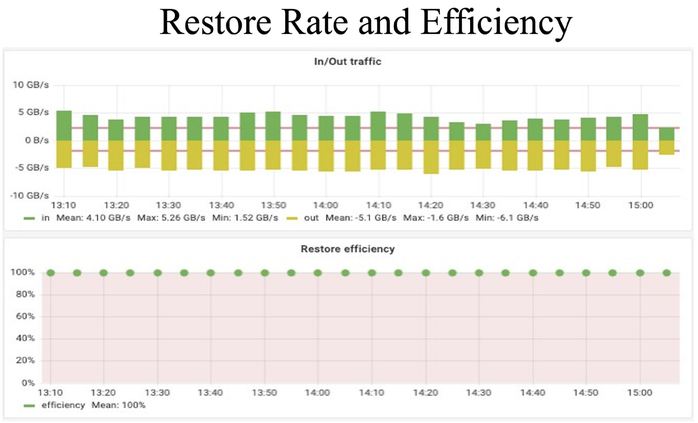
Benefits of ENDIT HPSSRetriever

- Performance improvements on pool hosts
  - Eliminates polling on pool hosts, resulting in minimal load even with a high number of requests
  - Enables dCache to handle a large number of active staging requests simultaneously
- Provides flexible control over the maximum concurrent PFTP threads on each pool
- Prevents duplicated requests in HPSS Batch
- Reduces stress on core dCache components (Pnfmanager, PoolManager,..) due to non-polling nature

Restore Performance



Restore Rate and Efficiency



Acknowledgements

We express our gratitude to the Nordic Data Grid Facility and the ENDIT project team (project website: https://github.com/neicnordic/endit) for providing the foundation for our customized ENDIT project. Their valuable insights and expertise have significantly contributed to the success of this work. Additionally, we extend our gratitude to TRIUMF for sharing their development and operational experiences from their ENDIT project. Their guidance has been invaluable in helping us bypass many obstacles, enabling us to rapidly achieve production quality.

U.S. DEPARTMENT OF ENERGY

Brookhaven National Laboratory
www.bnl.gov