Aric Werner
HPC Systems Engineer

Ben Kirk
HPC Consultant

Joey Mendoza
HPC Systems Engineer

HSM at NCAR

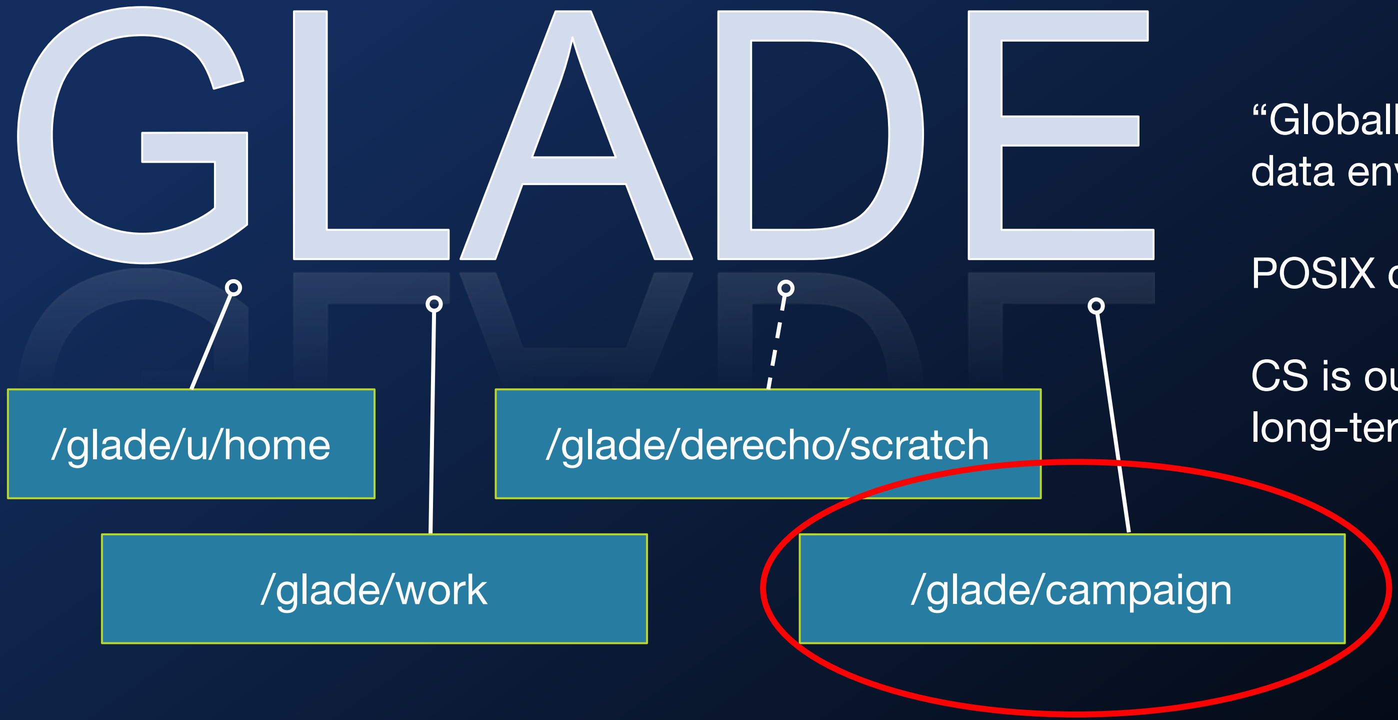
Increasing effective storage capacity with hierarchical storage management (HSM) for NCAR's Campaign Storage

November 22, 2024

Photo by NASA on Unsplash

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

Storage systems overview at NCAR



GLADE

/glade/u/home

/glade/derecho/scratch

/glade/work

/glade/campaign

“Globally-accessible data environment”

POSIX disk-based storage

CS is our largest long-term storage area

Storage systems overview at NCAR

Quasar Globus endpoint

Quasar

S3-compatible API

Stratus

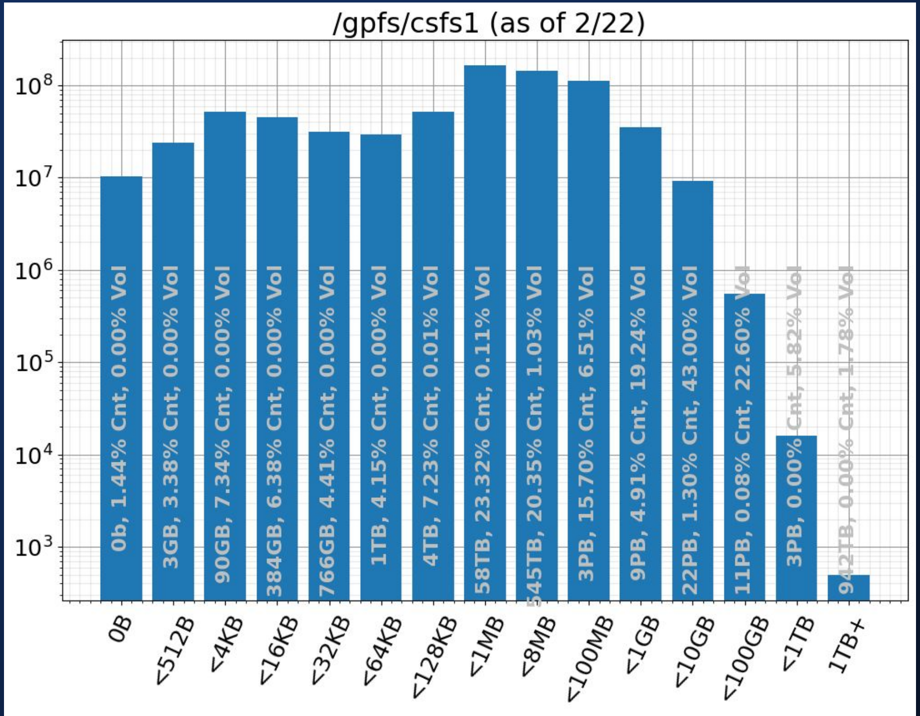
Tape storage with disk cache

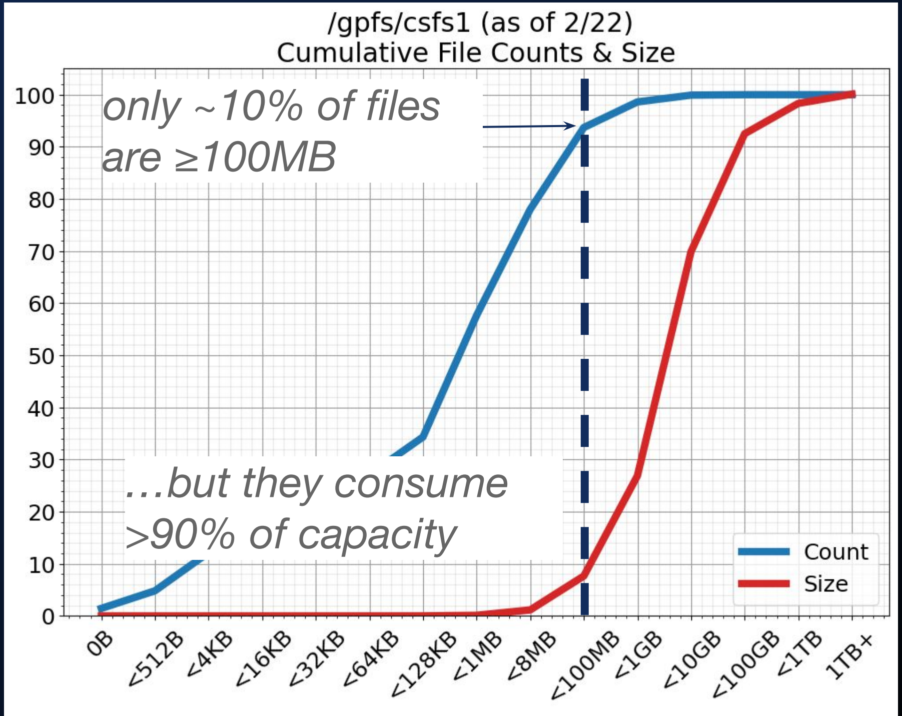
Globus access only

Object storage

Campaign Storage Filesystem Usage Analysis - are files large enough to move to tape?

>90% of capacity used by <10% files



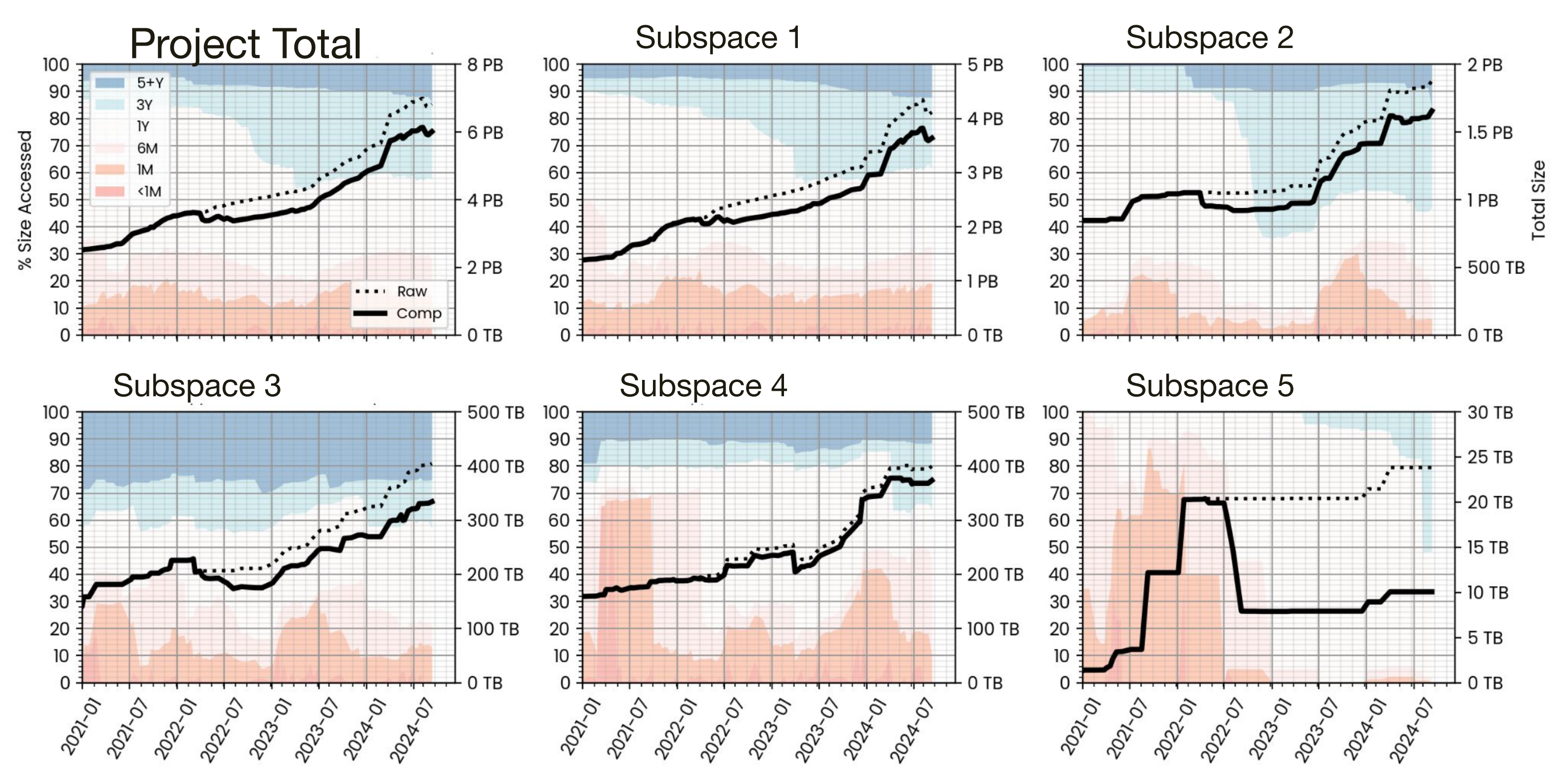


only ~10% of files are ≥100MB

…but they consume >90% of capacity

Campaign Storage Filesystem Usage Analysis - capacity limitations and prevalence of “cold” data

≳50% data (by volume) often not accessed in 3 or more years



Current capacity limitations and need for cold storage

Campaign Storage (CS)

- ~120 PB disk
- IBM Storage Scale

Quasar

- 30 PB tape library accessed via Globus
- 2 PB disk cache
- IBM Storage Archive

- High demand for more CS space with familiar POSIX access, but relatively low adoption of our tape library via Globus so far
- Budget won't allow for significant increases of CS capacity, so what can be done? Plenty of seemingly good candidate data for tape - set up automatic tape tiering for infrequently accessed data?
- We performed an analysis of CS to see how much "cold" data was being accessed and if the library could handle the associated workload → **it likely couldn't**

Current capacity limitations and need for cold storage

Campaign Storage (CS)

- ~120 PB disk
- IBM Storage Scale

Quasar

- 30 PB tape library accessed via Globus
- 2 PB disk cache
- IBM Storage Archive

- There was a need to develop:
  - a way to handle reading from tape in a **constrained** way
  - plus a more user-friendly, ideally POSIX interface to tape

Features of our solution

Campaign Storage (disk tier)

Large, inactive files migrated to tape

Files can be recalled from tape

Quasar (tape tier)

- glade_hsm script for users to move data to tape with minimal workflow disruption
  - Files get moved to a “COLD_STORAGE” directory and marked immutable to be migrated/stubbed later by migration policy
  - Prevents problems with user recalls (imagine a seemingly innocuous `grep *` on a huge directory) - and user access is not going to be linear
- Tape-sorted recall
  - Batch all recalls requests such that tape mounts are reduced, instead of letting arbitrary tape mounts and remounts occur
- Single vs. dual tape copies
  - Users can make this choice
- Frees up disk quota
  - This provides an incentive for users to move files to tape

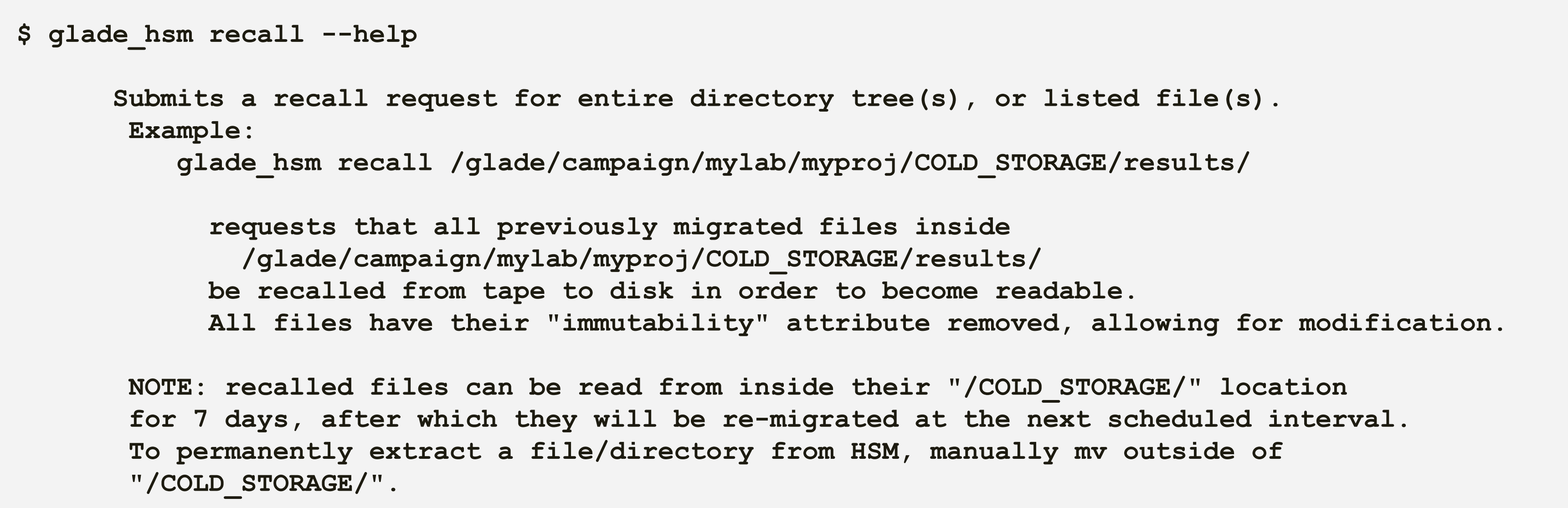
Current status of HSM at NCAR

- We've deployed a user-facing tool **`glade_hsm`** to several early users through a pilot program that has been running for about a year managing ~3PB.
  - Targeting "write once, read maybe" data sets kept for archival publication or similar reasons

```
Usage: glade_hsm <migrate|recall|status> <subcommand options> <dirname(s)>

$ glade_hsm migrate --help

    migrate:
       Relocates files/directories in preparation for HSM migration.
       [...]
     Example:

          glade_hsm migrate <--single-copy> /glade/campaign/mylab/myproj/results/

            will relocate
              /glade/campaign/mylab/myproj/results/ ->
              /glade/campaign/mylab/myproj/COLD_STORAGE/results/
```

Current status of HSM at NCAR

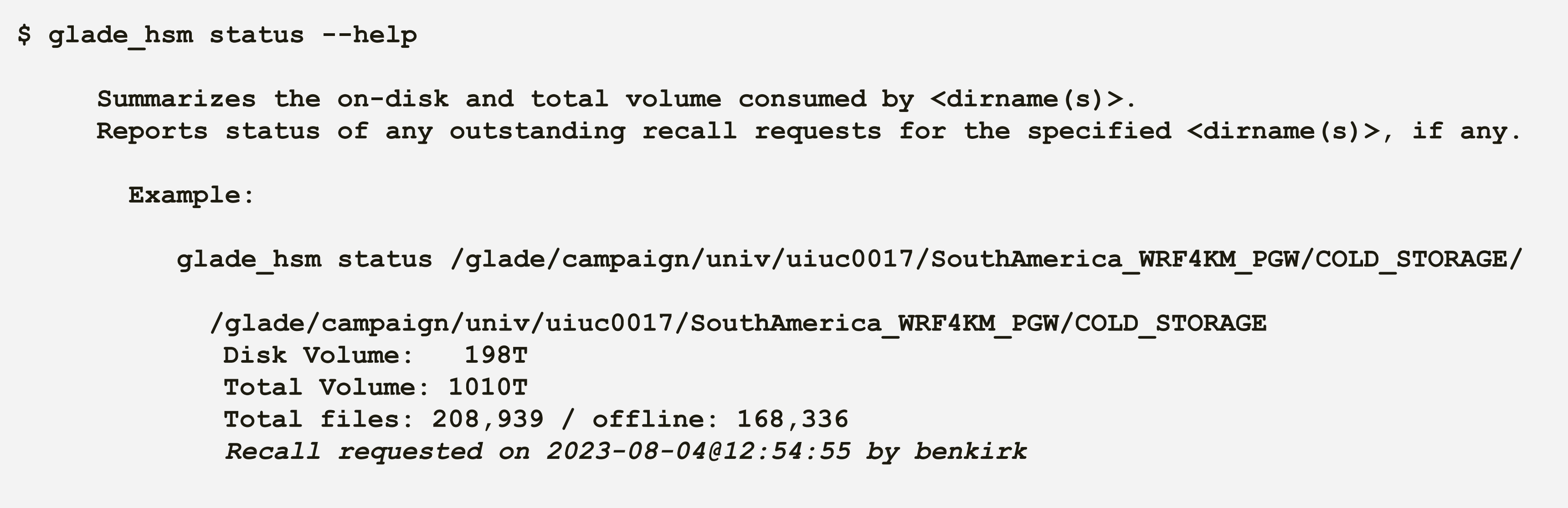
- We’ve deployed a user-facing tool **`glade_hsm`** to several early users through a pilot program that has been running for about a year managing ~3PB.
  - Targeting “write once, read maybe” data sets kept for archival publication or similar reasons



```
$ glade_hsm recall --help

      Submits a recall request for entire directory tree(s), or listed file(s).
       Example:
          glade_hsm recall /glade/campaign/mylab/myproj/COLD_STORAGE/results/

            requests that all previously migrated files inside
              /glade/campaign/mylab/myproj/COLD_STORAGE/results/
            be recalled from tape to disk in order to become readable.
            All files have their "immutability" attribute removed, allowing for modification.

       NOTE: recalled files can be read from inside their "/COLD_STORAGE/" location
       for 7 days, after which they will be re-migrated at the next scheduled interval.
       To permanently extract a file/directory from HSM, manually mv outside of
       "/COLD_STORAGE/".
```

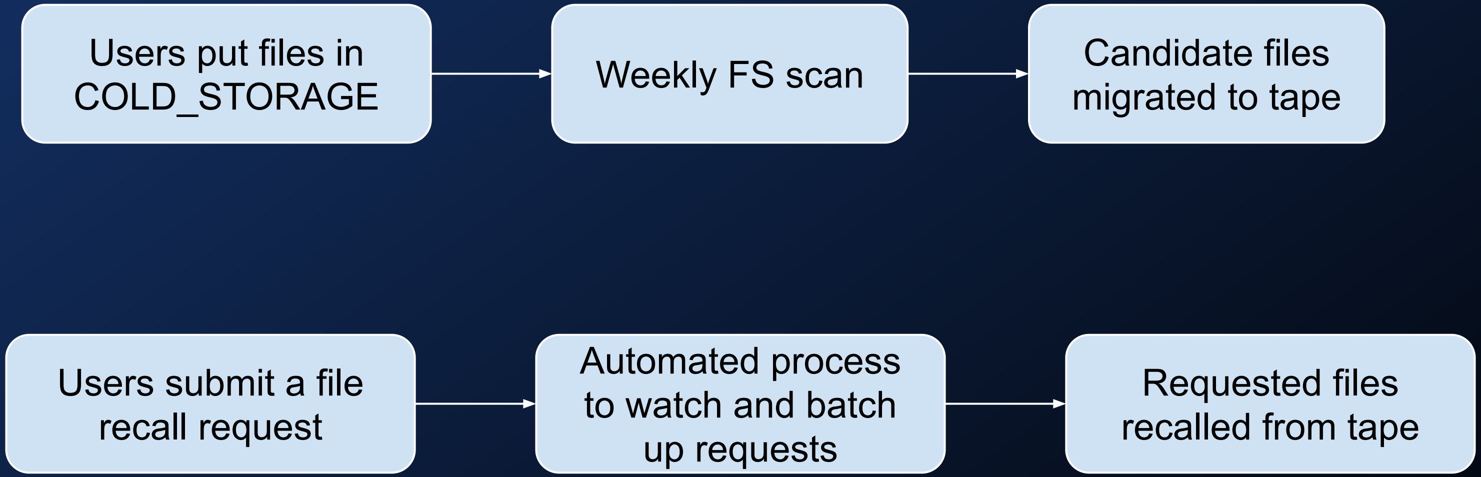
Current status of HSM at NCAR

- We've deployed a user-facing tool **`glade_hsm`** to several early users through a pilot program that has been running for about a year managing ~3PB.
  - Targeting "write once, read maybe" data sets kept for archival publication or similar reasons



```
$ glade_hsm status --help

     Summarizes the on-disk and total volume consumed by <dirname(s)>.
     Reports status of any outstanding recall requests for the specified <dirname(s)>, if any.

       Example:

          glade_hsm status /glade/campaign/univ/uiuc0017/SouthAmerica_WRF4KM_PGW/COLD_STORAGE/

            /glade/campaign/univ/uiuc0017/SouthAmerica_WRF4KM_PGW/COLD_STORAGE
             Disk Volume:   198T
             Total Volume: 1010T
             Total files: 208,939 / offline: 168,336
             Recall requested on 2023-08-04@12:54:55 by benkirk
```

Target workflow



Users put files in COLD_STORAGE

Weekly FS scan

Candidate files migrated to tape

Users submit a file recall request

Automated process to watch and batch up requests

Requested files recalled from tape

policy defines and external lists/pools

```
define(MB_ALLOCATED,(KB_ALLOCATED/1024.0))
define(GB_ALLOCATED,(KB_ALLOCATED/1048576.0))
define( DISPLAY_NULL, [COALESCE($1,'_NULL_')])
define( is_10m_old, (CURRENT_TIMESTAMP - MODIFICATION_TIME > INTERVAL '10' MINUTES))
define( is_7d_recalled, ((CURRENT_TIMESTAMP - EXPIRATION_TIME > INTERVAL '7' DAYS) OR EXPIRATION_TIME IS NULL))
define( is_immutable, MISC_ATTRIBUTES LIKE '%X%')
define( is_premigrated, (MISC_ATTRIBUTES LIKE '%M%' AND MISC_ATTRIBUTES NOT LIKE '%V%') )
define( is_migrated, (MISC_ATTRIBUTES LIKE '%V%'))
define( is_resident, (NOT MISC_ATTRIBUTES LIKE '%M%'))
define( file_exclude_list,
 (
  PATH_NAME LIKE '/gpfs/csfs1/.ltfsee/%'
  OR PATH_NAME LIKE '/gpfs/csfs1/.SpaceMan/%'
  OR PATH_NAME LIKE '/gpfs/csfs1/.mmSharedTmpDir/%'
  OR PATH_NAME LIKE '/gpfs/csfs1/.snapshots/%'
  OR NAME = 'dsmerror.log'
))

RULE EXTERNAL LIST 'set_immutable' EXEC '/var/mmfs/etc/set_immutable.pl'
RULE EXTERNAL LIST 'unset_immutable' EXEC '/var/mmfs/etc/set_immutable.pl' OPTS 'unset'
RULE EXTERNAL LIST 'list_immutable' EXEC '/var/mmfs/etc/cat.list'
RULE EXTERNAL LIST 'list_migrate' EXEC '/var/mmfs/etc/cat.list'
RULE EXTERNAL LIST 'list_all' EXEC '/var/mmfs/etc/cat.list'

RULE EXTERNAL POOL 'Campaign_HSM_Single'
  EXEC '/opt/ibm/ltfsee/bin/eeadm'
  OPTS '-p hsm_a@qplib01_test'
  SIZE(50971520)
RULE EXTERNAL POOL 'Campaign_HSM_Double'
  EXEC '/opt/ibm/ltfsee/bin/eeadm'
  OPTS '-p hsm_a@qplib01_test,hsm_b@qplib01_test'
  SIZE(50971520)
```

policy - migrate rules

```
RULE 'Set_All_Immutable' LIST 'set_immutable'
  WEIGHT(DIRECTORY_HASH)
  WHERE (PATH_NAME LIKE '/gpfs/csfs1/%/COLD_STORAGE/%' OR PATH_NAME LIKE '/gpfs/csfs1/%/COLD_STORAGE_SINGLE_COPY/%')
  AND NOT file_exclude_list
  AND is_10m_old
  AND is_7d_recalled
  AND NOT is_immutable


RULE 'Migrate_To_Tape_Double' MIGRATE
  WEIGHT(DIRECTORY_HASH)
  TO POOL 'Campaign_HSM_Double'
  WHERE PATH_NAME LIKE '/gpfs/csfs1/%/COLD_STORAGE/%'
  AND NOT file_exclude_list
  AND is_10m_old
  AND is_7d_recalled
  AND NOT is_migrated
  AND MB_ALLOCATED > 100
  AND GB_ALLOCATED < 19000
  AND NLINK == 1

RULE 'Migrate_To_Tape_Single' MIGRATE
  WEIGHT(DIRECTORY_HASH)
  TO POOL 'Campaign_HSM_Single'
  WHERE PATH_NAME LIKE '/gpfs/csfs1/%/COLD_STORAGE_SINGLE_COPY/%'
  AND NOT file_exclude_list
  AND is_10m_old
  AND is_7d_recalled
  AND NOT is_migrated
  AND MB_ALLOCATED > 100
  AND GB_ALLOCATED < 19000
  AND NLINK == 1
```

Future work and plans

- How do we handle file deletion from tape?
- Users need to know what data is cold, or they will be unable to select files to move to tape - assistance with cold data identification?

- The pilot program has been technically successful, but we haven't seen massive adoption yet
- As quota pressure increases due to lack of disk capacity growth, we hope to gain more adopters to relieve quota pressure
- Tape is typically lower-cost (with higher initial cost) than hard disk, and tape storage density is increasing faster than that of disk, so we don't see moving away from tape in the near future

Open source or other alternatives

- We've run GPFS for a number of years and already had the pieces available
  - IBM Spectrum Archive integrates well, of course
  - Powerful GPFS policy engine that we're taking advantage of
- Lustre has an HSM capability and could be a promising open-source option
  - We haven't looked into this, but it seems like the building blocks are there
  - Our solution could possible be translated to other filesystems
- Other proprietary solutions
  - HPSS
  - HPE DMF

Thank you

Contact information

Aric Werner
<aricw@ucar.edu>

Ben Kirk
<benkirk@ucar.edu>

Joey Mendoza
<jam@ucar.edu>