Data Retention Policy

Disclaimer

The user community and its sponsoring organizations are very diverse with differing requirements for data retention. The Linac Coherent Light Source (LCLS) cannot guarantee indefinite data archival. Users of LCLS are responsible for meeting the data management requirements of their home institutions and funding agencies. Once data have been provided to the user group, the user group is responsible for managing the long-term retention of their data. The ownership of data generated at LCLS is governed by the User Agreement in place between the user group and the facility. Refer to your User Agreement for more details.

Data Retention Practices

LCLS is committed to providing its users with their data in a timely and convenient fashion. Experiment data and metadata collected at LCLS may be stored at and retrieved from the facility for a period of ten years. The data retention practices described in this page are on a best effort basis, they depend on availability of the actual resources and may change at any time if these resources become unavailable. Historically we have been able to deliver on these lifetimes (with the exception of scratch where, twice, we had to delete data newer than 4 months), but we cannot guarantee them: available resources depend on the actual usage rate and available funding, which we do not fully control.

Policy by Folder

Space	Quota	Backup	Lifetime	Comment
xtc	None	Tape archive	4 months	Raw data
usrdaq	None	Tape archive	4 months	Raw data from users' DAQ systems
hdf5	None	Tape archive	4 months	Data translated to HDF5
scratch	None	None	4 months	Temporary data (lifetime not guaranteed)
results	4TB, 10K files	Tape backup	2 years	Analysis results ⭐
calib	None	Tape backup	2 years	Calibration data
User home	28GB	Disk + tape	Indefinite	User code (home in S3DF)
Tape archive	-	-	10 years	Raw data (xtc, hdf5, usrdaq)
Tape backup	-	-	Indefinite	User home, results and calib folder
Disk backup	-	-	Indefinite	Accessible under ~/.zfs/

⭐ *For older experiments this folder is called res instead of results*

Results space

- The *results* space is used to save the final results from the data processing of an experiment.
- The limits are 4TB and no more than 10,000 files. If you need to store many small files put them into a tar or zip file.
- If limits are exceeded the *results* folder will be disabled.
- Don't use results for intermediate output use *scratch* instead.

hdf5 and hdf5/smalldata folder

- the experiment is allowed to create files in hdf5/smalldata
- **the folder MUST contain only hdf5 files translated from xtc**
- non hdf5 files (logs, code, ...) will be deleted
- the hdf5 files are archived after they are 4 weeks old

Notes

- **Please do not store under the scratch folder data that you cannot recreate because this directory is not backed up and the oldest files on scratch may be deleted at any time to make space for data from new experiments.**
- For file cleanup in *scratch/* the **last access time** will be used for removing old files (Updated 2024-04).
  For files in results/ and calib/ the age is determined using **last modification time** of a file (not access time).
  For the xtc and hdf5 files the access time is used (see xtc/hdf5 cleanup).
- The tape archive (xtc, hdf5, usrdaq) and the tape backup (results, home) are fundamentally different:
  - In the **tape archive** the folders are frozen after the end of the experiments and their contents are stored on tape once.
  - In the **tape backup**, the system takes snapshots of the folders as appear at a given time. This implies that files which are deleted from disk are eventually, i.e. after a long enough time, also deleted from tape.
- Files under xtc and hdf5 can be restored from tape using the file manager tab in the electronic logbook. Files under home can be restored by the user following the instructions below. To restore data from results send an email to pcds-datamgt-l@slac.stanford.edu.
- For raw data the cleanup operations will affect all files, i.e. all streams and chunks, which make up one run, rather than individual files.
- After 2 years from the end of an experiment we'll remove the experiment from disk. At that point we'll take a snapshot of the results and calib folders and archive them to tape so that we can, upon request, restore an entire experiment back to disk.

- After 10 years we plan to remove the tapes with the archived raw data from the silos and store them in a safe environment.
- For questions regrading the data retention and data access send your question to: pcds-datamgt-l@slac.stanford.edu.

User Home

- **Please do not store large files under your home, this space is meant for code/scripts, documents, etc, not science data.**
- Users can check the used and available space under their home with a command like:

```
df -h ~<username>
```

Cleanup of xtc and hdf5 data

The raw data, the xtc and hdf5 files in the corresponding experiment folders, are purged from disk now and then. The minimum lifetime for new raw data is *four month* (see table above) and one month for runs that were restored from tape. Notice that runs that exceed the lifetime become eligible for purging but will not be automatically purged from disk.

Purging will remove all files that belong to a run (streams and chunks for xtc files) from disk. A few rules are applied for purging eligible runs:

- Purging is performed only if the free disk space is below a minimum threshold.
- Purging will stop if the free space is above a maximum threshold.
- The least recently accessed runs will be purged first.

The purging thresholds might vary depending on the size of a file system and its usage but typically are 5% and 10% (minimum/maximum threshold). Using these three rules we try to keep runs that are actively analyzed for as long as possible on disk and providing sufficient disk space for the ongoing experiment.

Restore of files

The xtc and hdf files are archived to tape and can be restored to disk in case runs were purged from disk. Restores are requested using the FileManager of the experiments eLog. A basic guide to the UI is described in Managing Files.

The requests are sent to a queue which is monitored by a process that will retrieve the files from tape. The restore time might vary from tens of minutes to many days depending on the amount of data to be restored but also how busy the tape system is. In particular if high throughput experiments are running the restore will take a backseat.

Rationale for Proposed Policy

Please see the children pages Version 1, Version 2 and Version 3 for a description of the evolution and rationale of the LCLS data retention policy.