

Recommendations for Intensity Frontier Experiment Use of DCache

Marc W. Mengel,
Stu Fuess

August 20, 2014

Abstract

DCache is large disk pool being used at Fermilab, which can act as a cache pool for data on tape, or as scratch disk. This document outlines various uses of dcache, and makes recommendations for how these uses should be performed, how permissions should be set, and how DCache file.family tags should be set in the various directories of DCache's /pnfs area. How DCache will map this configuration to DCache pools is beyond the scope of this document, up to the DCache administrators, and subject to change.

1 Conventions

In this document, we will use several conventions. We will give them in terms of the Hypothetical Experiment "hypot". Also, we will discuss general experiment user accounts as "username", so where you see "username" used, expect one per experiment user.

2 Use Cases

This document attempts to cover various use cases for DCache by experiments at Fermilab. Each use case will be performed via a particular directory or directories, which will have DCache file.family tags to indicate the usage.

- Raw DAQ data being archived to Enstore (/pnfs/hypot/raw)
- Production analysis data being archived to Enstore (/pnfs/hypot/*)
- Providing read access to enstore data for production and analysis (/pnfs/hypot/*)
- Historical archiving of files and directories (/pnfs/hypot/archive)
- Staging output data for later transfer to other storage (/pnfs/hypot/scratch/ifdh_stage)
- Dropbox disk for SAM File Transfer Services (/pnfs/hypot/fts/dropbox_*)
- Temporary storage for experiment user's analysis output (/pnfs/hypot/scratch/users/username)

3 Principles

Several principles are assumed here.

- Experiment raw/analyzed data written to tape will go through a SAM FTS, which will ensure it is recorded in the SAM database with suitable metadata. Users will not put experiment data files in tape backed areas directly.
- Historical archival storage (i.e. inactive home areas, old analysis projects, etc.) will be done by designated experiment librarians, at the request of experiment spokespeople. Again, normal users will not put files in tape backed areas.

3.1 Accounts

(At least) The following groups and accounts should be created for our Hypothetical Experiment:

Groups:

hypot general experiment group, all members of experiment should be in it.

hypotadm Administrative group – people in this group can do normal administrative actions, like adding user scratch areas and/or new dropbox areas, etc.

hypotofficial group including hypotadm, hypotsam, and possibly hypotpro; group ownership for official experiment "blessed" data.

Accounts:

hypotraw Account for data acquisition to run under - in 'hypot', 'hypotofficial' groups.

hypotana Account that "Role=Analysis" grid jobs will map to, in 'hypot' group.

hypotpro Account that "Role=Production" grid jobs will map to, in 'hypot', 'hypotofficial' groups.

hypotdat Account to own "official" data and software – in 'hypotofficial' group

hypotsam Account for SAM File Transfer servers, etc.– in 'hypotofficial', 'hypot' group

hypotadm Administrative account for meta-ownership, etc. – in 'hypotofficial', hypotadm, hypot groups.

Normal users will be in the hypot group, administrative users would be in hypotadm and hypotofficial as well.

These accounts and groups should all exist on the experiment administrative server, described below.

4 Experiment Administrative Server

There should be a node setup (probably a VM, possibly in fermicloud) named hypot-admin.fnal.gov, which should have the /pnfs/hypot area mounted without root-squash, and suitable designees from the experiment should be able to log into this system as root, to make appropriate permissions changes to the experiment pnfs area. This system will need to have a fixed IP address so that it can have special mount permissions.

Similarly, experiment disk areas on the bluearc should be exported to and mounted here without root-squash, so that suitable permissions changes can be made.

5 DCache areas

The experiment should have a /pnfs/hypot area in DCache, which will have the following areas, with separate areas used for separate purposes, as described above. These areas should have the indicated ownership and permissions, and have the indicated file_family tags to indicate to the DCache system how the files in that area are to be handled.

/pnfs/hypot/ Overall experiment area. Note that this may also be called:

- /pnfs/usr/hypot
- /pnfs/fnal.gov/usr/hypot

depending how you reach DCache.

file_family: hypot; Owner: hypotadm; Group: hypotadm; Perms: 775

scratch/ A directory level collecting all non tape-backed storage for the experiment.

Note that it is currently under discussion whether this should be configured so that it will fill up, and then need manual cleanup, or whether it should automatically purge old files to make room for new ones.

file_family: scratch ; Owner: hypotadm ; Group: hypotadm ; Perms: 775

scratch/users Should have a subdirectory for each user for scratch storage. Folks in hypotadm group should be able to create new directories here for new users, and then chown them to that user.

file_family: scratch (inherited); Owner: hypotadm ; Group: hypotadm ; Perms: 775

scratch/users/username This subdirectory should be created as part of adding new experiment users.

file_family: scratch (inherited); Owner: username ; Group: hypot ; Perms: 775

scratch/fts/ dropboxes for SAM file transfer services will go under here.

file_family: scratch (inherited) ; Owner: hypotadm; Group: hypotadm; Perms: 775

scratch/fts/dropbox_pro Dropbox for production output.

file_family: scratch (inherited); Owner: hypotpro; Group: hypotofficial; Perms: 775

scratch/fts/dropbox_ana Dropbox for analysis output.

file_family: scratch (inherited); Owner: hypotana; Group: hypot; Perms: 775

scratch/ifdh_stage/ directory for staging files to be copied back elsewhere (i.e. BlueArc).

file_family: scratch (inherited); Owner: hypotana; Group: hypot; Perms: 775

raw/ A directory level for raw data coming from the detector. If the experiment's raw files are small, it should also be configured for small file support in Enstore.

file_family: raw*; Owner: hypotraw; Group: hypotofficial; Mode: 775

archive/ A directory level for placing archives files that are being mothballed. (for example project areas of people who have left the experiment, etc.) We recommend people use the administrative server to pack up areas with "tar" and place the tarfiles here possibly with year by year subdirectories.

file_family: archive; Owner: hypotadm; Group: hypotadm; Mode: 775

mc/

data/

others/ Any other directories at this level should be tape-backed data areas. File family names should start with the experiment name, but should be set realizing that Enstore will group them on tapes by file family.
file_family: hypot* ; Owner: hypotsam ; Group: hypotofficial; Mode: 775

NOTE: If you have file families, libraries, etc. set at a lower level than /pnfs/hypot/directory, for example you set file families or libraries at /pnfs/hypot/production/dir1, etc. then the intermediate directory /pnfs/hypot/production should be group/owner hypotadm, to avoid the directory being rename-able by regular users.

6 GridFTP/SRM Mappings

To be able to use GridFTP and/or SRM transfers from DCache, our experiment needs to have the experiment accounts mapped in the DCache configuration, which is generally set to limit them to some particular subtree the dcache namespace.

We recommend requesting that all of the experiment accounts be mapped to the base /pnfs/hypot directory, so that software tools (i.e. ifdh cp) that want to make use of gridftp know what the mapping will be.

We recommend that all such gridftp mapping points have a pnfs sub-directory with symlinks to . and .. in it such that gsiftp url's with the full path work, so that in future clients can all use full paths, and mappings/restrictions at the gridftp layer can be otherwise arbitrary.