FNAL Storage and Data Transfer

Michael Kirby

Intensity Frontier Computing Summer School

18 June 2021

By Source, Fair use, https://en.wikipedia.org/w/index.php?curid=20680213

Overview of Storage and Data Movement

Questions

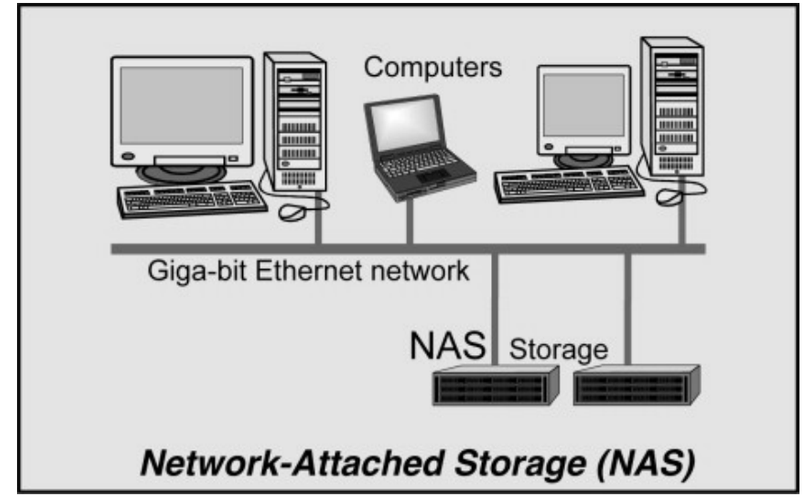
- What are the types and roles of DUNE's data volumes?
- What are the commands and tools to handle data?

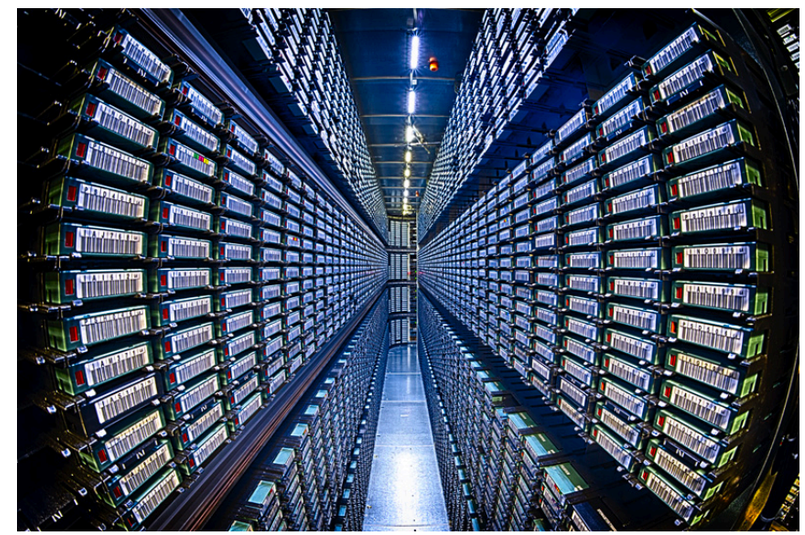
Objectives

- Understanding the data volumes and their properties
- Displaying volume information (total size, available size, mount point, device location)
- Differentiating the commands to handle data between grid accessible and interactive volumes

Introduction

- There are three types of storage volumes that you will encounter at Fermilab:
  - local hard drives
  - network attached storage
  - large-scale, distributed storage
- Each has it's own advantages and limitations, and knowing which one to use when isn't all straightforward or obvious.
- But with some amount of foresight, you can avoid some of the common pitfalls that have caught out other users.





Vocabulary

- What is immutable?

  A file is immutable means that once it is written to the volume it cannot be modified, only read, moved, or deleted. Not a good choice for code or other files you want to change.
- What is interactive or POSIX?

  Interactive volumes, or volume with POSIX access (Portable Operating System Interface Wikipedia) allow users to directly read, write and modify using standard commands, e.g. using bash language.
- What is meant by ‘grid accessible’?

  Volumes that are grid accessible require specific tool suites to handle data stored there. This will be explained in the following sections

Interactive Storage Volumes

- Home Area
- Locally mounted volumes
- Network Attached Storage (NAS)

Interactive Storage Volumes (on gpvms)

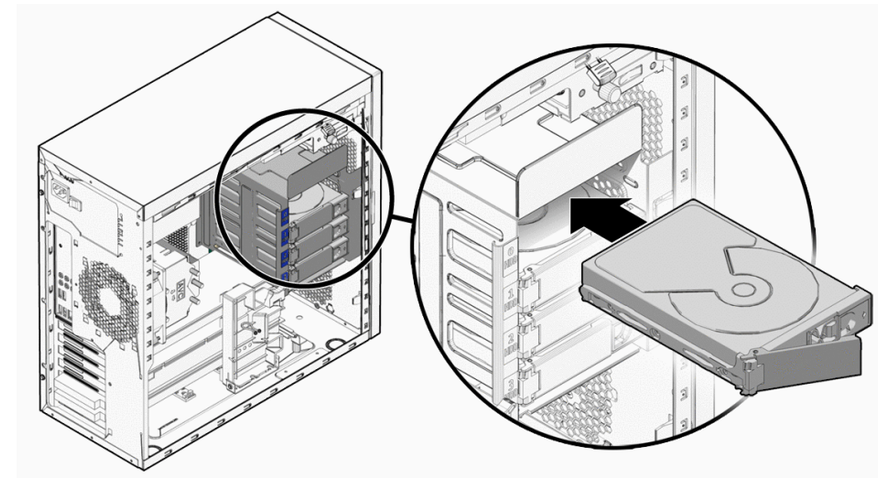
Home area is similar to the user's local hard drive but network mounted:

- access speed to the volume very high, on top of full POSIX access
- they NOT safe to store certificates and tickets
- not accessible as an output location from grid worker nodes
- not for code development (size of less than 2 GB)
- You need a valid Kerberos ticket in order to access files in your Home area
- Periodic snapshots are taken so you can recover

Interactive Storage Volumes (on gpvms)

Locally mounted volumes are local physical disks, mounted directly

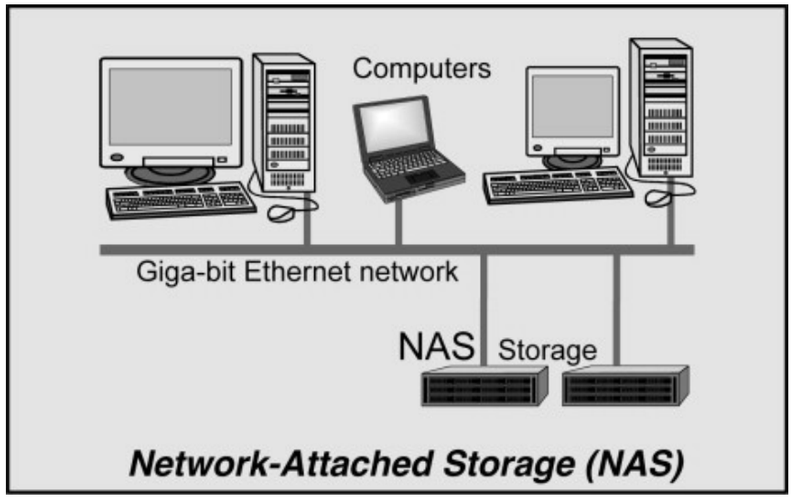
- mounted on the machine with direct links to the /dev/ location
- used as temporary storage for infrastructure services (e.g. /var, /tmp,)
- can be used to store certificates and tickets. These are saved there automatically with owner-read permission and other permissions disabled.
- usually very small and should not be used to store data files or for code development



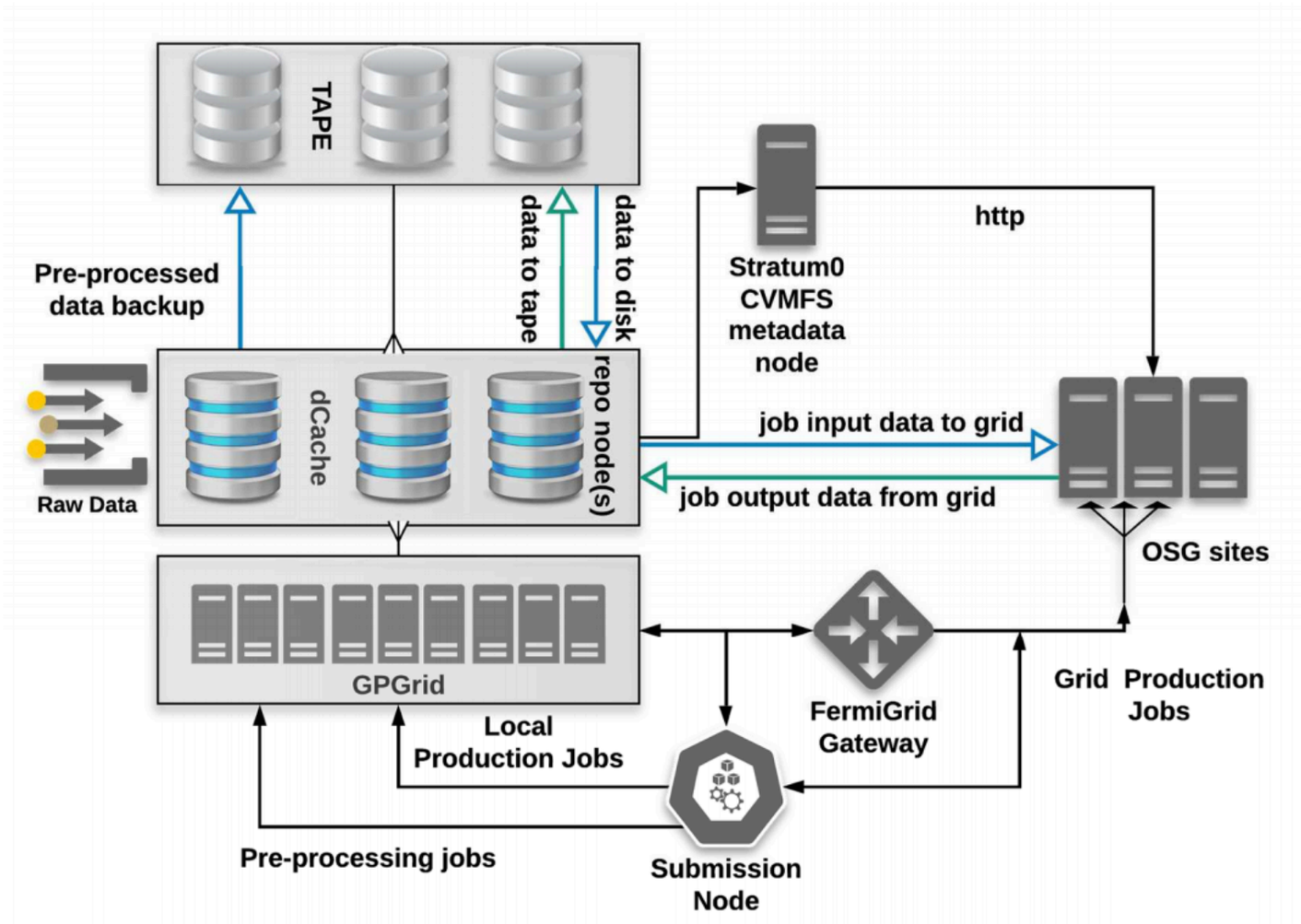
Interactive Storage Volumes (on gpvms)

Network Attached Storage (NAS) element behaves similar to a locally mounted volume.

- functions similar to services such as Dropbox or OneDrive
- fast and stable access rate
- volumes available only on a limited number of computers or servers
- not available to on larger grid computing
- /<exp>/app has periodic snapshots in /<exp>/app/.snapshot
- but /dune/data and /dune/data2 do not have snapshots



Grid-accessible storage volumes



Grid-accessible storage volumes at FNAL

- dCache+Enstore is used for large-scale, distributed storage with capacity for more than 100 PB of storage and O(10000) connections.
- Whenever possible, these storage elements should be accessed over xrootd (see next section) as the mount points on interactive nodes are slow and unstable. Here are the different dCache volumes:
- **Persistent dCache**: the data in the file is actively available for reads at any time and will not be removed until manually deleted by user
- **Scratch dCache**: large volume shared across all experiments. When a new file is written to scratch space, old files are removed in order to make room for the newer file.
- **Resilient dCache**: handles custom user code for their grid jobs, often in the form of a tarball. Inappropriate to store any other files here.
- **Tape-backed dCache**: disk based storage areas that have their contents mirrored to permanent storage on Enstore tape. Files are not available for immediate read on disk, but needs to be ‘staged’ from tape first.

Grid-accessible storage volumes at FNAL

	Quota/Space	Retention Policy	Tape Backed?	Retention Lifetime on disk	Use for	Path	Grid Accessible
Persistent dCache	No/~100 TB/exp	Managed by Experiment	No	Until manually deleted	immutable files w/ long lifetime	/pnfs/<exp>/persistent	Yes
Scratch dCache	No/no limit	LRU eviction - least recently used file deleted	No	Varies, ~30 days (*NOT* guaranteed)	immutable files w/ short lifetime	/pnfs/<exp>/scratch	Yes
Resilient dCache	No/no limit	Periodic eviction if file not accessed	No	Approx 30 days (your experiment may have an active clean up policy)	input tarballs with custom code for grid jobs (do NOT use for grid job outputs)	/pnfs/<exp>/resilient	Yes
Tape backed	dCache No/O(10) PB	LRU eviction (from disk)	Yes	Approx 30 days	Long-term archive	/pnfs/dune/…	Yes
NAS Data	Yes (~1 TB)/ 32+30 TB total	Managed by Experiment	No	Till manually deleted	Storing final analysis samples	/dune/data	No
NAS App	Yes (~100 GB)/ ~15 TB total	Managed by Experiment	No	Until manually deleted	Storing and compiling software	/dune/app	No
Home Area (NFS mount)	Yes (~10 GB)	Centrally Managed by CCD	No	Until manually deleted	Storing global environment scripts (All FNAL Exp)	/nashome/<letter>/<uid>	No

https://cdcvs.fnal.gov/redmine/projects/fife/wiki/Understanding_storage_volumes

Commands and Tools

- To find out what types of volumes are available on a node can be achieved with the command ‘df’.
- Use ‘-h’ for human readable format. It will list a lot of information about each volume (total size, available size, mount point, device location).

Bash

```bash
df -h
```

Exercise 1

From the output of the `df -h` command, identify:

1. the home area
2. the NAS storage spaces
3. the different dCache volumes

Fermilab

A puppy

Intensity Frontier Data Handling Client (ifdhc)

- a tool suite that facilitates selecting the appropriate data transfer method from many possibilities while protecting shared resources from overload
- here is an example of copying a file:

```
» source /cvmfs/fermilab.opensciencegrid.org/products/common/etc/setup
» setup ifdhc v2_5_4
» ifdh cp root://fndca1.fnal.gov:1094/pnfs/fnal.gov/usr/dune/tape_backed/dunepro/physics/full-reconstructed/2019/mc/out1/PDSPProd2/22/60/37/10/PDSPProd2_protoDUNE_sp_reco_35ms_sce_off_23473772_0_452d9f89-a2a1-4680-ab72-853a3261da5d.root /dev/null
```

https://cdcvs.fnal.gov/redmine/projects/ifdhc/wiki/Ifdh_commands

Note: red text above is a single file location

Intensity Frontier Data Handling Client (ifdhc)

- For reference: https://cdcvs.fnal.gov/redmine/projects/ifdhc/wiki/Ifdh_commands

Exercise 2

Using the ifdh command, complete the following tasks:

- create a directory in your dCache scratch area (/pnfs/dune/scratch/users/${USER}/) called “DUNE_tutorial_May2021”
- copy your ~/.bashrc file to that directory.
- copy the .bashrc file from your scrtach directory DUNE_tutorial_May2021 dCache to /dev/null
- remove the directory DUNE_tutorial_May2021 using “ifdh rmdir /pnfs/dune/scratch/users/${USER}/DUNE_tutorial_May2021” Note, if the destination for an ifdh cp command is a directory instead of filename with full path, you have to add the “-D” option to the command line. Also, for a directory to be deleted, it must be empty.

- Note for this tutorial, make the following changes
  - “DUNE_tutorial_May2021” to “FIFE_tutorial_June2021”
  - /pnfs/dune -> /pnfs/<your_experiment_here>/ (e.g. /pnfs/uboone/, /pnfs/nova)

Fermilab

Puppies

Fermilab

xrootd

- The eXtended ROOT daemon is software framework designed for accessing data from various architectures and in a complete scalable way (in size and performance).
- XRootD is most suitable for read-only data access. XRootD Man pages
- Issue the following commands and try to understand how the first command enables completing the parameters for the second command

```
» pnfs2xrootd /pnfs/dune/scratch/users/${USER}/
root://fndca1.fnal.gov:1094//pnfs/fnal.gov/usr/dune/scratch/users/kirby
» xrdfs root://fndca1.fnal.gov:1094/ ls /pnfs/fnal.gov/usr/dune/scratch/users/${USER}/
  FIFE_tutorial_June2021
```


Fermilab

Bringing it all together

Exercise 3

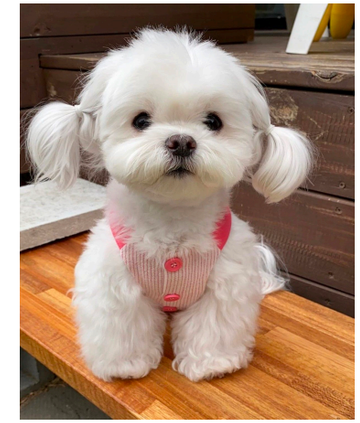
Using a combination of `ifdh` and `xrootd` commands discussed previously:

- Use `ifdh` locateFile to find the directory for this file `PDSPProd4a_protoDUNE_sp_reco_stage1_p1GeV_35ms_sce_off_43352322_0_20210427T162252Z.root`
- Use `pnfs2xrootd` to get the `xrootd` URI for that file.
- Use `xrdcp` to copy that file to `/dev/null`
- Using `xrdfs` and the `ls` option, count the number of files in the same directory as `PDSPProd4a_protoDUNE_sp_reco_stage1_p1GeV_35ms_sce_off_43352322_0_20210427T162252Z.root`

```
» ifdh locateFile PDSPProd4a_protoDUNE_sp_reco_stage1_p1GeV_35ms_sce_off_43352322_0_20210427T162252Z.root
enstore:/pnfs/dune/tape_backed/dunepro/protodune-sp/full-reconstructed/2021/mc/out1/PDSPProd4a/18/80/01/67(3910@fl8386l8)
» pnfs2xrootd /pnfs/dune/tape_backed/dunepro/protodune-sp/full-reconstructed/2021/mc/out1/PDSPProd4a/18/80/01/67
root://fndca1.fnal.gov:1094//pnfs/fnal.gov/usr/dune/tape_backed/dunepro/protodune-sp/full-reconstructed/2021/mc/out1/PDSPProd4a/18/80/01/67
```

Summary

- ifdh commands
- understanding storage volumes



Key Points

- Home directories are centrally managed by Computing Division and meant to store setup scripts or certificates.
- Network attached storage (NAS) /dune/app is primarily for code development.
- The NAS /dune/data is for store ntuples and small datasets.
- dCache volumes (tape, resilient, scratch, persistent) offer large storage with various retention lifetime.
- The tool suites idfh and XRootD allow for accessing data with appropriate transfer method and in a scalable way.