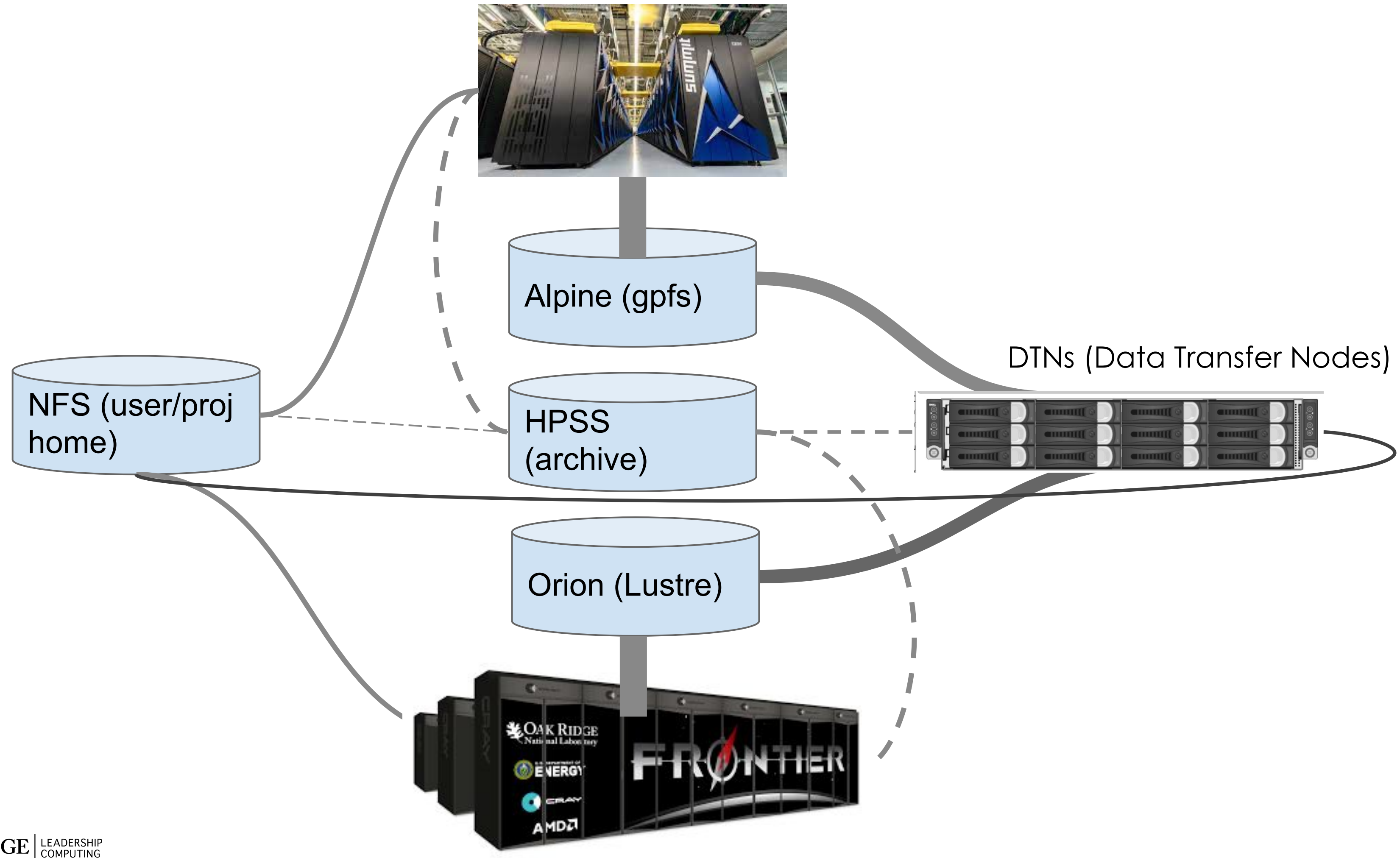
OLCF Storage and Orion Best Practices

Suzanne Parete-Koon NCCS HPC Engineer
Jesse Hanley, Senior HPC Linux Systems Engineer
5-31-23

ORNL is managed by UT-Battelle LLC for the US Department of Energy



Alpine (gpfs)

NFS (user/proj home)

HPSS (archive)

DTNs (Data Transfer Nodes)

Orion (Lustre)

A Storage Area for every Activity

User Centric

- **User Home: (NFS)** Long-term data for routine access that is unrelated to a project. Read/write from from Frontier compute nodes- but use Orion Lustre to launch/run jobs.
- **Member Work: (Orion/Alpine)** Short-term user data for fast batch-job access. Purged.
- **Member Archive: (HPSS)** Long-term project data for archival access that is not shared with other project members.

Project Centric

- **Project Home (NFS) :** Long-term project data for routine access that's shared with other project members. Read/write from from Frontier compute nodes- but use Orion Lustre to launch/run jobs.
- **Project Work: (Orion/Alpine)** Short-term project data for fast, batch-job access that's shared with other project members. Purged.
- **Project Archive: (HPSS)** Long-term project data for archival access that's shared with other project members.

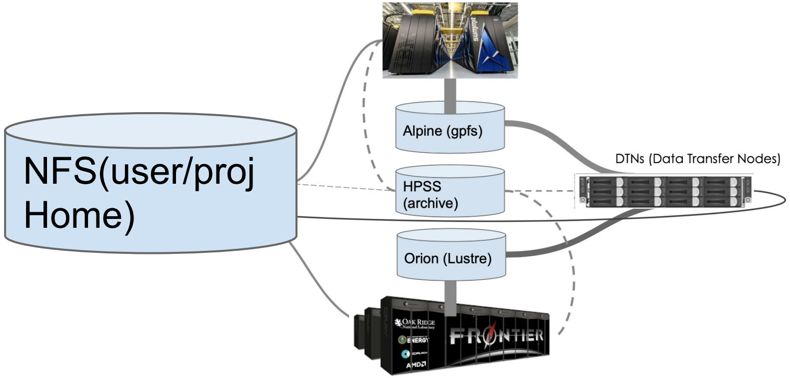
Areas for sharing between projects

- **World Work: (Orion/Alpine)** Short-term project data for fast, batch-job access that's shared with users outside your project. Purged. Only for Category 1 projects.
- **World Archive:(HPSS)** Long-term project data for archival access that's shared with users outside your project.

Note: Moderate Enhanced projects do not have access to HPSS.

Link to docs: https://docs.olcf.ornl.gov/data/index.html#data-storage-and-transfers

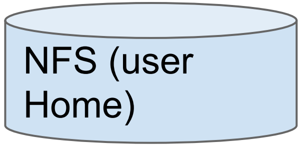
NFS Network File System



- User home: /ccs/home/$USER
  - User home is user-centric
- Project home: /ccs/proj/[projid]
  - Project-centric
- **Long-term** storage for your general data under home or related to project under proj
- Read/write from from Frontier compute nodes- but use Orion Lustre to launch/run jobs.
- **Not purged**
- **Quota** of 50GB (may request increase in well justified cases)
- There is an automated **backup**

Link to docs: https://docs.olcf.ornl.gov/systems/frontier_user_guide.html#nfs-filesystem

NFS Backups



NFS (user Home)

I deleted a file from my NFS, how do I recover it?

Answer: snapshots

Go to the .snapshot folder (ls will not show this folder):

```
[Summit ~]$ cd $HOME/.snapshot
[summit .snapshot]$ ls -l
total 2048
drwxr-xr-x 232 suzanne users 61440 Feb  2 14:04 daily.2023-02-03_0010
drwxr-xr-x 232 suzanne users 61440 Feb  7 13:09 hourly.2023-02-08_1605
drwxr-xr-x 232 suzanne users 61440 Feb  2 14:04 weekly.2023-02-05_0015
```

ORION

Orion is the largest and fastest single file POSIX namespace file system in the world.

- Orion is a Lustre filesystem
- Flash-based performance tier of 5,400 nonvolatile memory express (NVMe) devices providing 11.5 petabytes (PB) of capacity at peak read-write speeds of 10 TB/s
- A hard-disk-based capacity tier of 679 PB at peak read speeds of 5.5 TB/s and peak write speeds of 4.6 TB/s
- Flash-based metadata tier of 480 NVMe devices providing an additional capacity of 10 PB.

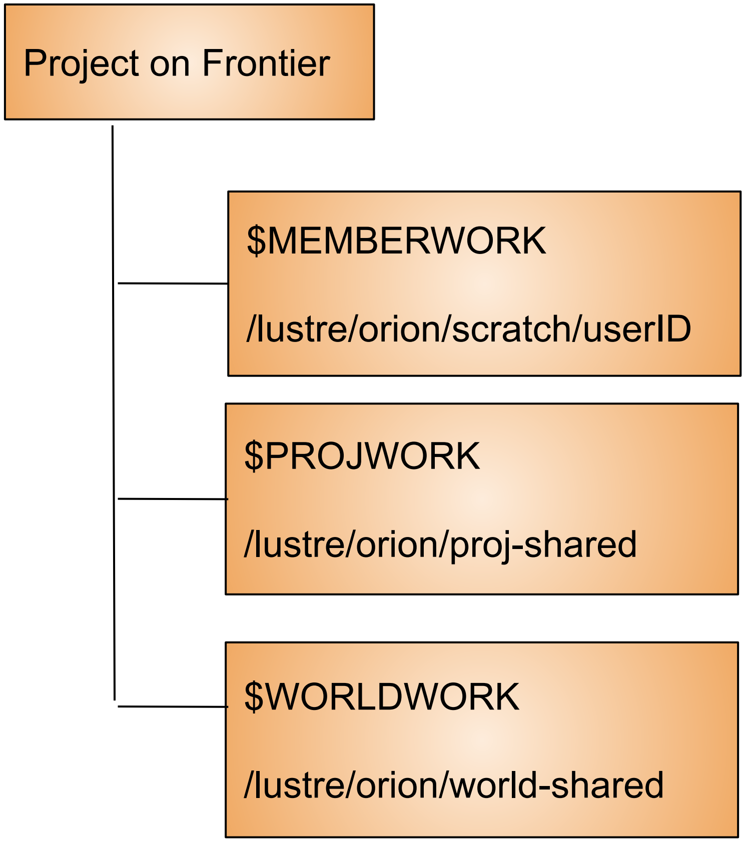
(*More about the specifics of these Lustre performance tiers in a moment*)

ORION

Orion is a Lustre filesystem

- Basic Lustre, in addition to other servers and components, is composed of Objects Storage Targets (OSTs) on which the data for files is stored. A file may be "striped" over multiple OSTs
- Striping provides the ability to store files that are larger than the space available on any single OST and allows a larger I/O bandwidth than could be managed by a single OST
- Orion has multiple performance tiers for storing different sizes of data, so the concept of striping is even more complex that what is described above.
- While users may control striping, OLCF has built tools to help automatically choose the most efficient striping pattern for most files.
- We recommend that users use the defalt striping unless writing very large single files in excess of 512 GB (See Jesse’s slides coming up

ORION


Project on Frontier
$MEMBERWORK
/lustre/orion/scratch/userID
$PROJWORK
/lustre/orion/proj-shared
$WORLDWORK
/lustre/orion/world-shared

MEMBERWORK:

- Short-term storage of user data related to the project but not shared

PROJWORK:

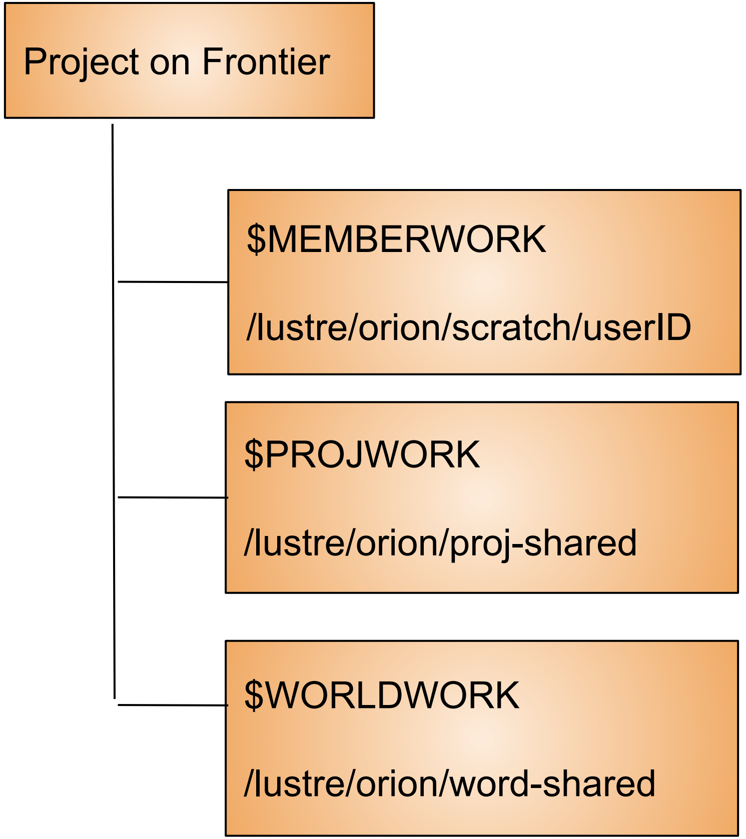
- Short-term storage of project data shared among the members of the project

WORLDWORK:

- Short-term storage of project data shared with OLCF users outside the project ; Only for Category 1 projects.

Note:These aliases on Andes and the DTNs will point to Alpine until further notice.

ORION



Project on Frontier

$MEMBERWORK

/lustre/orion/scratch/userID

$PROJWORK

/lustre/orion/proj-shared

$WORLDWORK

/lustre/orion/word-shared

SHORT TERM STORAGE

- **No backup**
- **Purged after 90 days**
- Short purge exemptions are available for well justified and well defined needs.
- Include
  - Amount of data
  - Time for exemption
  - Plan to reduce move and store data off of Lustre

Orion Storage Policy

Area	Path	Permission	Backups	Purged	On Compute Nodes
Member Work	/lustre/orion/[projid]/scratch/[userid]	700	No	90 days	Read/Write
Project Work	/lustre/orion/[projid]/proj-shared	770	No	90 days	Read/Write
World Work*	/lustre/orion/[projid]/world-shared	775	No	90 days	Read/Write

* Worldwork is only for Category 1 Projects

Link to docs: https://docs.olcf.ornl.gov/systems/frontier_user_guide.html#data-and-storage

HPSS

HPSS (archive)

- **Long-term** storage for large amounts of general data related to your project
- Access by htar and hsi from login nodes and DTNs, access by Globus using the “OLCF HPSS” Globus endpoint.
- HPSS is optimized for large files. Ideally, we recommend sending archives 768 GB or larger to HPSS.
  - If any of the individual files included in an htar are bigger than 68 GB size, then htar will fail, if there are more than 1 million files per archive, htar will fail
- Not purged
- Moderate Enhanced projects do not have access to HPSS.

Link to docs: https://docs.olcf.ornl.gov/data/index.html#hpss-data-archival-system

HPSS

HPSS
(archive)

Area	Path	Type	Permissions	Quota	Backup	Purged	On Compute Nodes
Member Archive	/hpss/prod/[projid]/users/$USER	HPSS	700	100 TB	No	No	No
Project Archive	/hpss/prod/[projid]/proj-shared	HPSS	770	100 TB	No	No	No
World Archive	/hpss/prod/[projid]/world-shared	HPSS	775	100 TB	No	No	No

Note: Moderate Enhanced projects do not have access to HPSS.

Link to docs: https://docs.olcf.ornl.gov/data/index.html#hpss-data-archival-system

Data Transfer Nodes

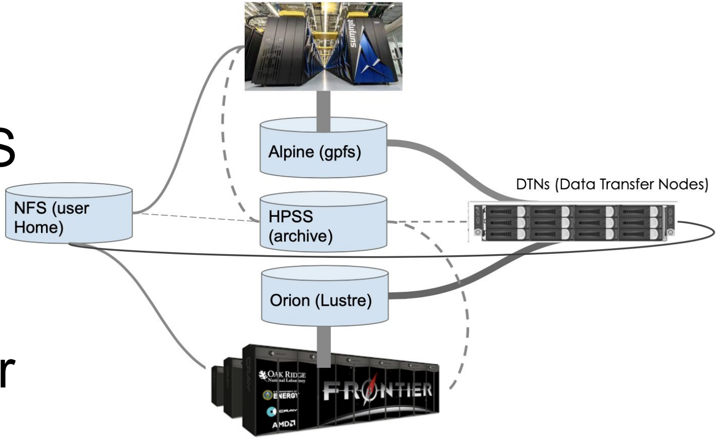
- The Data Transfer Nodes (DTNs) are hosts specifically designed to provide optimized data transfer between OLCF systems and systems outside of the OLCF network.
- Perform well on local-area transfers as well as the wide-area data transfers for which they are tuned.
- Access
  - ssh <username>@dtn.ccs.ornl.gov
  - Globus endpoint OLCF DTN

Data Transfer

- Frontier does not mount Alpine
- Summit does not mount Orion

There are a few ways you can move data between Alpine and Orion:

- We recommend that you use Globus and the DTNs as first choice (fastest)
- However, if you are already archiving restart files or initial data on HPSS, HPSS may be the most convenient path
- You can use the DTN or logins nodes to move small files from Alpine through User Home, but it will be slow.



Lustre Bug

There is currently a Lustre bug that can cause occasional single node crashes on the the DTN and Andes. We are working with HPE for a solution.

- This is the reason that Andes does not mount Orion yet.
- The DTNs do mount Orion because their primary purpose is data transfer and the crashes can be mitigated by transfer tools that handle restarts.
  - If you are using Globus on the DTN when a node crashes, your data transfer will pause and when the nodes comes back up, it will continue where it stopped.
  - For Rsync, try using --append-verify or --checksum
  - If you are using transfer tools that do not have options for interruption restarts, you will need to start your transfer over if the node crashes.

Globus

- Globus is a fast and reliable way to move files.
- It has a convenient Web-interface at globus.org that you log into with a username and password.
- Transfers are done by activating "endpoints"
  - Endpoints are portals where data can be moved using the Globus transfer
  - Activating the OLCF Globus endpoints is done using your OLCF User name and Token Code
  - Endpoints stay activated for hours or days so you don't need to enter your credentials for each transfer.
- Has a command-line Interface
  - https://docs.globus.org/cli/
  - https://docs.globus.org/cli/quickstart/

Link to examples and docs:
https://docs.olcf.ornl.gov/data/index.html#using-globus-to-move-data-to-orion

Globus

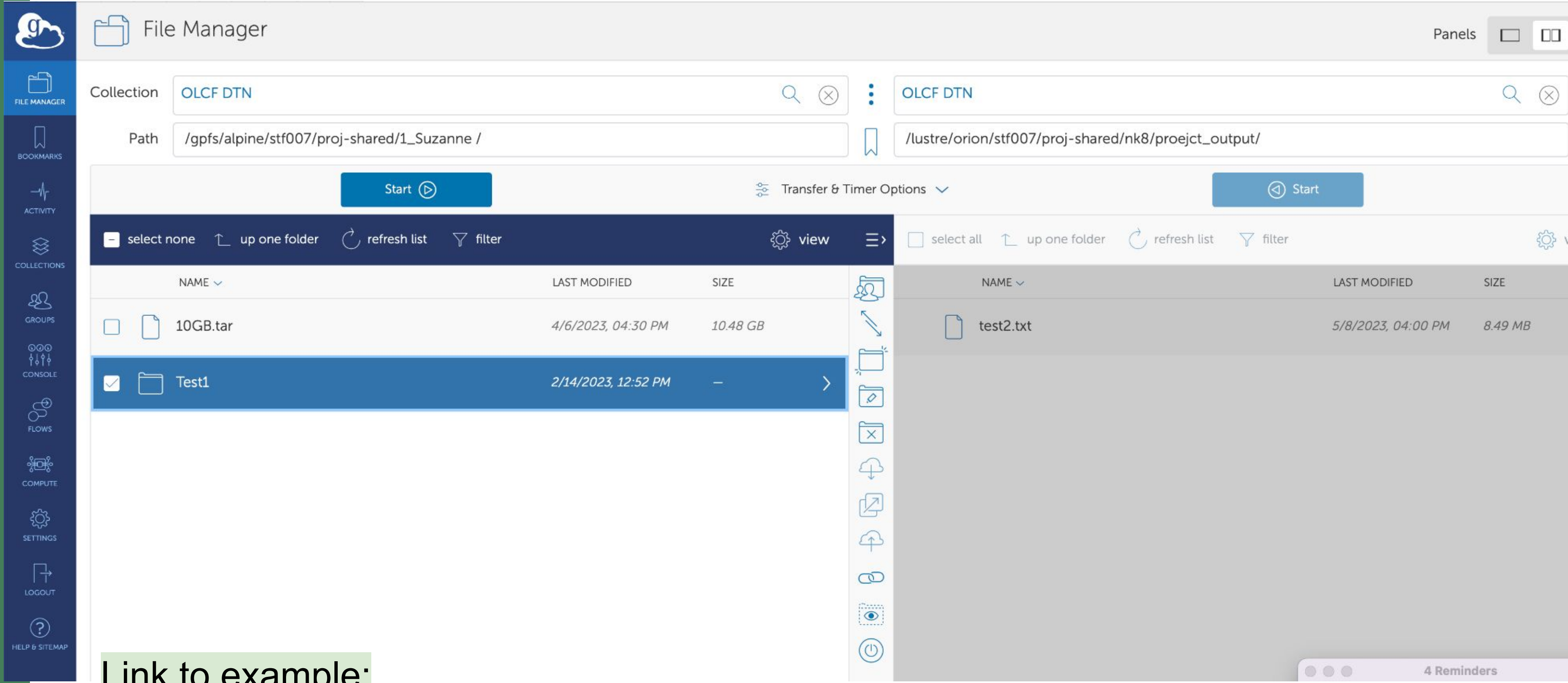
A few Globus Endpoints have been established for OLCF resources.

- OLCF DTN:
  - Provides access to User/Project Home areas as well as the Alpine filesystem and will provide access to the Orion filesystem
- OLCF HPSS
  - Provides access to the HPSS

By utilizing these endpoints you can transfer data between OLCF systems and you can use them with an external endpoint to move data outside of OLCF.

Note: Globus does not preserve file permissions. Files will arrive with User rw- group r-- and world r--. You will need to chmod to reset permissions so files will execute.

Globus



File Manager

Collection: OLCF DTN

Path: /gpfs/alpine/stf007/proj-shared/1_Suzanne /

Collection: OLCF DTN

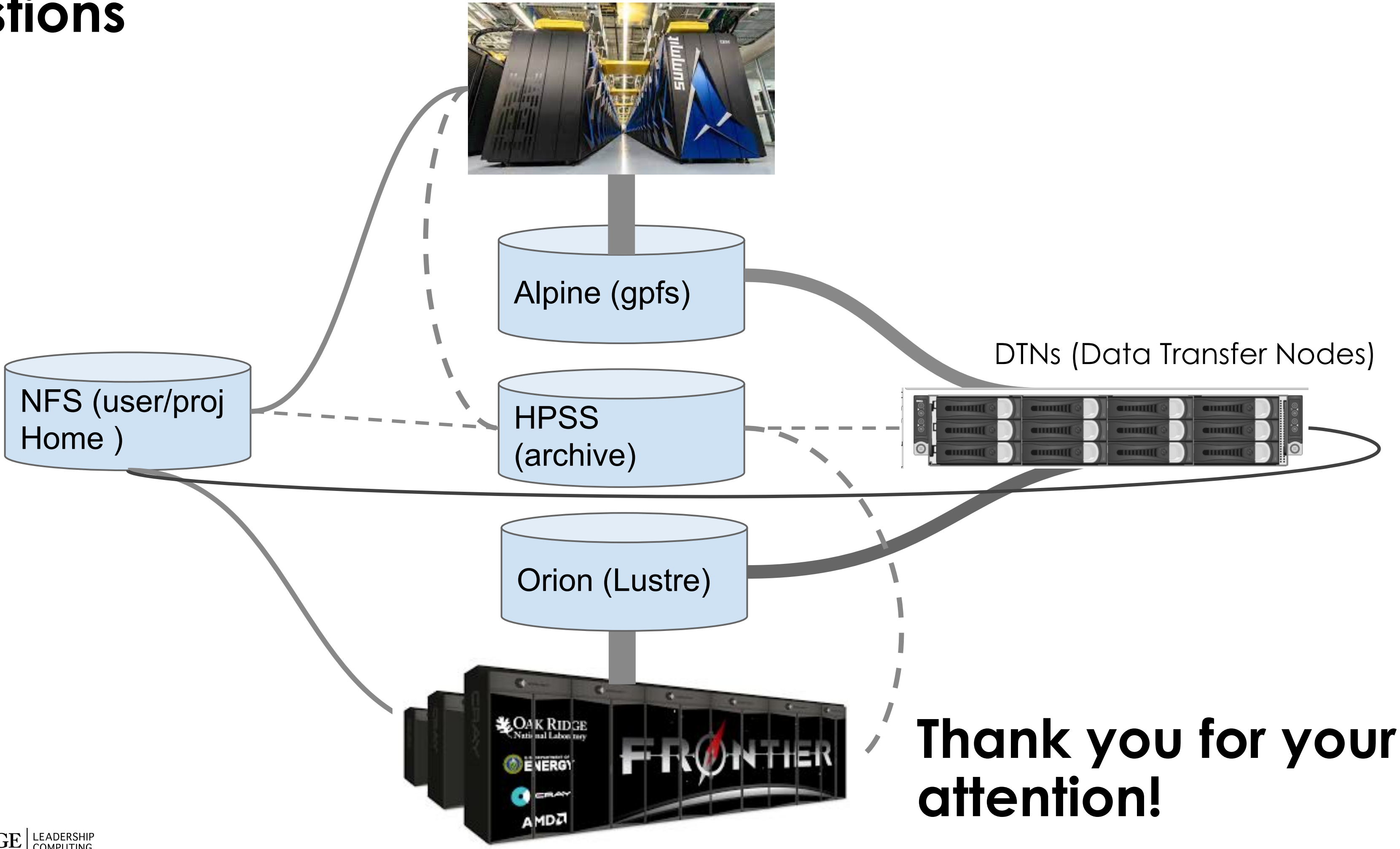
Path: /lustre/orion/stf007/proj-shared/nk8/proejct_output/

	NAME	LAST MODIFIED	SIZE
□	10GB.tar	4/6/2023, 04:30 PM	10.48 GB
☑	Test1	2/14/2023, 12:52 PM	–

NAME	LAST MODIFIED	SIZE
test2.txt	5/8/2023, 04:00 PM	8.49 MB

Link to example:
https://docs.olcf.ornl.gov/data/index.html#using-globus-to-move-data-to-orion

Questions

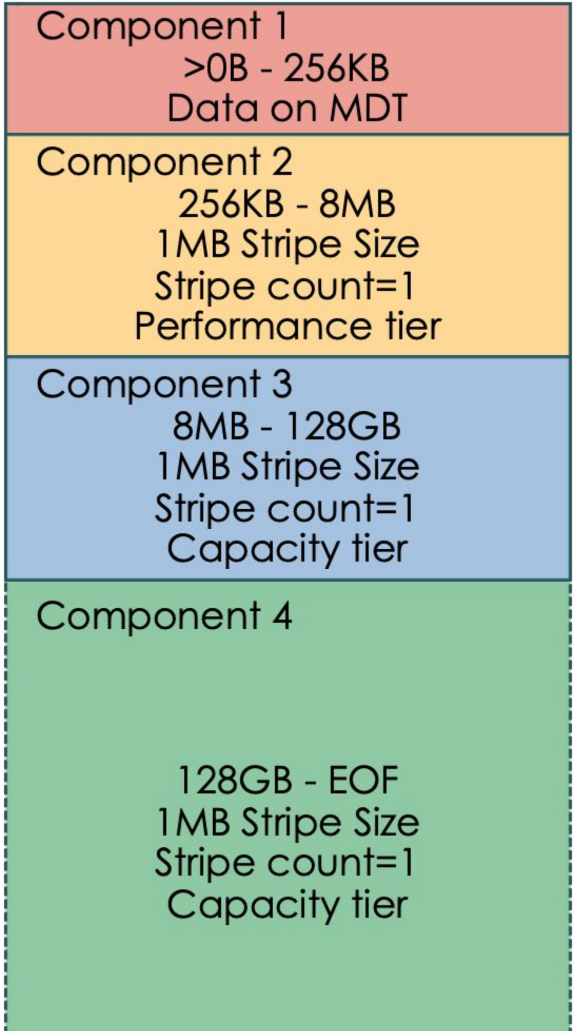


Thank you for your attention!

Default Namespace Layout

Orion uses 3 advanced Lustre features

- *Progressive File Layouts (PFL)*
  - The layout of the file changes as it grows
- *Data on MDT (DoM)*
  - A portion of the file’s content resides on the metadata target
- *Self-Extending Layouts (SEL)*
  - Dynamic PFL, which allows Lustre to avoid (to some degree) `ENOSPC` problems



Component 1
>0B - 256KB
Data on MDT

Component 2
256KB - 8MB
1MB Stripe Size
Stripe count=1
Performance tier

Component 3
8MB - 128GB
1MB Stripe Size
Stripe count=1
Capacity tier

Component 4
128GB - EOF
1MB Stripe Size
Stripe count=1
Capacity tier

Orion Recommendations

Some *sufficiently large single-shared-file workloads* may benefit from explicit striping; please contact help@olcf.ornl.gov

Size	**Stripe Command**
512 GB+	`lfs setstripe -c 8 -p capacity -S 16M`
1 TB+	`lfs setstripe -c 16 -p capacity -S 16M`
8 TB+	`lfs setstripe -c 64 -p capacity -S 16M`
16 TB+	`lfs setstripe -c 128 -p capacity -S 16M`

Potential tooling in development to assist

Storage Usage

```
> lfs quota -h /lustre/orion/
      Disk quotas for usr <uname> (uid <uid>):
            Filesystem  used   quota   limit   grace   files   quota   limit   grace
       /lustre/orion/  10.46G       0k    0k    -     7283        0     0     -
      uid <uid> is using default block quota setting
      uid <uid> is using default file quota setting
      Disk quotas for grp <gname> (gid <gid>):
            Filesystem  used   quota   limit   grace   files   quota   limit   grace
       /lustre/orion/  10.46G       0k    0k    -     7286        0     0     -
      gid <gid> is using default block quota setting
      gid <gid> is using default file quota setting
```

Storage Usage - Continued

Previous command can be augmented

- Part of multiple groups?
  - Specify using `-g <gname|gid>`
  - Example: `lfs quota -g abc123 /lustre/orion/`
- See `man lfs-quota`

Data Considerations

- Directory Structure
  - OLCF provides a top-level structure (`proj-shared`, …)
  - Consider specific subdirectories for collaborative efforts
- Naming Conventions
  - Meaningful file and directory names avoid having to open a file to determine content
  - Well-patterned naming can decreasing onboarding issues
- File Sizes and File Count
  - For best results, try to limit directory entries to no more than ~100k
- Data Lifecycle Management
  - Orion is a scratch file system
  - Ensure critical data is replicated to more durable resources, like HPSS

Orion Data Retention

- Files that have not been accessed or modified within 90 days are subject to purge
- Though often a requested feature, it is usually infeasible to provide a list of purge-eligible files for users (i.e., daily purgeable report/email)
- Special requests for purge exemptions can be submitted with a summary and justification

Darshan

- The `darshan-runtime` is now part of `DefApps` and is loaded by default on Frontier
- Allows users to profile their application's I/O
- Logs available to user in `/lustre/orion/darshan/frontier/<year>/<mm>/<dd>`
- Tooling provided via `darshan-util` modulefile

Questions?