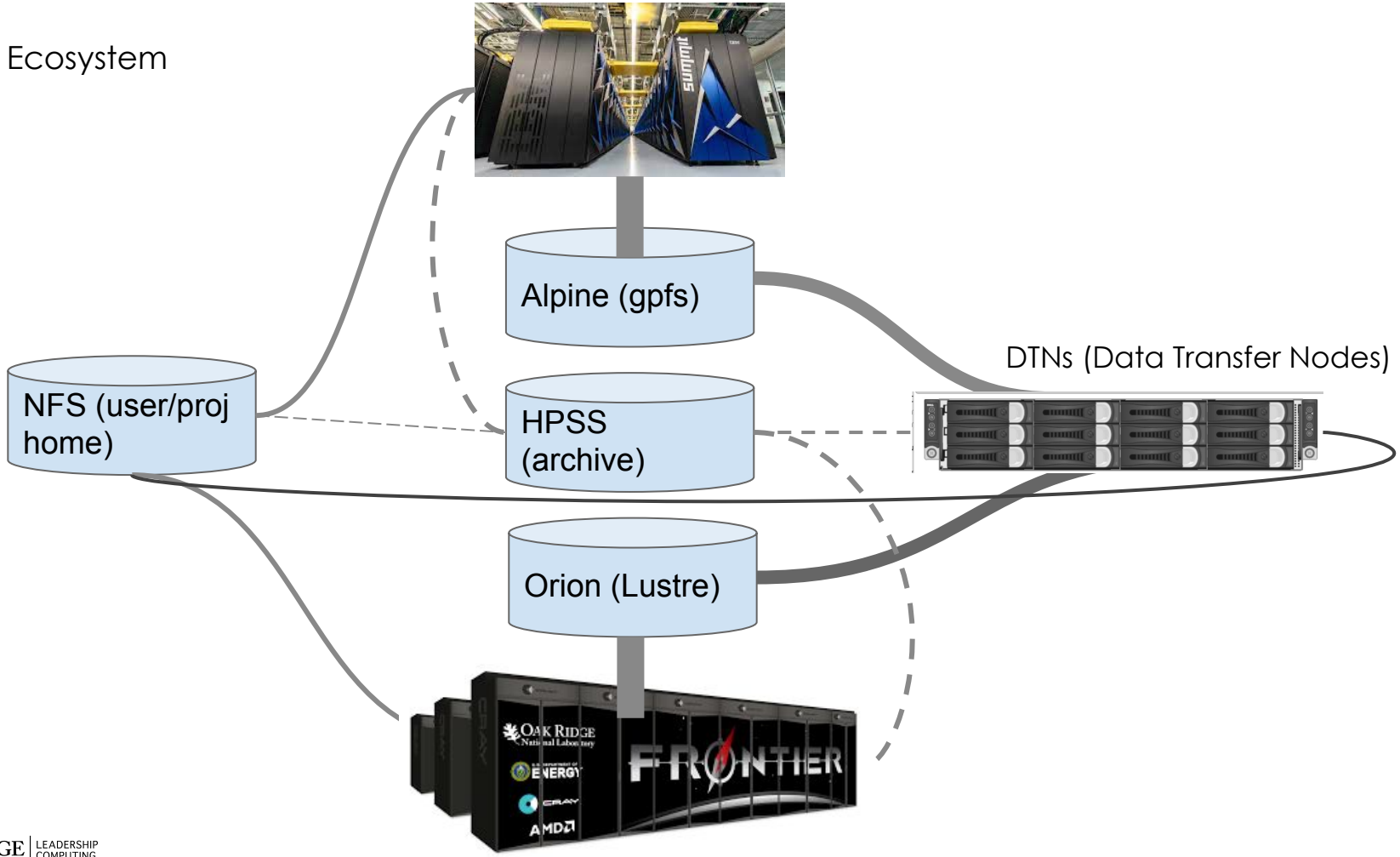
OLCF Data Transfer and Storage Best Practices

Suzanne Parete-Koon NCCS HPC Engineer

8-23-23

Data Ecosystem



Data Management

OLCF systems generate lots of data very quickly; projects should develop a data strategy *as soon as possible*. (It's easier to fix things with 100 files than with 100,000!)

Some things to consider:

- How are files/directories shared among project members?
  - Where will project members store data?
  - What file attributes (permissions, group, etc.) are needed?
- What happens when someone leaves the project?
- What happens when the project ends?
  - Where does the data need to go?
  - How much data is there, who's moving it, and how long will it take?
- Data on parallel file systems, Orion and Atlas, are purged after 90 days.

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.

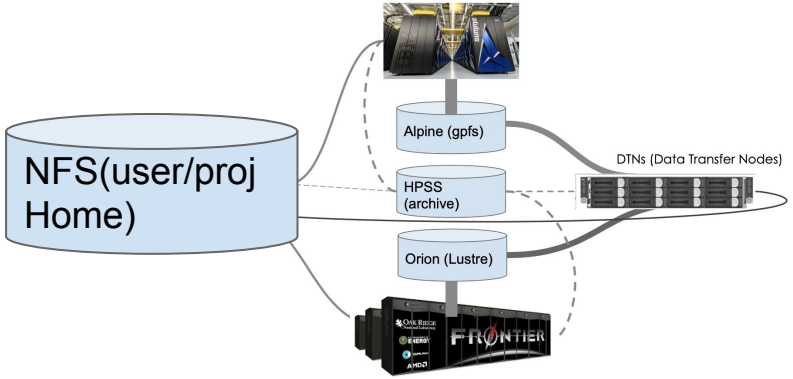
Areas for sharing between projects

- **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

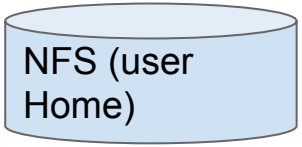


- User home: /ccs/home/$USER
  - User home is user-centric
- Project home: /ccs/proj/[projid]
  - Project-centric
- **Long-term** storage for your general data under home or related to project under proj
- Read/write from from Frontier compute nodes- but use Orion Lustre to launch/run jobs.
- **Not purged**
- **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

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.

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

Orion Recommendations

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

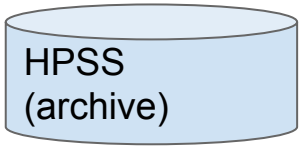
Size	Stripe Command
512 GB+	`lfs setstripe -c 8 -p capacity -S 16M`
1 TB+	`lfs setstripe -c 16 -p capacity -S 16M`
8 TB+	`lfs setstripe -c 64 -p capacity -S 16M`
16 TB+	`lfs setstripe -c 128 -p capacity -S 16M`

Potential tooling in development to assist

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

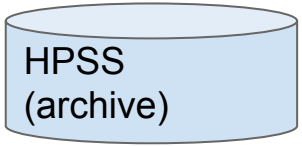
HPSS



- **Long-term** storage for large amounts of general data related to your project
- Not purged
- Moderate Enhanced projects do not have access to HPSS.
- Do not use HPSS as your Alpine/Orion transfer conduit unless it is already part of your data workflow and you have a data management plan

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

HPSS



- Access to HPSS is by htar and hsi from login nodes and DTNs, and by Globus using the “OLCF HPSS” Globus endpoint.
  - If using Globus with HPSS, please tar directories with large numbers of files first before transfer.
    - You risk filling the cache
    - HPSS/ Globus interface restarts interrupted transfers at the beginning
- 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
- If you have millions of files break them up into tars or htars with less than 1 million files
- If you have a several files larger than 68 GB, use Globus

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

HPSS: htar example https://docs.olcf.ornl.gov/data/index.html#htar

To move data from Summit/Alpine to the project shared area of HPSS:

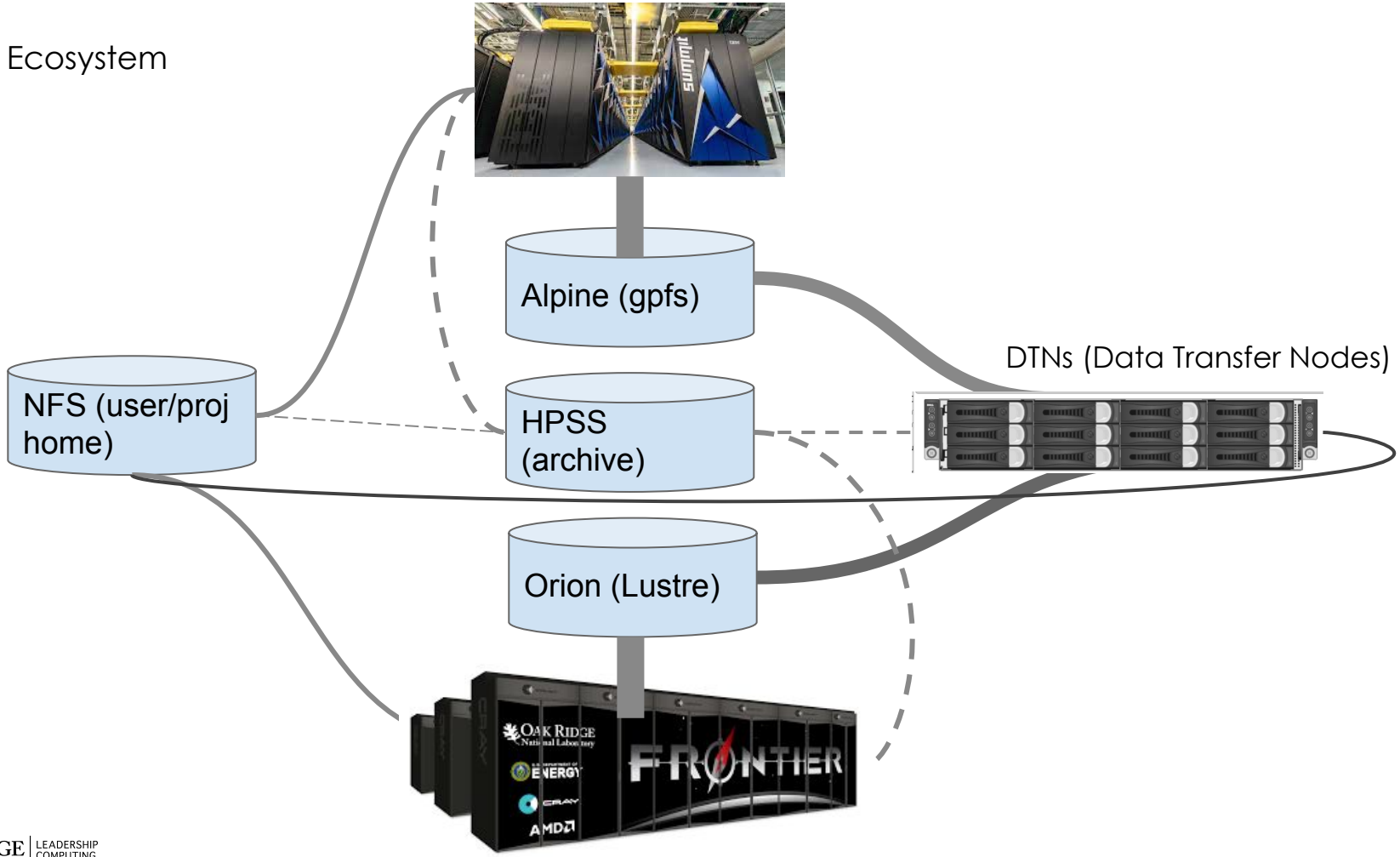
```
Summit> htar -cvf /hpss/prod/stf007/proj-shared/Test1.tar Test1

creating HPSS Archive file /hpss/prod/stf007/proj-shared/Test1.tar
HTAR: a  Test1/
. . .
HTAR: a   /tmp/HTAR_CF_CHK_4042346_1676312676
HTAR Create complete for /hpss/prod/stf007/proj-shared/Test1.tar. 10485767168 bytes
written for 10 member files, max threads: 3 Transfer time: 15.901 seconds (659.440
MB/s) wallclock/user/sys: 16.198 6.593 7.431 seconds
HTAR: HTAR SUCCESSFUL
```

To move data from HPSS to Frontier/Orion

```
Frontier> htar -xvf /hpss/prod/stf007/proj-shared/Test1.tar Test1
.  .  .
wallclock/user/sys: 25.243 0.368 4.898 seconds
```

Data Ecosystem



Alpine

GPS parallel file system attached to Summit

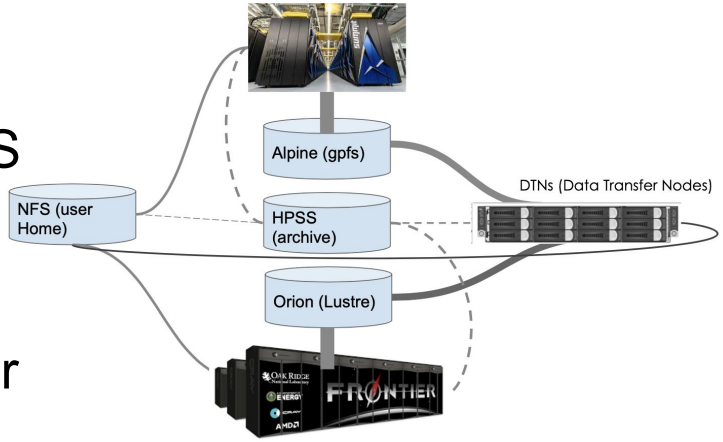
- OLCF Alpine will be Decommission
- Alpine will become read-only on December 19, 2023
- The DTNs mount the new Orion filesystem and all projects with access to Alpine have now been granted access to the Orion filesystem.
- Encourage all teams to start migrating and/or deleting data from the Alpine filesystem now.
- Any data remaining on the Alpine filesystem after December 31, 2023 will truly be unavailable and not recoverable
- More details on the Alpine decommission timeline can be found on https://docs.olcf.ornl.gov/systems/2023_olcf_system_changes.html

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.



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.
  - 2 100 GbE connections to ESnet
  - 1 40 GBE connection to internet
  - 1 FDR IB connection to each storage resource
- 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

Basic command line tools for transfers: SCP

Please use the DTN (ssh <username>@dtn.ccs.ornl.gov)

Sending a file to OLCF:

```
scp yourfile $USER@dtn.ccs.ornl.gov:/path/
```

Retrieving a file from OLCF:

```
scp $USER@dtn.ccs.ornl.gov:/path/yourfile .
```

Sending a directory to OLCF

```
scp -r yourdirectory $USER@dtn.ccs.ornl.gov:/path/
```

- https://docs.olcf.ornl.gov/data/index.html#command-line-terminal-tools

Basic command line tools for transfers: rsync

Please use the DTN (ssh <username>@dtn.ccs.ornl.gov)

Sync a directory named mydir from your local system to the OLCF

```
rsync -avz mydir/ $USER@dtn.ccs.ornl.gov:/path/
```

Sync a directory from the OLCF to a local directory

```
rsync -avz  $USER@dtn.ccs.ornl.gov:/path/dir/ mydir/
```

where:

- a is for archive mode
- v is for verbose mode
- z is for compressed mode

https://docs.olcf.ornl.gov/data/index.html#command-line-terminal-tools

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 the Orion filesystem
- OLCF HPSS
  - Provides access to the HPSS
  - Bundle your files if you can with TAR or ZIP on a DTN node, then transfer using globus. Larger transfers stream better to HPSS and recall better from tape. Globus does not have a utility for doing this automatically.

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.

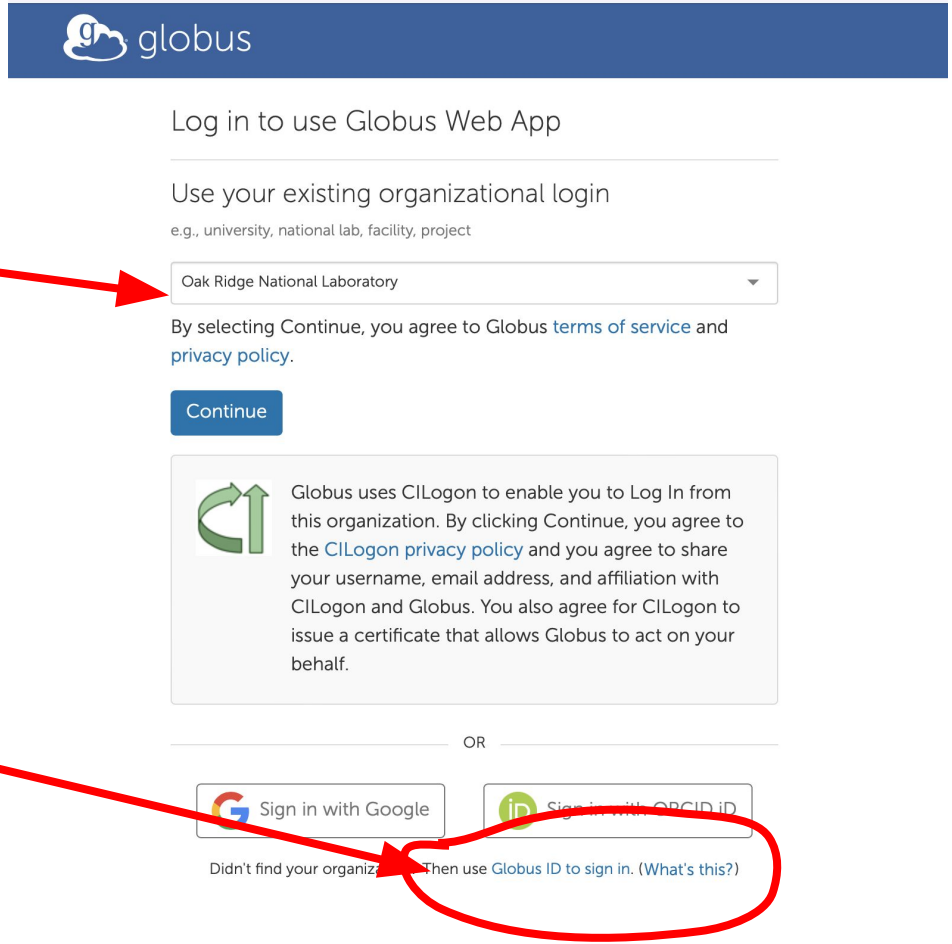
Note: Globus does not preserve file permissions. Files will arrive with User rw- group r-- and world r--. You will need to chmod to reset permissions so files will execute.

Globus example

- Go to https://www.globus.org and log in

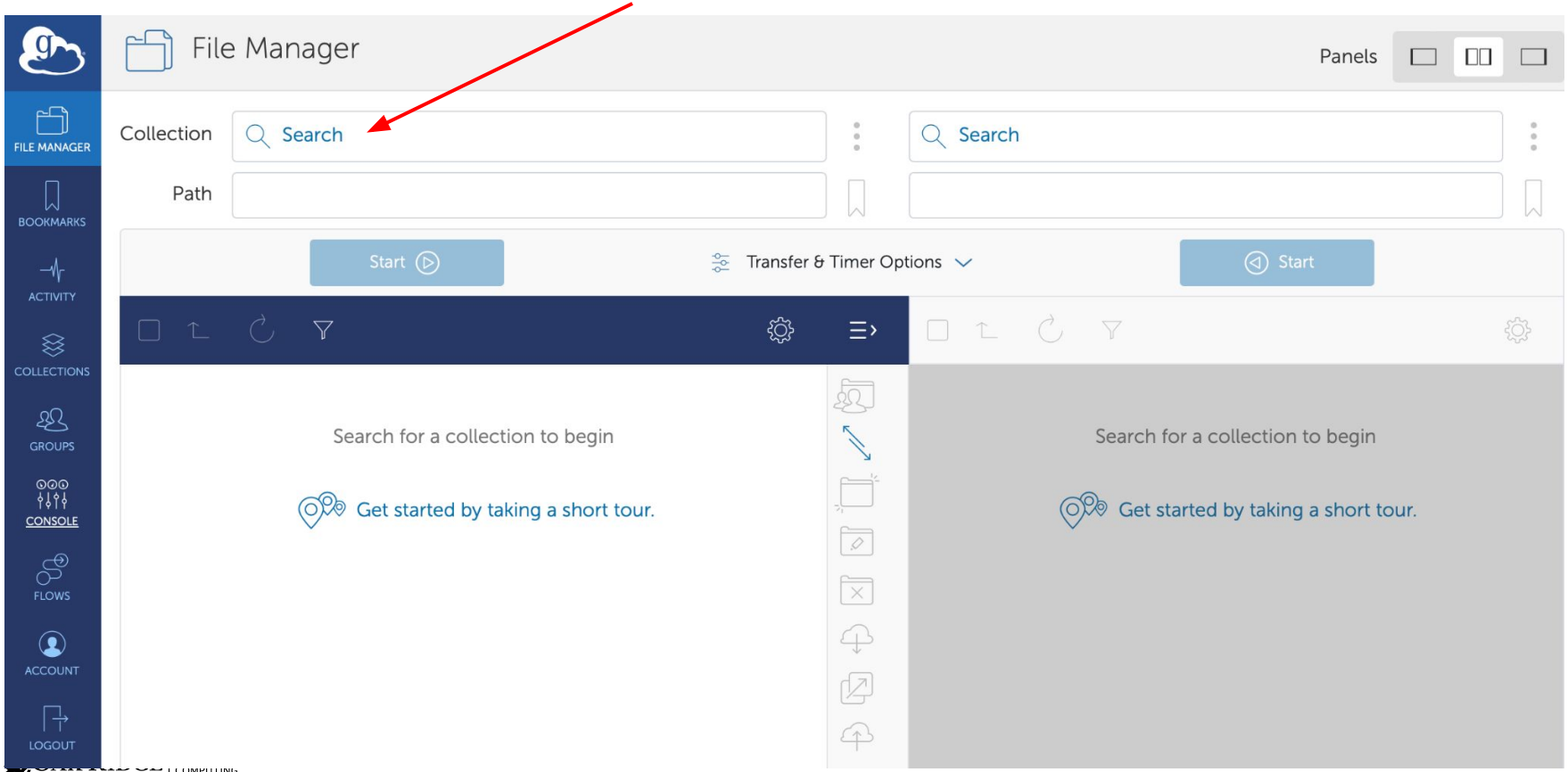
Globus example

- Select the organization that you belong to
- If you don't work for ORNL, do not select ORNL
- If your organization is not in the list, create a Globus account



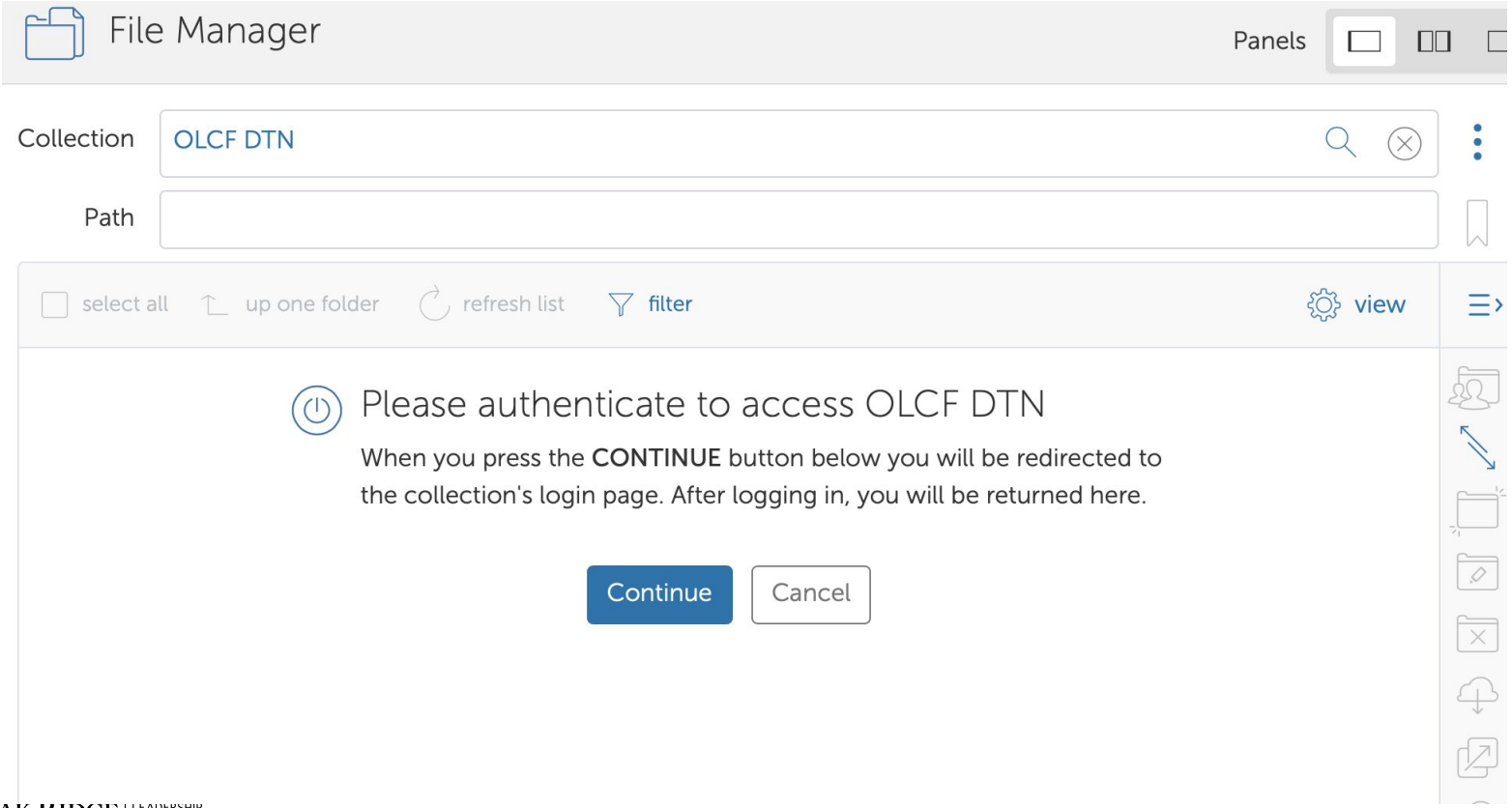
Globus example

- Search for the endpoint OLCF DTN



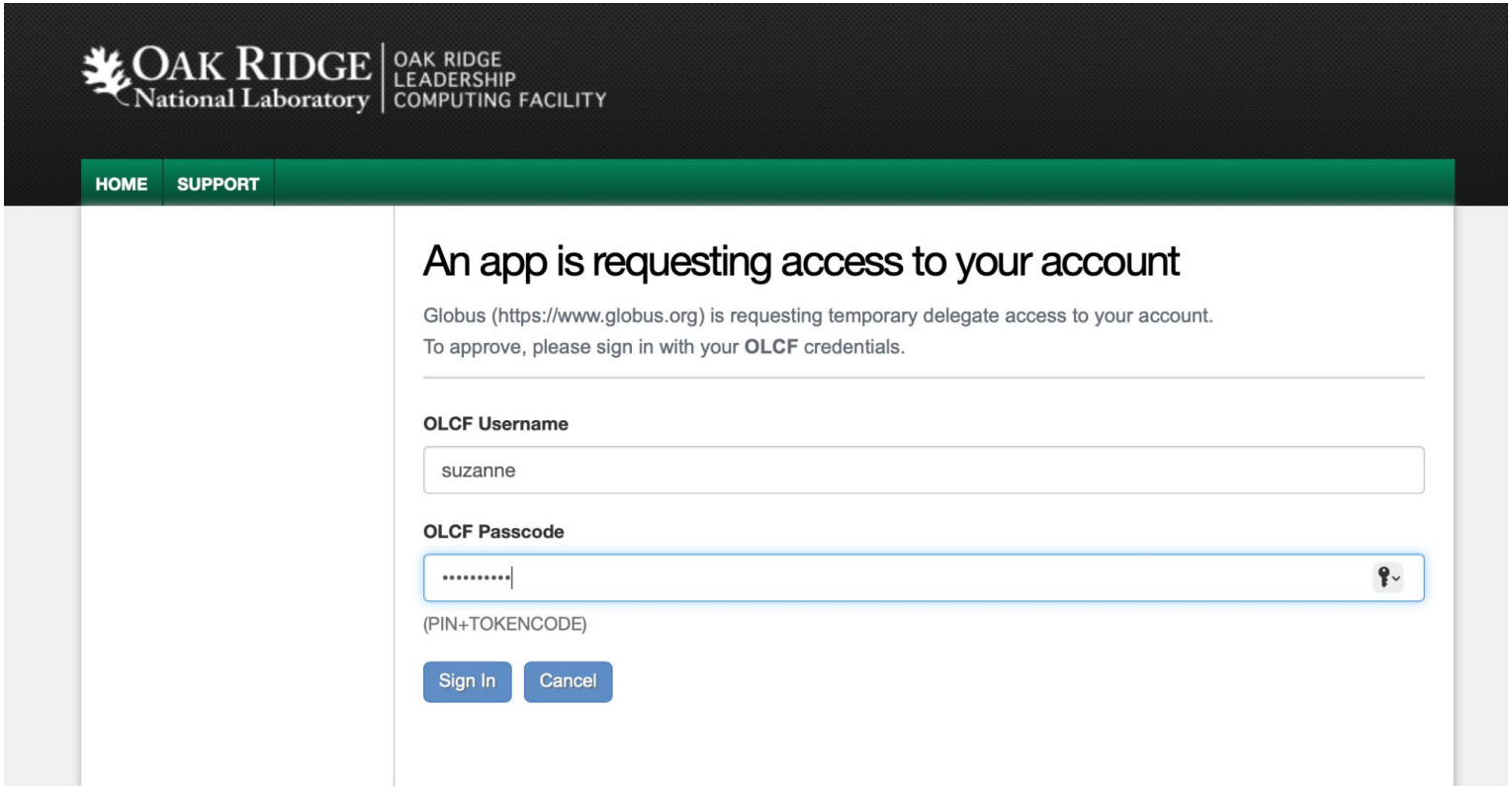
Globus example

- Activate the OLCF DTN endpoint with you OLCF credentials



Globus example

- Activate the OLCF DTN endpoint with your OLCF credentials



Globus example

Enter the desired paths

Select the file(s) you want to transfer.

File Manager

Panels

FILE MANAGER
BOOKMARKS
ACTIVITY
COLLECTIONS
GROUPS
CONSOLE
FLOWS
COMPUTE
SETTINGS
LOGOUT
HELP & SITEMAP

Collection: OLCF DTN

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

Collection: OLCF DTN

Path: /lustre/orion/stf007/proj-shared/nk8/1_Suzanne /

Start

Transfer & Timer Options

Start

view

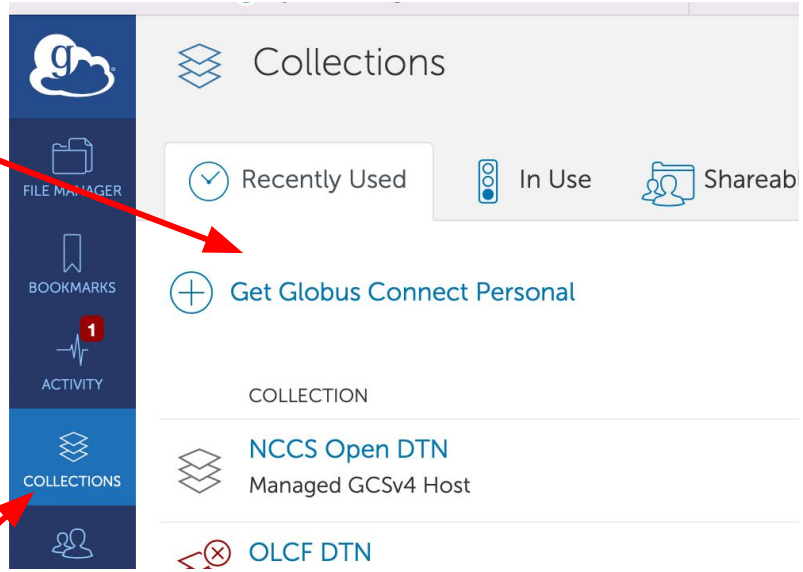
NAME	LAST MODIFIED	SIZE
10GB.tar	4/6/2023, 04:30...	10.48 GB
Test1	7/17/2023, 01:28...	–
Test2	7/17/2023, 01:42...	–
Test2.tar	7/17/2023, 03:0...	85.89 GB

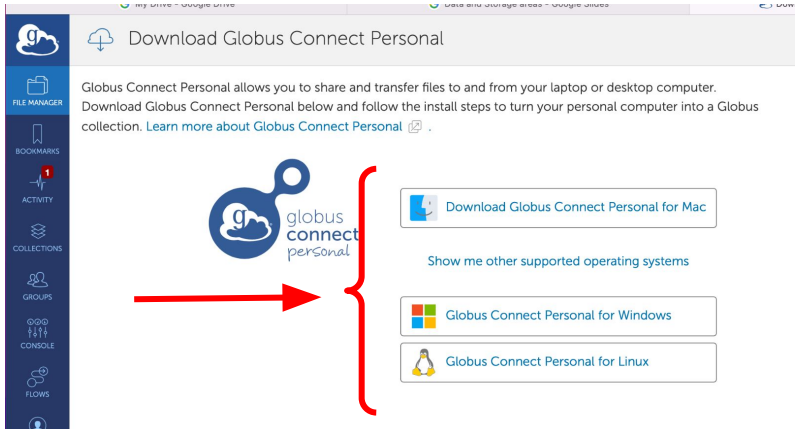
Share
Transfer or Sync to...
New Folder
Rename
Delete Selected
Download
Open
Upload
Get Link
Show Hidden Items
Manage Activation

view

LAST MODIFIED	SIZE
5/15/2023, 01:...	10.48 GB
5/15/2023, 01:...	–

Globus endpoint for your laptop





1. Go to Collections
2. Click "Get Globus Connect Personal"

3. Download the version for your machine and follow the given instructions

4. Once installed, globus must be running and your laptop must be open for the transfer to happen

5. Don't expect to see the same transfer speed to/from your laptop as you see when you use endpoint on DTNs

Globus CLI

- Has a command-line Interface
  - https://docs.globus.org/cli/
  - https://docs.globus.org/cli/quickstart/
- You must install globus CLI to use it. If you install it in your project home area on NFS (/ccs/proj/*) your whole project will be able to use it. Use a cray-python venv or conda for installation

For Python see: https://www.olcf.ornl.gov/wp-content/uploads/2-16-23_python_on_frontier.pdf

Example installation for project stf007 on Frontier using cray-python :

```
$ module load cray-python
$ python3 -m venv /ccs/proj/stf007/globus_cli
$ source /ccs/proj/stf007/globus_cli/bin/activate
$ pip install globus-cli
```

Globus CLI

Example for project stf007:
Use Web interface to active the OLCF DTN (stays activated for 3 days)

```
$ source /ccs/proj/stf007/globus_cli/bin/activate

$ globus login
(may ask you to use the browser interface to get a code to log in.)

$ globus endpoint search 'OLCF'
ID                                      | Owner                  | Display Name
--------------------------------------- | ---------------------- | ---------------------
70a7ea3e-1fb1-11e7-bc36-22000b9a448b | olcf@globusid.org      | NCCS Open DTN
ef1a9560-7ca1-11e5-992c-22000b96db58 | olcf@globusid.org      | OLCF DTN
ac9ea984-dd7f-11e6-9d11-22000a1e3b52 | olcf@globusid.org      | OLCF HPSS

$ olcfdtn=ef1a9560-7ca1-11e5-992c-22000b96db58

$ globus transfer $olcfdtn:/ccs/home/suzanne/tompacc.F90 $olcfdtn:/lustre/orion/stf007/proj-shared/nk8/tompacc.F90
```

Globus CLI

- OLCF is running Globus Connect Server version 4
  - It means that we use the globus rules for “endpoints” rather than collections on the CLI interface.
  - ID of the endpoint is used for management and data transfer.
  - Globus v5 and above have distinct functions for endpoints and collections.
  - See: https://docs.globus.org/cli/collections_vs_endpoints/

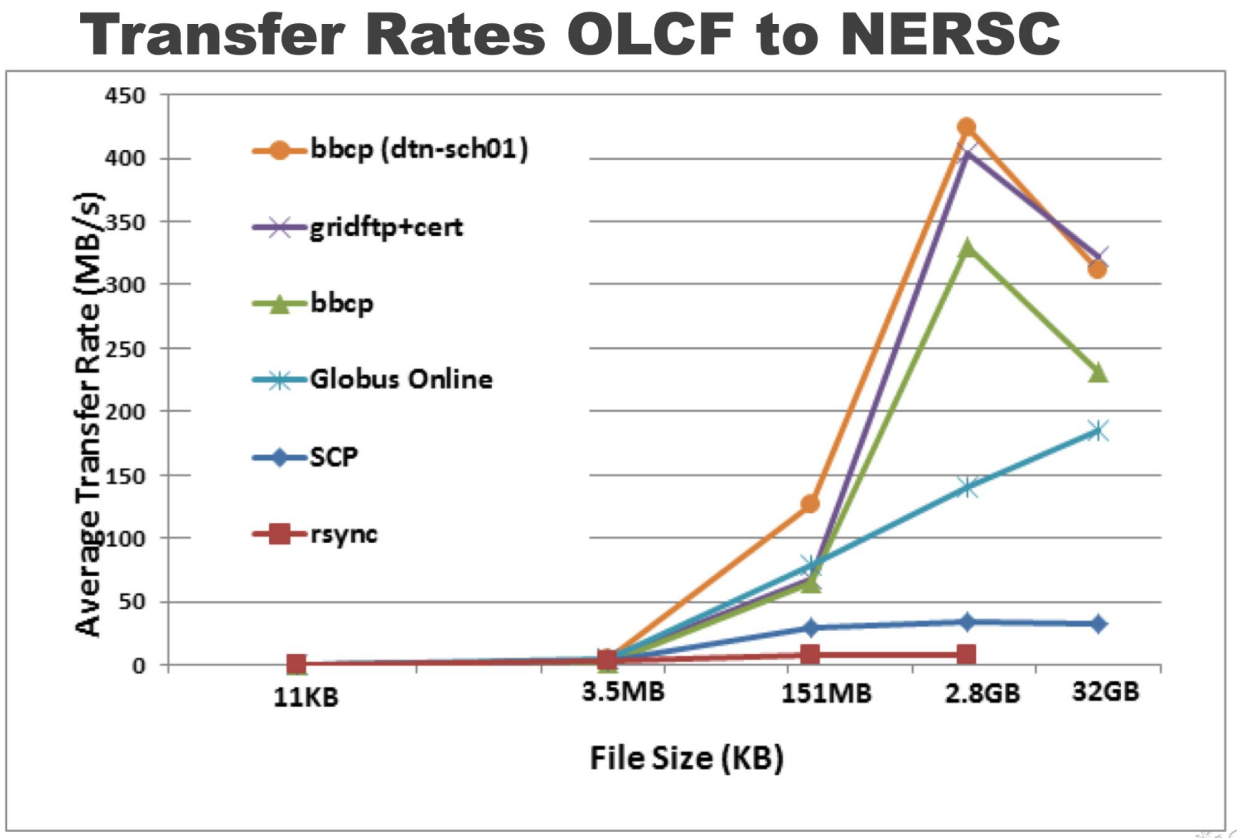
Globus Speed for different file distributions

Alpine to Orion Transfers 7-17-23

Files	Time (s)	Effective Speed MB/s
one 8.5G file	74	114
three 8.5GB files	103	249
a folder of 10 8.5 GB files	64	1037
1 85GB tar file	252	341

- Globus is a parallel transfer so it gives a faster transfer per byte for many small files at once.
- Unless you are transferring to/from HPSS -then send tar files for best results.

Globus Compared to other tools



Questions

Alpine (gpfs)

NFS (user/proj Home)

HPSS (archive)

DTNs (Data Transfer Nodes)

Orion (Lustre)

Thank you for your attention!