Transferring Your Data

Brandon Reyes

- *email: brre2566@colorado.edu*
- *RC Homepage: https://www.colorado.edu/rc*

Slides available on GitHub:

Summer_Camp_2023/Day_Two/Transferring_your_data/slides

Outline

- Ways to access your data
- Data transfer using the command line
- Data transfer using Open OnDemand
- Data transfer using Globus
- Sharing Data
- Getting A Petalibrary Allocation

Accessing Data on RC Resources

- When you use RC resources the data is not on your local machine
- Ways to access the data from your local machine
  - Command line (a variety of tools)
  - Open OnDemand (straightforward GUI interface)
  - Globus (GUI interface with some set up required)

Access through the Command Line

- If you don’t need a *fancy* GUI
- Provides a larger variety of tools
  - SCP
  - SFTP
  - RSYNV
  - RCLONE
  - SSHFS
  - SMB
- The tools provided can improve your data workflow (more on this later)

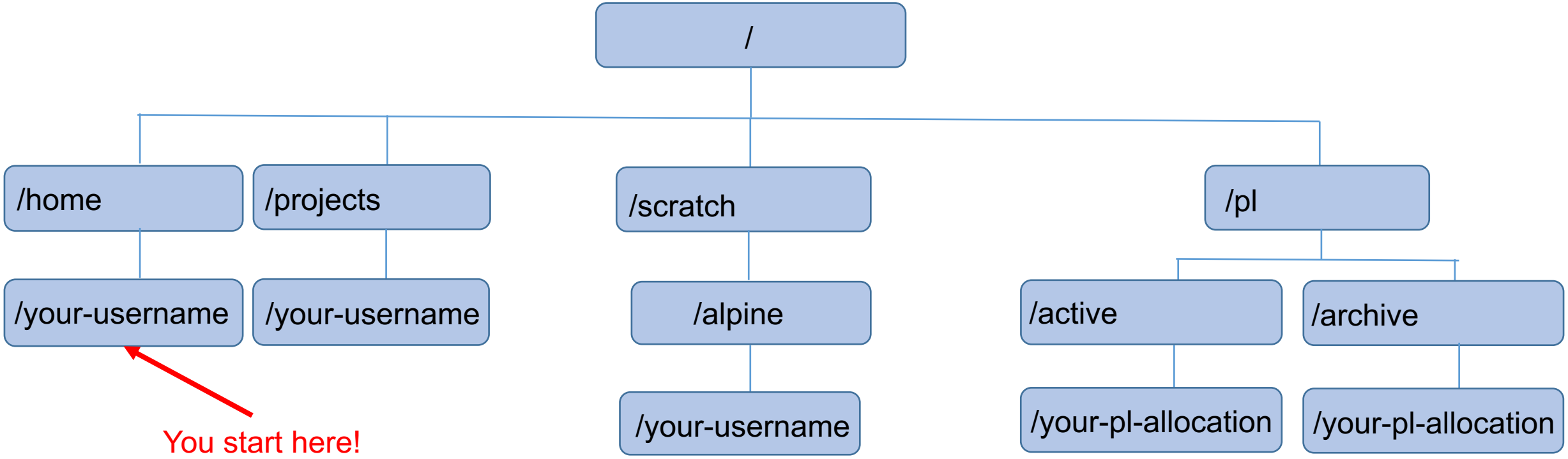
General Filesystem Structure

/home (2GB)	/projects (250GB)	/scratch/alpine (10TB)
• Small important data • Backed up frequently • Not for sharing files or job output	• Medium sized important data • Software • Can be shared with others • Backed up, but less frequently • Not for job output	• Large data • Can be shared with others • Fast Data transfer to compute nodes • Not backed up! • Purged after 90 days!
Filesystem documentation: https://curc.readthedocs.io/en/latest/compute/filesystems.html		

Let’s get on a login node!

```
ssh <your-username>@login.rc.colorado.edu
```

RC Filesystem Map



Basic Navigation Commands

- Change directories

```
cd <relative-or-full-path>
```

- List contents of a directory

```
ls <optional-path>
```

- Print current working directory

```
pwd
```

RC endpoints

Endpoint – one of the two file transfer locations i.e., it is either the source or the destination we want to copy data from or to.

- For data on RC resources, we have two endpoints
  - The login* nodes
    - Only use for small transfers!!

```
<your-username>@login.rc.colorado.edu
```

  - Data transfer nodes (DTNs)

```
<your-username>@dtn.rc.int.colorado.edu
```

    - CSU

```
<your-username>@dtn.rc.colorado.edu
```

RC Data transfer nodes (DTNs)

- Command line use of DTNs only available if you are on CU Boulder or CSU's network or VPN
- Dedicated nodes for transferring data
  - Faster transfers
  - More stable transfers
- Suitable for
  - Large and frequent transfers
  - Automated (passwordless) transfers
    - Only for CU Boulder folks
- Cannot ssh into the DTNs!

Command line option - SCP

SCP (Secure Copy Protocol) is a command line tool to transfer files/directories to, from, or between remote locations.

- Simple, but useful!
- Copying a local file to RC resources using a login node:

```
scp file1 <username>@login.rc.colorado.edu:<remote-path>
```

- Copying a directory from RC resources to local path via a DTN:

```
scp -r <username>@dtn.rc.int.colorado.edu:<path-to-directory> <local-path>
```

Command line option - SFTP

SFTP (Secure File Transfer Protocol) a command line tool that is similar to SCP, but provides an sftp session where both the local and remote filesystems are available

- Slightly more advanced than SCP
- Useful for multiple file/directory transfers
- Starting a SFTP session on a local machine

```
sftp <username>@login.rc.colorado.edu
```

- Demo time!

Command line options (2)

- rsync
  - `rsync` is a popular Linux utility for updating changed files to a remote filesystem.

```
rsync -v file1 <username>@login.rc.colorado.edu:<remote-path>
```

  - Useful when working on a file on both remote and local machines with modifications that need to be updated
  - Flags:

```
-v    # verbose mode
-r    # recursive (directory)
-t    # sync based off timestamp
-c    # sync changed files based on content
-a    # archive mode
```

Command line options (3)

- rclone
  - `rclone` is a command line program to manage files on cloud storage. It is a feature rich alternative to cloud vendors' web storage interfaces. Over 40 cloud storage products support rclone including S3 object stores, business & consumer file storage services, as well as standard transfer protocols. Rclone has powerful cloud equivalents to the unix commands rsync, cp, mv, mount, ls, ncdu, tree, rm, and cat.

```
rclone copy rclonetest.csv aws_s3:testbucket/
```

  - https://curc.readthedocs.io/en/latest/compute/data-transfer.html#rclone

Command line options (4)

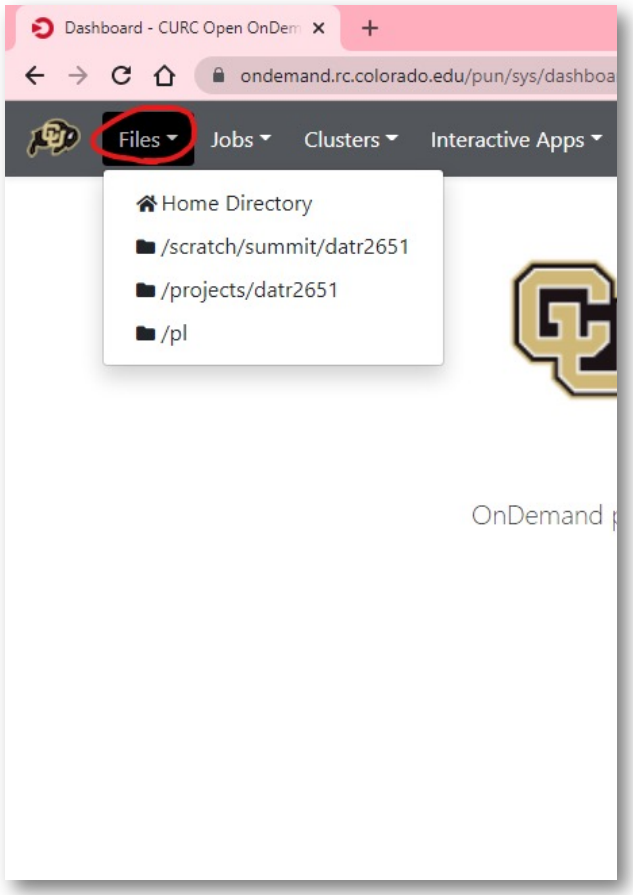
- sshfs
  - Mount a remote directory to a local Unix operating system!
  - Mac and Linux Exclusive:

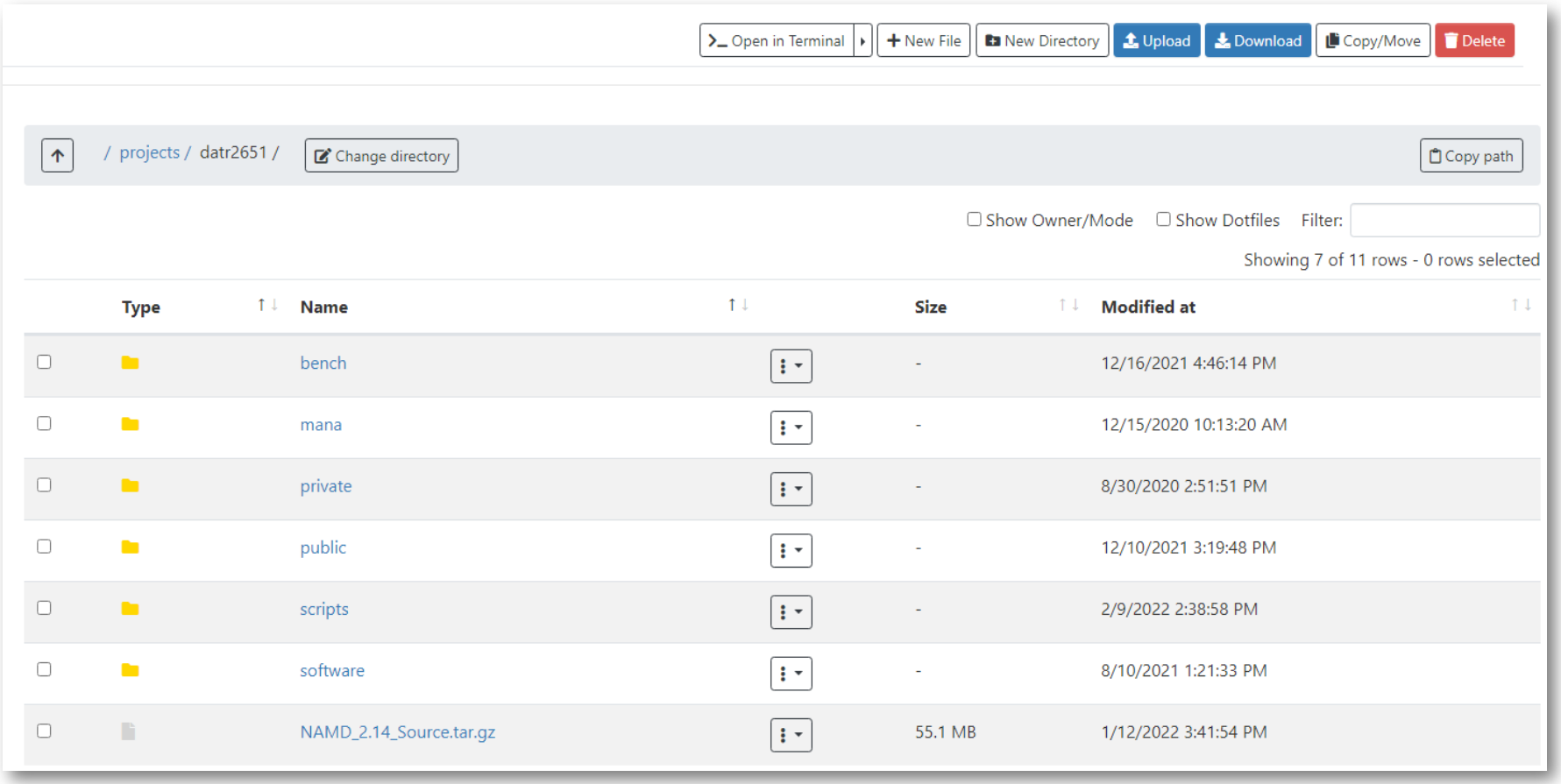
```
sshfs <username>@login.rc.colorado.edu:<path> <local-mountpoint>
```

- SMB mounting
  - Universal mounting protocol that is built into every operating system
  - Contact RC to get this set up!

OpenOnDemand

- GUI approach to RC Resources!
  - http://ondemand.rc.colorado.edu/
- File management
  - Create, Delete, Move, and Rename
- File transfers
  - Upload and Download
- Good for managing your files and conducting small file transfers without interacting with a command line.


Dashboard - CURC Open OnDem
ondemand.rc.colorado.edu/pun/sys/dashboa
Files
Jobs
Clusters
Interactive Apps
Home Directory
/scratch/summit/datr2651
/projects/datr2651
/pl
OnDemand


Open in Terminal
New File
New Directory
Upload
Download
Copy/Move
Delete
/ projects / datr2651 /
Change directory
Copy path
Show Owner/Mode
Show Dotfiles
Filter:
Showing 7 of 11 rows - 0 rows selected
Type
Name
Size
Modified at
bench - 12/16/2021 4:46:14 PM
mana - 12/15/2020 10:13:20 AM
private - 8/30/2020 2:51:51 PM
public - 12/10/2021 3:19:48 PM
scripts - 2/9/2022 2:38:58 PM
software - 8/10/2021 1:21:33 PM
NAMD_2.14_Source.tar.gz 55.1 MB 1/12/2022 3:41:54 PM

Globus

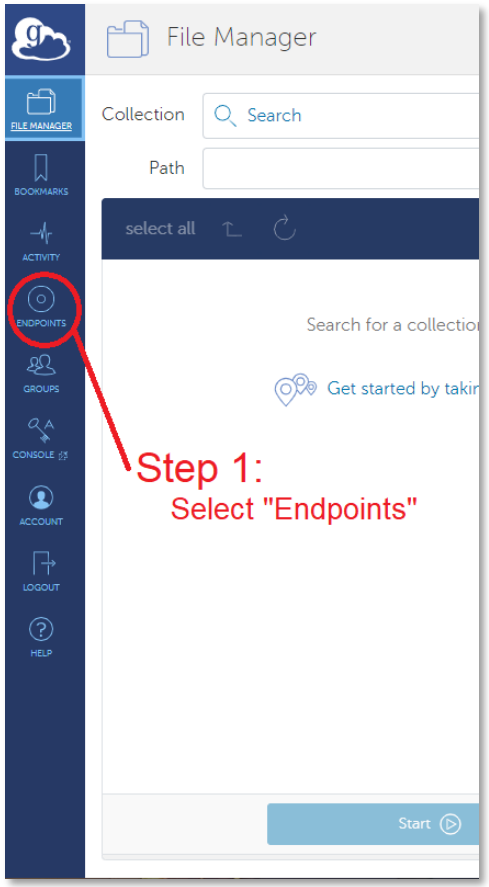
- Globus
  - By far the most stable and recommended way for data transfers
  - Fast transfers
  - Transfers continue if a user disconnects
  - Web GUI option or Globus Connect Personal
- Demo:

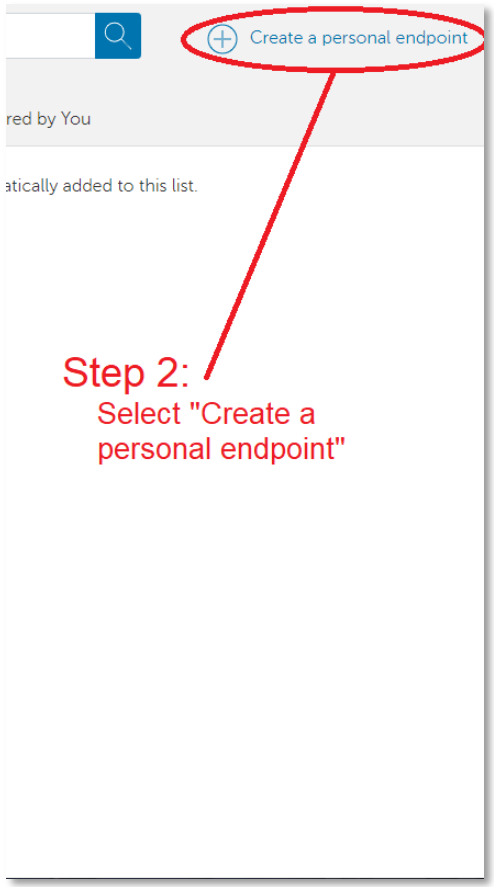
Globus Demo (1)

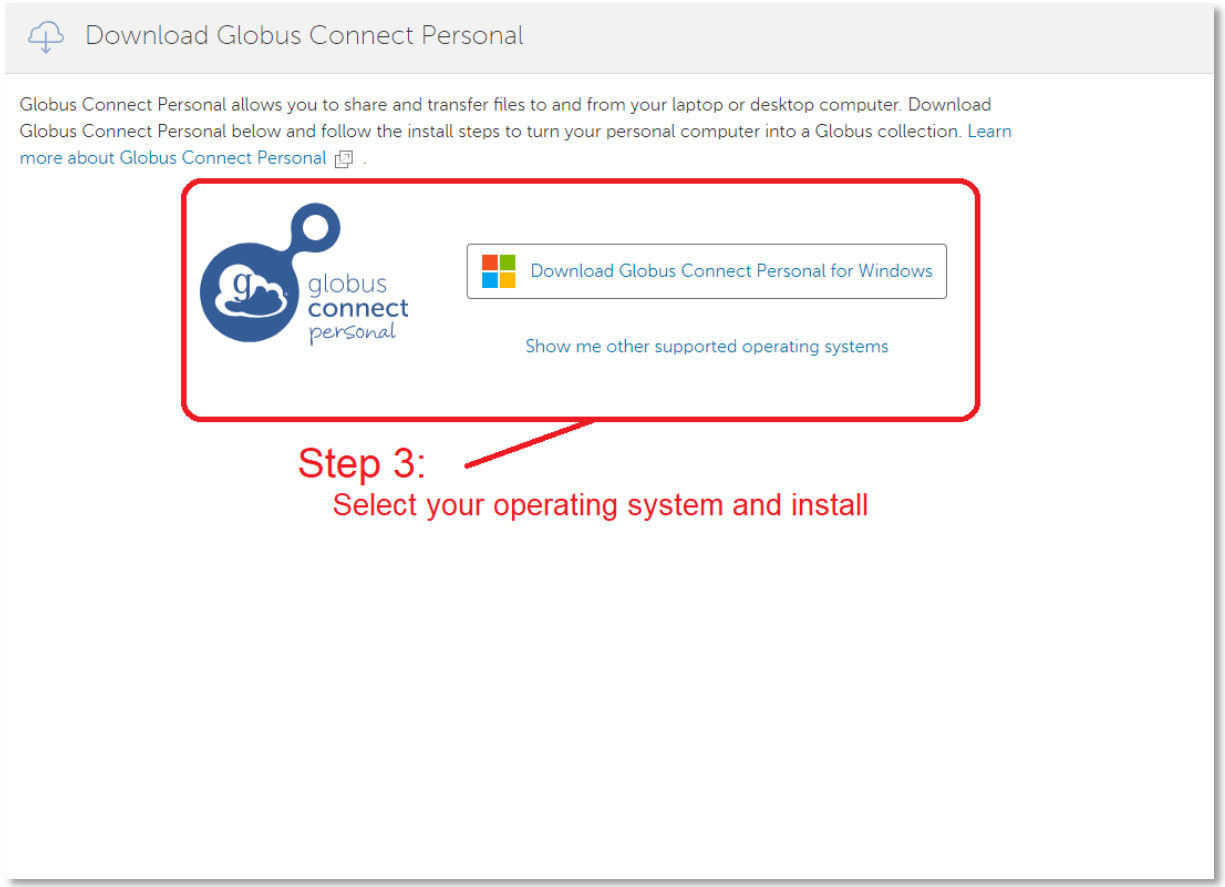
- Globus login is simple and quick: https://app.globus.org
  1. Select University of Colorado at Boulder under the dropdown menu
  2. Login with your CU credentials
  3. Continue with onscreen prompts until you are brought to the Globus WebGUI
- Installing a Globus Endpoint on your local machine
  1. Navigate down to Endpoints on the sidebar
  2. Click create an endpoint on the top right of the page
  3. Select your operating system and download the installer
  4. Follow the prompts on the installer and complete the installation

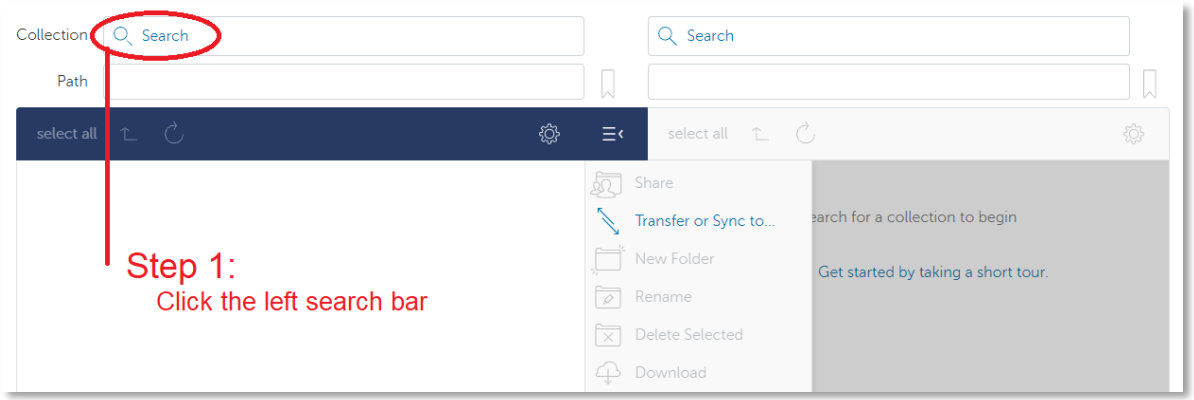
Globus Demo (2)

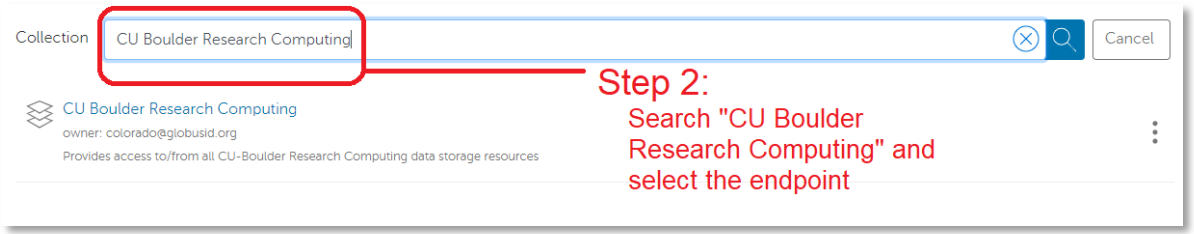
- Transferring Files can be done through the GUI
- From the File Manager tab:
  1. Click the “Two Panel” view button at the top right.
  2. Click the top left Search bar.
  3. Search “CU Boulder Research Computing” and select the end point.
  4. Sign into Research Computing’s Endpoint
  5. Click the right search bar
  6. On the ‘Your Collections’ tab, choose the endpoint you created
  7. Transfer your files!

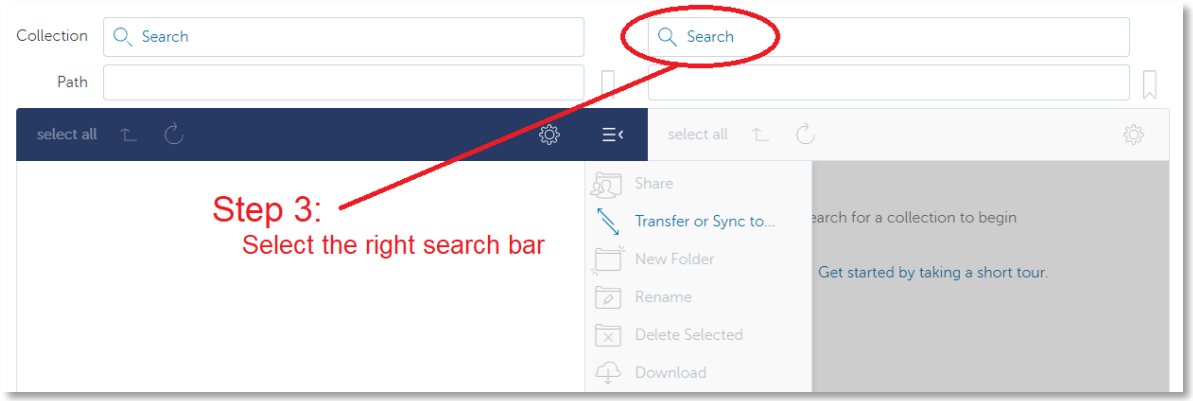

File Manager
FILE MANAGER
BOOKMARKS
ACTIVITY
ENDPOINTS
GROUPS
CONSOLE
ACCOUNT
LOGOUT
HELP
Collection
Search
Path
select all
Search for a collectio
Get started by takir
Start
Step 1:
Select "Endpoints"

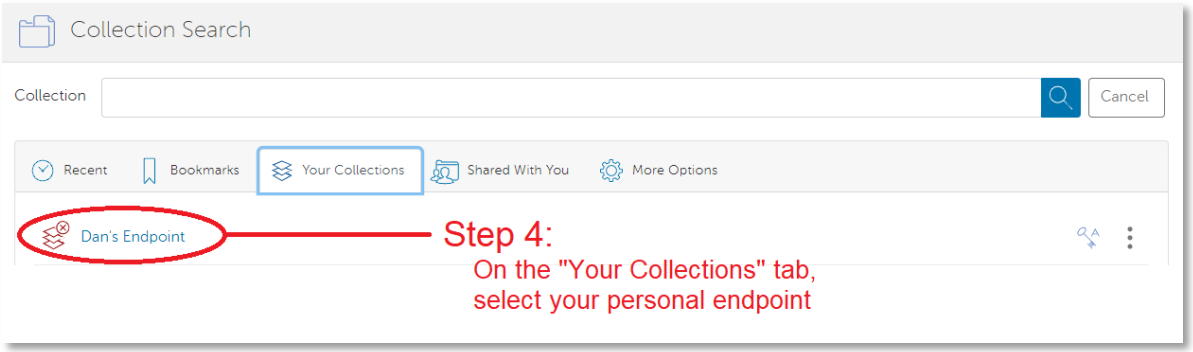

Create a personal endpoint
red by You
atically added to this list.
Step 2:
Select "Create a
personal endpoint"


Download Globus Connect Personal
Globus Connect Personal allows you to share and transfer files to and from your laptop or desktop computer. Download Globus Connect Personal below and follow the install steps to turn your personal computer into a Globus collection. Learn more about Globus Connect Personal .
globus
connect
personal
Download Globus Connect Personal for Windows
Show me other supported operating systems
Step 3:
Select your operating system and install


Collection
Search
Path
select all
Share
Transfer or Sync to...
New Folder
Rename
Delete Selected
Download
Get started by taking a short tour.
Step 1:
Click the left search bar


Collection
CU Boulder Research Computing
Cancel
CU Boulder Research Computing
owner: colorado@globusid.org
Provides access to/from all CU-Boulder Research Computing data storage resources
Step 2:
Search "CU Boulder Research Computing" and select the endpoint


Collection
Search
Path
select all
Share
Transfer or Sync to...
New Folder
Rename
Delete Selected
Download
Get started by taking a short tour.
Step 3:
Select the right search bar


Collection Search
Collection
Cancel
Recent
Bookmarks
Your Collections
Shared With You
More Options
Dan's Endpoint
Step 4:
On the "Your Collections" tab, select your personal endpoint

Sharing Data

- Other RC Users
  - To share files that you own with other RC users, contact RC with a list of users you would wish to allow access
    - RC will place the chosen users in the owner's group
    - The owner can then set up permissions in the space
  - On-premise collaborators can also access Petalibrary files with Globus Shared Endpoints
- Off-premise collaborators
  - Off-premise collaborators can only access Petalibrary files through Globus Shared Endpoints

Unix Groups

- Unix Groups
  - 3 Levels of permissions:
    - User
    - Group
    - Other
  - All users have a group associated with their username
  - Permissions can be set for an individual file with the `chmod` command

```
chmod g+rx file.exe
```

Globus Shared Endpoints

- Globus offers ‘shared endpoints’ which don’t require a user to have an account with RC.
- RC provides this capability for easy access of Data.
- Petalibrary exclusive!
- Generates a shared collection that can be accessed with a link.
  - Can assign various permissions to specific users or all users withing Globus
  - More information on here: https://docs.globus.org/how-to/share-files/

Data Publishing with Petalibrary

- Using Globus shared endpoints can be a great way to publish your data while maintaining the convenience of having it Petalibrary.

Example: https://scholar.colorado.edu/concern/datasets/9593tw13k

Petalibrary Notes

- *curc-quota* – Research Computing tool to monitor disk usage.
  - Provides detailed summary of your core storage
  - Provides detailed summary of scratch space on compile and compute nodes
  - Also lists current capacity of all Petalibrary allocations you have access to

```
[userXXXX@login12 ~]$ curc-quota
```

- .cstats- usage statistics file for an allocation

```
cat /pl/active/<allocation_name>/.cstats
cat /pl/active/rcops/.cstats
```

Note: Confidential Data is unsupported and *should not be stored on Petalibrary*!

Thank you!

- Please fill out the survey:
- Contact information: rc-help@colorado.edu