Transferring your data to & from CU Boulder Research Computing

August 13, 2024

Andy Monaghan
Contributors: Brandon Reyes, Layla Freeborn, Trevor Hall

Research Computing
UNIVERSITY OF COLORADO BOULDER

Learning Objectives

- Ways to access your data
- Filesystem structure of CURC resources
- Data transfer with OnDemand “Files” application
- Data transfer using Globus
  - Web-based interface
  - Local Globus endpoint
  - Connecting local and CURC endpoints

Acronyms and Terms

- Acronyms
  - HPC – “High Performance Computing” (“supercomputing”)
  - CURC – “CU Boulder Research Computing”
  - RMACC – “Rocky Mountain Advanced Computing Consortium
  - GUI – “Graphical user interface”

- Terms
  - OnDemand – Browser-based gateway to CURC resources
  - Globus – Browser based file transfer service
  - Endpoint – the source or destination of a specific Globus file transfer (“Collection”)
  - PetaLibrary – CU’s file storage service for large data sets
  - ACCESS -- National Science Foundation (NSF) supercomputing Program
  - XSEDE – Previous NSF supercomputing program

Accessing Data on CURC Resources

- When you use CURC resources the data is not on your local machine.
- Ways to access/transfer the data from/to your local machine:

  - Command line (a variety of tools – we won't cover today)
    - https://curc.readthedocs.io/en/latest/compute/data-transfer.html
  - **Open OnDemand (straightforward GUI) – files < 1 GB**
  - **Globus (GUI with some set up required) – any file size**

General Filesystem Structure

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

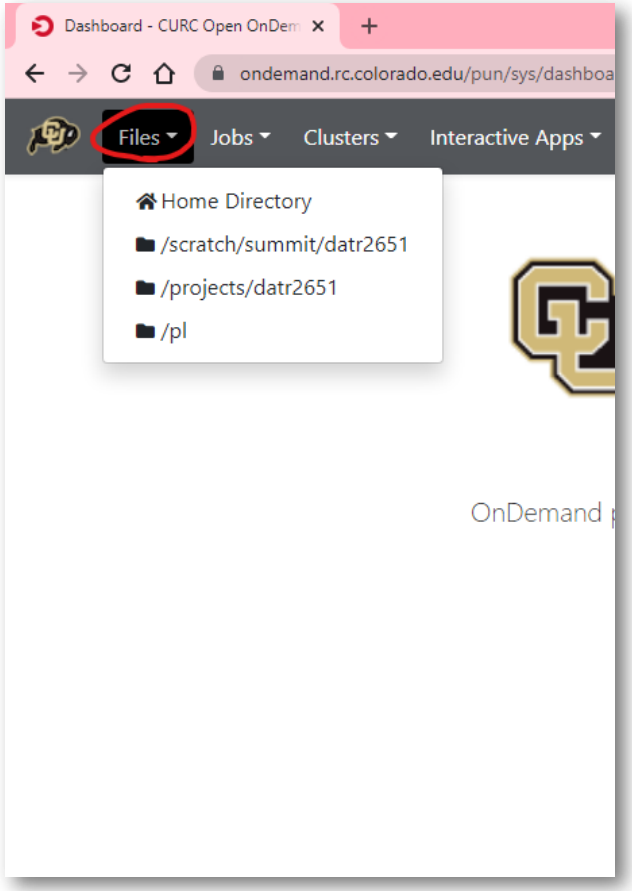
PetaLibrary is also available via Globus!

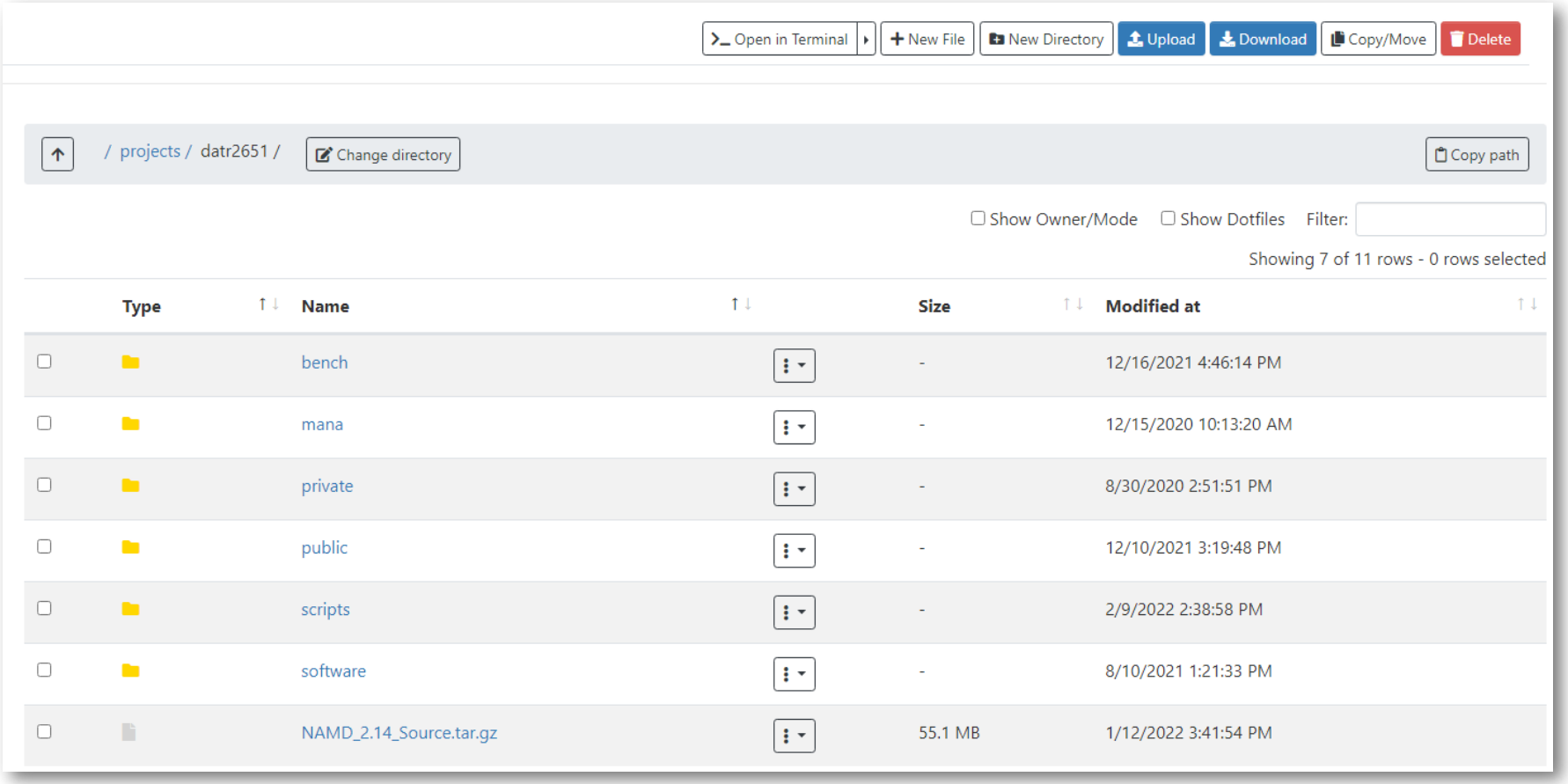
Filesystem documentation: https://curc.readthedocs.io/en/latest/compute/filesystems.html

GUI option - Open OnDemand

- No command line required!
  - http://ondemand.rc.colorado.edu/
  - http://ondemand-rmacc.rc.colorado.edu/
- File management
  - Create, Delete, Move, and Rename
- File transfers
  - Upload and Download

Open OnDemand “Files” Interface





GUI option - Globus

Globus is a service that allows for users to reliably move, share, and discover data

- Command line version is also available
- Our recommended way to transfer data
  - Stable and fast data transfers
  - Transfers continue if a user disconnects
  - Web GUI or Globus Connect Personal GUI
- Supported on all major operating systems
  - Works well with cloud storage providers

Globus Demo

- Logging into the web-based interface for Globus
  - Provides a nice GUI for managing files on CURC resources
- Installing a Globus Endpoint on your local machine
  - Allows you to interact with files on your local machine via Globus

Globus Web-based interface

Globus login is simple and quick

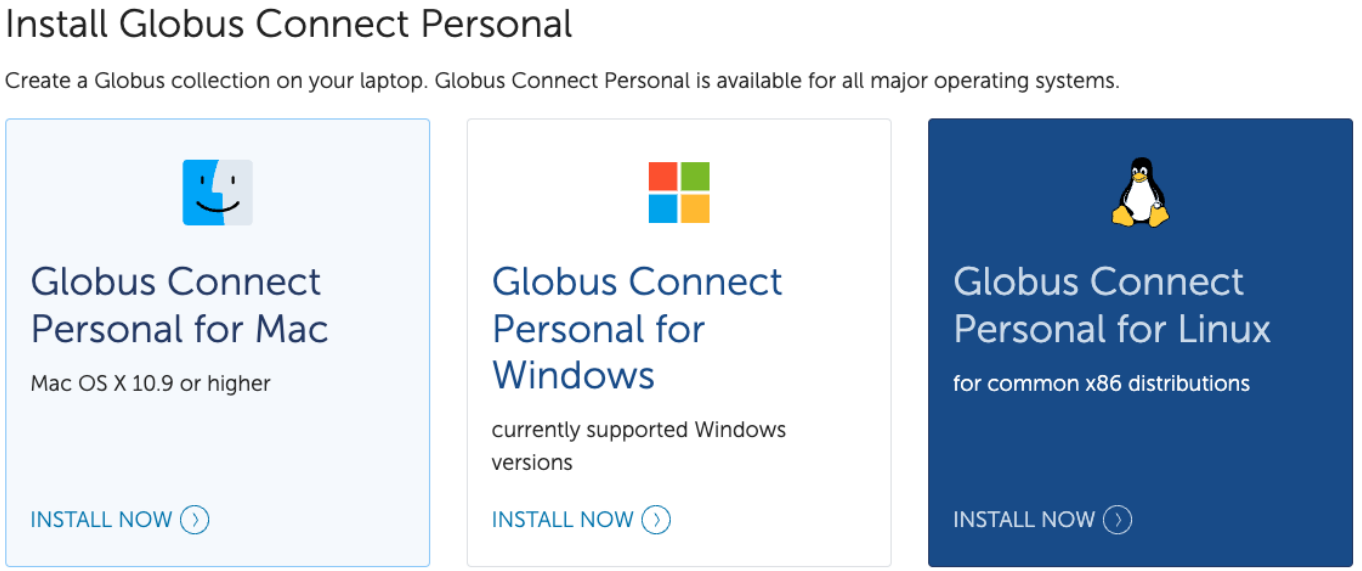
1. Navigate to https://app.globus.org
   - CU Boulder users
     - Select “University of Colorado at Boulder” in the dropdown menu
   - CSU users
     - Select “Colorado State University”
   - AMC and RMACC users
     - Select “ACCESS CI (formerly XSEDE)”
2. Login with your credentials
3. Continue with onscreen prompts until you are brought to the Globus Web GUI

Globus Web GUI

Let’s take a look!

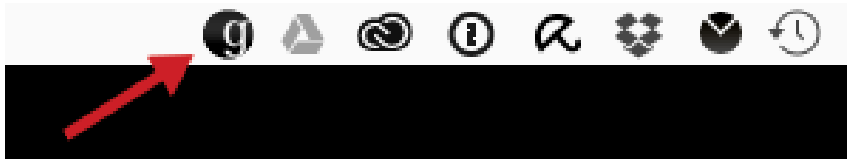
Local Globus Endpoint

- Required if you want to transfer data between your machine and CURC resources
- Navigate to https://www.globus.org/globus-connect-personal



Local Globus Endpoint-Installed

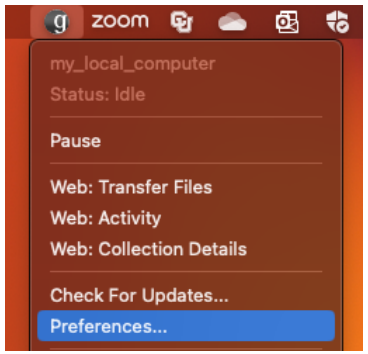
- For “Collection Name” you can use anything name you would like
  - You will use this as a reference name later, so please remember it!
- For Mac and Windows, it is easy to see that Globus has been installed:



- For Linux, at the end of the installation, the setup will exit with a message stating it was successfully set up and the main Globus Connect Personal application will launch

Local Globus Endpoint-Configure

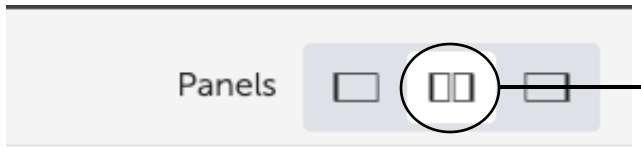
- Once the Globus endpoint has been installed on your computer, you then need to configure it.
- On Mac and Windows follow the “Configuration” section
- On Linux follow the “Running” section
- During Configuration you will need to let Globus know what directories or files it can have access to



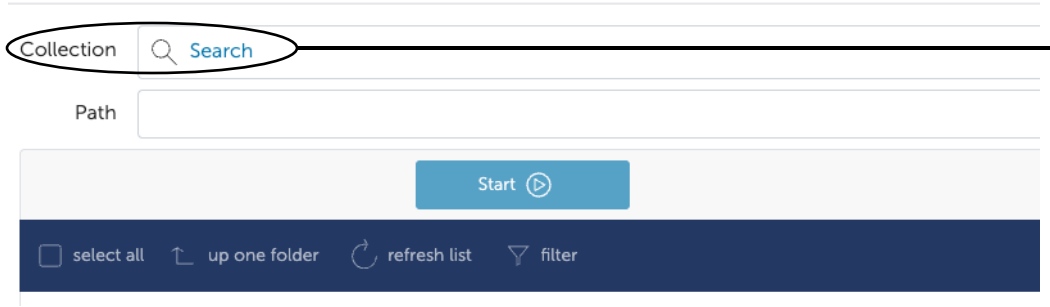
Select the “Access” tab and add folders you want to be accessible by Globus

Connecting local endpoint to CURC resources

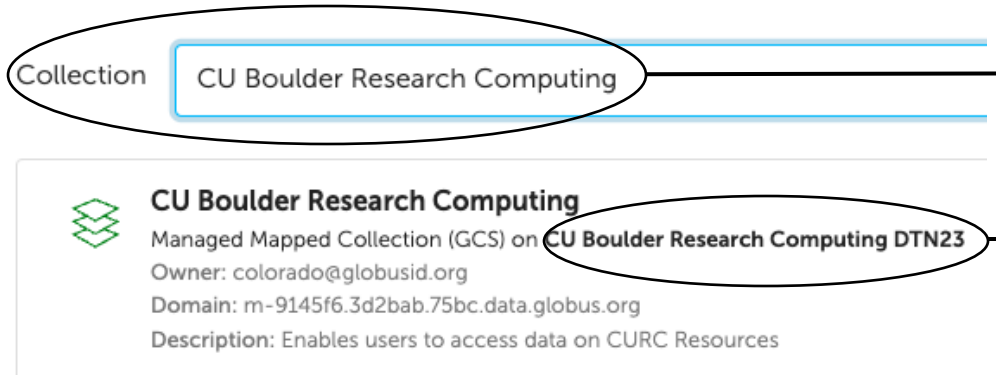
Globus



Step 1: Select “set two pane” in the upper right corner



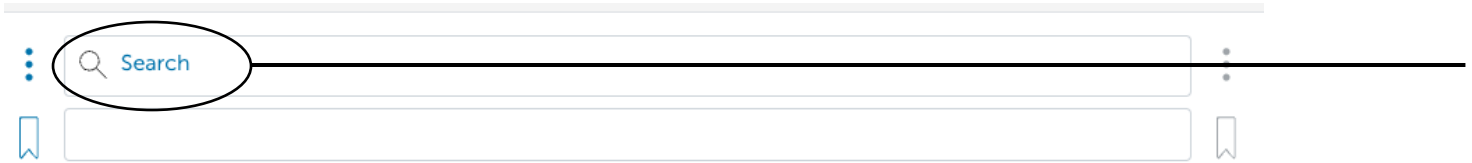
Step 2: In the left pane click the search bar



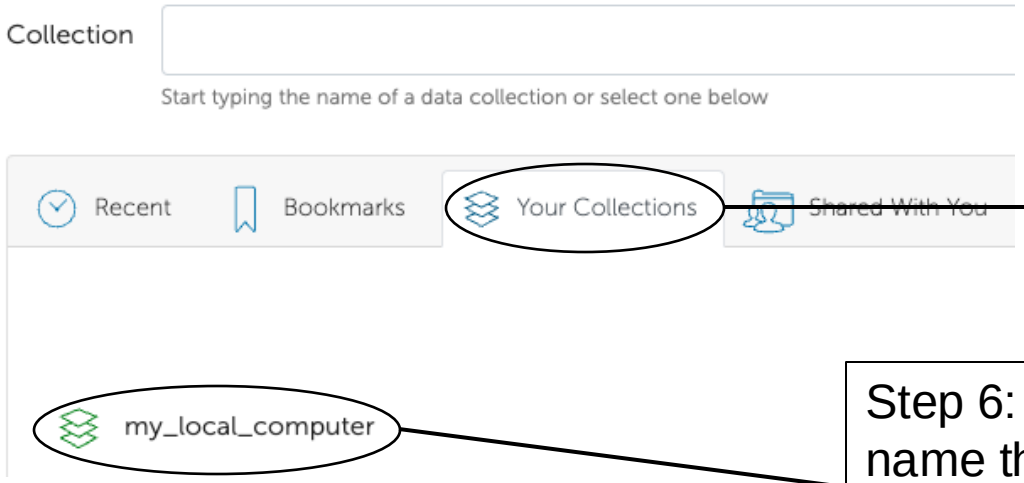
Step 3: Search “CU Boulder Research Computing”

Make sure you are using DTN23!

Globus



Step 4: In the right pane, click on “Search”



Step 5: Select the “Your Collections” tab

Step 6: Click the collection name that **you** chose when you created your local Globus endpoint

Advanced topic: Globus Shared Endpoints

- Globus offers ‘shared endpoints’, which allow you to share your data with external collaborators (i.e. you can share data with people who don't have CURC accounts)
- CURC provides this capability, however, it is only available for PetaLibrary customers
- Shared Endpoints generate a shared collection that can be accessed with a link
  - For an example, see https://scholar.colorado.edu/concern/datasets/9593tw13k
  - You can assign various permissions to specific users or all users within Globus
  - More information is provided at https://docs.globus.org/how-to/share-files/

Review of what we have covered

- Ways to access your data
- Filesystem structure of CURC resources
- Data transfer using Globus
  - Web-based interface
  - Local Globus endpoint
  - Connecting local and CURC endpoints

Questions?

- Website: www.rc.colorado.edu
- Documentation: https://curc.readthedocs.io
- Helpdesk: rc-help@colorado.edu