The most important resource on Blue Waters: Web Portal (bluewaters.ncsa.illinois.edu) user guide:

1. Mouse over

2. Click on “User Guide”
Don’t waste time figuring stuff out; submit a ticket

- Send email to help+bw@ncsa.Illinois.edu
- OR submit through the portal
- Don’t spend more than a day working on something.
  - Maybe even no more than half a day
Data Management on Blue Waters

- Where data lives on Blue Waters
  - Lustre
  - Nearline (tape) (granularity)
- Getting data on/off Blue Waters
  - Globus (GUI, CLI)
- Running jobs
- Archiving data to Nearline
  - (if you HAVE to)
- Retrieving data from Nearline
  - Preparing data for outside transport
  - DELETING data OFF of Nearline
- Pushing data off of Blue Waters
Questions about the process

- What questions do I need to find answers to in order to do this task effectively?
- Documentation may have some answers
- My workflow may CHANGE some of the answers
Players in data movement and layout

- **login nodes (3)**
- **Online (mounted):**
  - /scratch
  - /projects
  - /u (home)
- **compute nodes**
- **Nearline (tape) file systems:**
  - /projects,
  - /~ (home)
- **ie mover nodes (64)**
- **outside**
- **MPI app**
During your Blue Waters work:

- **Nearline (tape) file systems:**
  - `/projects`
  - `/~/ (home)`

- **Online (mounted):**
  - `/scratch`
  - `/projects`
  - `/u (home)`

- **Login nodes (3)**
- **IE mover nodes (64)**
- **Compute nodes**

Data flows from outside to login nodes, then to compute nodes and MPI app. Data is also transferred between nearline and online storage.
When your Blue Waters work finishes

- **login nodes**: (3)
- **compute nodes**
- **Online (mounted)**:
  - /scratch
  - /projects
  - /u (home)
- **Nearline (tape)**
  - file systems:
    - /projects,
    - /~ (home)
- **outside**
- **MPI app**
- **ie mover nodes**: (64)
Where data lives:
Blue Waters file system topology

• Online Lustre (disk) volumes (mounted on login, MOM, compute nodes, accessible via Globus)
  • home directory
  • /projects
  • /scratch

• Nearline (tape) volumes (accessible via Globus only)
  • home directory (distinct & separate from online home)
  • /projects (distinct & separate from online projects)*
Lustre

- All mounted file systems are on Lustre (home, /projects, /scratch)
- Every file has a “stripe count”
Lustre

- All mounted file systems are on Lustre (home, /projects, /scratch)
- Every file has a “stripe count"
- Striping is MANUAL
What is file striping in Lustre?

stripe count 1 file

OST
OST
OST
OST

stripe count 2 file

OST
OST
OST
OST
OST
OST
OST
How do I set stripe count?

- lfs setstripe --c 4 file_to_set.dat
- lfs setstripe --c 4 /dir/to/set/
Lustre general striping rules

• (BW /scratch): At least one stripe per 10-100 GB of ultimate file size to spread the files among many OSTs
  • (remember—stripe is fixed once the file is created and cannot be changed without copying the file)
• Match access patterns if you can (see section on application topology)
• With all that, pick the smallest stripe count that matches everything else
Stripe Count Inheritance

- A file’s stripe count is permanent
- A file inherits the stripe count from the containing directory AT CREATION TIME
  - You can use “touch” to set a file’s stripe characteristics before it’s created
- mv PRESERVES a file’s stripe characteristics
- the only way to change a file’s stripe count is to COPY it to a new file (first making sure the target file has the correct characteristics)
Lustre striping questions

• How big are my files?
• How many ranks will be writing to output files at the same time?
• Can I arrange files to help striping considerations (big files in different directories than small files)
Online $\rightarrow$ Nearline
(mostly don’t do this on BW any more)

- Both act like file systems, copy files with Globus GUI or Globus CLI

- HOWEVER:
  - Many small files store easily at the end of tapes
  - your file collection becomes fragmented
  - retrieval (copying from Nearline $\rightarrow$ Online) must mount dozens or hundreds or more tapes; very slow or impossible
Moving data between Online and Nearline (data granularity is CRITICAL; next slide)

- **Online**
  - /scratch
  - /projects
  - /u (home)

- **Nearline**
  - file systems
  - /projects,
  - /~/ (home)

- **ie mover nodes** (64)

- **login nodes** (3)

- **compute nodes**

- **MPI app**

- **Globus Control**

- **outside**
Data Granularity is CRITICAL for successful use of nearline

- Nearline (tape) has a virtual file system; it *acts* like a disk file system
- BUT
- Files are grouped onto tapes to maximize storage efficiency and COMPLETELY IGNORES considerations for retrieval efficiency
- Very many files and/or very small files tend to fragment your file collection across dozens or hundreds of tapes
Package files BEFORE moving to Nearline

- Moving off-site is BETTER (given short remaining life of Blue Waters)
- Delete Nearline data AS SOON as you’re done with it (good in general, critical for Blue Waters)
How to tar (or otherwise package) files and directories

• You can use tar in a one-node job script
• Example job script:
  #!/bin/bash
  #PBS stuff
  aprun –n 1 tar cvf /path/to/archive.tar /path/to/target/dir/
Getting data on (and off) Blue Waters

- Use Globus
  - Good!
    - Asynchronous
    - Parallel
    - Free auto-retries
  - HOWEVER
    - Errors are ignored; you must monitor
    - You must maintain access credentials
Monitoring Globus

• Periodically look at AVERAGE TRANSFER RATE of your transfers
Long-distance file copying via Globus

- Transfers files in “chunks” of 64 files at a time (regardless of size)
- Groups of small files transfer very slowly because of Globus transfer latency
- Transfer data in larger files, or package (or tar) small files into larger archive files BEFORE transferring over network
Data Ingest to Blue Waters: Use Globus; data movement by dedicated mover nodes

- **User data**: ie mover nodes (64)
  - **Online**: /scratch /projects /u (home)
  - **Nearline file systems**: /projects, /~/ (home)

**Login nodes (3)**

**outside**

**compute nodes**

**MPI app**

**Globus Control**
Questions to ask about long-distance data transfers

• How big of files is my data grouped in NOW?
• What file size range is reasonable in its current location?
• What file size range is reasonable at its destination? (is that the same as previous question?)
• What file size range will transfer most quickly?
Blue-Waters-specific questions

• Are my files less than 10 GB?
• Do I have more than 1000 files to transfer?
• (if either is yes, maybe re-group files)
Transfer overview page that covers Globus
https://bluewaters.ncsa.illinois.edu/data-transfer-doc
Getting to Globus GUI

1. Mouse over

2. Click on “Data”
Getting to Globus GUI

Data Transfer

Click

Moving Data

NCSA recommends using Globus Online for data transfer. Currently, there are two endpoints supported: ncsa#BlueWaters and ncsa#Nearline. The ncsa#BlueWaters endpoint transfers data to the Blue Waters Lustre file system and the ncsa#Nearline endpoint accesses the Blue Waters Nearline storage system. For each endpoint, the Globus Online environment will choose the “least busy” server for each transfer to distribute the data movement load so as to move the data as efficiently as possible. The Nearline storage system automatically migrates data to a tape subsystem. Use the ncsa#BlueWaters endpoint to transfer data into and out of your home and project spaces. Use the ncsa#Nearline endpoint to transfer data to/from the Nearline tape system.

Helpful Tips

- Do not use scp/sftp/globus-url-copy on the login nodes to transfer data. Performance will be far below that of Globus Online and the network load will impact other users.
- Use Globus Connect if you need to transfer files with your local office machine.
Globus GUI
Farther down: Globus Python-based CLI

Command Line Interface (CLI)

The ssh-based CLI was deprecated August 1, 2018. We are packaging the globus-cli on Blue Waters as guided by Globus. Please check back for updates.

Self-deployment of Python-based Globus CLI

To self-deploy the new Python-based Globus CLI, follow these 7 steps:

1. module load bqpy
2. virtualenv ~/globus-cli-virtualenv
3. source ~/globus-cli-virtualenv/bin/activate
4. pip install globus-cli
5. deactivate
6. export PATH="$PATH:~/globus-cli-virtualenv/bin"
7. globus login

You are now logged into Globus and can initiate queries and transfers. For example find the Blue Waters endpoint:

```bash
> globus endpoint search BlueWaters
ID | Owner | Display Name
---|-------|-------------
0f9900ef-6004-11e5-ba40-220000092c6c | ncsa@globusid.org | ncsa@BlueWaters
```

You can then activate an endpoint and list content:

```bash
> globus endpoint activate 0f9900ef-6004-11e5-ba40-220000092c6c
The endpoint could not be auto-activated.
```

This endpoint supports the following activation methods: web, oauth, delegate proxy.
python/Globus CLI (see portal)

- scriptable
  
  usage example:
  
  module load bwpy
  virtualenv "${HOME}/.globus-cli-virtualenv"
  source "${HOME}/.globus-cli-virtualenv/bin/activate"
  pip install globus-cli
  deactivate
  export PATH="$PATH:${HOME}/.globus-cli-virtualenv/bin"
  globus login
  globus endpoint activate d59900ef-6d04-11e5-ba46-22000b92c6ec
  globus ls -l d59900ef-6d04-11e5-ba46-22000b92c6ec:${HOME}

Please see https://docs.globus.org/cli/ for more commands and examples
new BW wrapper for python/Globus (forthcoming)

```
python transferHelperInstaller.py
export PYTHONPATH=/path/to/python/helper
ipython

import globusTransferHelper
hlp=globusTransferHelper.GlobusTransferHelper()
hlp.<TAB>
```

(lists function completions)

```
BWkey=hlp.EP_BLUEWATERS
hlp.ls(BWkey, <path>)
```

• will live here:

https://git.ncsa.illinois.edu/bw-seas/globustransferhelper
Globus accounts (no matter how you access Globus)

- You will have one Globus account
- You will *link* that Globus account to any organizational account that you need write access to ("NCSA" for Blue Waters)
- From then on you can log into Globus using just the linked account credentials
Globus Endpoints

- Globus transfers files between “endpoints”
- permanent endpoints:
  - ncsa#BlueWaters (for BW Online File Systems)
  - ncsa#Nearline (for BW Nearline tape system)
  - XSEDE TACC stampede2
- You can create temporary Globus endpoints with “Globus Connect Personal” for transferring data to/from personal machines
Tools to NOT use on login nodes for data staging on and off BW

• rsync
• tar
• scp
• sftp

• on the login nodes are ok….for SMALL directories of code that take a short time to download

• login nodes are SHARED resources. Beating up a login node spoils that login node for many other people too.
Why sftp, ftp, scp use shared resources on logins and slow things down for everyone
Running Your Jobs: data best practices

• Read and write to /scratch
  • hundreds of OSTs (as opposed to dozens for /projects and home)
  • Much larger and more capable file system metadata server than /projects or home
Running jobs: Data Access Patterns

- N ranks, 1 file, 1 reader/writer (file contents distributed via MPI)
- N ranks, N files, N reader/writers: each rank reads/writes its own file
  - this is Ok up to medium scale
  - slows down at large scale
- N ranks, 1 file, N readers/writers: ranks write to one file with offset
  - manually manage writing stride, OR
  - IO libraries: HDF, netcdf
Scale limits for large simulations

• as one-file-per-rank simulations scale up, they may hit limits for the maximum number of files to have open
• as one-file-many-ranks simulations scale up, they may hit effective limits on file locking
Questions for large code runs

• How many files does my code read/write?
• Are the inputs and outputs on appropriate file systems, and are those directories configured appropriately?
• Have I revisited these questions after increasing scale/run length/file size?
Specific hint for Blue Waters → TACC

• NCSA and TACC want you to be able to move your data efficiently
• There are knobs to turn and buttons to push to make transfers faster and more efficient
• For that help to apply to YOUR transfers, you must specifically ask for help (open a ticket)
If it’s not working, if you can’t figure out it, if you’re confused--

• SUBMIT A TICKET!
  • Ask questions. We may know a quick clever solution