BLUE WATERS

SUSTAINED PETASCALE COMPUTING

June 6, 2019

Data Management and Best Practices for Data Movement

Craig Steffen

BW SEAS (User Support) Team

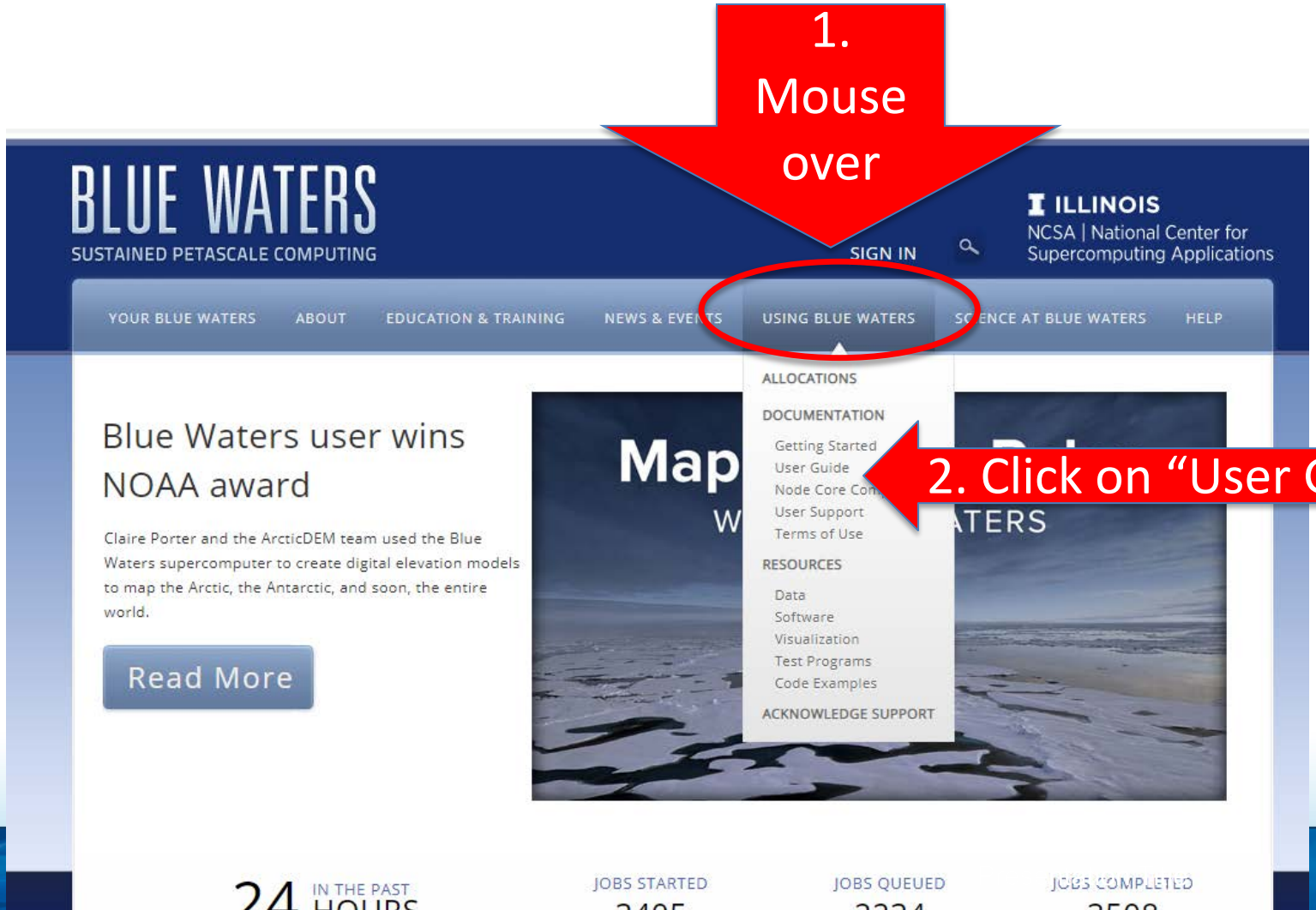
NCSA

NSF

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

CRAY

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

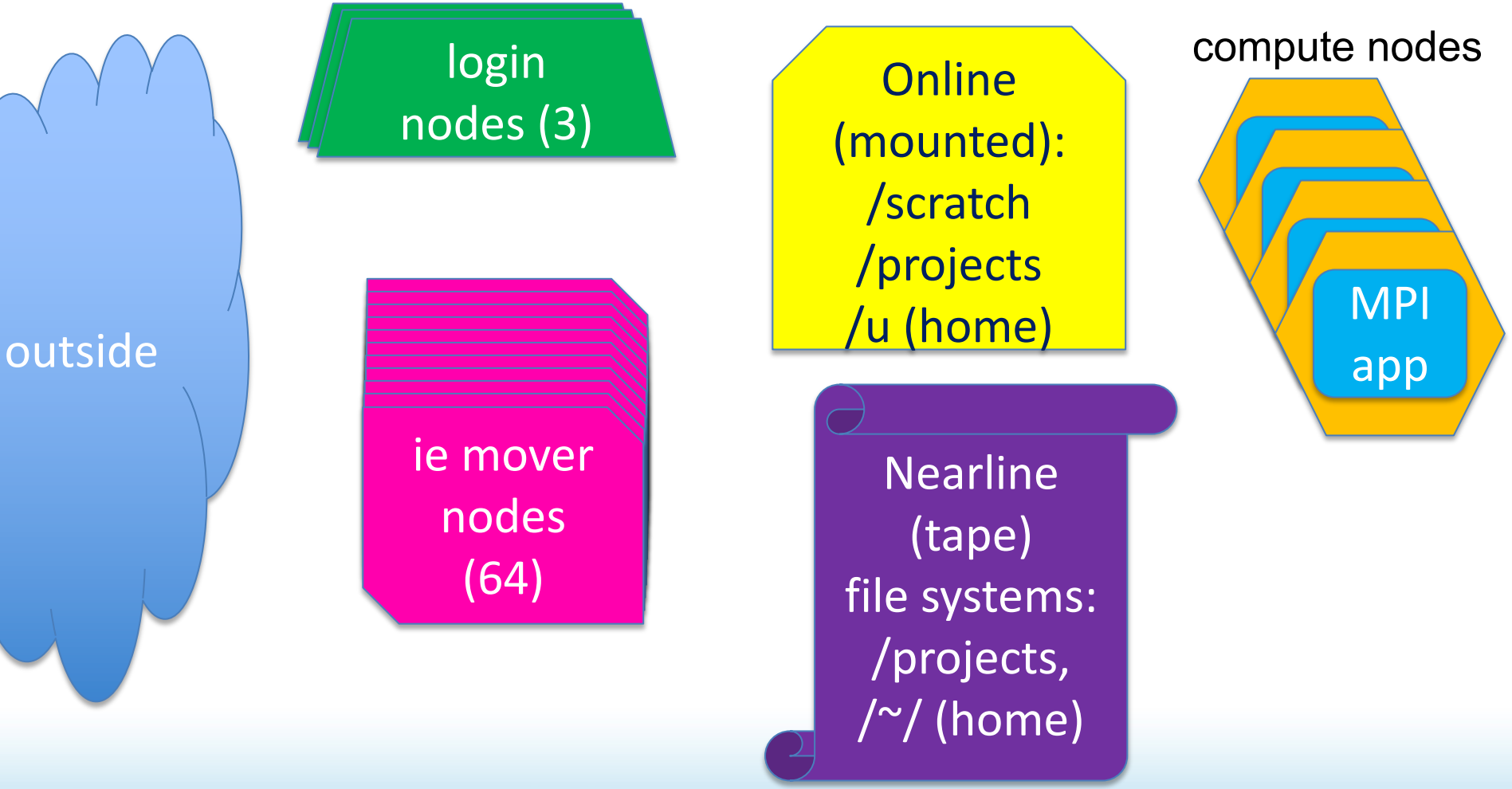
Data Management on Blue Waters

- 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

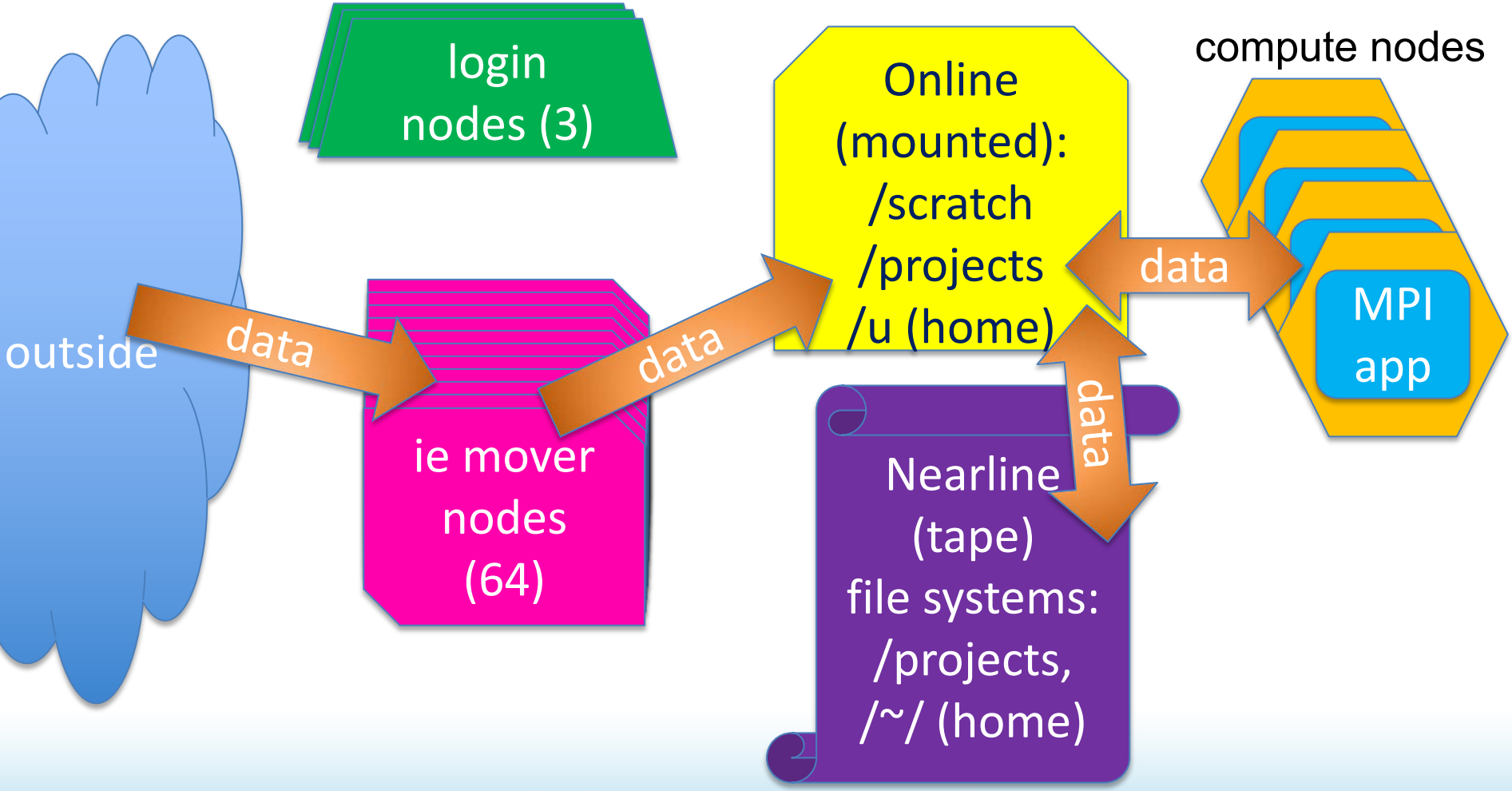
Questions about the process

- 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

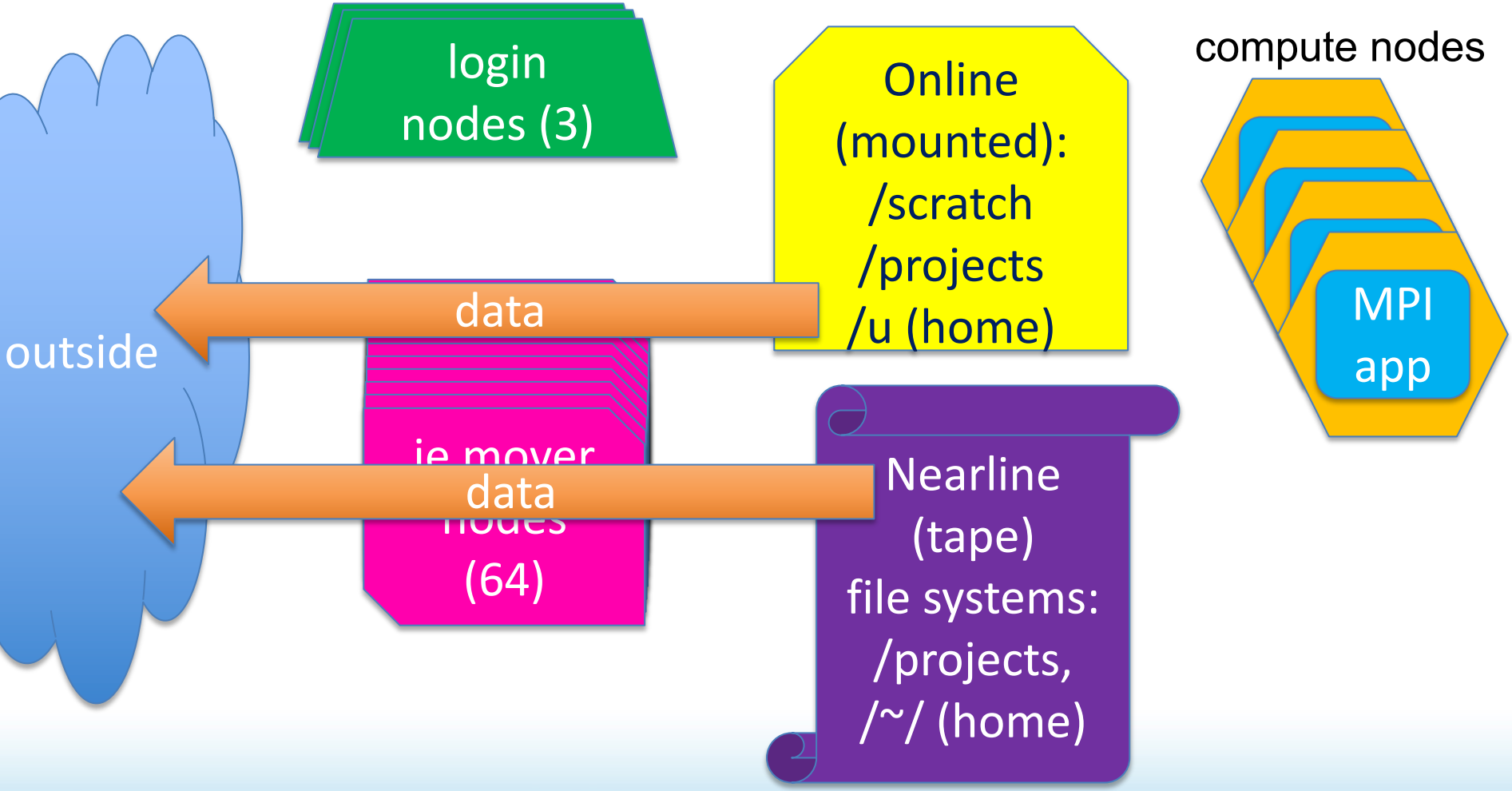
Players in data movement and layout



During your Blue Waters work:



When your Blue Waters work finishes



Where data lives: Blue Waters file system topology

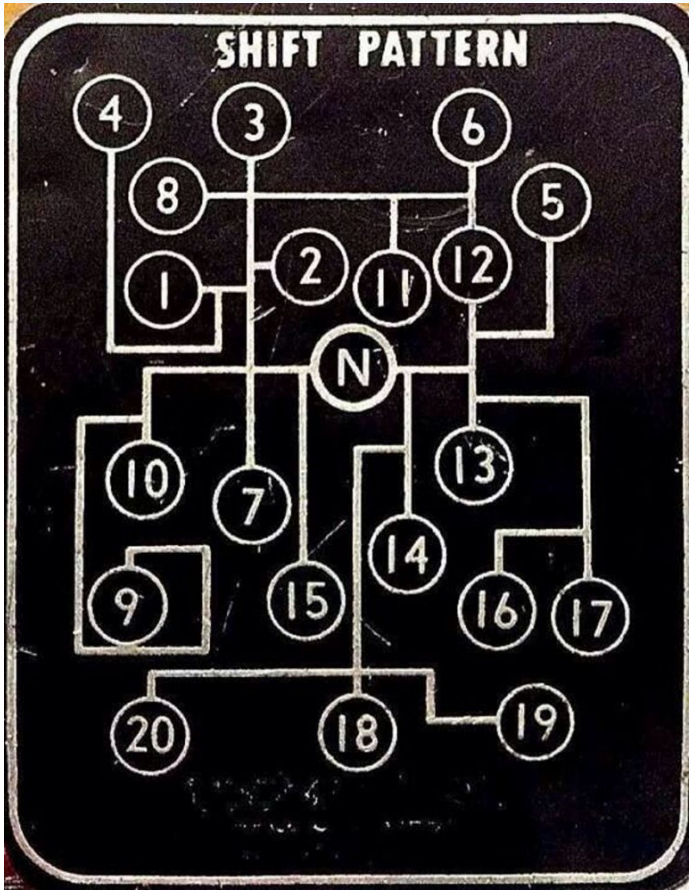
- Online Lustre (disk) volumes (mounted on login, MOM, compute nodes, accessible via Globus)
  - home directory
  - /projects
  - /scratch
- Nearline (tape) volumes (accessible via Globus only)
  - home directory (distinct & separate from online home)
  - /projects (distinct & separate from online projects)*

Lustre

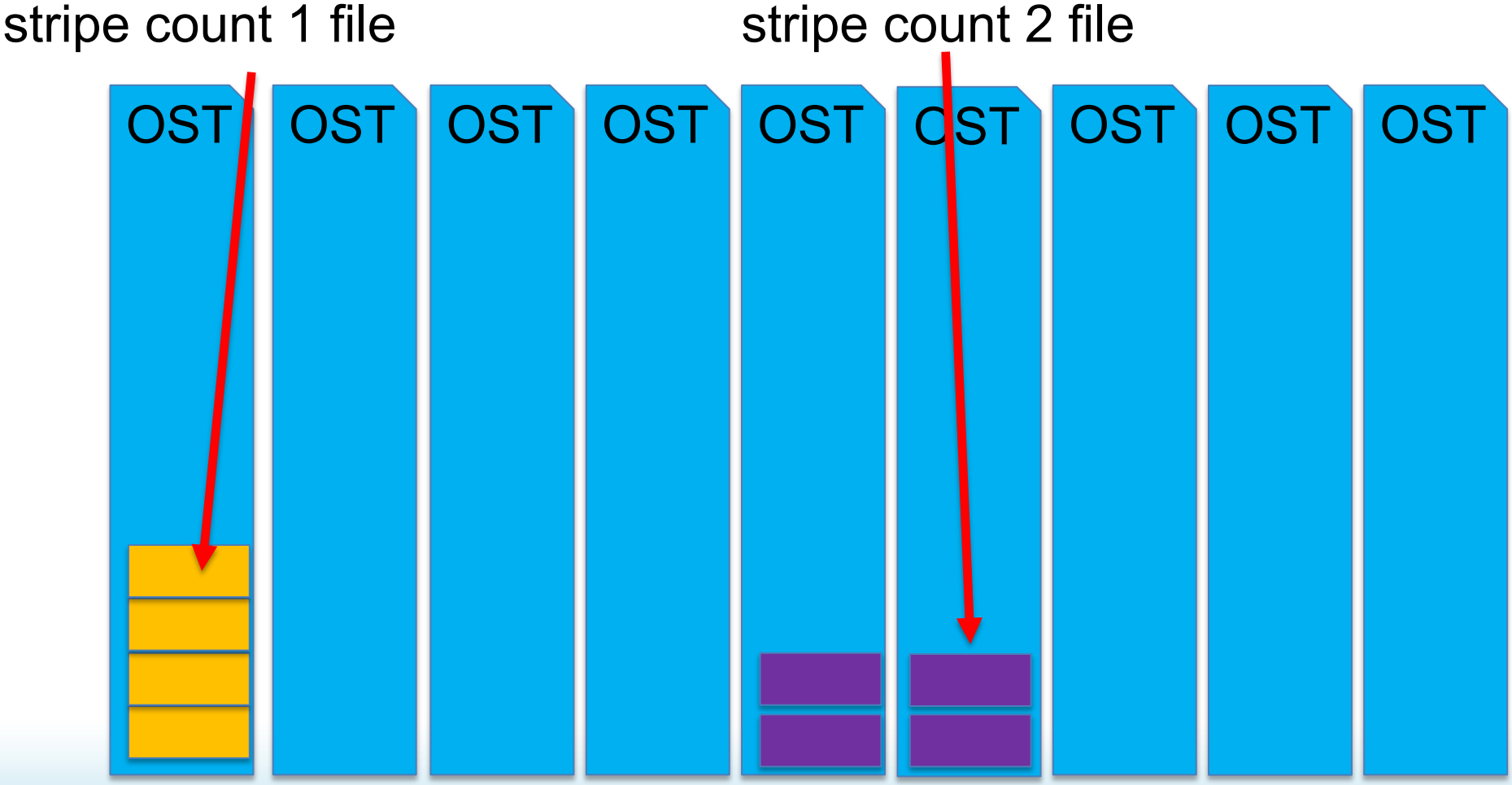
- All mounted file systems are on Lustre (home, /projects, /scratch)
- Every file has a "stripe count"

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?



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)

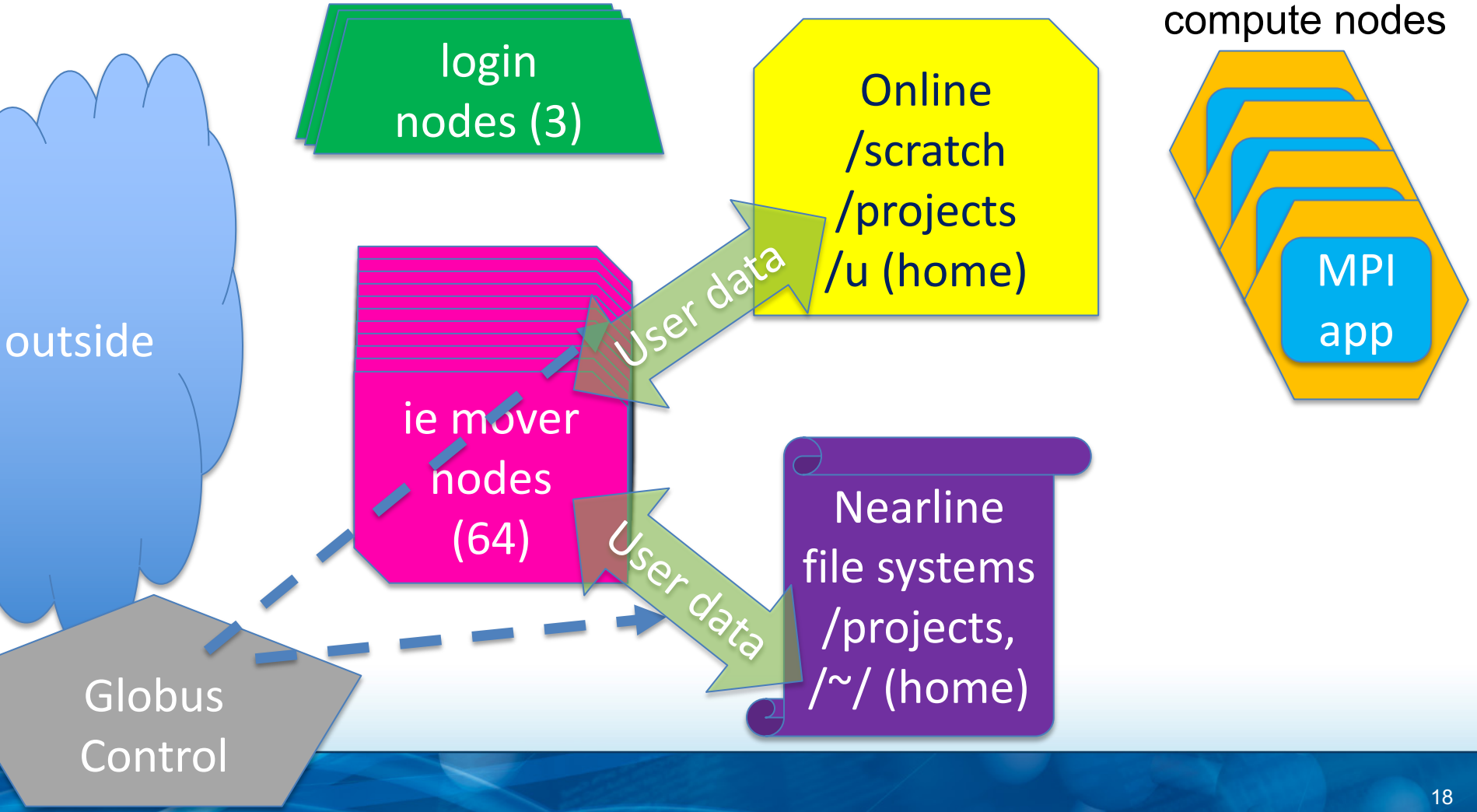
Lustre striping questions

- 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 → 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 → 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)



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

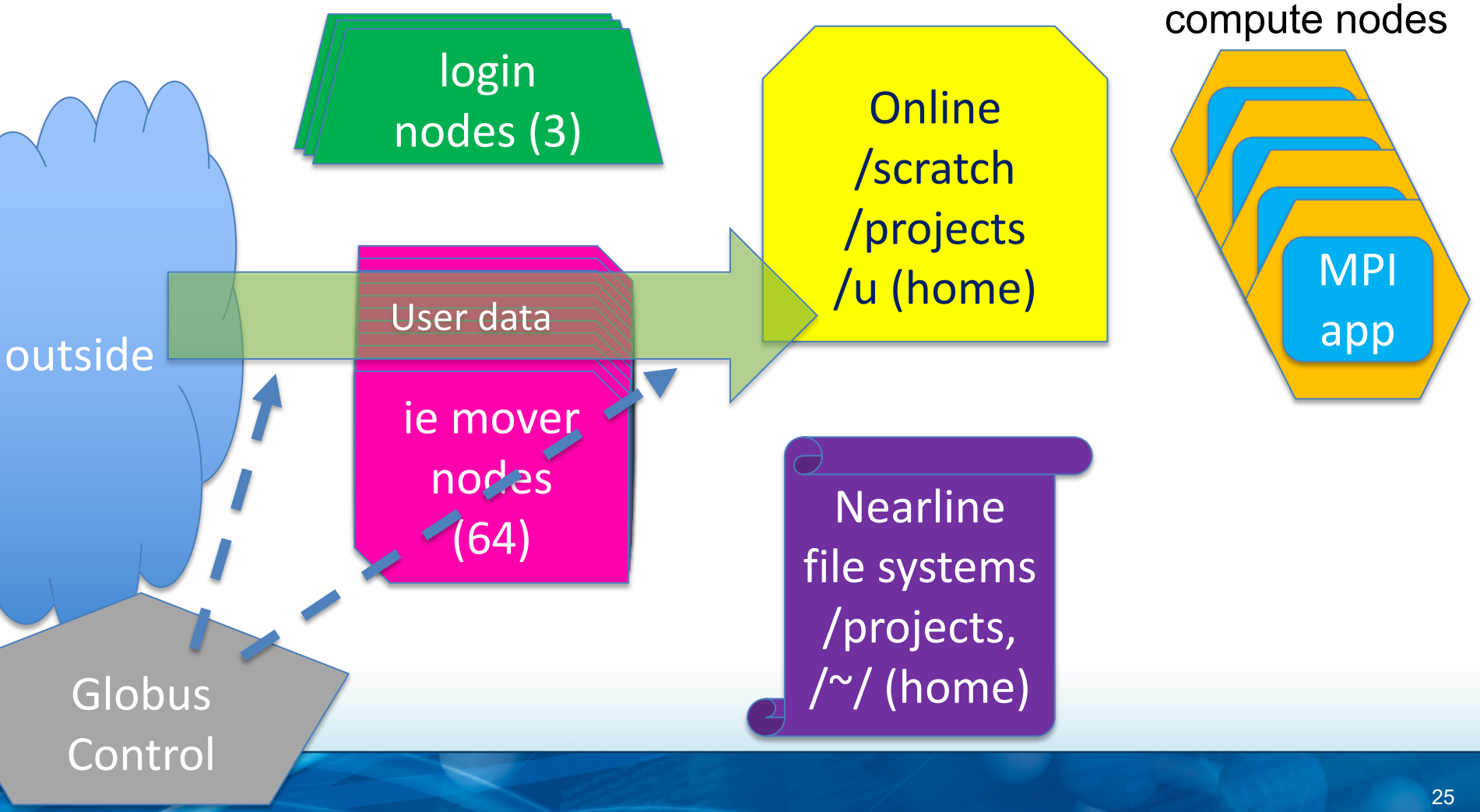
Monitoring Globus

- 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



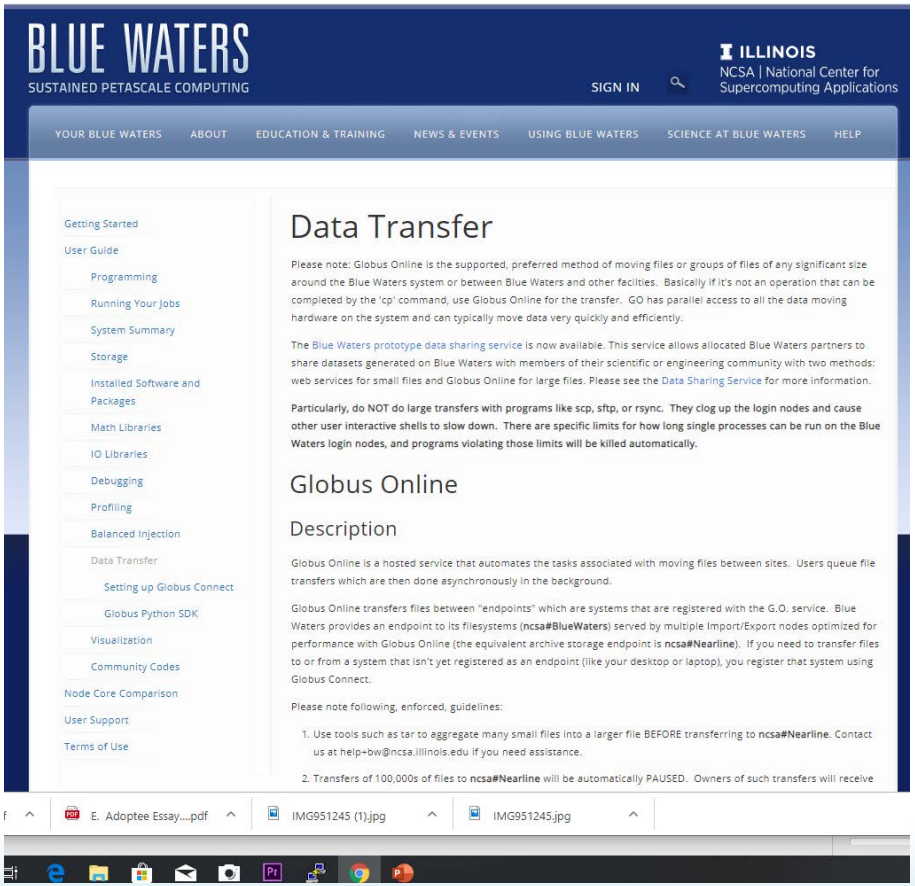
Questions to ask about long-distance data transfers

- 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

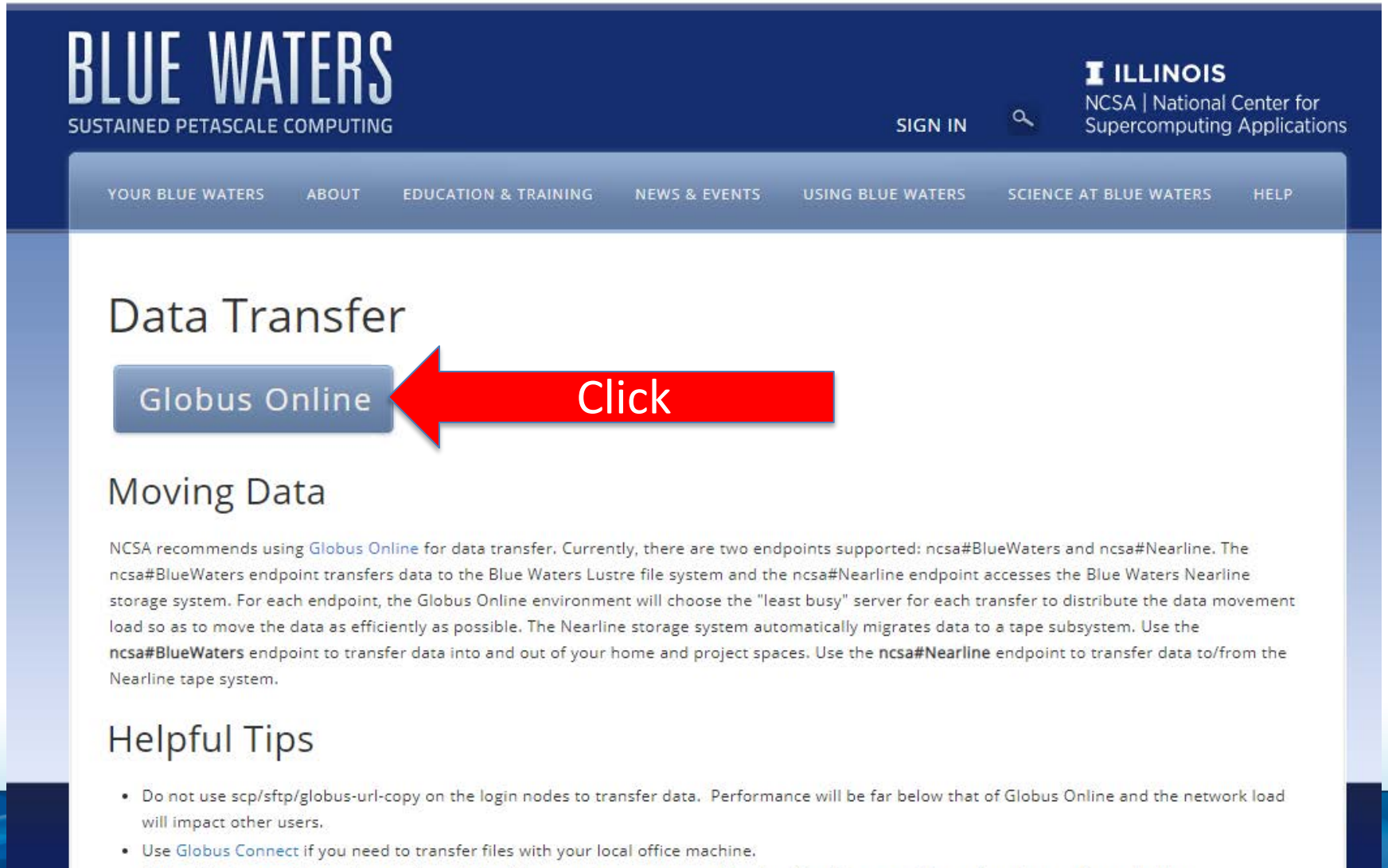
- Are my files less than 10 GB?
- Do I have more than 1000 files to transfer?
- (if either is yes, maybe re-group files)

Transfer overview page that covers Globus https://bluewaters.ncsa.illinois.edu/data-transfer-doc

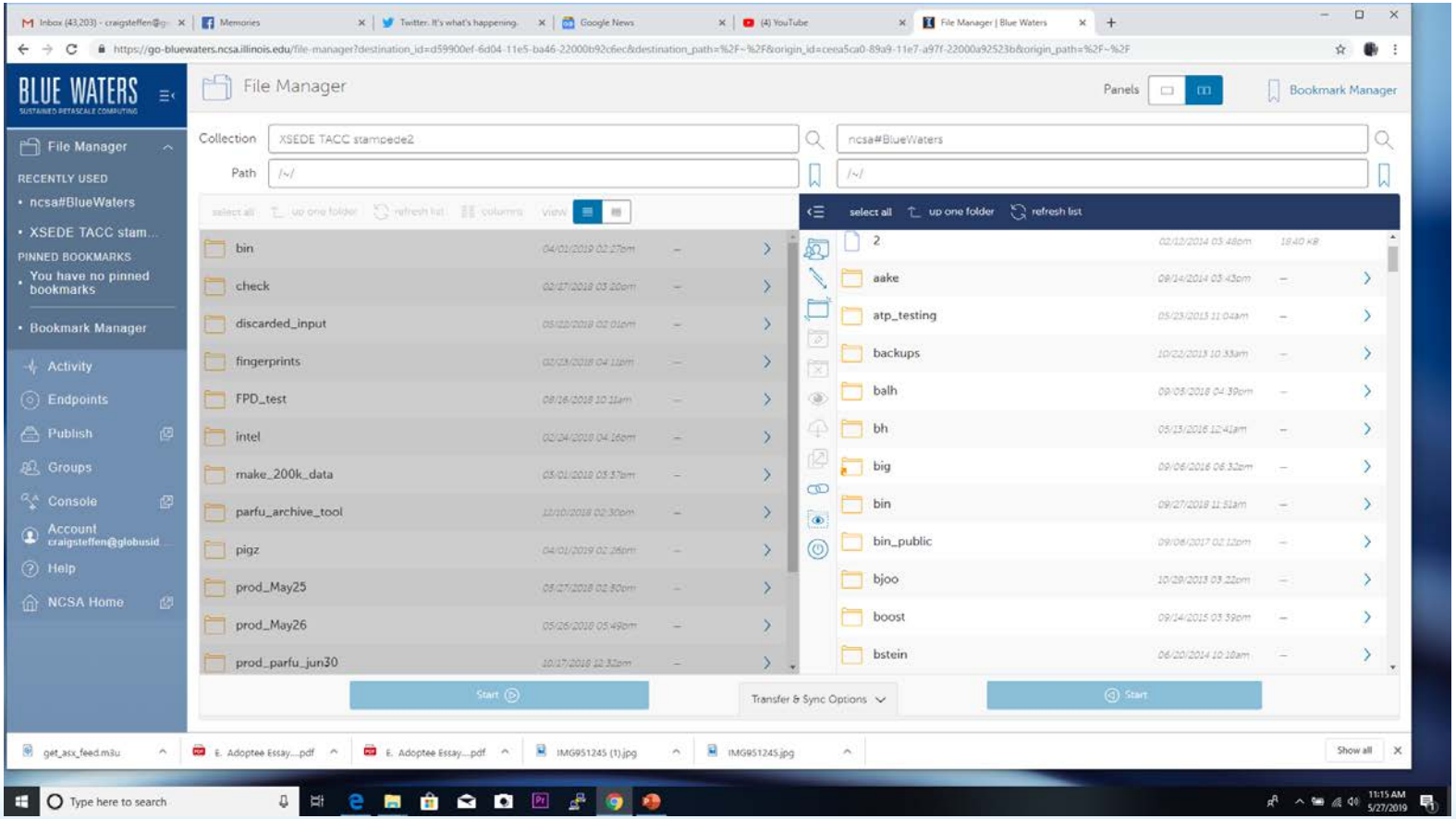


Getting to Globus GUI

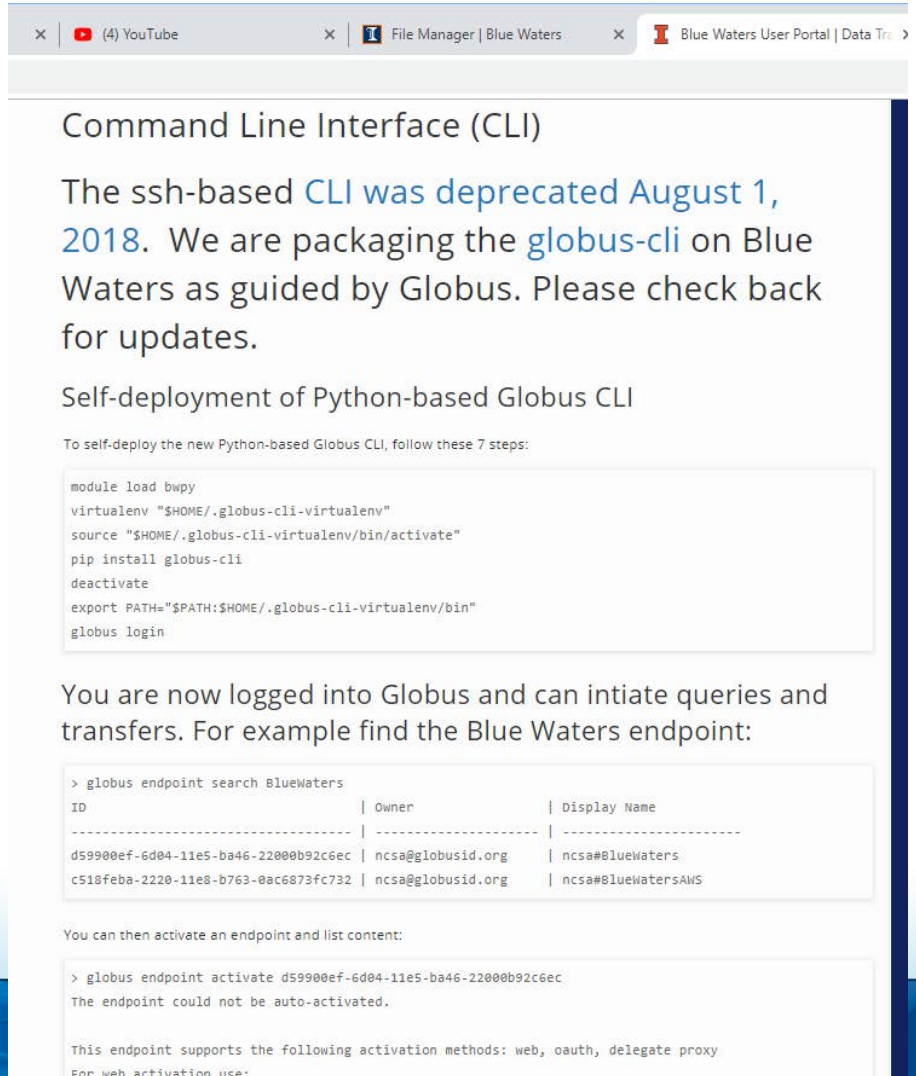
Getting to Globus GUI



Globus GUI


BLUE WATERS
SUSTAINED PETASCALE COMPUTING
File Manager
Panels
Bookmark Manager
Collection
XSEDE TACC stampede2
ncsa#BlueWaters
Path
/~/
File Manager
RECENTLY USED
ncsa#BlueWaters
XSEDE TACC stam...
PINNED BOOKMARKS
You have no pinned bookmarks
Bookmark Manager
Activity
Endpoints
Publish
Groups
Console
Account
craigsteffen@globusid...
Help
NCSA Home
select all
up one folder
refresh list
bin
check
discarded_input
fingerprints
FPD_test
intel
make_200k_data
parfu_archive_tool
pigz
prod_May25
prod_May26
prod_parfu_jun30
2
aake
atp_testing
backups
balh
bh
big
bin
bin_public
bjoo
boost
bstein
Start
Transfer & Sync Options
Start
Show all
Type here to search

Farther down: Globus Python-based CLI



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

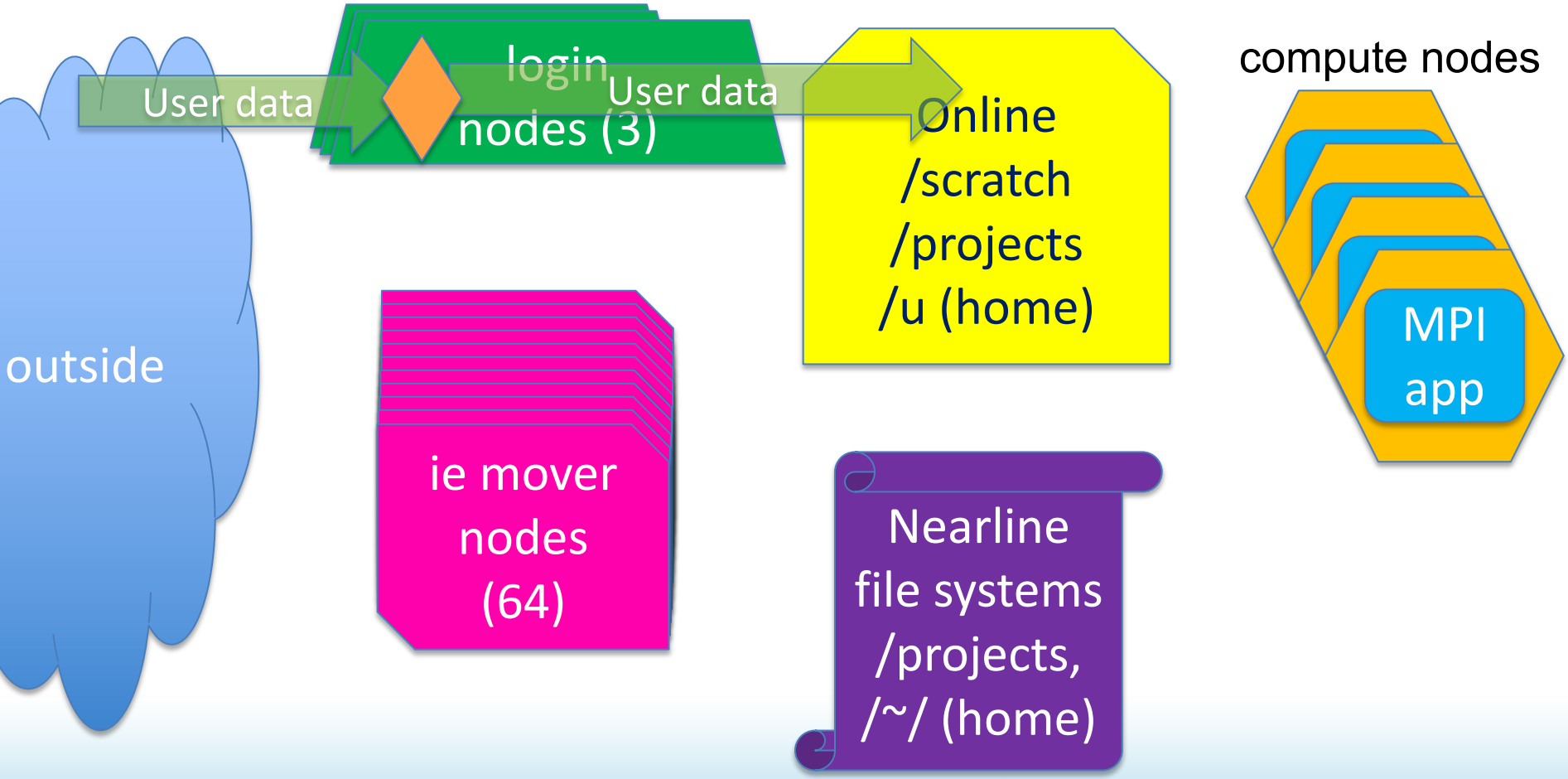
Globus Endpoints

- 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