

BLUE WATERS

SUSTAINED PETASCALE COMPUTING

June 6, 2019

Data Management and Best Practices for Data Movement

Craig Steffen

BW SEAS (User Support) Team



GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION

CRAY®

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

1.

Mouse
over

The screenshot shows the Blue Waters web portal interface. At the top left is the 'BLUE WATERS SUSTAINED PETASCALE COMPUTING' logo. On the right, it says 'ILLINOIS NCSA | National Center for Supercomputing Applications'. A navigation bar contains links: 'YOUR BLUE WATERS', 'ABOUT', 'EDUCATION & TRAINING', 'NEWS & EVENTS', 'USING BLUE WATERS', 'SCIENCE AT BLUE WATERS', and 'HELP'. The 'USING BLUE WATERS' link is circled in red. A red arrow labeled '1. Mouse over' points to this link. A dropdown menu is open under 'USING BLUE WATERS', listing categories: 'ALLOCATIONS', 'DOCUMENTATION', 'RESOURCES', and 'ACKNOWLEDGE SUPPORT'. Under 'DOCUMENTATION', the 'User Guide' link is highlighted with a red arrow labeled '2. Click on "User Guide"'. On the left side of the page, there is a news article titled 'Blue Waters user wins NOAA award' with a 'Read More' button. At the bottom, there are statistics: '24 IN THE PAST HOURS', 'JOBS STARTED 2405', 'JOBS QUEUED 2324', and 'JOBS COMPLETED 2500'.

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

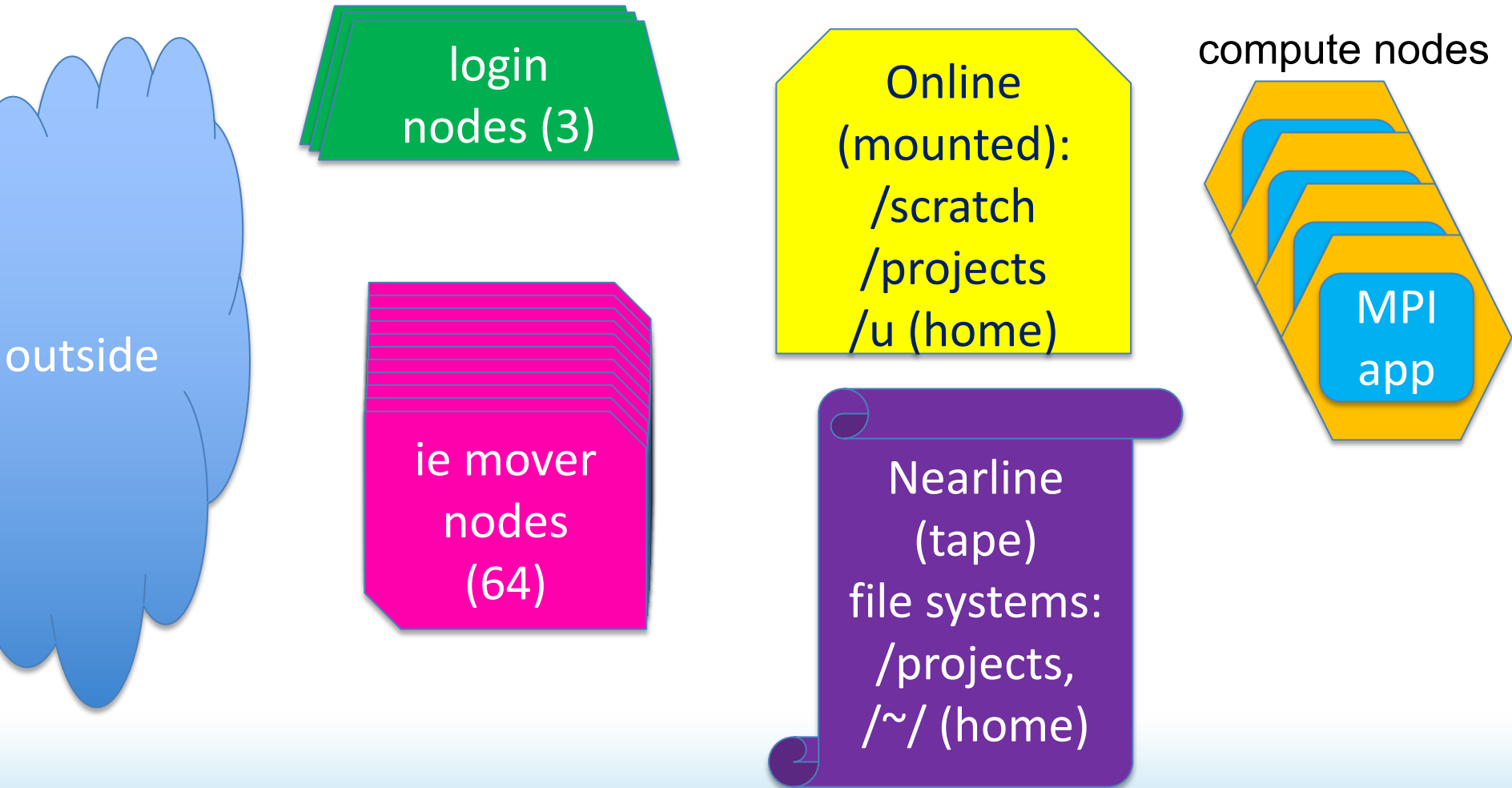
Data Management on Blue Waters

- 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

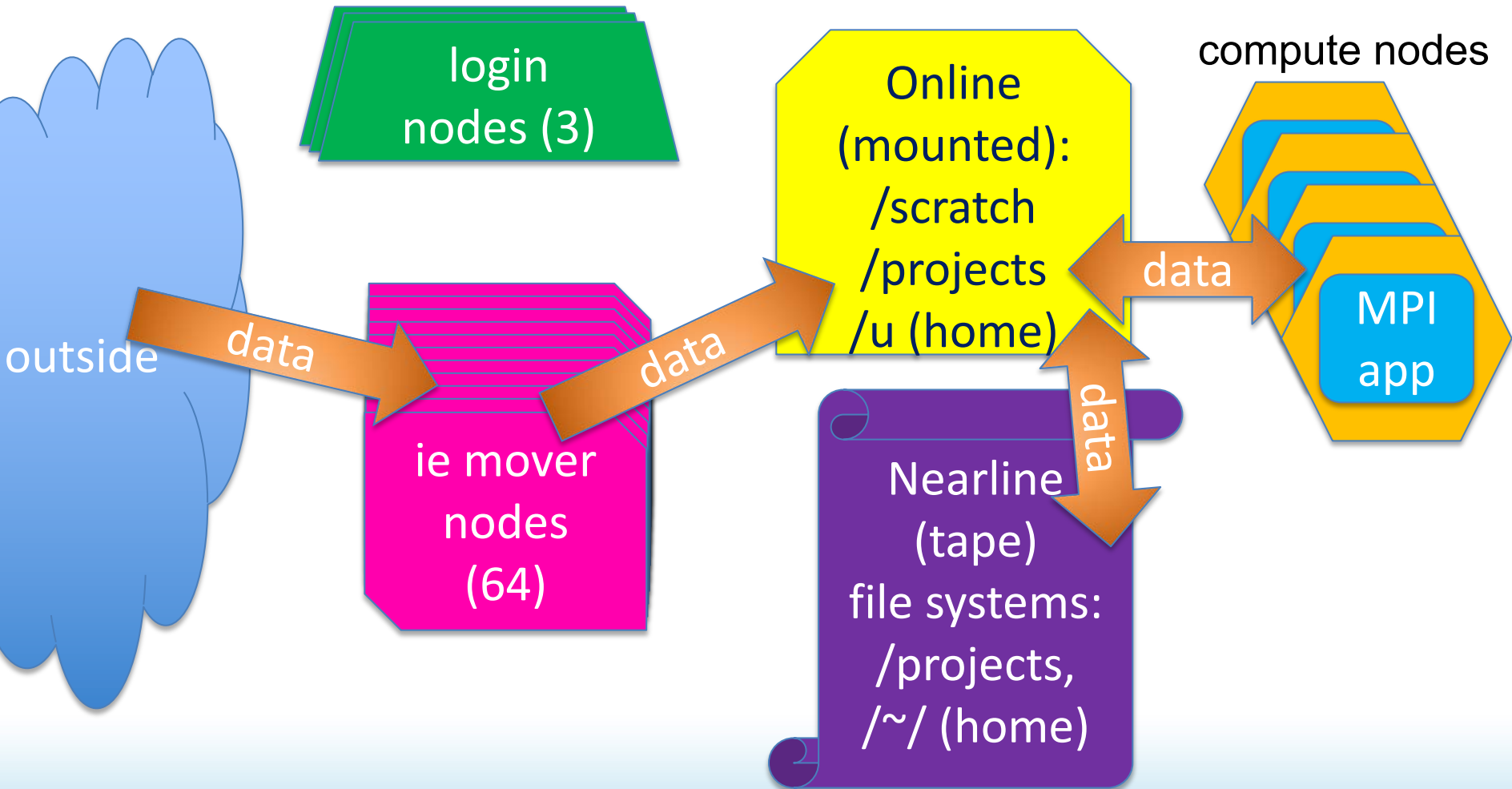
Questions about the process

- 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

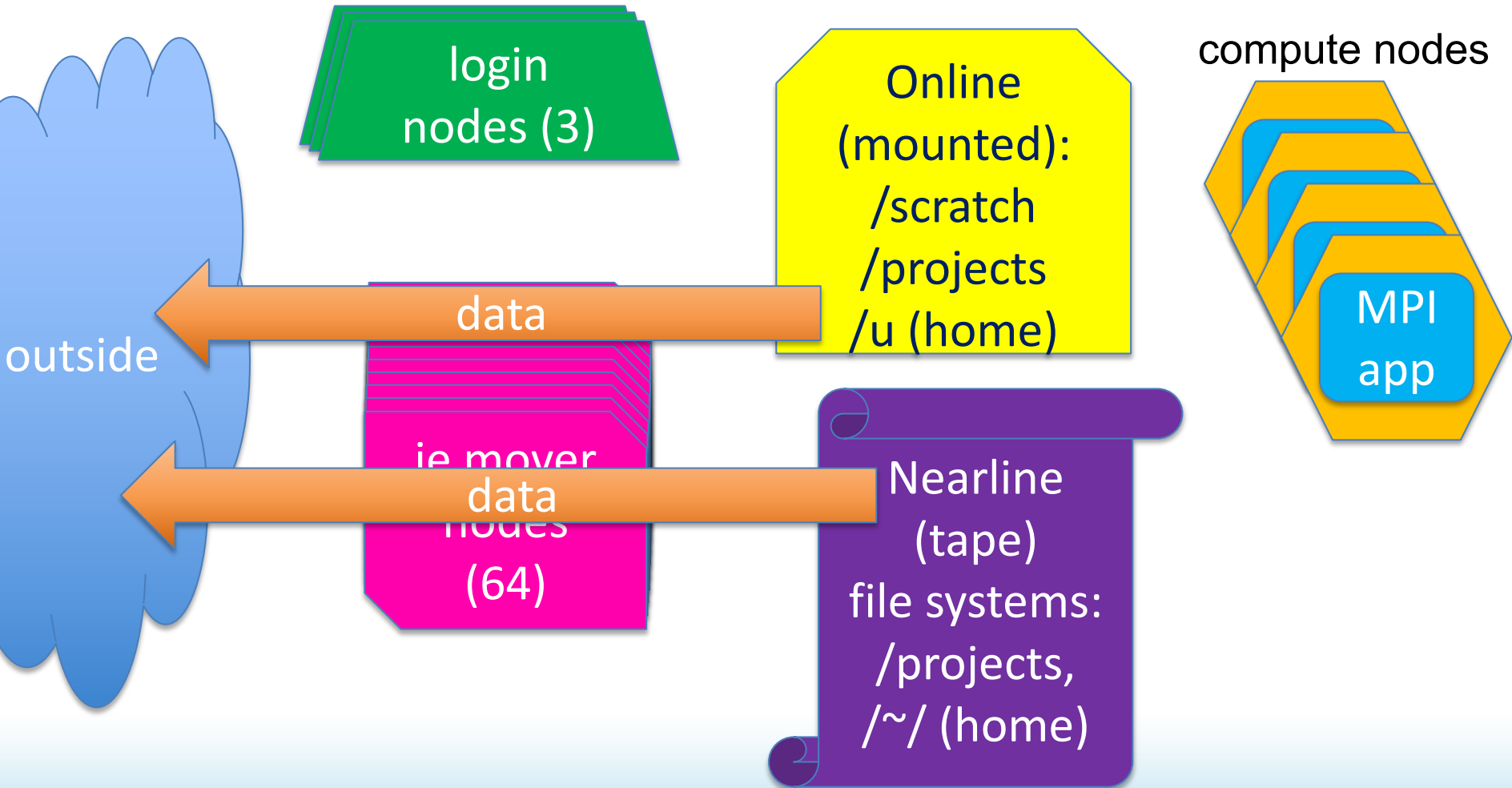
Players in data movement and layout



During your Blue Waters work:



When your Blue Waters work finishes



Where data lives: Blue Waters file system topology

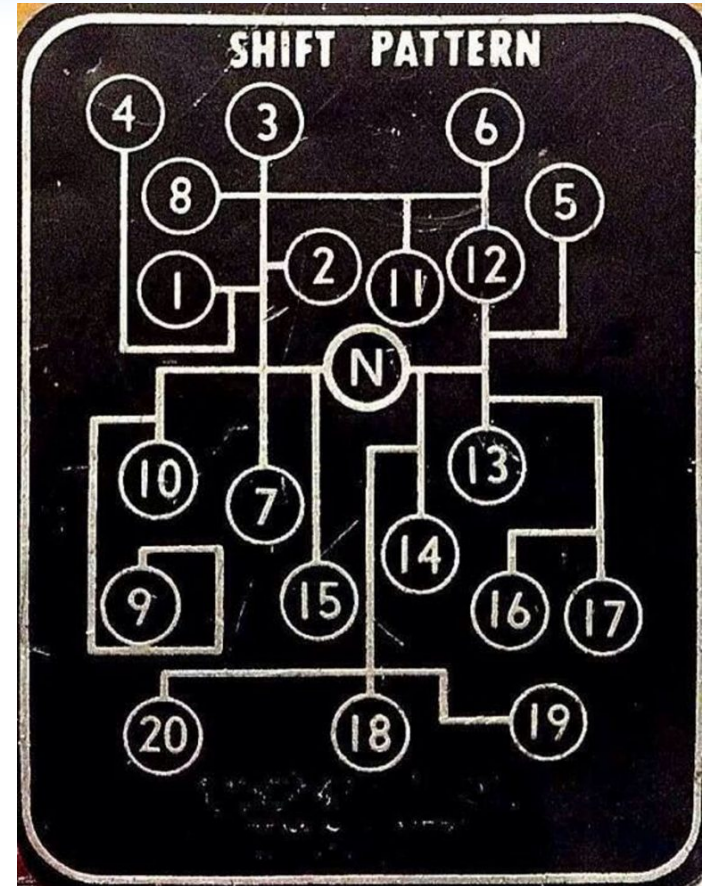
- Online Lustre (disk) volumes (mounted on login, MOM, compute nodes, accessible via Globus)
 - home directory
 - /projects
 - /scratch
- Nearline (tape) volumes (accessible via Globus only)
 - home directory (distinct & separate from online home)
 - /projects (distinct & separate from online projects)*

Lustre

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

Lustre

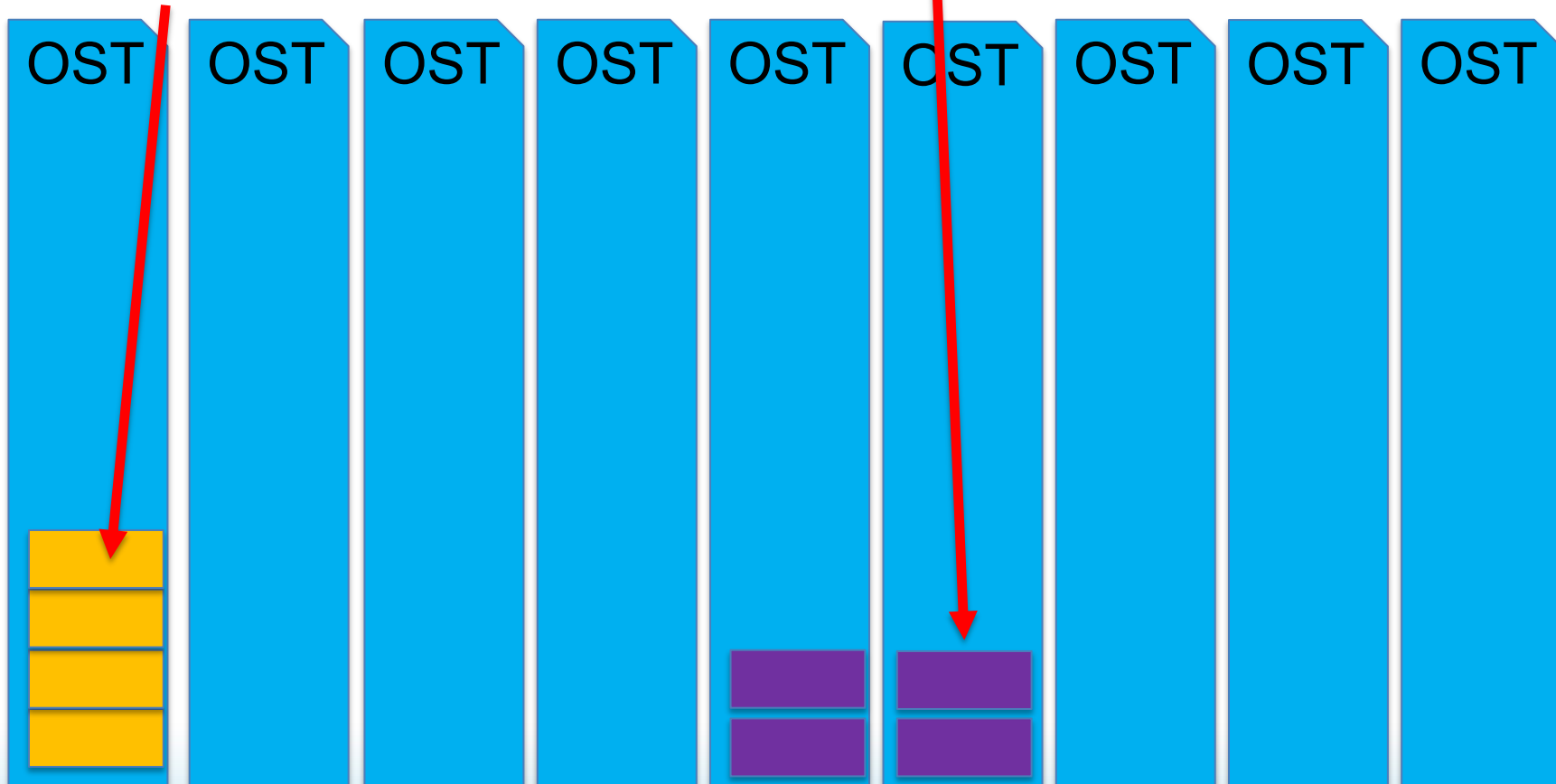
- All mounted file systems are on Lustre (home, /projects, /scratch)
- Every file has a “stripe count”
- striping is MANUAL



What is file striping in Lustre?

stripe count 1 file

stripe count 2 file



How do I set stripe count?

- `ifs setstripe -c 4 file_to_set.dat`
- `ifs 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)

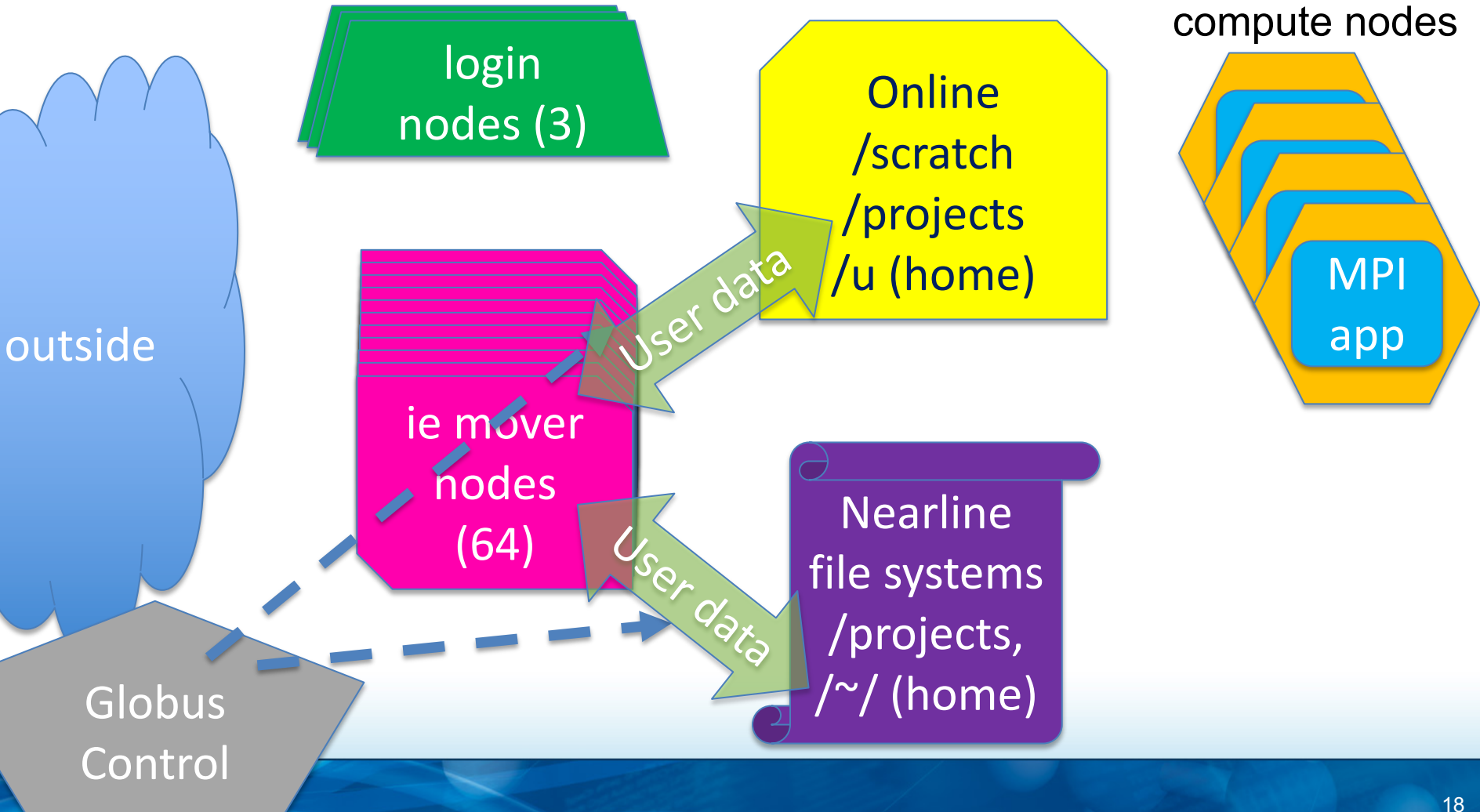
Lustre striping questions

- 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

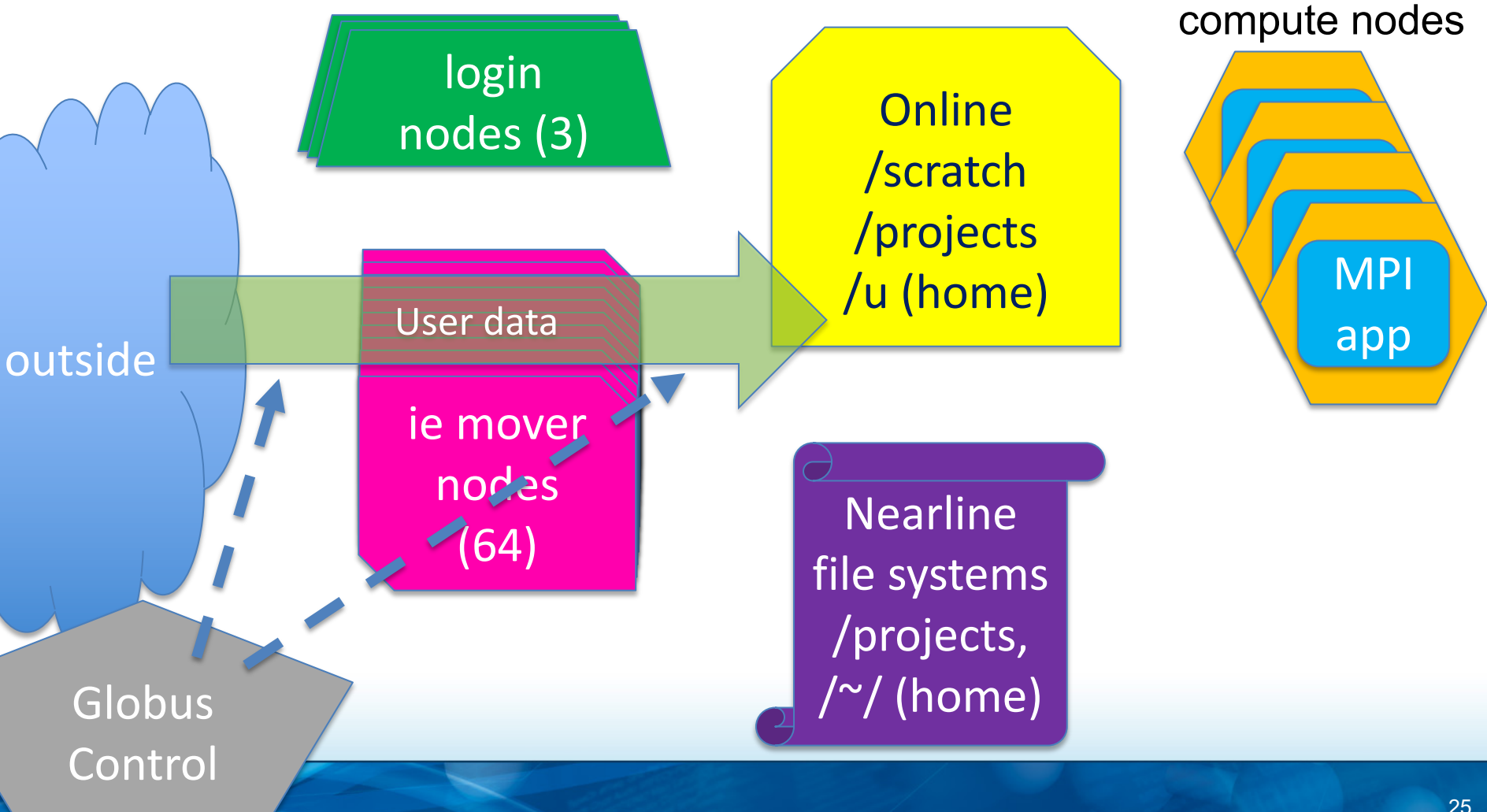
Monitoring Globus

- 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) <https://bluewaters.ncsa.illinois.edu/data-transfer-doc>

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

ILLINOIS
NCSA | National Center for Supercomputing Applications

SIGN IN

YOUR BLUE WATERS ABOUT EDUCATION & TRAINING NEWS & EVENTS USING BLUE WATERS SCIENCE AT BLUE WATERS HELP

Getting Started
User Guide
Programming
Running Your Jobs
System Summary
Storage
Installed Software and Packages
Math Libraries
IO Libraries
Debugging
Profiling
Balanced Injection
Data Transfer
Setting up Globus Connect
Globus Python SDK
Visualization
Community Codes
Node Core Comparison
User Support
Terms of Use

Data Transfer

Please note: Globus Online is the supported, preferred method of moving files or groups of files of any significant size around the Blue Waters system or between Blue Waters and other facilities. Basically if it's not an operation that can be completed by the 'cp' command, use Globus Online for the transfer. GO has parallel access to all the data moving hardware on the system and can typically move data very quickly and efficiently.

The Blue Waters prototype data sharing service is now available. This service allows allocated Blue Waters partners to share datasets generated on Blue Waters with members of their scientific or engineering community with two methods: web services for small files and Globus Online for large files. Please see the Data Sharing Service for more information.

Particularly, do NOT do large transfers with programs like scp, sftp, or rsync. They clog up the login nodes and cause other user interactive shells to slow down. There are specific limits for how long single processes can be run on the Blue Waters login nodes, and programs violating those limits will be killed automatically.

Globus Online

Description

Globus Online is a hosted service that automates the tasks associated with moving files between sites. Users queue file transfers which are then done asynchronously in the background.

Globus Online transfers files between "endpoints" which are systems that are registered with the G.O. service. Blue Waters provides an endpoint to its filesystems (ncsa#BlueWaters) served by multiple Import/Export nodes optimized for performance with Globus Online (the equivalent archive storage endpoint is ncsa#Nearline). If you need to transfer files to or from a system that isn't yet registered as an endpoint (like your desktop or laptop), you register that system using Globus Connect.

Please note following, enforced, guidelines:

1. Use tools such as tar to aggregate many small files into a larger file BEFORE transferring to ncsa#Nearline. Contact us at help-bw@ncsa.illinois.edu if you need assistance.
2. Transfers of 100,000s of files to ncsa#Nearline will be automatically PAUSED. Owners of such transfers will receive

Getting to Globus GUI

1.
Mouse
over

BLUE WATERS

SUSTAINED PETASCALE COMPUTING

SIGN IN



ILLINOIS
NCSA | National Center for
Supercomputing Applications

YOUR BLUE WATERS ABOUT EDUCATION & TRAINING NEWS & EVENTS **USING BLUE WATERS** SCIENCE AT BLUE WATERS HELP

Blue Waters user wins NOAA award

Claire Porter and the ArcticDEM team used the Blue Waters supercomputer to create digital elevation models to map the Arctic, the Antarctic, and soon, the entire world.

Read More

ALLOCATIONS

DOCUMENTATION

Getting Started
User Guide
Node Core Comparison
User Support
Terms of Use

RESOURCES

Data
Software
Visualization
Test Programs
Code Examples

ACKNOWLEDGE SUPPORT

2. Click on "Data"

24 IN THE PAST
HOURS

JOBS STARTED
3405

JOBS QUEUED
2234

JOBS COMPLETED
3508

Getting to Globus GUI

The screenshot shows the Blue Waters website interface. At the top left is the Blue Waters logo and tagline. To the right are logos for the University of Illinois, NSF, NCSA, and the Great Lakes Consortium for Petascale Computation, along with the CRAY logo. Below the logos is a navigation bar with links: YOUR BLUE WATERS, ABOUT, EDUCATION & TRAINING, NEWS & EVENTS, USING BLUE WATERS, SCIENCE AT BLUE WATERS, and HELP. On the right side of the navigation bar, there is a 'SIGN IN' link and a search icon. Below the navigation bar, the main content area features a 'Data Transfer' section. A button labeled 'Globus Online' is highlighted with a red arrow pointing to it from the right, with the word 'Click' written on the arrow. Below this is a 'Moving Data' section with a paragraph of text explaining NCSA's recommendation for using Globus Online for data transfer. At the bottom, there is a 'Helpful Tips' section with a list of two items.

BLUE WATERS

SUSTAINED PETASCALE COMPUTING

ILLINOIS
NCSA | National Center for
Supercomputing Applications

SIGN IN

YOUR BLUE WATERS ABOUT EDUCATION & TRAINING NEWS & EVENTS USING BLUE WATERS SCIENCE AT BLUE WATERS HELP

Data Transfer

Globus Online ← Click

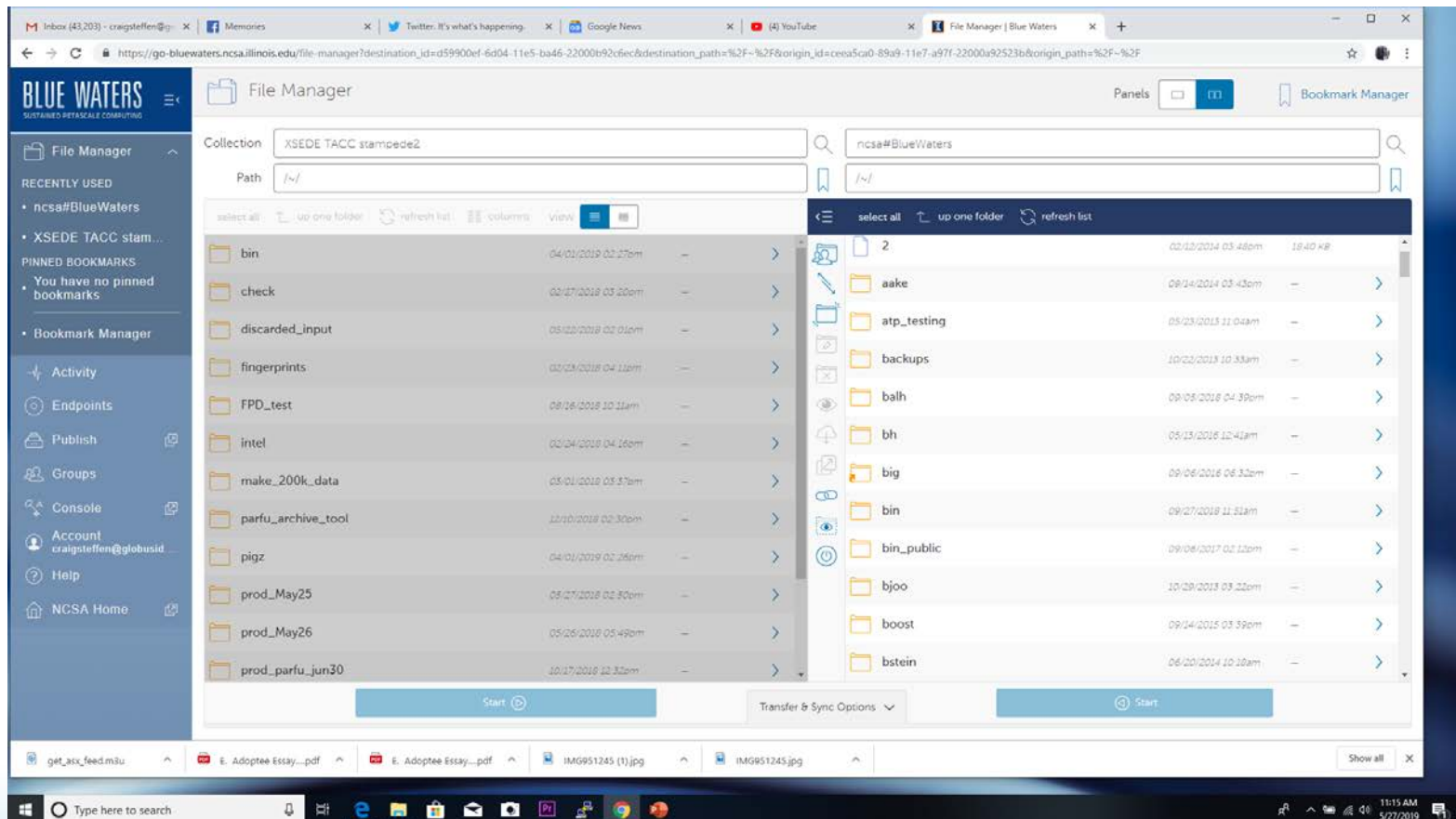
Moving Data

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:

```
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
```

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

```
> globus endpoint search Bluewaters
```

ID	Owner	Display Name
d59900ef-6d04-11e5-ba46-2200b92c6ec	nlsa@globusid.org	nlsa#Bluewaters
c518Feba-2220-11e8-b763-0ac6873fc732	nlsa@globusid.org	nlsa#BluewatersAWS

You can then activate an endpoint and list content:

```
> globus endpoint activate d59900ef-6d04-11e5-ba46-2200b92c6ec
The endpoint could not be auto-activated.
```

This endpoint supports the following activation methods: web, oauth, delegate proxy
For web activation use:

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

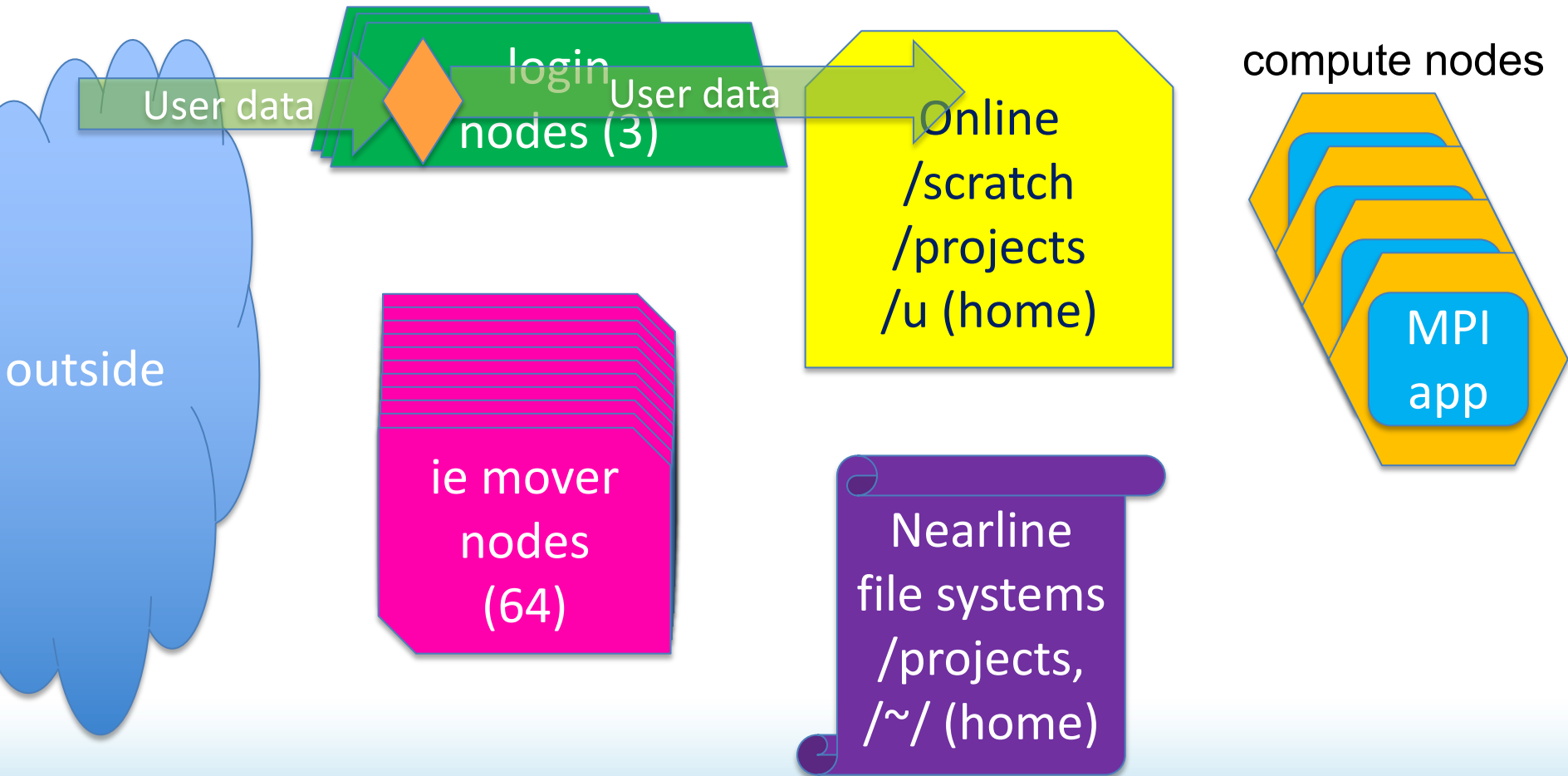
Globus Endpoints

- 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