BLUE WATERS

SUSTAINED PETASCALE COMPUTING

February 11, 2015

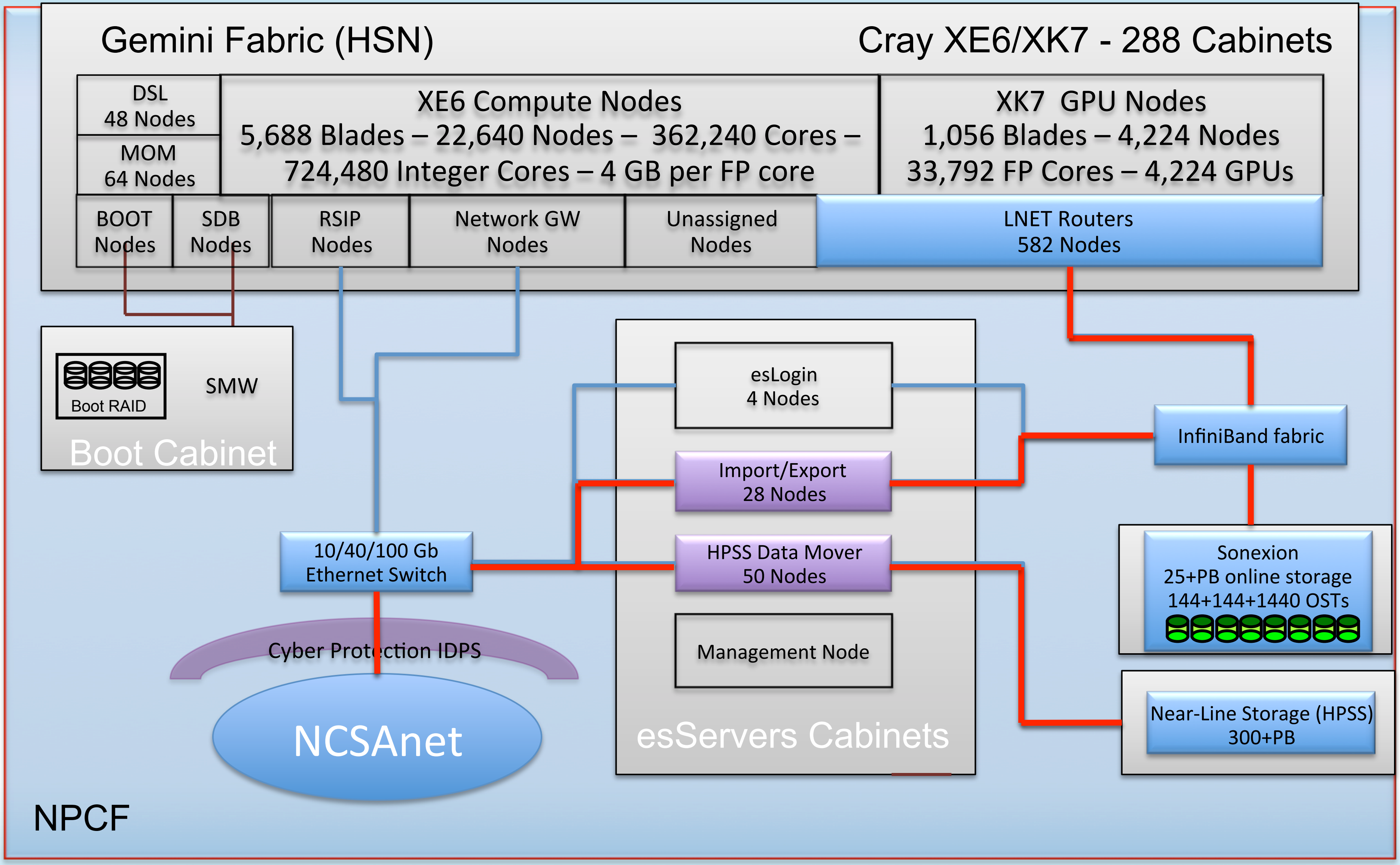
# Data: Creation, Management, Transfer, and Sharing

Ryan Mokos

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

# Outline

- Storage system overview
  - File system sizes, quotas, and structure
- Data creation
  - Maximizing IO performance
  - Lustre characteristics including striping
- Data management
  - File sizes, number of files, and compression
- Data transfer
  - Globus Online for transferring files
- Data sharing
  - Active Project Data Share Plan
  - Community Data Share Plan



Gemini Fabric (HSN)

Cray XE6/XK7 - 288 Cabinets

DSL
48 Nodes

MOM
64 Nodes

XE6 Compute Nodes
5,688 Blades – 22,640 Nodes – 362,240 Cores – 724,480 Integer Cores – 4 GB per FP core

XK7 GPU Nodes
1,056 Blades – 4,224 Nodes
33,792 FP Cores – 4,224 GPUs

BOOT Nodes

SDB Nodes

RSIP Nodes

Network GW Nodes

Unassigned Nodes

LNET Routers
582 Nodes

Boot RAID

SMW

Boot Cabinet

esLogin
4 Nodes

Import/Export
28 Nodes

HPSS Data Mover
50 Nodes

Management Node

esServers Cabinets

InfiniBand fabric

Sonexion
25+PB online storage
144+144+1440 OSTs

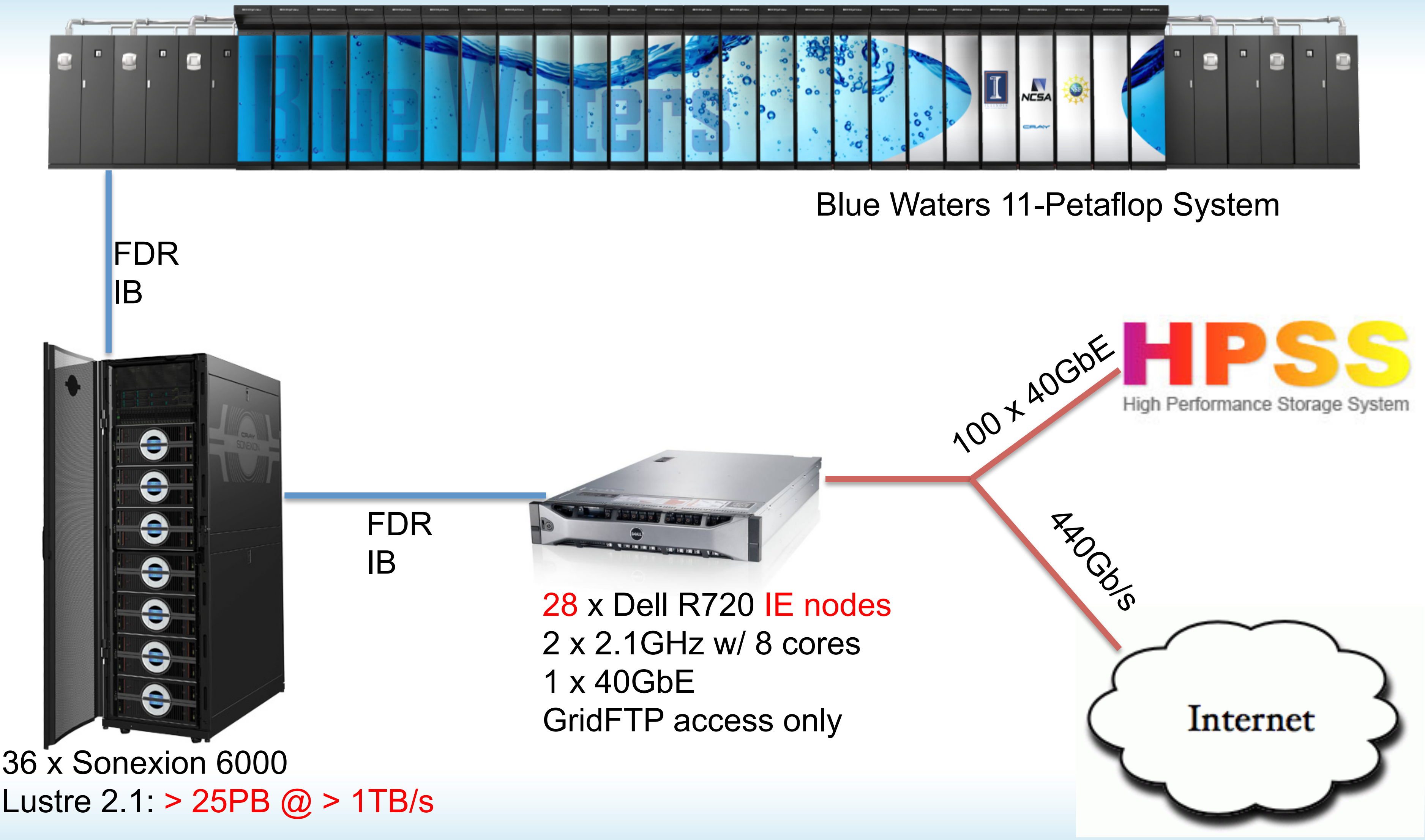
Near-Line Storage (HPSS)
300+PB

10/40/100 Gb Ethernet Switch

Cyber Protection IDPS

NCSAnet

NPCF



Blue Waters 11-Petaflop System

FDR IB

FDR IB

36 x Sonexion 6000
Lustre 2.1: > 25PB @ > 1TB/s

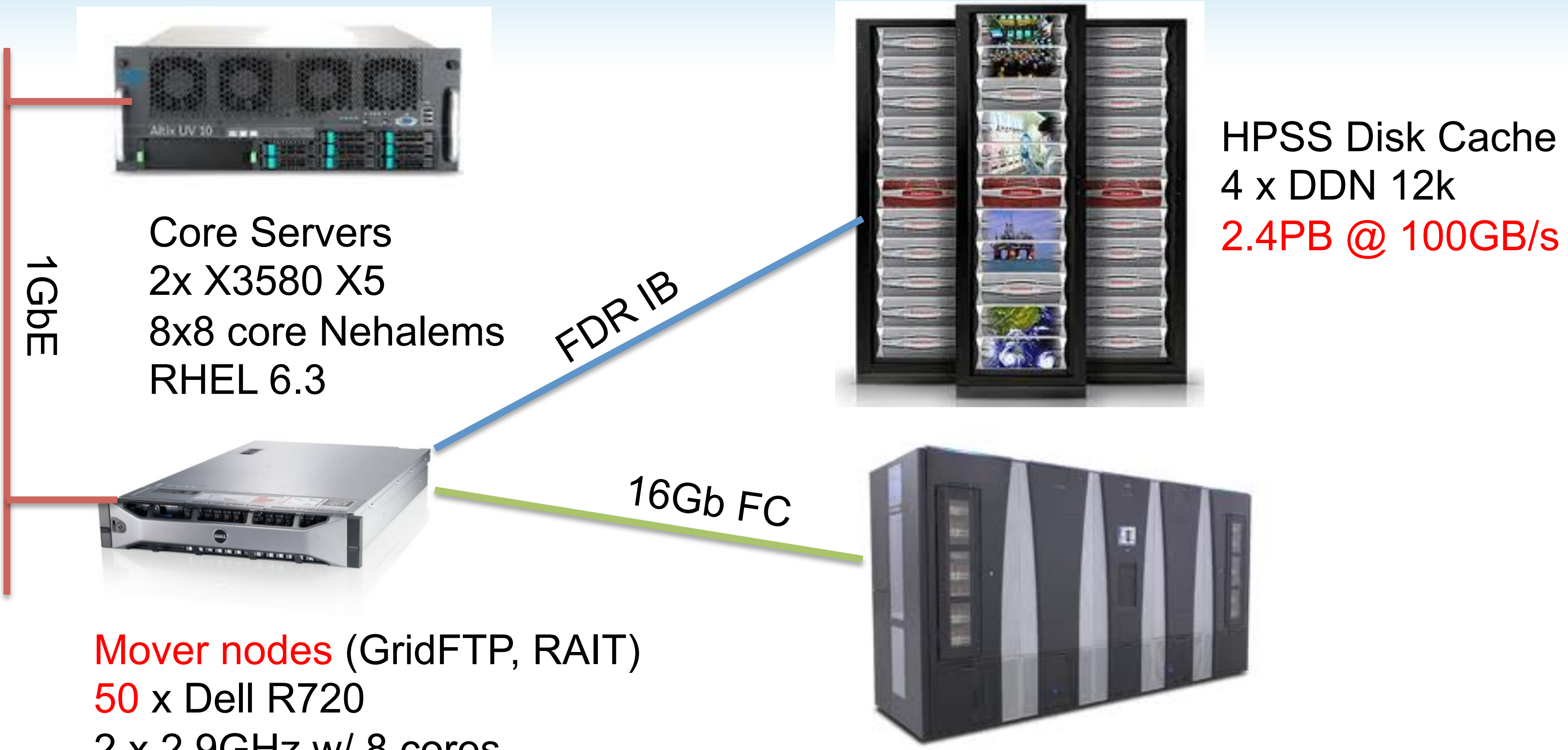
28 x Dell R720 IE nodes
2 x 2.1GHz w/ 8 cores
1 x 40GbE
GridFTP access only

100 x 40GbE

HPSS
High Performance Storage System

440Gb/s

Internet



Core Servers
2x X3580 X5
8x8 core Nehalems
RHEL 6.3

HPSS Disk Cache
4 x DDN 12k
2.4PB @ 100GB/s

Mover nodes (GridFTP, RAIT)
50 x Dell R720
2 x 2.9GHz w/ 8 cores
2 x 40GbE (Bonded)
RHEL 6.3
GridFTP access only

6 x Spectra Logic T-Finity
12 robotic arms
360PB in 95580 slots
366 TS1140 Jaguars @ 240MB/s

# Lustre (Online) Storage Summary

| Filesystem | Total Usable Space | Quota | OSTs (Object Storage Target) | Backed Up | Purge Policy |
|---|---|---|---|---|---|
| /home | 2.2 PB | 1 TB / user | 144 | Daily | No |
| /projects | 2.2 PB | 5 TB / group | 144 | Daily | No |
| /scratch | 22 PB | 500 TB / group | 1440 | No | 30-day |

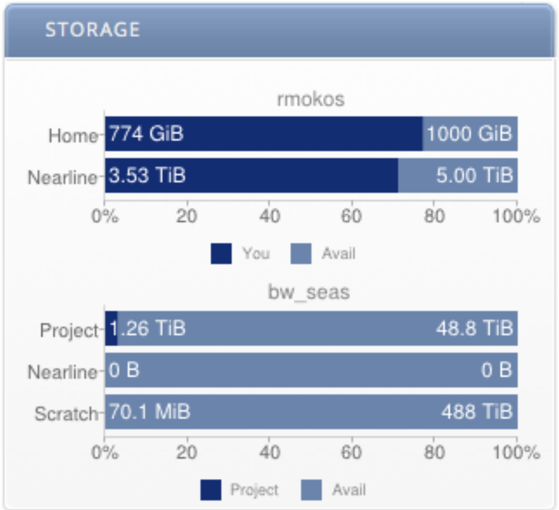
- All filesystems visible from compute nodes
    - Can run application from any filesystem, but scratch highly preferred, especially for heavy I/O
- /home and /projects backed up daily – saved for 30 days
- /scratch
    - 30-day purge policy
    - Not backed up – archive checkpoints and results regularly

# HPSS (Nearline) Storage Summary

- Tape Capacity: 300 PBs (*usable*)
  - Robotic tape libraries => large seek time
- Disk cache: 1.6 PB (*usable*)
- Bandwidth: 100 GB/sec (*sustained*)
  - 50 mover nodes
- Designed for large files (multi-GB+)
- Same /home and /projects directory structure as Lustre
- 5 TB user quota, 50 TB group quota
- GridFTP access only – no ssh

# Quotas

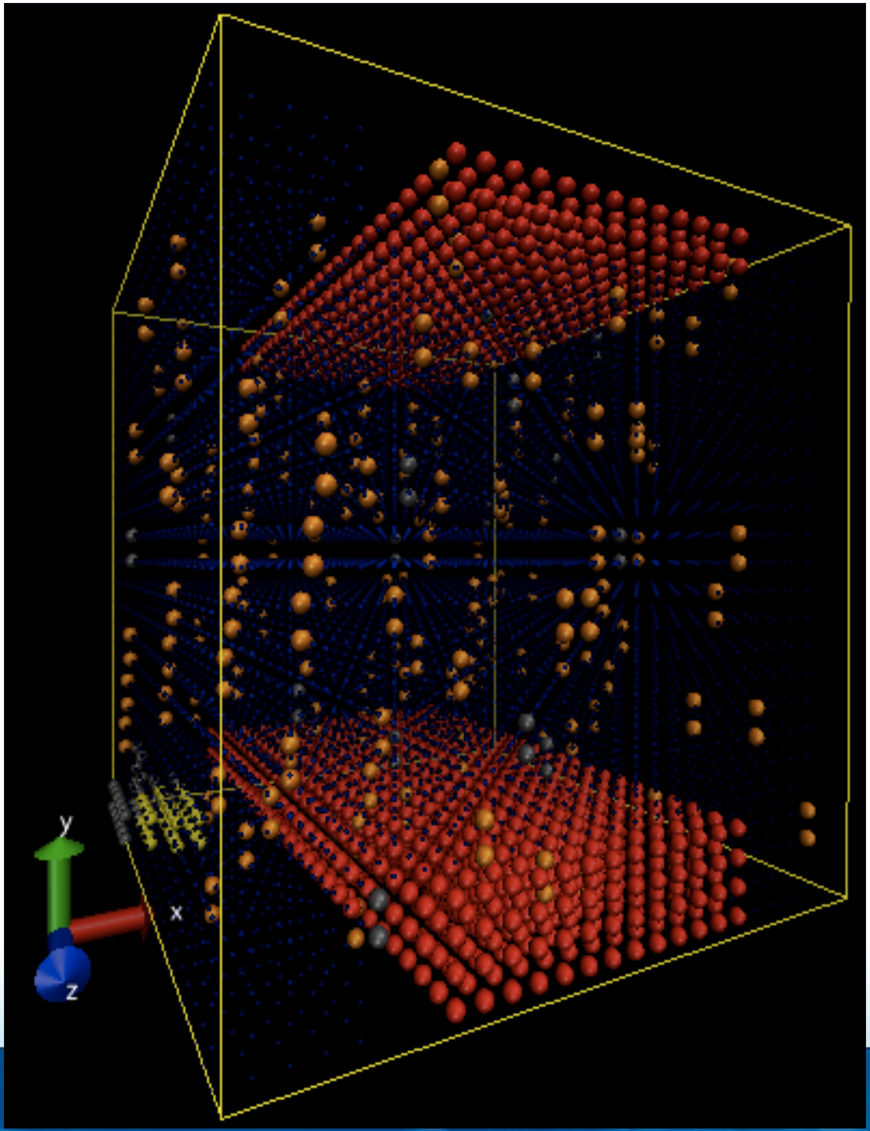
- Check quota via
  - Command line: `quota`
  - Blue Waters portal under “Your Blue Waters” tab



- Submit a ticket to request exceptions/changes to storage policies (e.g., quota increase)

# Data Creation - Considerations

- IO through LNET routers
  - Via Gemini network
  - LNETs = orange spheres
    - Scattered throughout torus
- Object Storage Targets (OSTs)
  - 1440 for /scratch
  - 144 each for /home and /projects
  - ~14 TB / OST
- Meta Data Server (MDS)
  - Only 1 per Lustre filesystem



# Data Creation – Best Practices

- Distribute IO throughout the system
- Balance between file size and number of files
  - Writing
    - Avoid creating hundreds of thousands of tiny files
    - Avoid creating a single huge file for a large job
    - One file per node is often a good starting point
  - Reading
    - Avoid reading in hundreds of thousands of files
    - Avoid having all processes read the same file
      - Consider read followed by broadcast
- Lustre striping
  - Default – no striping
  - Stripe large files across multiple OSTs
- Create subdirectories – avoid putting 100k files in one directory

# Data Creation - Resources

- Several IO libraries available
  - NetCDF, HDF5, etc.
- Darshan library
  - IO profiling
  - Module loaded by default
- Blue Waters portal pages
  - IO libraries: https://bluewaters.ncsa.illinois.edu/io-libraries
  - Lustre striping: https://bluewaters.ncsa.illinois.edu/storage
    - “Application IO on Blue Waters” presentation: https://bluewaters.ncsa.illinois.edu/c/document_library/get_file?uuid=d70b1e4f-c733-495e-833c-85d4d36c0ae9&groupId=10157
  - Darshan: https://bluewaters.ncsa.illinois.edu/darshan

# Data Management

- Tar up large numbers of small files
  - Multi-threaded ptar module available for compression while tarring (`tar -z …`)
  - Aim for hundreds of MBs to tens of GBs range for apps with a lot of data
- Perform longer/CPU intensive file operations on compute nodes (single-node jobs)
  - E.g., large tars, file copies, etc.
  - Eases burden on login nodes – faster for everyone
  - Run in interactive CCM (Cluster Compatibility Mode)
    - `qsub –I –lnodes=1:ppn=32:xe –lgres=ccm`
    - Portal: https://bluewaters.ncsa.illinois.edu/cluster-compatibility-mode

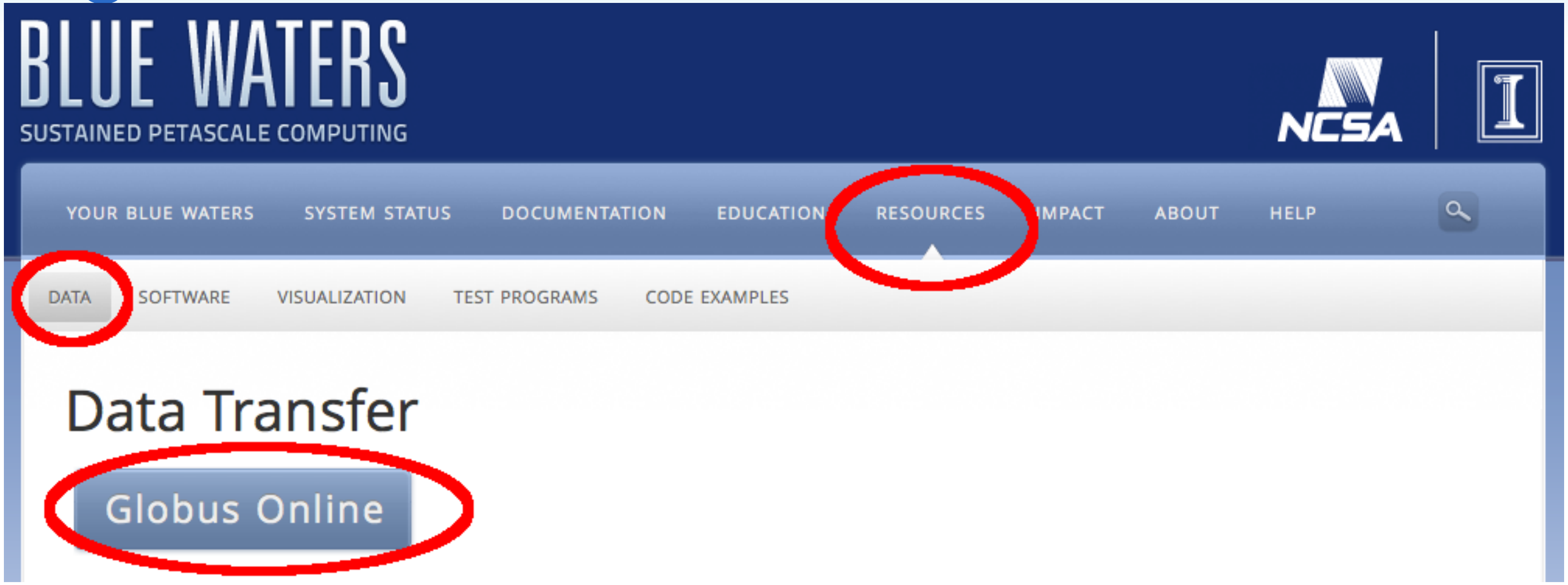
# Data Transfers

- Lustre ⇔ Lustre
  - /home, /projects, /scratch all independent
    - `mv` between file systems = `cp` + delete
  - Serial with `cp`: ~1.5 Gbits/s
- Lustre ⇔ Nearline
  - Must use GridFTP client (Globus, UberFTP)
- Blue Waters ⇔ other sites
  - Use GridFTP client (Globus)
  - **Do not use `scp`, `rsync`, or `sftp`**
    - Hard on login nodes and slower

# Globus Online (GO)

- GO – GridFTP client
- Parallel file transfer
  - Up to 20 files in flight per transfer
  - Up to 3 simultaneous transfers
- GO uses mover nodes
  - Lightens load on logins
- Web GUI and Command Line Interface (CLI)
- GUI vs. CLI
  - Limit on viewable files in GUI
  - CLI cumbersome but more powerful

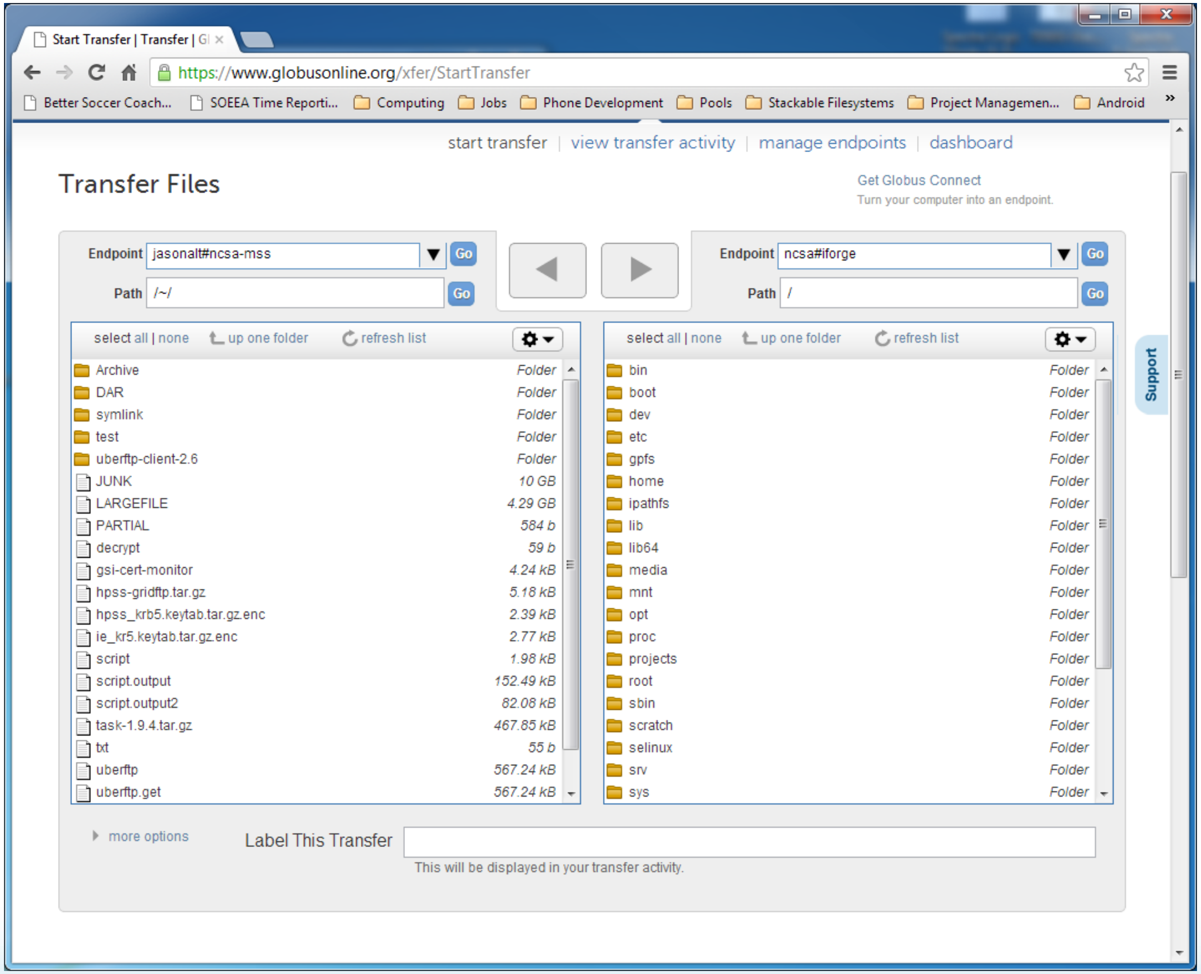
# Using Globus Online GUI



- BW Portal
  - Documentation: https://bluewaters.ncsa.illinois.edu/data-transfer-doc
  - GO access: https://bluewaters.ncsa.illinois.edu/data
- Use Globus Connect to create local endpoints for your own computer/cluster

# Globus Online Web GUI

- BW endpoints
  - ncsa#BlueWaters
  - ncsa#Nearline
- Advantages
  - Easy transfers
    - Select src/dest
    - Select files/dirs
    - Click arrow
  - Simple option selection
- Limitations
  - Some parameters inaccessible
  - 10k file max listing
  - Sometimes < full concurrency


Start Transfer | Transfer | G
https://www.globusonline.org/xfer/StartTransfer
Better Soccer Coach...
SOEEA Time Reporti...
Computing
Jobs
Phone Development
Pools
Stackable Filesystems
Project Managemen...
Android
start transfer | view transfer activity | manage endpoints | dashboard
Get Globus Connect
Turn your computer into an endpoint.
Support

Transfer Files

Endpoint jasonalt#ncsa-mss Go
Path /~/ Go

Endpoint ncsa#iforge Go
Path / Go

select all | none   up one folder   refresh list

| | |
|---|---|
| Archive | Folder |
| DAR | Folder |
| symlink | Folder |
| test | Folder |
| uberftp-client-2.6 | Folder |
| JUNK | 10 GB |
| LARGEFILE | 4.29 GB |
| PARTIAL | 584 b |
| decrypt | 59 b |
| gsi-cert-monitor | 4.24 kB |
| hpss-gridftp.tar.gz | 5.18 kB |
| hpss_krb5.keytab.tar.gz.enc | 2.39 kB |
| ie_kr5.keytab.tar.gz.enc | 2.77 kB |
| script | 1.98 kB |
| script.output | 152.49 kB |
| script.output2 | 82.08 kB |
| task-1.9.4.tar.gz | 467.85 kB |
| txt | 55 b |
| uberftp | 567.24 kB |
| uberftp.get | 567.24 kB |

select all | none   up one folder   refresh list

| | |
|---|---|
| bin | Folder |
| boot | Folder |
| dev | Folder |
| etc | Folder |
| gpfs | Folder |
| home | Folder |
| ipathfs | Folder |
| lib | Folder |
| lib64 | Folder |
| media | Folder |
| mnt | Folder |
| opt | Folder |
| proc | Folder |
| projects | Folder |
| root | Folder |
| sbin | Folder |
| scratch | Folder |
| selinux | Folder |
| srv | Folder |
| sys | Folder |

more options   Label This Transfer

This will be displayed in your transfer activity.

# GO CLI (Command-Line Interface)

- Advantages
  - Powerful – access to all features and parameters
  - Can use commands in scripts
  - Full concurrency
- Disadvantages
  - Takes a little time to learn
  - Verbose
- Transfer example:

  -
    ```
    ssh cli.globusonline.org "transfer -- \
      ncsa#BlueWaters/scratch/sciteam/<username>/a_file \
      ncsa#Nearline/u/sciteam/<username>/a_file"
    ```

# CLI Usage

- Either `ssh` into cli.globusonline.org or include `ssh cli.globusonline.org` at the beginning of each command
- Transfers
  - Use `transfer` command on individual files or on entire directories with `-r`
  - Check transfers with `status` command
  - Use `cancel` to stop a transfer
- Basic file system commands: `ls, mkdir`
- For examples, see the BW Portal
- For a complete listing and man pages, `ssh` into cli.globusonline.org and type “help”

# Optimizing Globus Online Performance

- On-site (Lustre | Nearline ⇔ Lustre | Nearline)
  - Expected performance
    - Up to ~25-30 Gbits/s for 1 transfer
    - Up to ~70-80 Gbits/s for 3 simultaneous transfers
  - Optimum conditions
    - File size of several GB+
    - 40+ files per transfer
- Offsite
  - Expected performance up to a few Gbits/s

# Globus Online Transfer Options

- Checksums
    - Verify file integrity after transfer and resend if mismatch
    - On by default – performance hit but best to leave it on
- Synchronization
    - Only transfer new or changed files
- See “more options” near the bottom left of GUI transfer window

# UberFTP

- UberFTP: GridFTP client developed at NCSA
- Completely separate from GO
- Command-line interface to Nearline
- Interactive sessions take place on Blue Waters mover nodes
  - Simpler - no endpoints to specify like with GO CLI
    - E.g. ncsa#Nearline/<path>
- What GO can't do: `chmod, chgrp`
- Avoid using for file transfers
- More info: https://bluewaters.ncsa.illinois.edu/nearline

# Other Filesystem Notes

- Moving files on Nearline
  - GO transfers through GUI and CLI only copy
  - Two methods
    - GO CLI `rename`
    - UberFTP `rename`
- Deleting large numbers of files takes a long time with GO
  - 1,000,000-file `rm -r`
    - ~10-15 minutes on /scratch
  - 1,000,000-file GO delete
    - ~40 minutes on /scratch
    - ~17 hours on Nearline

# Rules of Thumb for Data Transfers

- Don't use `scp, rsync, or sftp`
- Use UberFTP for `chmod` and `chgrp`
- Use GO CLI for scripts and moving files with `rename`
- Otherwise use GO GUI
- For onsite transfers
  - Transfer large files (several GB+)
  - Transfer 40 up to a few tens of thousands of files

# Data Sharing

- Active Project Data Share Plan – for active project allocations
  - May share from either Lustre or Nearline project storage
  - Shared files count toward quota
- Community Data Share Plan – for groups without an active allocation
  - Shared from Nearline only
  - Still being implemented
- Globus Online interface
  - Any data set size, but especially large sets
  - Access control
- Web interface
  - Only for small data sets (< few hundred files; file size < 4 GB)
  - Fully public – no access control
- More info on the portal: https://bluewaters.ncsa.illinois.edu/data-sharing