BLUE WATERS

SUSTAINED PETASCALE COMPUTING

December 4, 2013

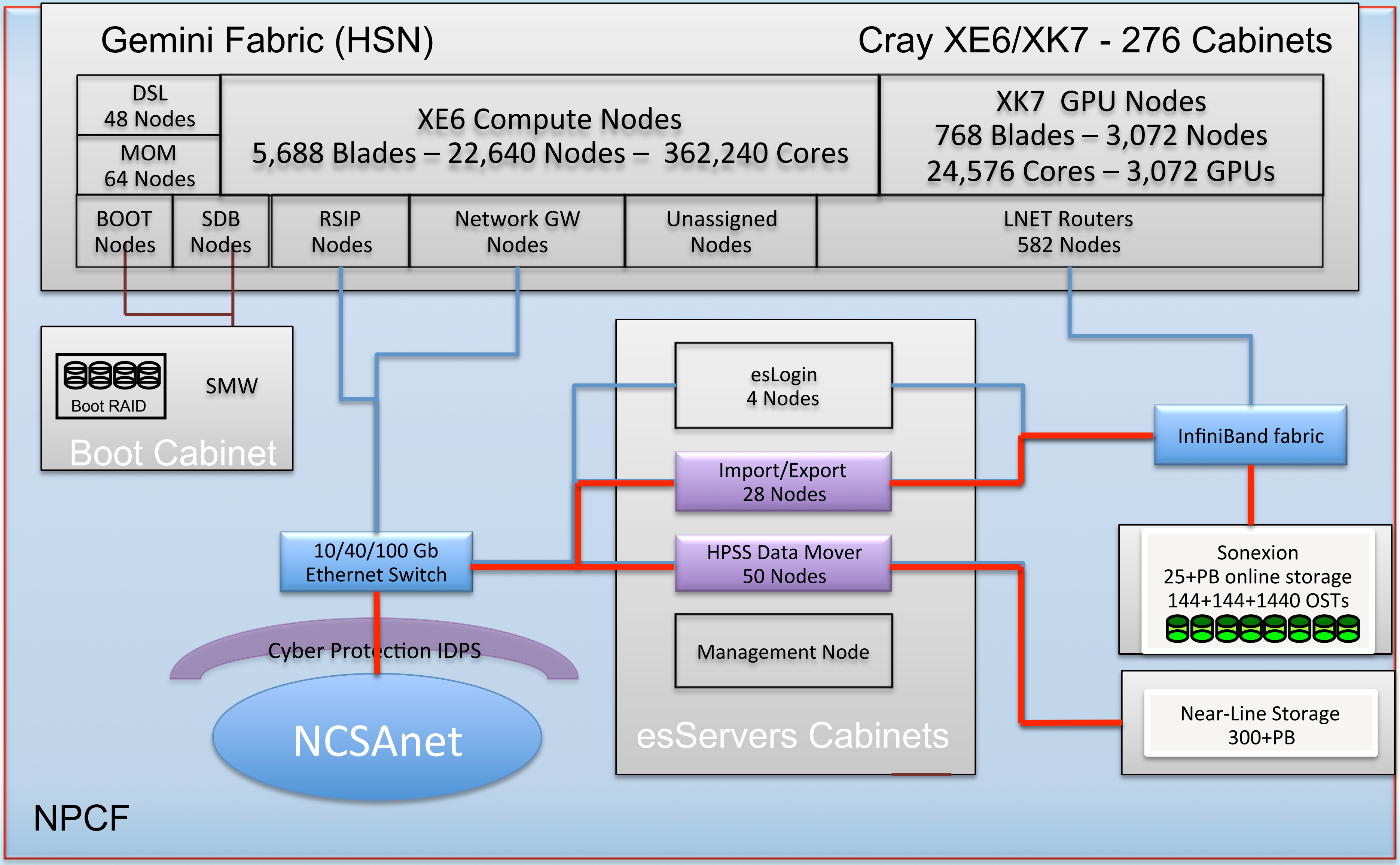
Data Management Best Practices

Ryan Mokos

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

Outline

- Overview of Nearline system (HPSS)
  - Hardware
  - File system structure
- Data transfer on Blue Waters
- Globus Online (GO) interface
  - Web GUI
  - Command-Line Interface (CLI)
- Optimizing data transfers
  - Transfer parameters
  - Transfer rates
  - Transfer errors



Gemini Fabric (HSN)

Cray XE6/XK7 - 276 Cabinets

DSL
48 Nodes

MOM
64 Nodes

XE6 Compute Nodes
5,688 Blades – 22,640 Nodes – 362,240 Cores

XK7 GPU Nodes
768 Blades – 3,072 Nodes
24,576 Cores – 3,072 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

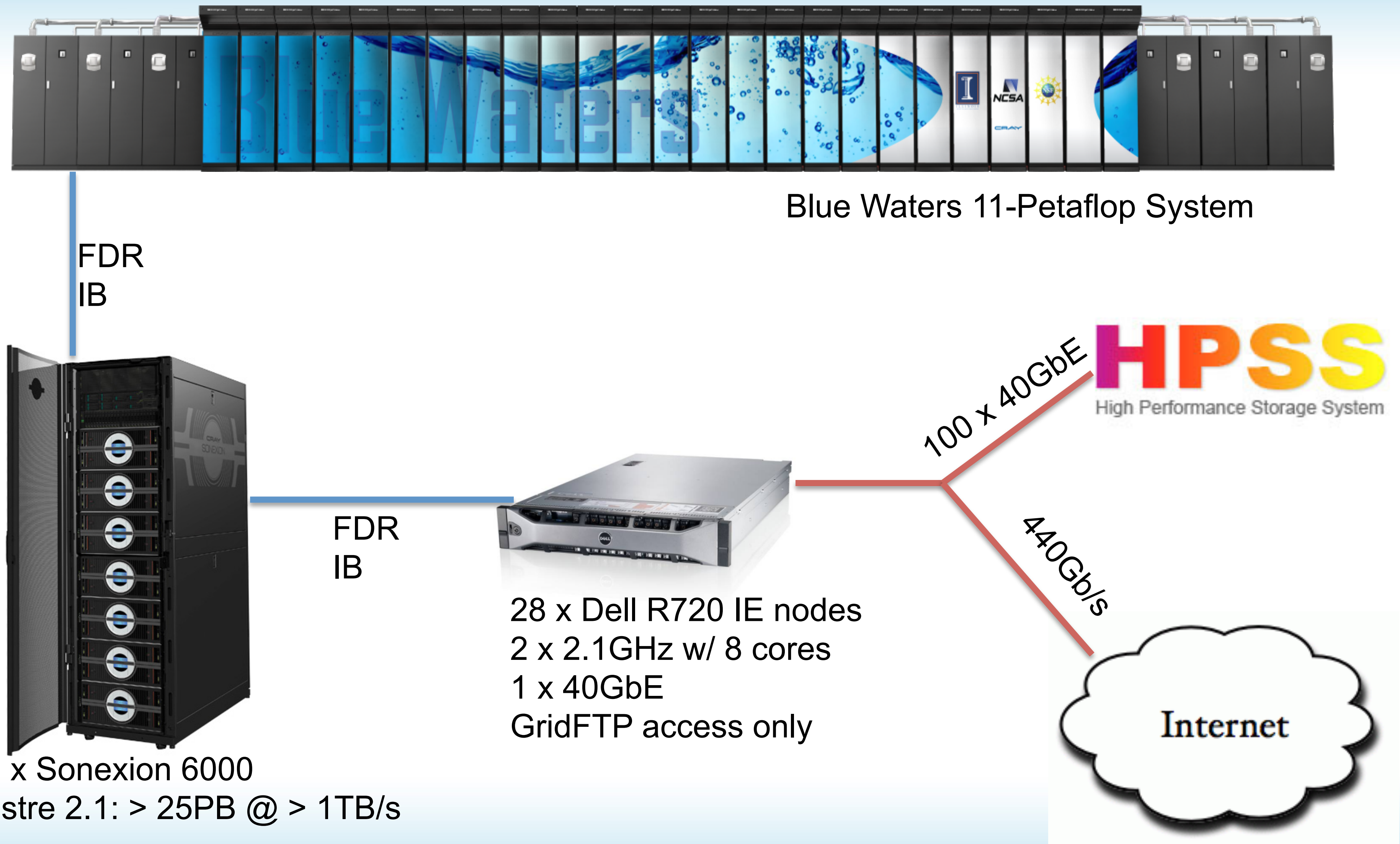
Near-Line Storage
300+PB

10/40/100 Gb
Ethernet Switch

Cyber Protection IDPS

NCSAnet

NPCF



Blue Waters 11-Petaflop System

FDR IB

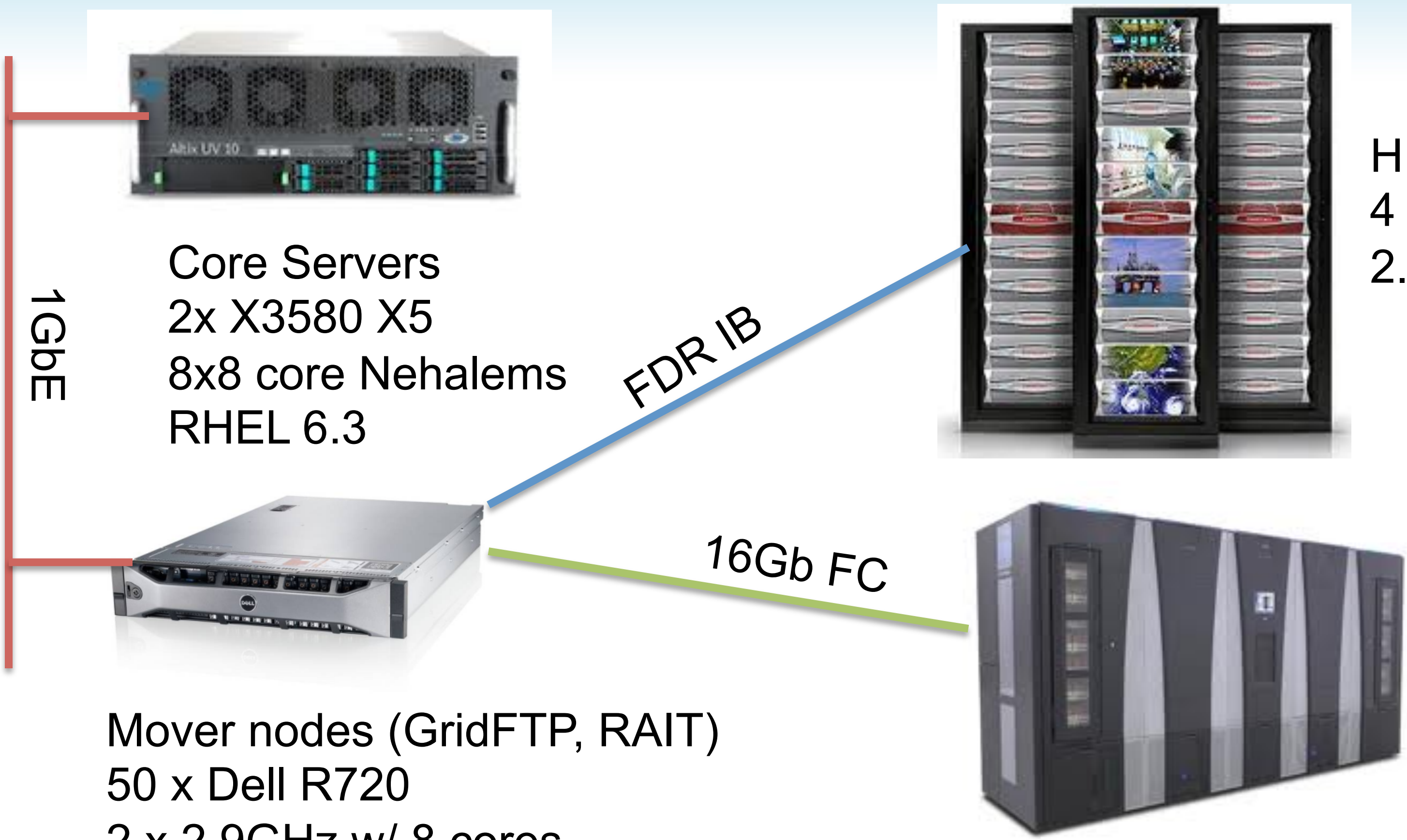
36 x Sonexion 6000
Lustre 2.1: > 25PB @ > 1TB/s

FDR IB

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

100 x 40GbE

440Gb/s



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

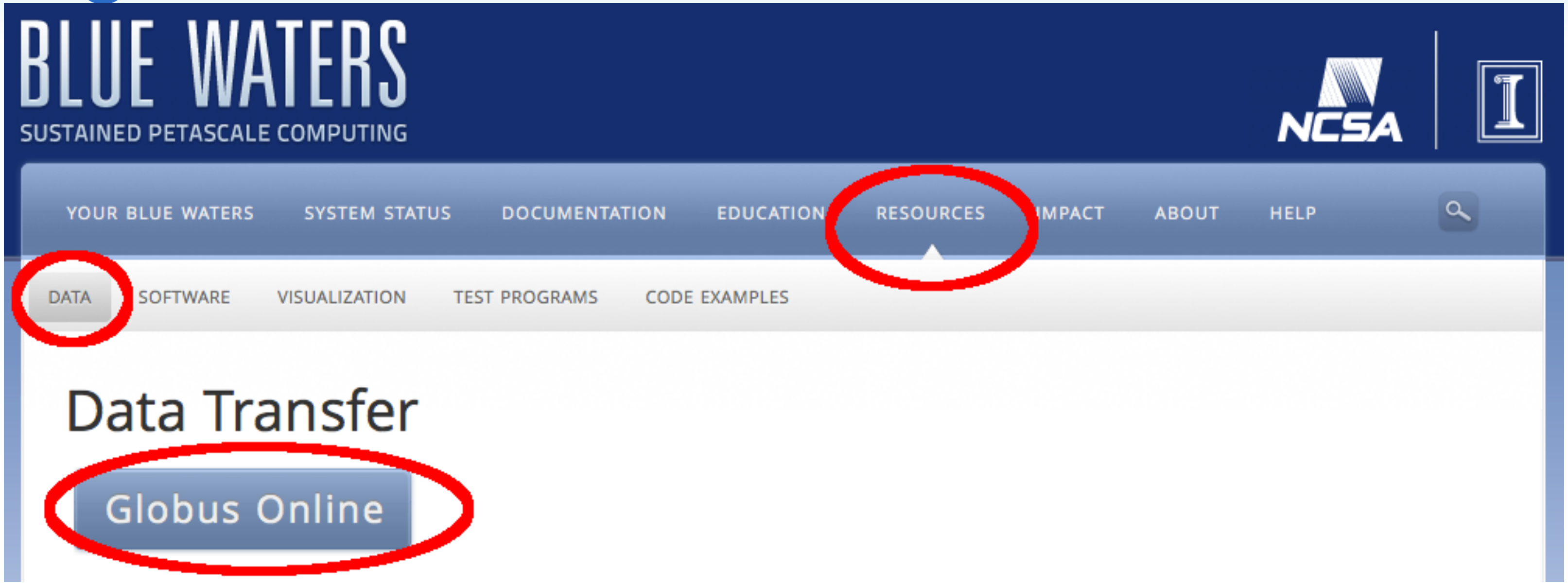
HPSS File System Structure

- Your home directory
  - /u/sciteam/<username> (same as Lustre)
  - Default quota of 5TB; can not be increased
- Your project directories
  - /projects/sciteam/<psn (e.g., jn0)> (same as Lustre)
  - Default quota of 50TB; can be increased with a request through the Blue Waters ticket system
- No purge policy! Data stays for the life of your project

Data Transfer on Blue Waters

- BW Lustre ⇔ HPSS
    - Use GO (Globus Online)
    - Cannot use scp and sftp
- BW (Lustre, HPSS) ⇔ Outside world
    - Use GO
    - Can use scp, sftp, and rsync but DON'T!
        - Impacts login node performance
        - Slower than GO
- BW Lustre ⇔ BW Lustre
    - Using cp is ok
    - GO is faster for multiple large files
        - Example: copying 50 1-GB files from /scratch to /home
            - cp: 244 sec.
            - GO: 39 sec.

Using Globus Online



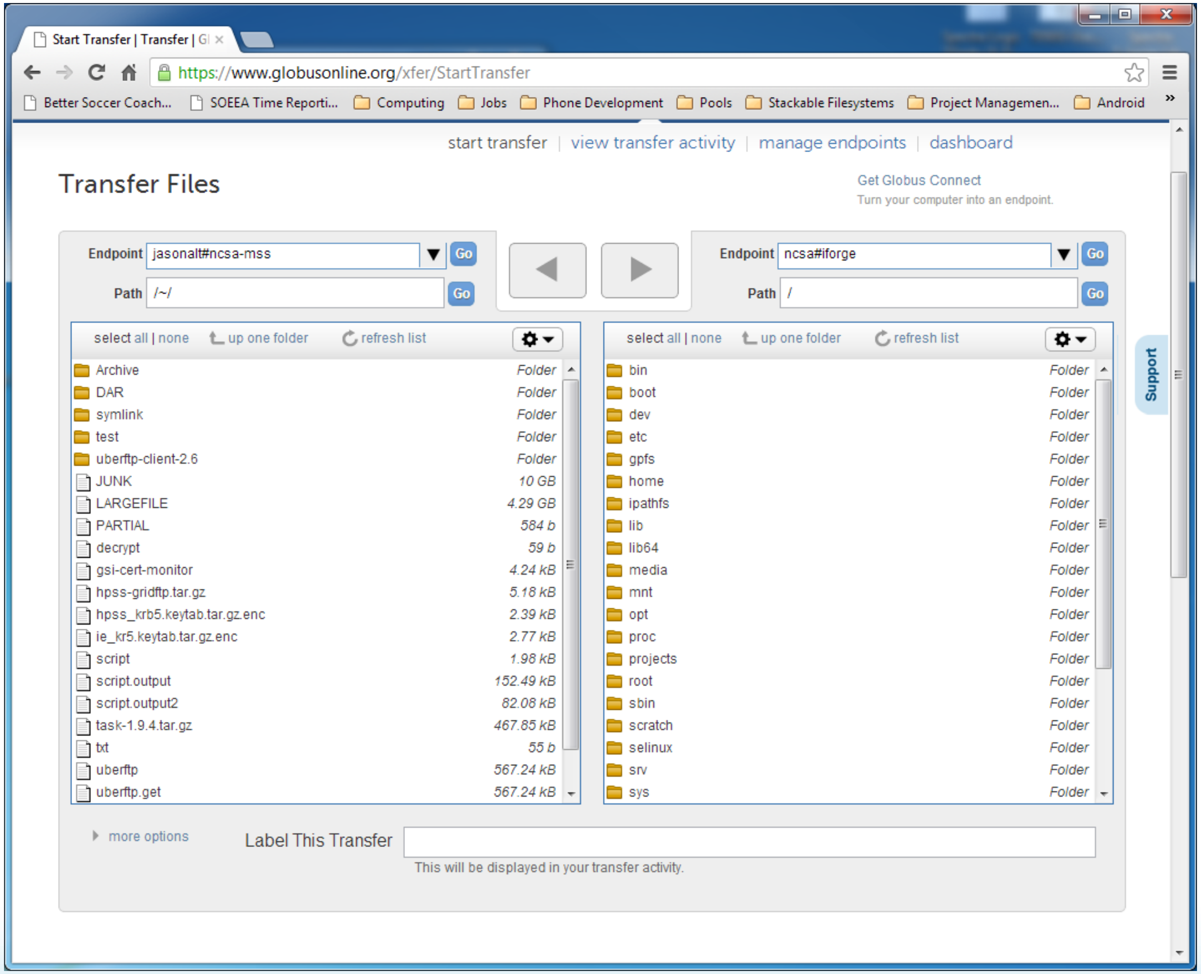
BLUE WATERS
SUSTAINED PETASCALE COMPUTING

Data Transfer

- 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
  - 100k 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

Transfer Files

Get Globus Connect
Turn your computer into an endpoint.

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

Endpoint: ncsa#iforge Go
Path: / Go

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

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”

Moving HPSS Files

- Important note: transfer commands (GUI- and CLI-based) only copy files
- To move files, use the CLI "rename" command (example on BW Portal)
- Files cannot be moved using the GO GUI

Optimizing Transfers

- GUI does pretty well, but CLI can sometimes get better results
- Transfer large files (GB+ range)
- But also transfer lots of files to take advantage of concurrency
  - Max concurrency 20 files/transfer * max 3 active transfers = up to 60 files in flight

CLI-Only Transfer Parameters

- Format: ssh cli.globusonline.org “transfer <parameters> -- <src> <dest>”
- --perf-p <num>
  - Parallelism level (data streams/control channel)
  - Valid values: 1-16
- --perf-cc <num>
  - Concurrency (number of control channels; i.e., number of files in flight)
  - Valid values: 1-16
  - Default on BW to HPSS: 20, but only see ~12
- --perf-pp <num>
  - Pipeline depth (files in flight/control channel)
  - Valid values: 1-32

Recommendations for BW ⇔ HPSS Parameters for GB-Sized Files

- Don't set --perf-p (parallelism)
- Set --perf-cc 16 (concurrency = files in flight)
- Set --perf-pp 1 (pipeline depth)
- Important note: there's a minimum queue length of 2 events, meaning you need at least 2x your concurrency in files or you won't get full concurrency
  - E.g., need >= 32 files to get 16 files in flight with --perf-cc set to 16
- Play with settings for remote sites

Transfer Rates

- Rates calculated by GO are for entire transfer, including initialization and checksum verification, if applicable
  - Checksum approximately halves the total rate
  - Whole file is transferred, then checksum is computed
- BW ⇔ HPSS for GB+ files
  - Single file transfer rate: ~2-3 Gbits/sec raw (1-1.5 Gbits/sec with checksum enabled)
  - We've seen aggregate transfer rates (16 files in flight, each file 10s of GB) up to ~36 Gbits/sec raw (18 Gbits/sec with checksum)
- Other sites for GB+ files
  - BW ⇔ Kraken and BW ⇔ Gordon: ~0.9-1.3 Gbits/sec with checksum

Transfer Errors

- Highly recommend using checksums, which are on by default for both the GUI and CLI
- Errors are infrequent but do occur
  - My testing: 1,352 50-GB transfers, 20 errors
  - Tend to occur in bursts

Other Notes

- Lustre striping
  - When transferring to BW, files inherit the stripe settings of the directory in which they're placed (unless the file is so big that it requires a higher stripe setting, in which case it's adjusted higher)
- Slow staging on HPSS tape
  - Intelligent staging in the works
  - One case: concurrency of only 2 when transferring from tape (files in the 10s of GB); 16 when transferring from HPSS disk
  - Lesson: avoid writing many many files to HPSS

Summary

- Use GO for all transfers to and from both BW and HPSS (not scp, sftp, or rsync)
- GO web GUI is simple; CLI is more powerful
- Balance large file size and large number of files to optimize transfers
  - Try to transfer files of at least 1 GB
- Store large files on HPSS; avoid many small files
  - Tar up files if necessary
    - Single-compute-node jobs recommended for large tar tasks
- Use checksums
- Ask for support: help+bw@ncsa.illinois.edu