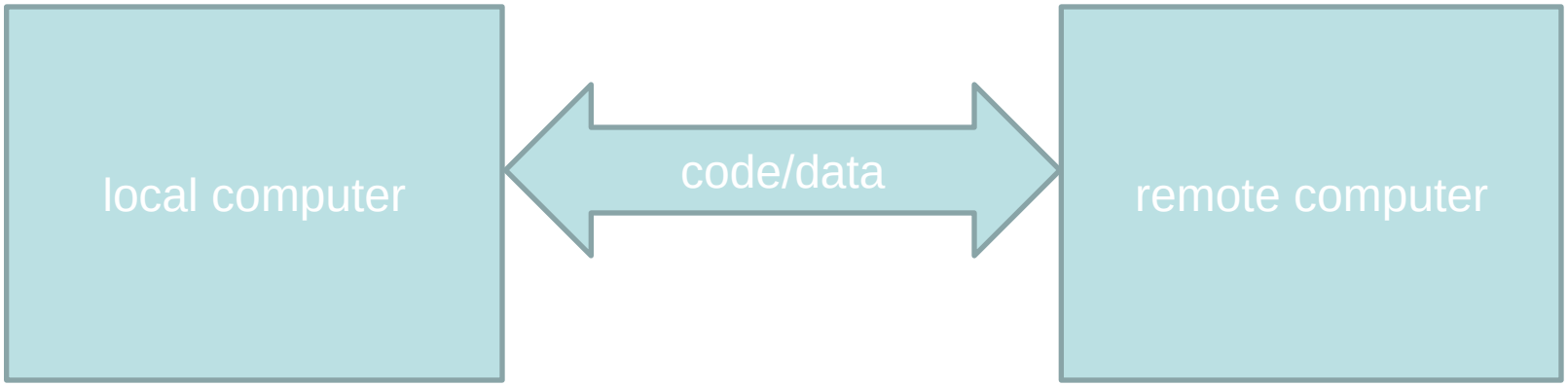
Cornell University
Center for Advanced Computing

Data Transfer

Adam Brazier – *brazier@cornell.edu*
Computational Scientist
Cornell University Center for Advanced Computing (CAC)

Data Transfer-how do I move my data from here to there?



- Needs to be a secure transfer
- Speed becomes important as the amount of data increases

Data storage options on Stampede

File system	Total Size	User Quota	Short cut	Backup Policy	Purpose
$HOME cwd at login	524TB	5GB	cdh	nightly	store source code; build executables
$WORK	450TB	1TB	cdw	none	store large files
$SCRATCH	8.5PB	none	cds	purged after 10 days	store temporary files
/tmp on each compute node	80GB	none		Purged after job completes	store files during job processing

Archival Storage on Stampede

Ranch (http://www.tacc.utexas.edu/user-services/user-guides/ranch-user-guide)

- Mass storage server called Ranch (ranch.tacc.utexas.edu) with 50 TB of online storage; 60 PB of offline tape storage; not backed up
- Uses Sun's Storage Archive Manager File system to move files in and out of a tape archival system
- Tar files before moving to Ranch; works best with large files (< 10 GB)
- Running jobs cannot access Ranch directly
- Files on tape need to be "staged" before attempting to access them

Data Transfer Software

- Easy secure transfer for small files (~15 MB/s)
  - SCP (secure copy protocol)
  - SFTP (Secure File Transfer Protocol) like SCP, but has browsing capability
  - rsync--only copies parts of files or directories that differ between machines

- Transfers using GridFTP protocol or similar
  - GUI Interface
    - Globus Online
  - Command Line Interface (~125 MB/s)
    - lftp command-line client allows parallel streams and supports FTPS
    - Globus Online CLI
    - Globus-url-copy

Data Transfer for Small Files--Linux

- SCP—requires password for every transfer

  local -> remote computer

  [local] $ scp localBig userName@stampede.tacc.utexas.edu:/path/to/project/directory

  remote -> local computer

  [local] $ scp userName@stampede.tacc.utexas.edu:big localBig

- SFTP—requires password for initial connection

  [local] $ sftp stampede.tacc.utexas.edu

  local -> remote computer

  put big

  remote -> local computer

  get big

Data Transfer with RSYNC—Linux Only

- Copies only those parts of a file that have changed, making it significantly faster and more efficient than other ssh transfers

  rsync source.c userName@stampede.tacc.utexas.edu:/path/to/project/directory

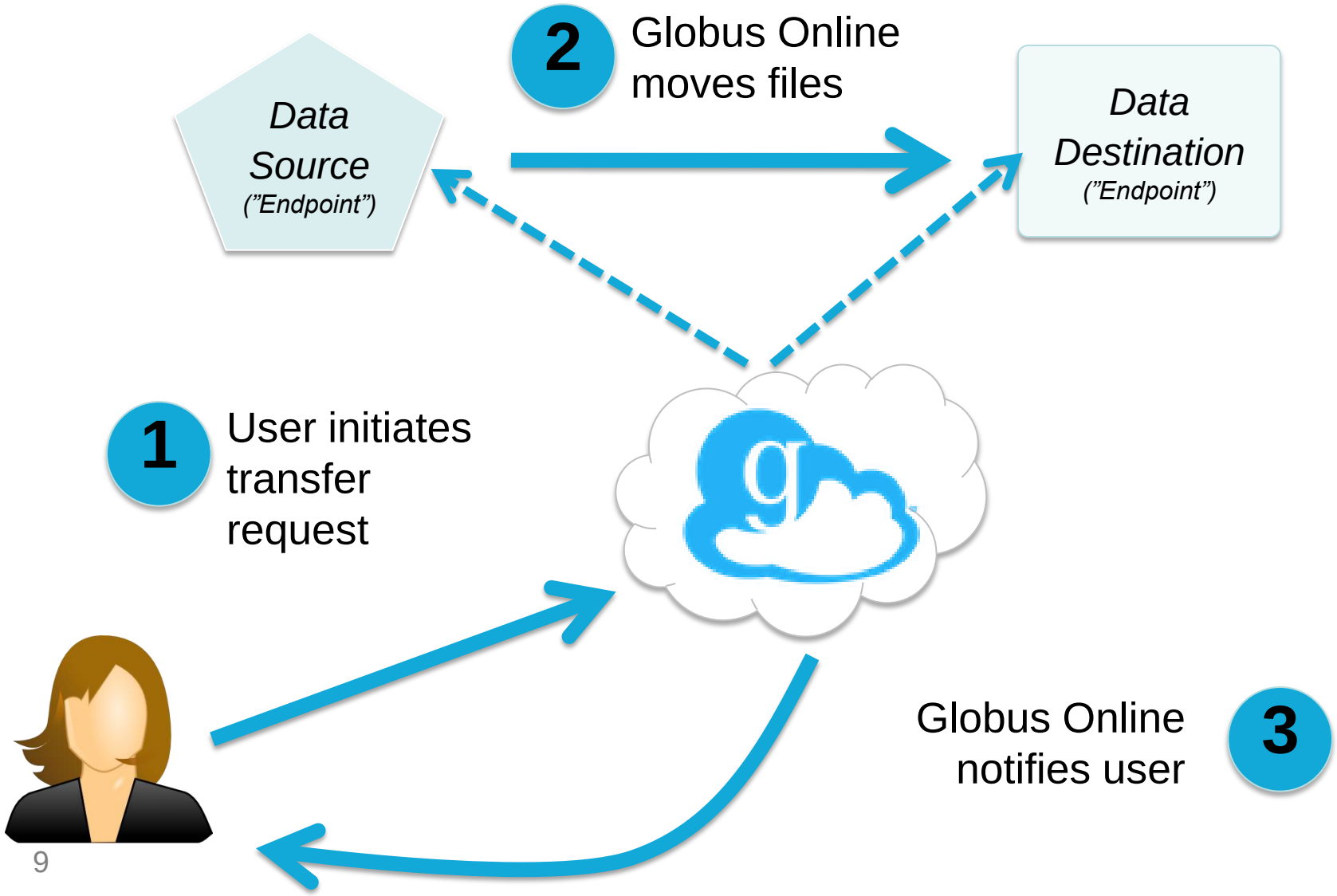
- Directory changes can also be copied recursively with rsync

  rsync –avtr ./Source userName@stampede.tacc.utexas.edu:/path/to/project/NewSource
  - Options
    - -a archive mode preserves symbolic links, devices, attributes, permissions, ownerships, etc
    - -t keeps modification times
    - -v verbose increases the information displayed during transfer
    - -r transfers the files recursively

Data Transfer for Small Files--Windows

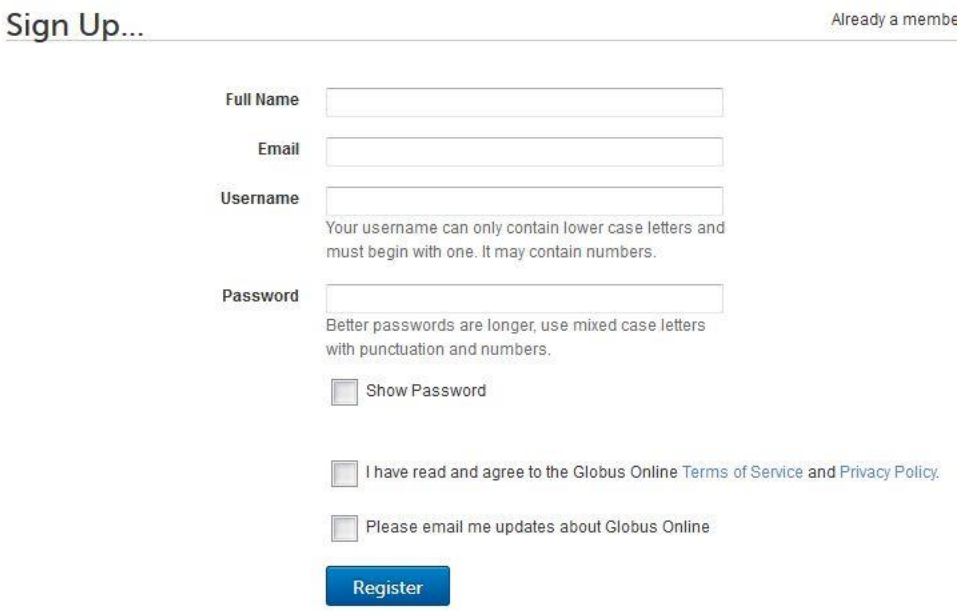
- There are a number of SCP and SFTP clients for windows
  - **Putty for both SCP and SFTP** (http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html)
  - **WinSCP for drag-and-drop SCP (can use stored PuTTY sessions), SFTP** (http://winscp.net)
  - **FileZilla for SFTP, FTPS** (https://filezilla-project.org/)
  - **FireFTP (Firefox plugin) for SFTP, FTPS** (https://fireftp.net/)
- The syntax of the commands is the same for Windows and Linux

How It Works


2
Globus Online moves files
Data Source ("Endpoint")
Data Destination ("Endpoint")
1
User initiates transfer request
Globus Online notifies user
3

Globus

- Get a Globus account https://www.globusonline.org/SignUp



Globus

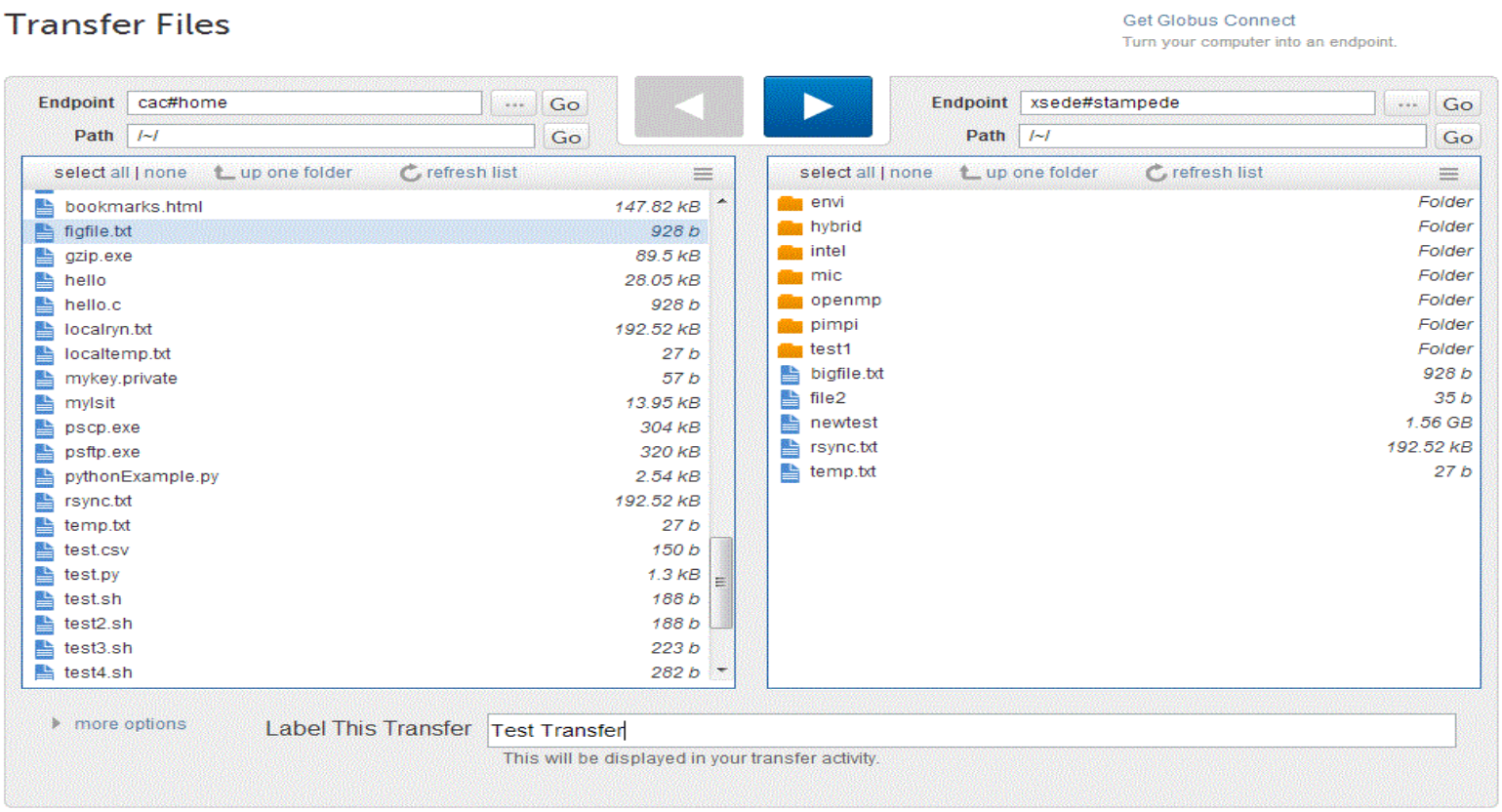
- Install Globus Connect https://www.globusonline.org/globus_connect/

  Available for Linux, Windows, Mac OS X

- Use Globus Online https://www.globusonline.org/dashboard/Main

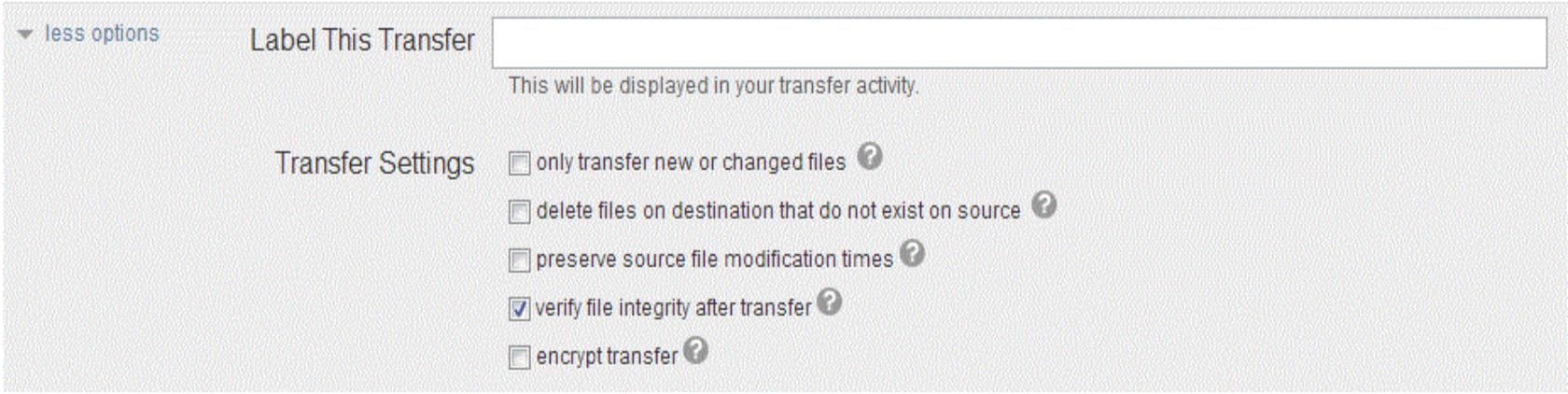
Globus

- Transfer files https://www.globusonline.org/xfer/StartTransfer



Globus

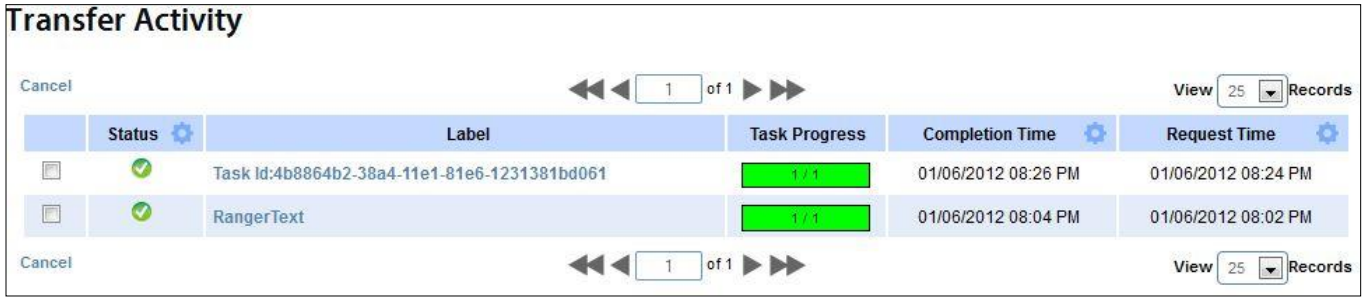
- Options (expand “more options” in transfer dialog)



  - “only transfer new or changed files” operates like rsync
  - “encrypt transfer” will slow transfer, but can be important for certain types of data

Globus

- Check on file transfers https://www.globusonline.org/xfer/ViewTransfers



Email Notification:

Task ID : c30dc1b2-389a-11e1-81e6-1231381bd061
Task Type : TRANSFER
Status : SUCCEEDED
Request Time : 2012-01-06 20:02:40Z
Deadline : 2012-01-07 20:02:39Z
Completion Time : 2012-01-06 20:04:14Z
Total Tasks : 1
Tasks Successful : 1
Tasks Canceled : 0
Tasks Failed : 0
Command : API 0.10 GO
Label : RangerText
Files : 1
Files Skipped : 0
Directories : 0
Bytes Transferred: 104857600
Bytes Checksummed: 0
MBits/sec : 8.924

Globus Online CLI

- Create a Globus Online Account

  No need to download Globus Client Software

- Enable globus account for ssh

  add SSH public key https://www.globusonline.org/account/ManageIdentities

- ssh to cli.globusonline.org

  ssh username@cli.globusonline.org

- Transfer files using globus scp

  scp –D xsede#stampede:file.txt cac#home:newfile.txt

  use the –D option to run the transfer in the background

https://www.globusonline.org/usingcli/

https://www.globusonline.org/beyondbasics/

Globus-url-copy

- Transfer between sites with GridFTP servers or via a 3rd party
- Preferred method for transferring files between XSEDE sites (including to and from Ranch)
- Necessary steps for transferring files on XSEDE
  - module load globus
  - Grid-proxy-info (check for a valid proxy)
  - myproxy-logon (if you don't have a valid proxy)
- Syntax for transferring files—can be incorporated in a script

  globus-url-copy gsiftp://sourceURL gsiftp://destinationURL
- XSEDE GridFTP server name without the ":2811"

  https://www.xsede.org/web/guest/data-transfers#table12

Tips

- Small files will transfer faster with scp or sftp than those using GridFTP protocol
- Globus Online and Globus Online CLI have the same transfer rates
  - If you are transferring a large number of files, tar them; aim for a tar ball < 10GB
- Cornell-related Globus advice at www.cac.cornell.edu/wiki/
- When updating files, use rsync or the similar option in Globus Online
- Data transfer is resource intensive
  - limit simultaneous transfers to 3 or less
  - only one globus-url-copy should be active at a time
  - avoid using the recursive (-r) flag with large transfers
- Beware of cross platform issues with filenames
  - avoid spaces in the names
  - Linux is case sensitive and Windows is not

References

- TACC User Guides
  - https://www.xsede.org/web/guest/tacc-stampede
  - https://www.xsede.org/tacc-ranch
  - https://www.xsede.org/software/globus

- Globus
  - http://www.globusonline.org
  - support@globusonline.org