TSCC Bootcamp: Introduction to Accessing and Running Jobs on the TSCC System
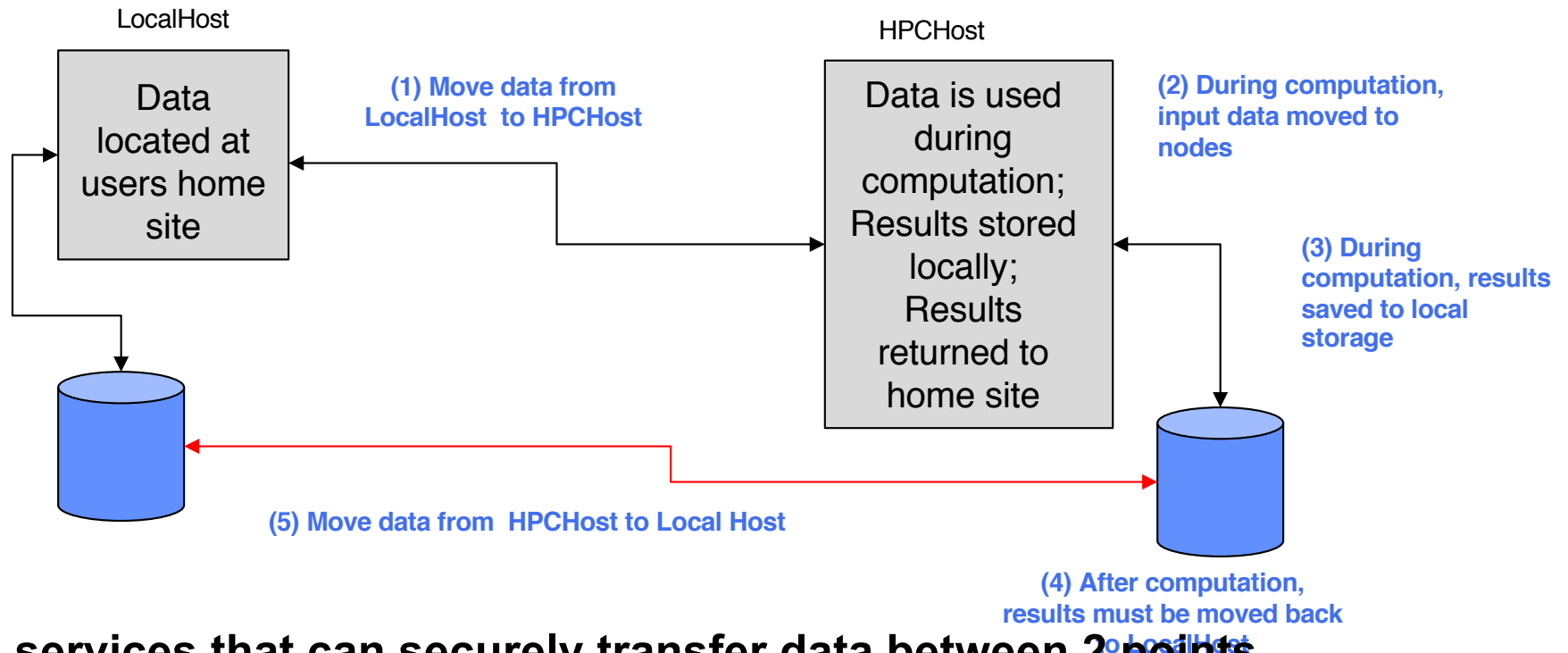
Moving Data

Moving Data To/From TSCC

LocalHost

Data located at users home site

(1) Move data from LocalHost to HPCHost

HPCHost

Data is used during computation; Results stored locally; Results returned to home site

(2) During computation, input data moved to nodes

(3) During computation, results saved to local storage

(5) Move data from HPCHost to Local Host

(4) After computation, results must be moved back to LocalHost

- **Need services that can securely transfer data between 2 points.**
  - Hosted on compute nodes, storage devices,

- **Most common commands are SCP, SFTP, wget**
  - Apps run from command line, from scripts, or other apps

Factors Affecting TSCC Data Transfer

- **Network bandwidth:**
  - CPU network: 10GbE (QDR InfiniBand optional)
  - Home-to-work: TCP 10 MB/sec, Cox optical ~ 400MB/sec
  - Time to xfer 1 TB with Cox optical fiber to SDSC:

    Txfer ~ (1000 GB)*(8 Bits/Byte)/(400 MBits/sec) ~ 20,000 secs ~ 5.5 hours

- **IO hardware on node, storage, servers**
- **Data size and type**
- **Other activity on node, network, servers**

Ways to Move Data

- **Direct Transfers:**

  - Secure copy protocol (SCP): secure transfer of computer files between a local host and a remote host or between two remote hosts.

  - SSH File Transfer Protocol (SFTP): SSH + remote file system management functionality, from within a shell

  - GNU wget: used for retrieving files using HTTP, HTTPS, FTP and FTPS. Best for web page scraping.

  - GUI Apps: Fetch, Filezilla

  - Many other applications (IBM aspera, ….)

- **3rd party file transfers:**

  - move data between endpoints (services); independent of requester.

  - Globus IO:

- **TSCC Data Mover Node**: tscc-dm1.sdsc.edu

  - need to login to move data from external location

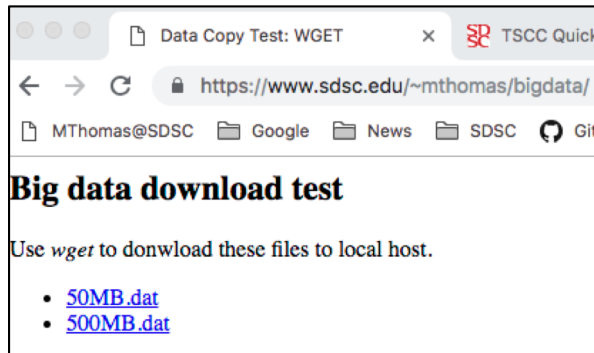# SCP Example: Copy from Laptop to TSCC System

Syntax:

scp [options] username1@source_host:directory1/filename1
username2@destination_host:directory2/filename2

```
[mthomas@gidget:~/working/bigdatafiles]
[mthomas@gidget:~/working/bigdatafiles] ll 50*.dat
-rw-r--r--  1 mthomas  staff  500000000 Sep 29  2015 500MB.dat
-rw-r--r--  1 mthomas  staff   50000000 Sep 29  2015 50MB.dat
[mthomas@gidget:~/working/bigdatafiles] scp 50MB.dat tscc-login.sdsc.edu:bigdata:50MB.dat
50MB.dat                                        100%   48MB  23.8MB/s   00:02
[mthomas@gidget:~/working/bigdatafiles]
[mthomas@gidget:~/working/bigdatafiles] scp 500MB.dat tscc-login.sdsc.edu:bigdata:500MB.dat
500MB.dat                                       100%  477MB  20.7MB/s   00:23
[mthomas@gidget:~/working/bigdatafiles] []
```

```
[mthomas@tscc-login2 bigdata]$ pwd
/home/mthomas/bigdata
[mthomas@tscc-login2 bigdata]$ ls -al
total 537325
drwxr-xr-x 2 mthomas sdsc            4 Jan 25  2019 .
drwxr-xr-x 5 mthomas sdsc           16 Jan 25 19:32 ..
-rw-r--r-- 1 mthomas sdsc 500000000 Jan 25  2019 500MB.dat
-rw-r--r-- 1 mthomas sdsc  50000000 Jan 25 19:32 50MB.dat
[mthomas@tscc-login2 bigdata]$ 
```

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UC San Diego

Using GNU wget to Transfer files from Web site to TSCC

Globus Data Transfer

- **3rd Party file transfer**
- **Service provider hosts "endpoints"**
- **Client installs local software or visits website.**
- **Client submits request to transfer files from A→B**
  - Globus gets to work moving/monitoring transfer of data
- **Client goes and gets coffee.**
- **Faster than SCP/SFTP**

transfer

secure endpoint

2 Globus moves the data for you

secure endpoint

A

B

1 You submit a transfer request

3 Globus notifies you once the transfer is complete

https://www.globus.org/data-transfer