Data Movement and Storage

Drew Dolgert
and previous contributors

Data Intensive Computing

Location

Viewing

Manipulation

Storage

Movement

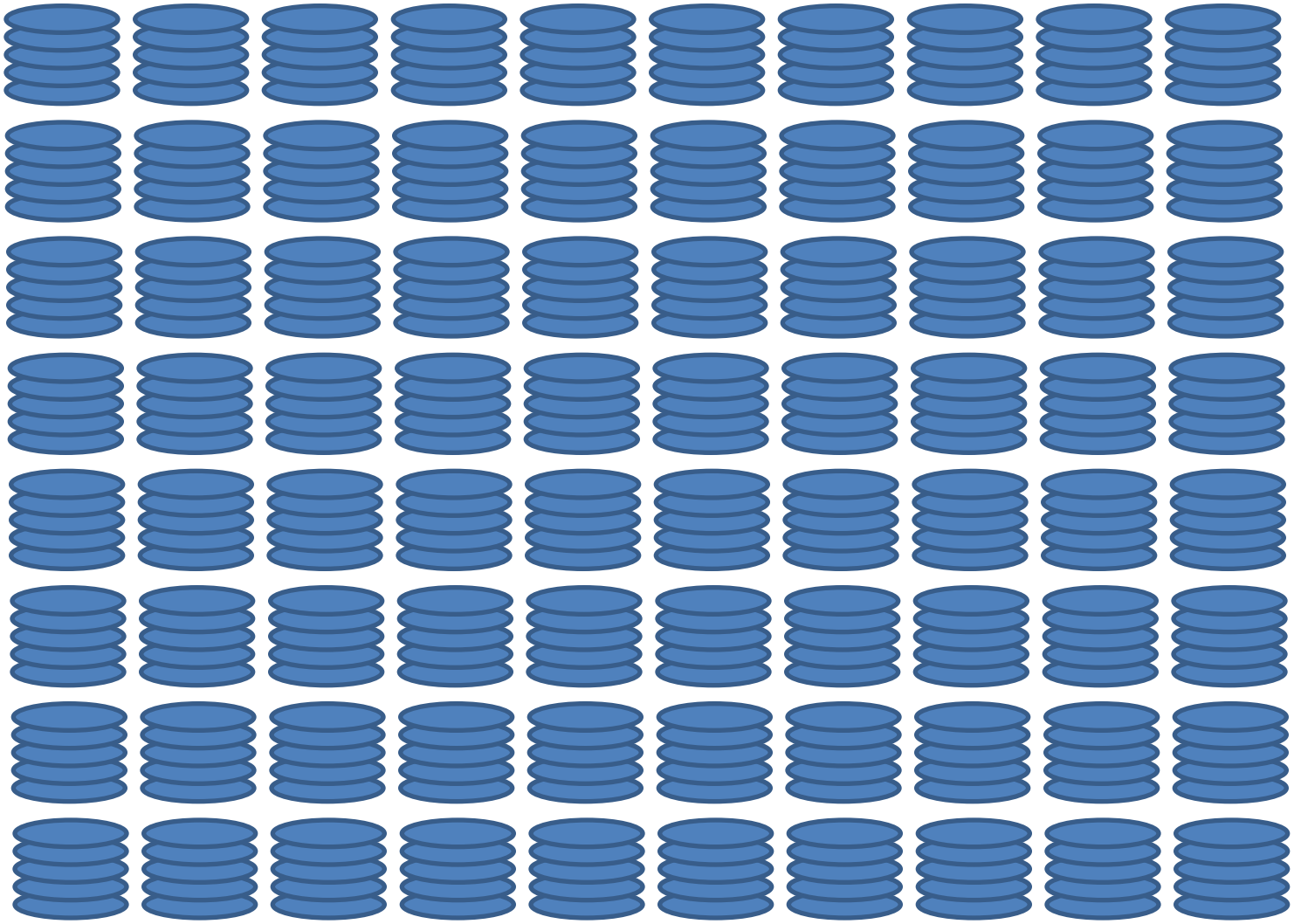
Sharing

Interpretation

$HOME

$WORK

$SCRATCH



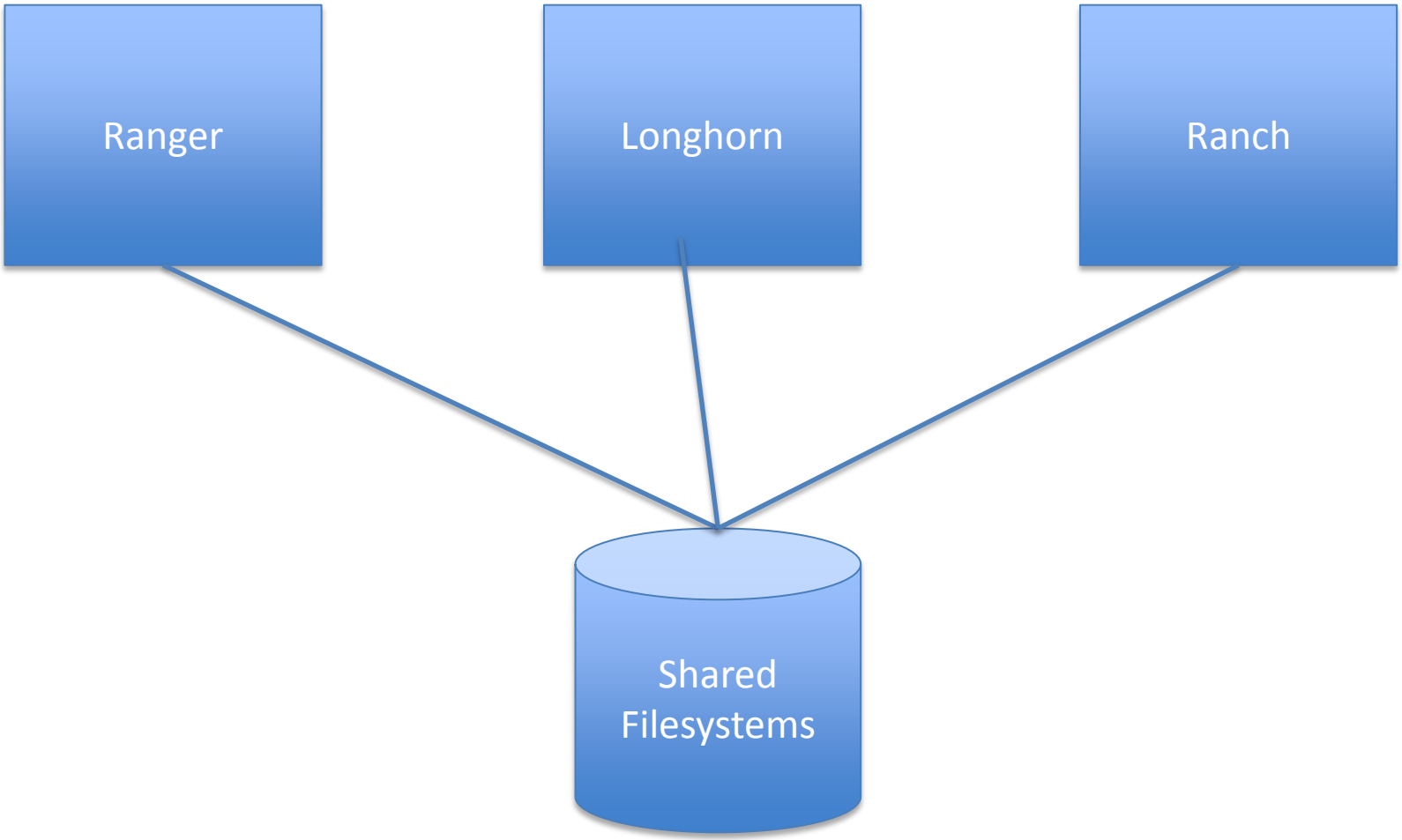
72 is a Lot, Right?

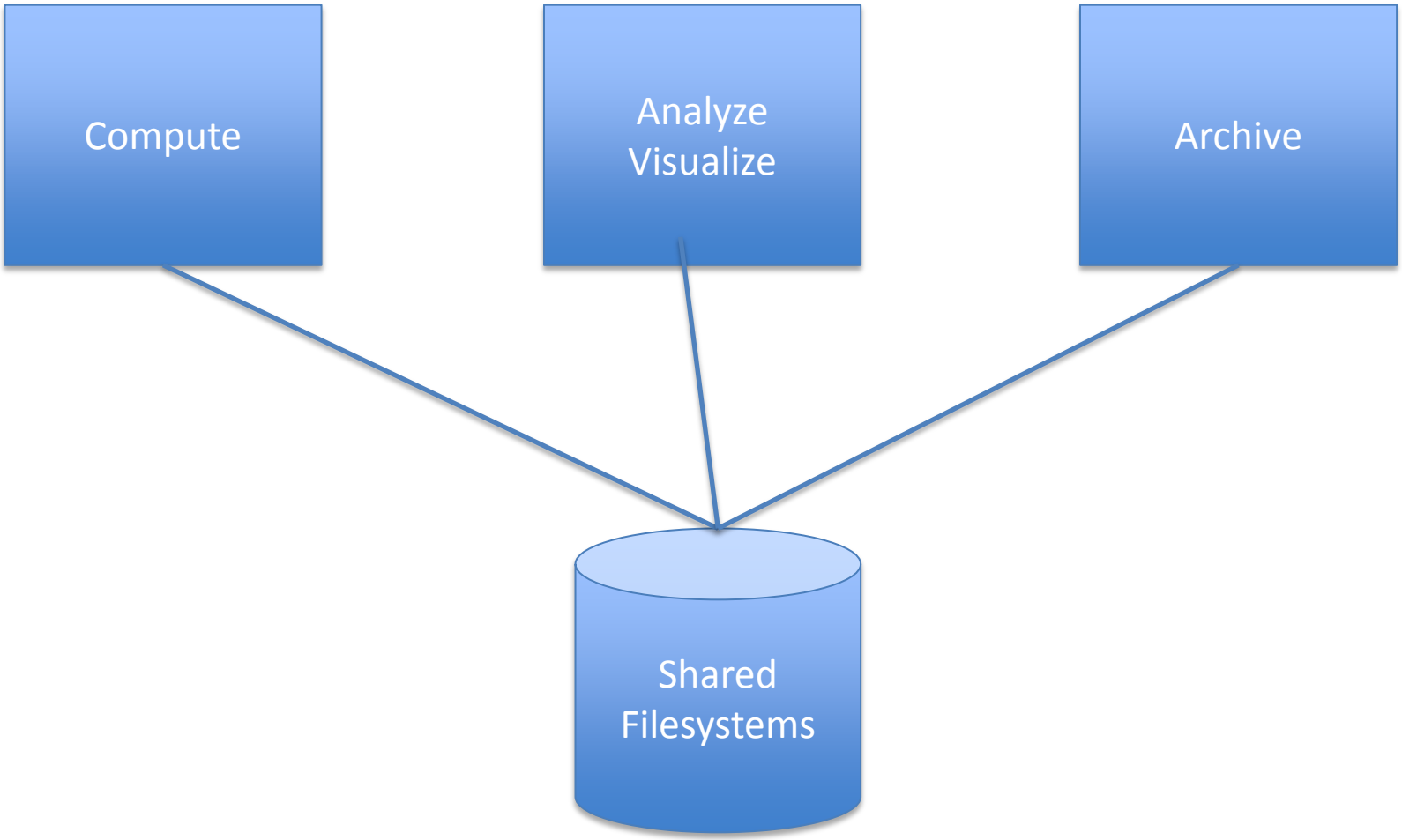
- 25-50 GB/s
- No scratch space on nodes.

- What could go wrong?

Ranch

- Ranch is TACC's long term archival storage
  - Sun StorageTek Mass Storage (1 -10 PB)
- Ranger and Spur have access to Ranch
  - [rsh|ssh] ${ARCHIVER}
- Files should be tar-ed prior to moving to Ranch, but compression is not necessary (and probably slower)
  - scp work.tar ${ARCHIVER}:${ARCHIVE}/work
- bbcp – faster movement
  - Multistream copy with fast compression built in.
  - bbcp –r < data >${ARCHIVER}:$ARCHIVE/data
- Retrieval from long-term storage
  - ssh $ARCHIVER stage "data work"
  - rcp $ARCHIVER:"data work"


Ranger
Longhorn
Ranch
Shared
Filesystems


Compute
Analyze
Visualize
Archive
Shared
Filesystems


Got it!

Basic file transfer

- SCP (secure copy protocol) is available on any POSIX machine for transfering files.

```
naw47@varushka bin]$ scp ~/oretools_svg.xpi ranger.tacc.utexas.edu:~/oretools.xpi
oretools_svg.xpi                              18% 1824KB   1.8MB/s   00:04 ETA
```

  - scp myfile.tar.gz remoteUser@ranger.tacc.utexas.edu:remotePath
  - scp remoteUser@ranger.tacc.utexas.edu:~/work.gz localPath/work.gz
- SFTP (secure FTP) is generally available on any POSIX machine and is roughly equivelant to SCP, just with some added UI features. Most notable, it allows browsing:

```
naw47@varushka bin]$ sftp consultrh5
Connecting to consultrh5...
sftp> cd stuff
sftp> lcd ../
sftp> put file
```

Basic file transfer

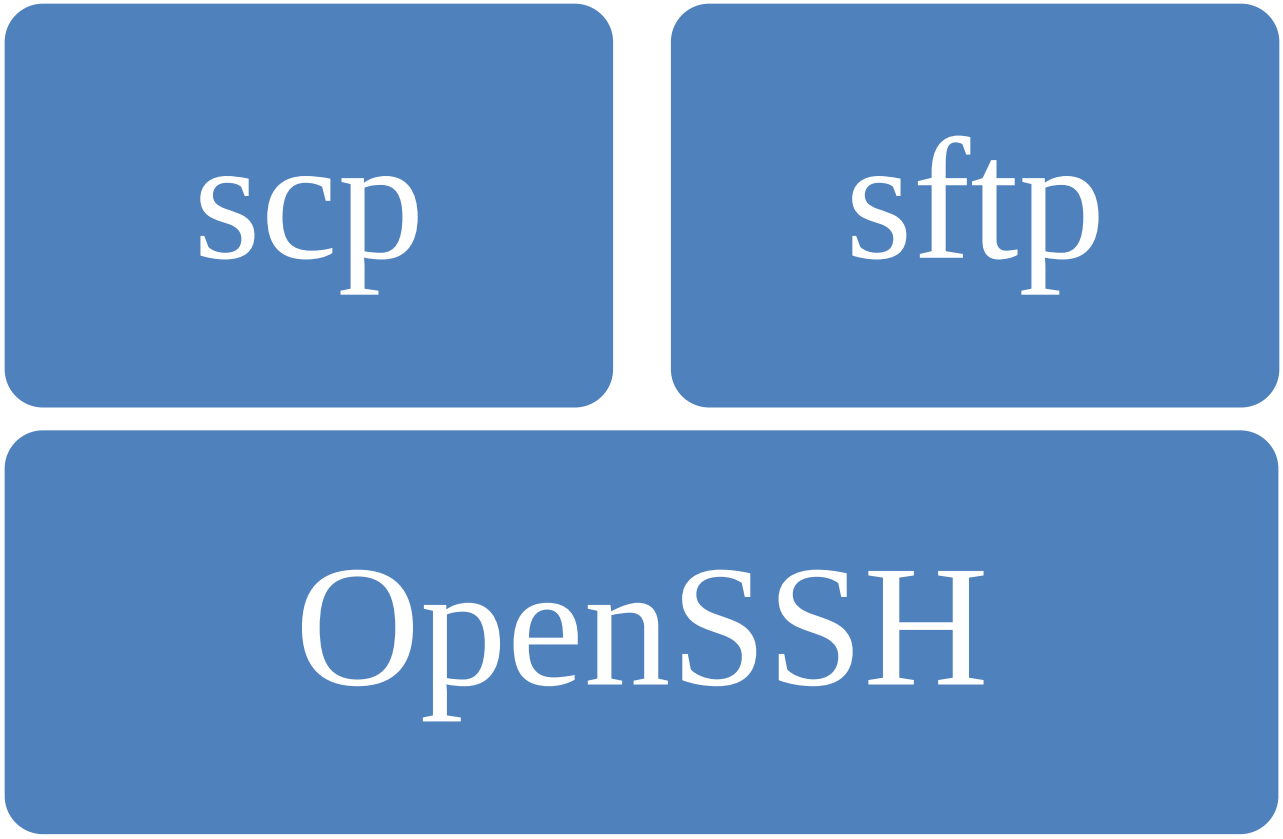
- On most Linux systems, scp uses sftp, so you're likely to see something like this:

Command	Filesize	Transfer Speed
scp	5 MB	44 MB/s (10 sec)
sftp	5 MB	44 MB/s
scp	5 GB	44 MB/s (2:00)
sftp	5 GB	44 MB/s (2:00)

- The CW is that sftp is slower than scp and this may be true for your system, but you're likely to see the above situation.

scp from to

user@machine.domain.edu:path


scp
sftp
OpenSSH

Lab: Get Good with SCP

How Much Time Do You Have?

File Size	10 Gbpx	54 Mbps
1 GB	1 sec	2.5 min
1 TB	~17 min	2.5 min
1 PB	~12 days	~5 years

Globus toolkit

- Install the globus client toolkit on your local machine and setup a few environment variables.

```
#GLOBUS Teragrid single sign-on stuff
GLOBUS_LOCATION=$HOME/globus
MYPROXY_SERVER=myproxy.teragrid.org
MYPROXY_SERVER_PORT=7514
export GLOBUS_LOCATION MYPROXY_SERVER MYPROXY_SERVER_PORT
. $GLOBUS_LOCATION/etc/globus-user-env.sh
```

- Acquire a proxy certificate and then you have a temporary certificate which will allow you to ssh/scp/sftp without re-entering a password.

```
naw47@varushka bin]$ myproxy-logon -T -l nwoody
Enter MyProxy pass phrase:
A credential has been received for user nwoody in /tmp/x509up_u16777502.
Trust roots have been installed in /home/gfs01/naw47/.globus/certificates/.
naw47@varushka bin]$ gsiscp ~/file.big ranger.tacc.utexas.edu:~/file.big
file.big                                      70%  311MB  14.8MB/s   00:08 ETA
```

UberFTP

- UberFTP is an interactive GridFTP-enabled client that supports GSI authentication and parallel data channels.
- UberFTP is to globus-url-copy what sftp is to scp
  - GSI authentication means that once you've acquired a proxy certificate from the myproxy server, you won't need to provide a password again.
  - Parallel data channels means the client opens multiple FTP data channels when transferring files, but all are controlled through a single control channel, hopefully increasing the speed.
  - UberFTP and globus-url copy also support third party transfers, which means you can transfer from a remote site to another remote site (provided they all accept the current proxy certificate).

UberFTP options

- UberFTP options are set by opening the interactive console and typing the commands.
- Parallel N
  - Set the number of parallel data connections to move your data.
  - Setting to 16 doesn't make it 16x faster, increase with high network traffic
- tcpbuf BYTES
  - Set the size of the TCP buffer used in the transfer
  - In range of 2-8 MB, decrease with network traffic, recommend leaving at system default (tcpbuf 0)
- TEST!

UberFTP example

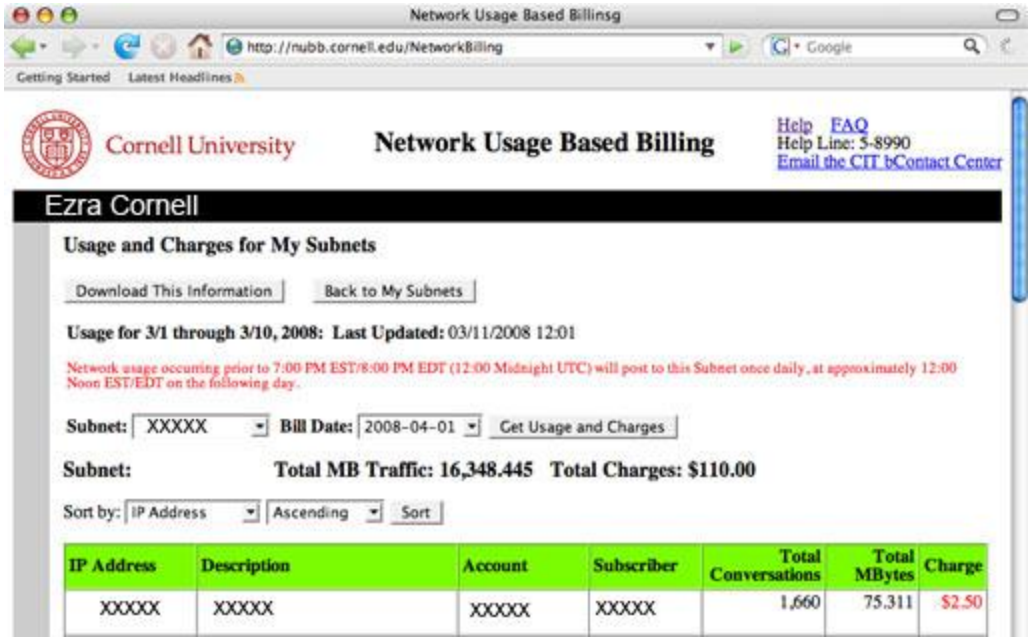
- Moving a 450 MB file from a workstation on a gigabyte connection to ranger with variable numbers of data channels.

```
naw47@varushka bin]$ uberftp ranger.tacc.utexas.edu
220 login3.ranger.tacc.utexas.edu GridFTP Server 2.8 (gcc64, 1217607445-63) [Gl
bus Toolkit 4.0.8] ready.
230 User tg801871 logged in.
UberFTP> parallel
Using 1 parallel data chanels for extended block transfers
UberFTP> put file.big
file.big: 457651136 bytes in 20.379396 Seconds (21.416 MB/s)
UberFTP> parallel 8
Using 8 parallel data chanels for extended block transfers
UberFTP> put file.big
file.big: 457651136 bytes in 15.107727 Seconds (28.889 MB/s)
UberFTP> parallel 16
Using 16 parallel data chanels for extended block transfers
UberFTP> put file.big
file.big: 457651136 bytes in 14.162568 Seconds (30.817 MB/s)
UberFTP>
```

The Seriously Out-of-date Map

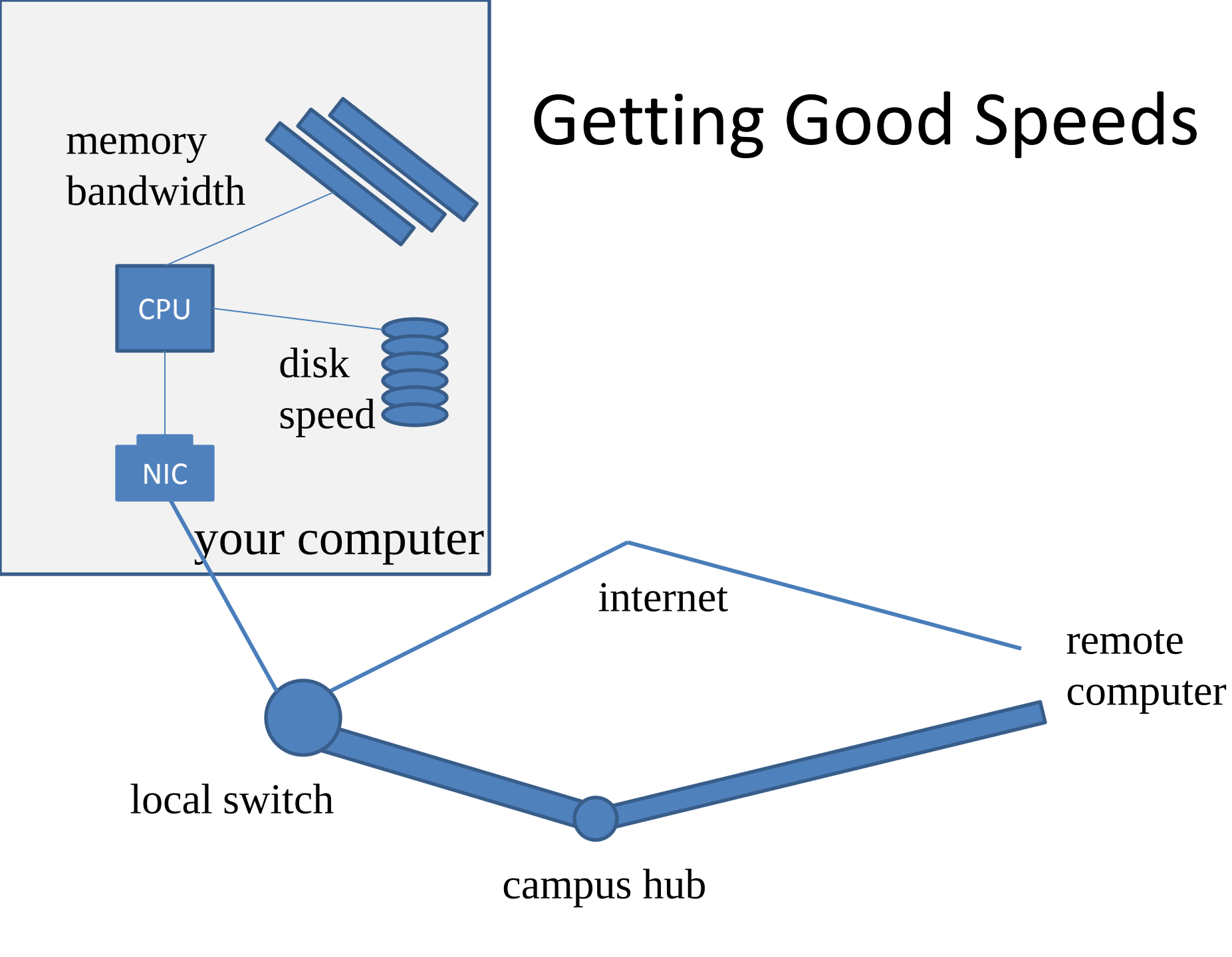
Are You on the Map?

- No NUBB charges.
- Access to 10 Gb connection on campus.
- Access to 10 Gb connection from country.
- Then test it.
  - Network ops help
  - Talk with provider



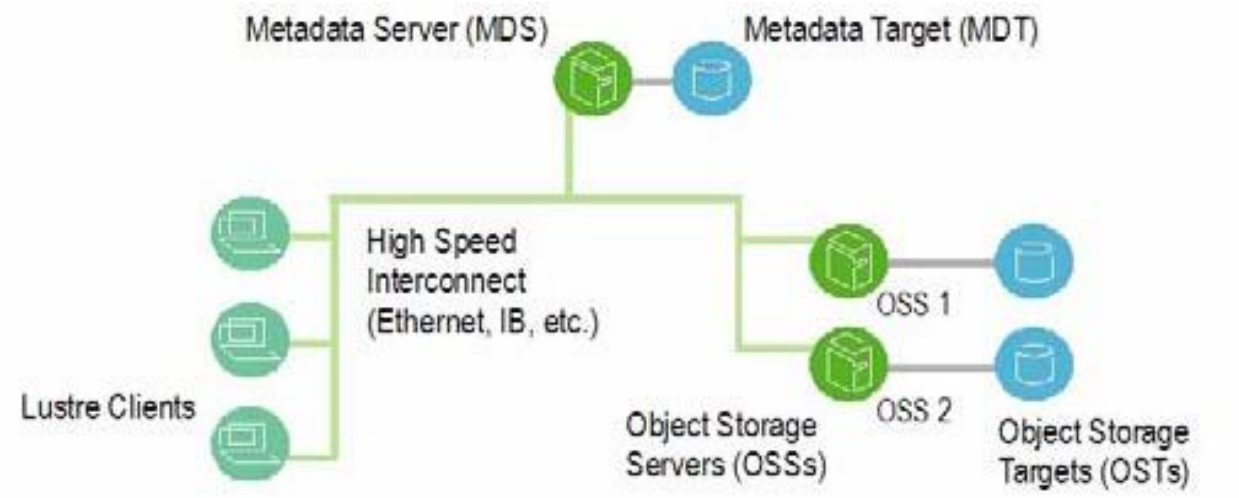
- Third-party file transfers
- Combine computation and image generation
- Remote visualization
- Apply smart filters to generated data

Getting Good Speeds



Lustre

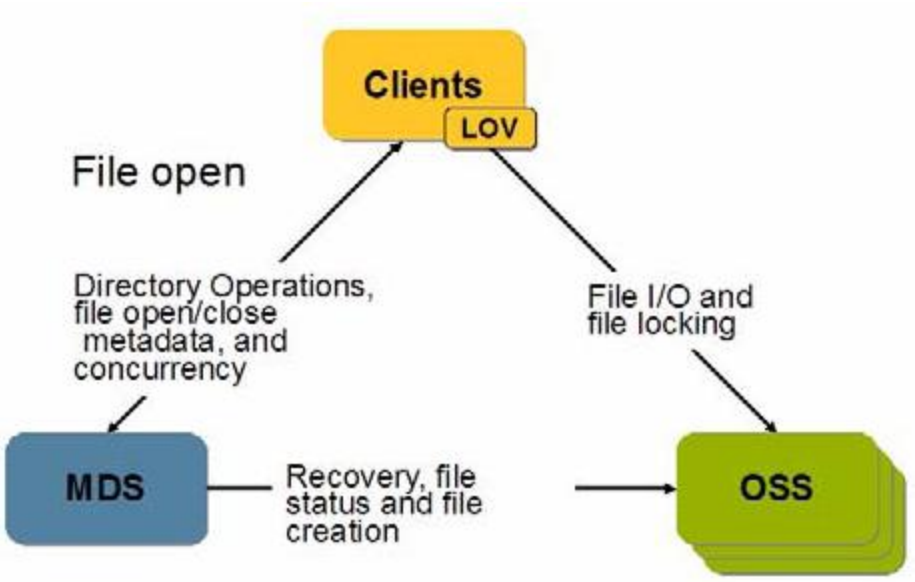
- All Ranger filesystems are Lustre, which is a globally available distributed file system.
- The primary components are the MDS and OSS nodes, OSS contain the data, MDS contains the filename to object map



Lustre Operations manual: http://manual.lustre.org/images/8/86/820-3681_v15.pdf

Lustre

- The client (you) must talk to both the MDS and OSS servers in order to actually use the Lustre system.
- Actual File I/O goes to the OSS, opening files, directory listings, etc go to the MDS.
- The client doesn't have to care, the Lustre file system simply appears like any other large volume that would be mounted on a node.



Lustre

- The Lustre filesystem scales with the number of OSS's available.
- Ranger provides 72 Sun I/O nodes, with an achievable data rate of something like 50GB/s, but this speed is being split by all users of the system.
- Fun comparison:
  - 500 MB file, on my workstation using 2 disks in a striped RAID array.
  - Same file, on Ranger, copying from $HOME to $SCRATCH
  - Lustre scales to multiple nodes reading/writing!

Workstation local copy

```
naw47@varushka ~]$ time cp file.big file2.big

real    0m1.580s
user    0m0.053s
sys     0m1.468s
```

Ranger Lustre copy

```
login4% time cp $HOME/file.big $SCRATCH/file.big
0.000u 3.020s 0:03.46 87.2%     0+0k 0+0io 0pf+0w
login4% time cp $HOME/file.big $HOME/file1.big
0.000u 2.220s 0:02.81 79.0%     0+0k 0+0io 0pf+0w
```

Lab: Striping Lustre

http://www.flickr.com/photos/musebrarian/3231408047/

http://www.flickr.com/photos/kenmccown/3174273793/

http://www.flickr.com/photos/squeakywheel/478967864/

http://www.flickr.com/photos/kruggg6/107764366/

http://www.flickr.com/photos/robbaldwin-photography/4094297085/

Source: U.S. Department of Commerce, National Oceanic and Atmospheric Administration [1]

http://www.flickr.com/photos/amagill/3367543296/

http://www.flickr.com/photos/johncohen/55582632/