

Introduction to GACRC Storage Environment

Georgia Advanced Computing Resource Center

University of Georgia

Zhuofei Hou, HPC Trainer

zhuofei@uga.edu

Outline

- What is GACRC?
- Overview of Linux Commands
- GACRC Storage Environment
- Data Transferring
- Snapshot and Backup
- Best Practice Suggestions from GACRC

What is GACRC?

Who Are We?

- Georgia **A**dvanced **C**omputing **R**esource **C**enter
- Collaboration between the Office of Vice President for Research (**OVPR**) and the Office of the Vice President for Information Technology (**OVPIIT**)
- Guided by a faculty advisory committee (GACRC-AC)

Why Are We Here?

- To provide computing hardware and network infrastructure in support of high-performance computing (**HPC**) at UGA

Where Are We?

- <http://gacrc.uga.edu> (Web) <http://wiki.gacrc.uga.edu> (Wiki)
- <http://gacrc.uga.edu/help/> (Web Help)
- https://wiki.gacrc.uga.edu/wiki/Getting_Help (Wiki Help)

Overview of Linux Commands

- Folder Navigating
- File Copying and Moving
- File Compression and Packaging
- Disk Storage and Filesystem

Overview of Useful Linux Commands

➤ Folder Navigating

`pwd`: Print the absolute path of your current directory : `pwd`

`cd`: Change current directory : `cd ..`, `cd /`, `cd /home/yourHome`

➤ File Copying and Moving

`cp`: Copy files : `cp file1 file2`, `cp file1 ./subDir`

`mv`: Rename or move files : `mv file1 file2`,
`mv file1 file2 ./subDir`

Overview of Useful Linux Commands

➤ File Compression and Packaging

`gzip`: Compress files with GNU Zip

`gzip file` → Compress *file* to create *file.gz*. Original *file* is deleted

`gunzip`: Uncompress GNU Zip files

`gunzip file.gz` → Uncompress *file.gz* to create *file*. Original *file.gz* is deleted.

Overview of Useful Linux Commands

➤ File Compression and Packaging

`tar`: Pack multiple files and directories into a single file for *transport*, optionally *compressed*

```
tar -cvf myarchive.tar ./myDir
```

➔ Create package

```
tar -tvf myarchive.tar
```

➔ List contents

```
tar -xvf myarchive.tar
```

➔ Extract package

```
tar -czvf myarchive.tar.gz ./myDir
```

➔ Create & Compress

```
tar -tzvf myarchive.tar.gz
```

➔ List contents

```
tar -xzvf myarchive.tar.gz
```

➔ Uncompress & Extract

Overview of Useful Linux Commands

➤ Disk Storage and Filesystem

`ls`: List the contents (files and subdirectories) of a directory

`ls -l` ➔ Long listing including file attributes

`ls -h` ➔ Print file sizes in KB, MB, and GB, instead of bytes

`ls -a` ➔ List all files, including hidden files whose names begin with a dot

`du`: Measure the disk space occupied by files and directories

`du -h` ➔ Measure the size of current directory and all its subdirectories

`du -h file1 file2` ➔ Measure the size of two files

Overview of Useful Linux Commands

➤ Disk Storage and Filesystem

df: Report on all mounted filesystems with the size, used space, and free space

df -h ➔ Print human-readable output, and choose the most appropriate unit for each size

Filesystem	Size	Used	Avail	Use%	Mounted on
/dev/mapper/VolGroup01-LogVol_root	99G	14G	84G	15%	/
devtmpfs	16G	0	16G	0%	/dev
tmpfs	16G	2.4M	16G	1%	/run
/dev/sda1	486M	59M	402M	13%	/boot
/dev/mapper/VolGroup01-LogVol_home	493G	86G	406G	18%	/home

GACRC Storage Environment

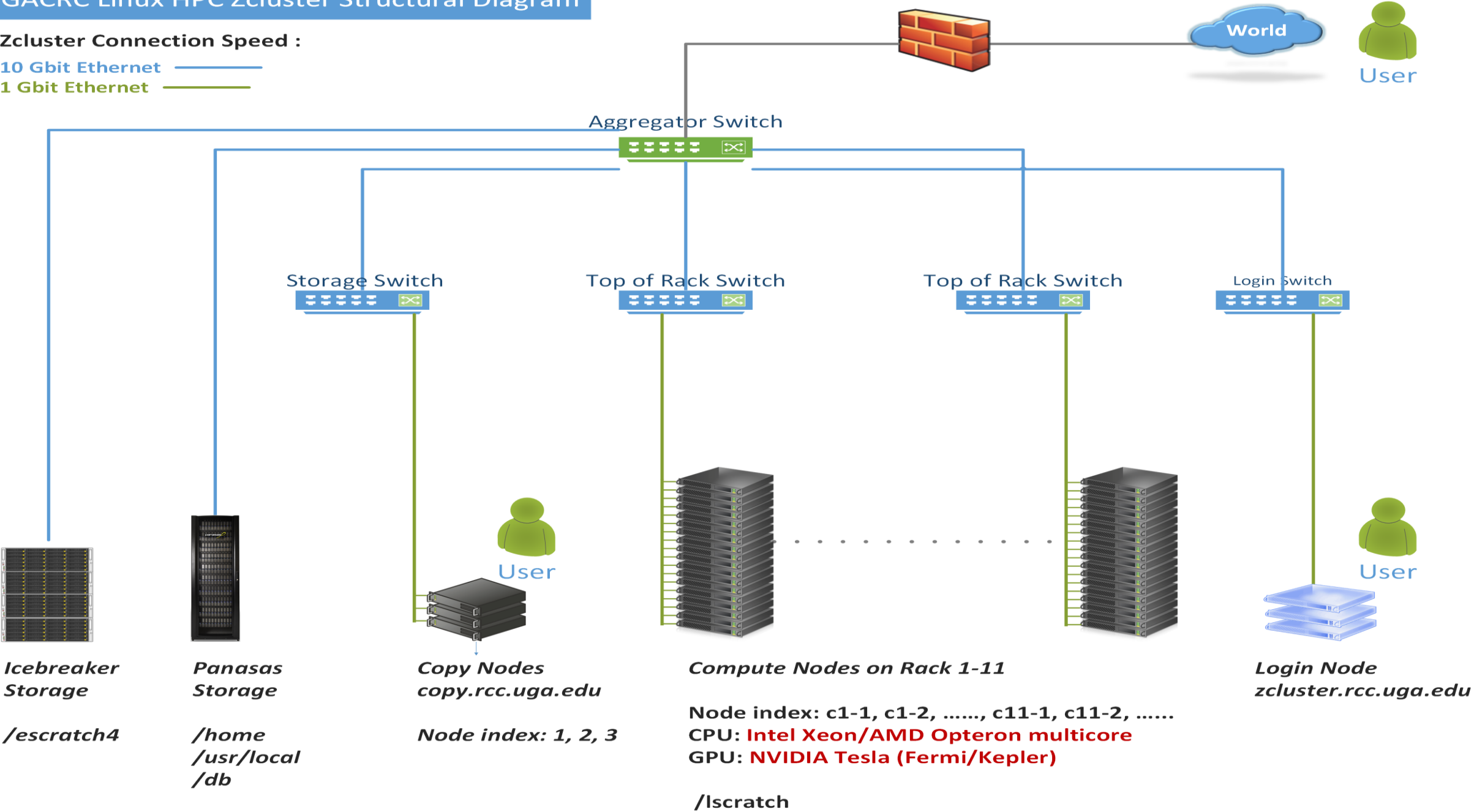
- zcluster Storage Environment
- Sapelo Storage Environment
- GACRC Storage Environment

GACRC Linux HPC Zcluster Structural Diagram

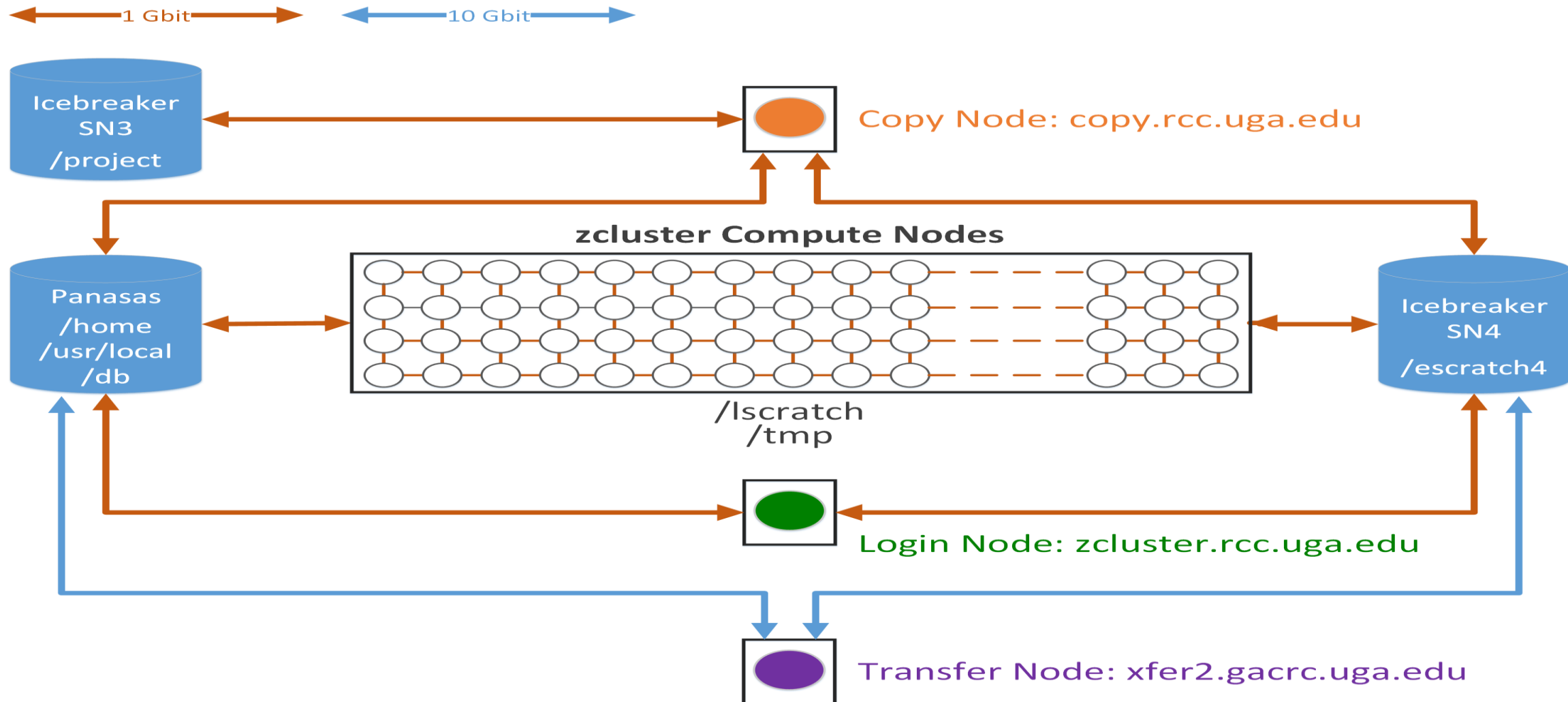
Zcluster Connection Speed :

10 Gbit Ethernet

1 Gbit Ethernet



zcluster Storage Environment



zcluster Storage Environment

Filesystem	Role	Quota	Accessible from	Intended Use	Notes
/home/abclab/username	Home	100GB	zcluster.rcc.uga.edu (Login) copy.rcc.uga.edu (Copy) Interactive nodes (Interactive) xfer2.gacrc.uga.edu (Transfer) compute nodes (Compute)	Highly static data being used frequently	Snapshots
/escratch4/username	Scratch	4TB		Temporarily storing large data being used by jobs	Auto-deleted in 37 days
/lscratch/username	Local Scratch	18 ~ 370GB	Individual compute node	Jobs with heavy disk I/O	User to clean up

- Note:
1. /usr/local : Software installation directory
/db : bioinformatics database installation directory
 2. To login to [Interactive](#) nodes, use [qlogin](#) from [Login](#) node

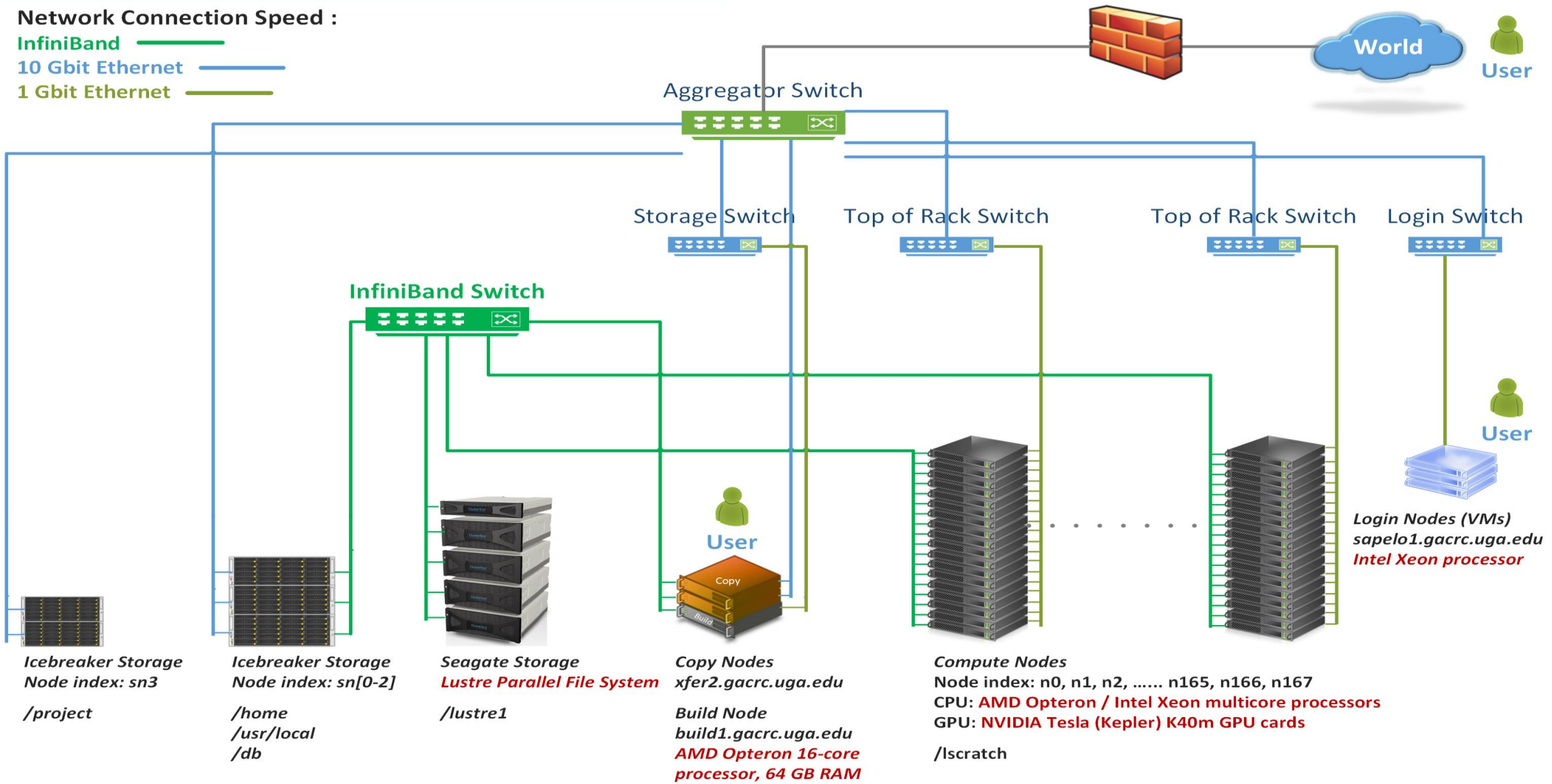
zcluster Storage Environment

Main Function	Related Filesystem	On/From-Node
Login Landing	/home/abclab/username (Home)	Login
Batch Job Submitting	/home/abclab/username (Home) /escratch4/username (Scratch)	Login or Interactive
Interactive Job Running	/home/abclab/username (Home) /escratch4/username (Scratch)	Interactive
Data Archiving , Compressing and Transferring	/home/abclab/username (Home) /escratch4/username (Scratch)	Copy or Transfer
Job Data Temporarily Storing	/lscratch/username (Local Scratch) /escratch4/username (Scratch)	Compute

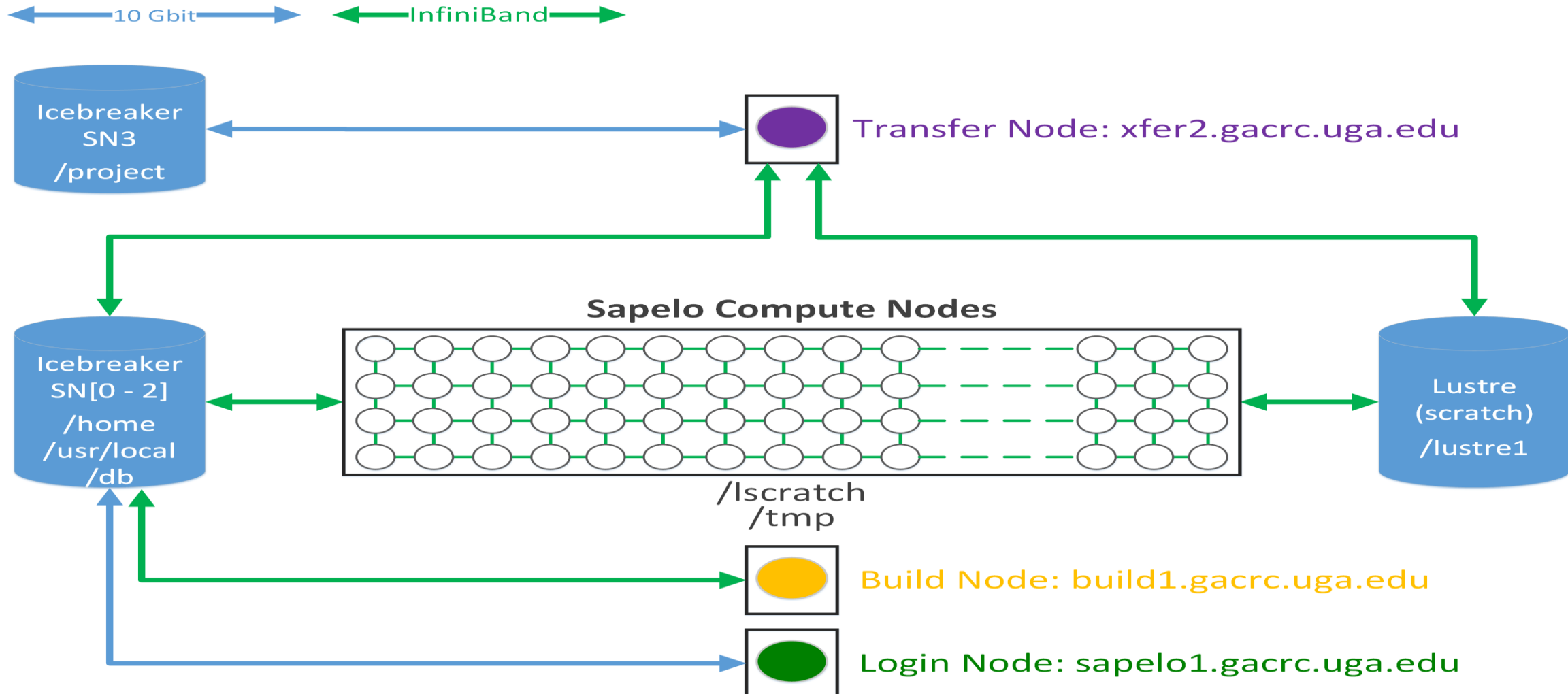
The New GACRC Linux HPC Cluster Structural Diagram

Network Connection Speed :

- InfiniBand
- 10 Gbit Ethernet
- 1 Gbit Ethernet



Sapelo Storage Environment



Sapelo Storage Environment

Filesystem	Role	Quota	Accessible from	Intended Use	Notes
/home/username	Home	100GB	sapelo1.gacrc.uga.edu (Login) Interactive nodes (Interactive) xfer2.gacrc.uga.edu (Transfer) build1.gacrc.uga.edu (Build) compute nodes (Compute)	Highly static data being used frequently	Snapshots
/lustre1/username	Scratch	No Limit	Interactive nodes (Interactive) xfer2.gacrc.uga.edu (Transfer) compute nodes (Compute)	Temporarily storing large data being used by jobs	Auto-moved to /project if 30 days no modification
/lscratch/username	Local Scratch	250GB	Individual compute node	Jobs with heavy disk I/O	User to clean up
/project/abclab	Storage	Variable	xfer2.gacrc.uga.edu (Transfer)	Long-term data storage	Group sharing possible

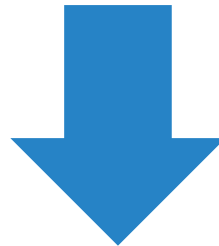
- Note:
1. /usr/local/apps : Software installation directory
/db : bioinformatics database installation directory
 2. To login to Interactive nodes, use `qlogin` from Login node

Sapelo Storage Environment

Main Function	Related Filesystem	On/From-Node
Login Landing	/home/username (Home)	Login
Batch Job Submitting	/home/username (Home)	Login or Interactive
	/lustre1/username (Scratch)	Interactive
Interactive Job Running	/home/username (Home) /lustre1/username (Scratch)	Interactive
Data Archiving , Compressing and Transferring	/home/username (Home) /lustre1/username (Scratch)	Transfer
Job Data Temporarily Storing	/lscratch/username (Local Scratch) /lustre1/username (Scratch)	Compute
Code Compilation	/home/username (Home)	Build

GACRC Storage Environment

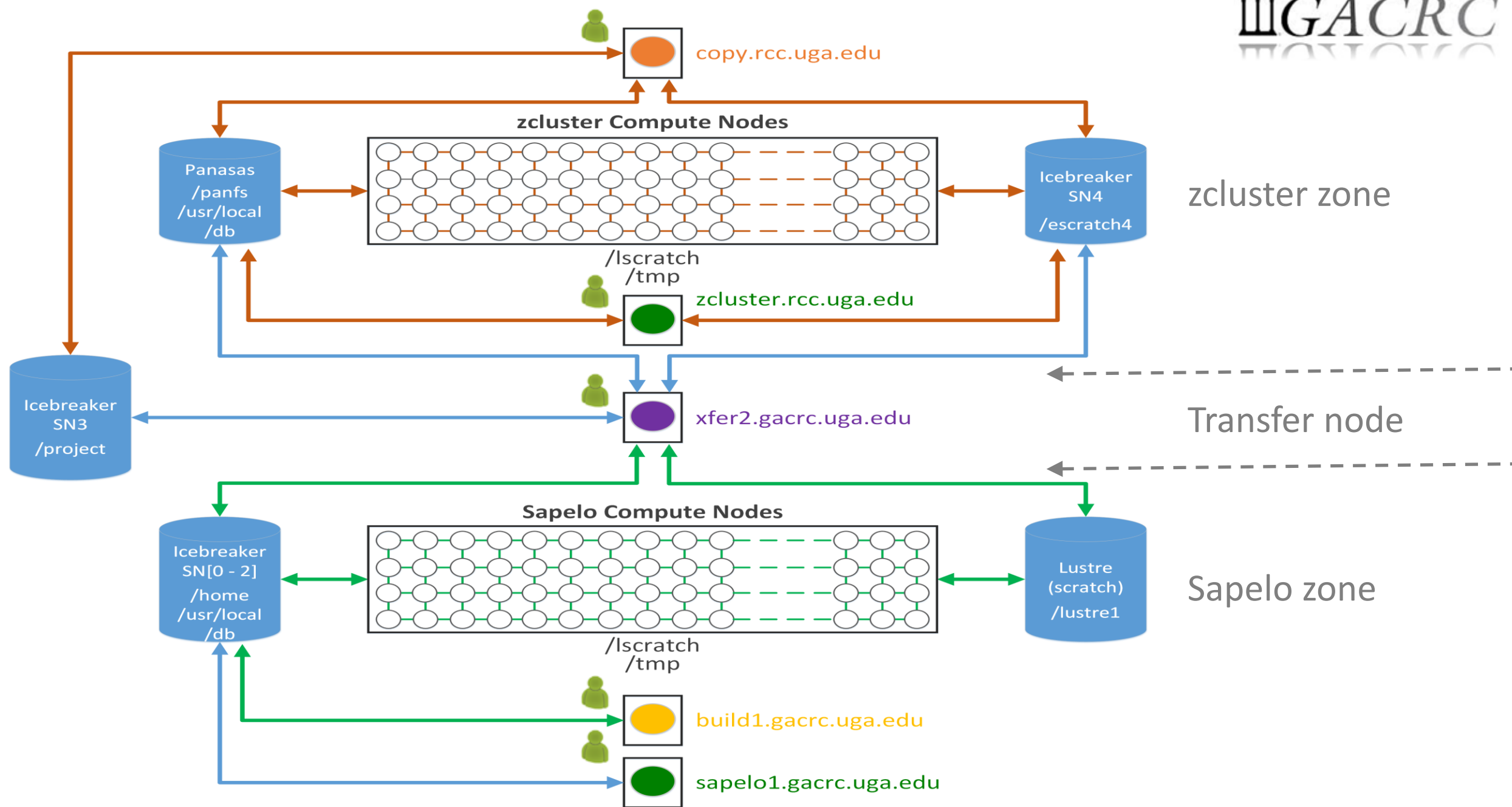
Next Page



1 Gbit

10 Gbit

InfiniBand



GACRC Storage Environment

What you should know about **xfer2**:

- ✓ **Transfer node** b/w zcluster and Sapelo + **Copy node** of Sapelo
- ✓ Home directory on **xfer2** = Home directory on **Login** of Sapelo : `/home/username`
- ✓ File systems on **xfer2**:
 - `/home/username` : Sapelo home
 - `/panfs/pstor.storage/home/abclab/username` : zcluster home
 - `/lustre1/username` : Sapelo scratch
 - `/escratch4/username` : zcluster scratch
 - `/project/abclab` : long-term archival storage
- ✓ Most file systems on **xfer2** are **auto-mounted** upon **the first time full-path access**, e.g.,
`cd /lustre1/username`. The command `ls` and TAB auto-completion may not work if the file system has not been mounted.

Data Transferring

- b/w two filesystems on zcluster
- b/w two filesystems on Sapelo
- b/w local and GACRC Storage
- b/w GACRC zcluster and Sapelo
- b/w Internet and GACRC Storage
- Refer to https://wiki.gacrc.uga.edu/wiki/Transferring_Files

Data Transferring b/w two filesystems on zcluster

- Transfer interactively:
 - ✓ Login to **Copy**
 - ✓ Use **cd** to change directory
 - ✓ Use **cp** or **mv** to copy or move data
- Transfer by copy queue:
 - ✓ Create copying job submission script: **copy.sh**, e.g.:

```
#!/bin/bash
cd ${HOME}
cp -r dataDir /project/abclab/username
```
 - ✓ Submit to copyq: **qsub -q copyq copy.sh**

Data Transferring b/w two filesystems on Sapelo

- /lustre1 scratch is visible on **xfer2** or **Interactive**, NOT on **Login**!
- Transfer interactively on **xfer2**:
 - ✓ Login to **xfer2**
 - ✓ Use **cd** to change directory
 - ✓ Use **cp** or **mv** to copy or move data

Data Transferring b/w local and GACRC Storage

- zcluster users:
 - ✓ Use **Copy**
 - ✓ Linux/Mac OS X machine: *scp*, *sftp*, or *FileZilla*
 - ✓ Windows machine: *SSH file Transfer*, *FileZilla*, or *WinSCP*
- Sapelo users:
 - ✓ Use **xfer2**
 - ✓ Linux/Mac OS X machine: *scp*, *sftp*, or *FileZilla*
 - ✓ Windows machine: *SSH file Transfer*, *FileZilla*, or *WinSCP*

Data Transferring b/w GACRC zcluster and Sapelo

- All users having zcluster and Sapelo accounts:
 - ✓ Login to **xfer2**
 - ✓ Filesystems on **xfer2**:
 - /home/username : Sapelo home
 - /panfs/pstor.storage/home/abclab/username : zcluster home
 - /lustre1/username : Sapelo scratch
 - /escratch4/username : zcluster scratch
 - /project/abclab : long-term archival storage
 - ✓ Use **cd** to change directory
 - ✓ Use **cp** or **mv** to copy or move data

Data Transferring b/w Internet and GACRC Storage

- zcluster users: Login to **Copy** (copy.rcc.uga.edu)
- Sapelo users: Login to **xfer2** (xfer2.gacrc.uga.edu)
- Use command **wget** or **curl** to download software from internet, e.g.,

```
wget http://www.ebi.ac.uk/ena/data/view/SRR1183952
```

```
Curl -OL http://www.ebi.ac.uk/ena/data/view/SRR1183952
```

Snapshot

- Only **homes** on zcluster and Sapelo are snapshotted!

*Note: home is for **highly static** data being used frequently*

- Snapshots are **completely invisible**, **read-only**, and **moment-in-time**.
- **4 daily** ones and **1 weekly** one are maintained.
- Snapshots are **eating up** your Sapelo home **100GB**, if there are frequent data modifications in home.

Backup

- Backup environment has not been implemented by GACRC yet.
- In the future, file systems to be included in GACRC Backup:

Zcluster /home

Sapelo /home

Sapelo /project

Best Practice Suggestions from GACRC

1. From **scratch** (Sapelo /lustre1 or zcluster /esratch4), instead of from home, to submit your batch jobs or run your interactive jobs!

Question: How to submit batch jobs from scratch?

1) From Sapelo /lustre1:

Method 1: Login to **Interactive** (qlogin) → `cd /lustre1/username` → submit job

Method 2: Login to **Login** → Put `cd /lustre1/username` in job submission script → submit job

2) From zcluster /escratch4:

Method 1: Login to **Login** → `cd /escratch4/username` → submit job

Method 2: Login to **Interactive** (qlogin) → `cd /escratch4/username` → submit job

Best Practice Suggestions from GACRC

2. **Clean Up Files** that are not needed from scratch
3. **Move Files** from scratch to /project for long-term storage
4. **Compress Files**, especially text files in /project, to save space



Please Do NOT Park Your Data in Scratch!

Otherwise, whole system scratching performance will be affected, and your and others' job will be affected!

Thank You!