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

What is GACRC?

Who Are We?

- **G**eorgia **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 (**OVPIT**)
- 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)

GACRC Users September 2015

Colleges & Schools	Depts	PIs	Users
Franklin College of Arts and Sciences	14	117	661
College of Agricultural & Environmental Sciences	9	29	128
College of Engineering	1	12	33
School of Forestry & Natural Resources	1	12	31
College of Veterinary Medicine	4	12	29
College of Public Health	2	8	28
College of Education	2	5	20
Terry College of Business	3	5	10
School of Ecology	1	8	22
School of Public and International Affairs	1	3	3
College of Pharmacy	2	3	5
	40	214	970
Centers & Institutes	9	19	59
TOTALS:	49	233	1029

GACRC Users September 2015

Centers & Institutes	PIs	Users
Center for Applied Isotope Study	1	1
Center for Computational Quantum Chemistry	3	10
Complex Carbohydrate Research Center	6	28
Georgia Genomics Facility	1	5
Institute of Bioinformatics	1	1
Savannah River Ecology Laboratory	3	9
Skidaway Institute of Oceanography	2	2
Center for Family Research	1	1
Carl Vinson Institute of Government	1	2
	19	**59**

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

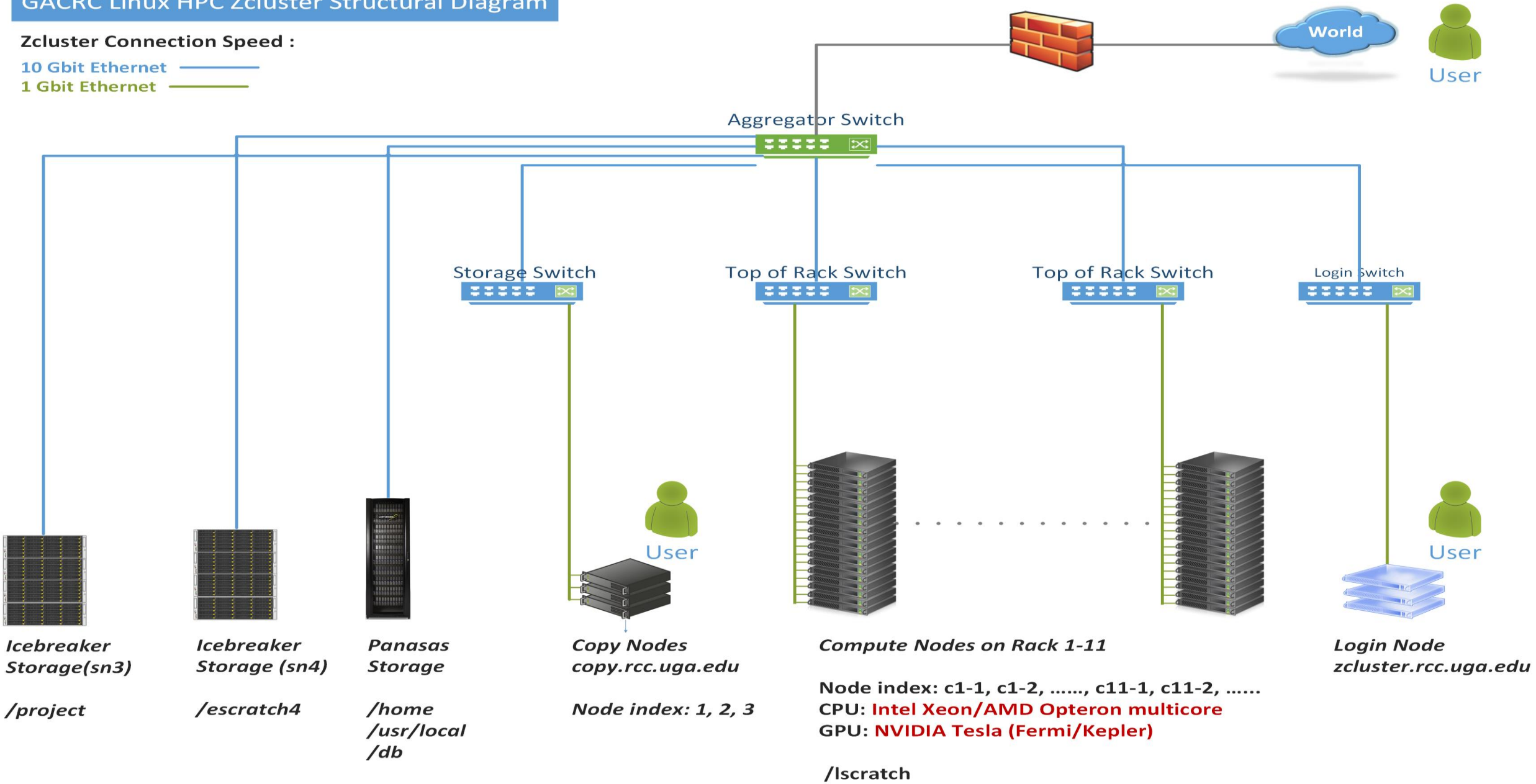
  `df -h` ➔ Print human-readable output, and choose the most appropriate unit for each size

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

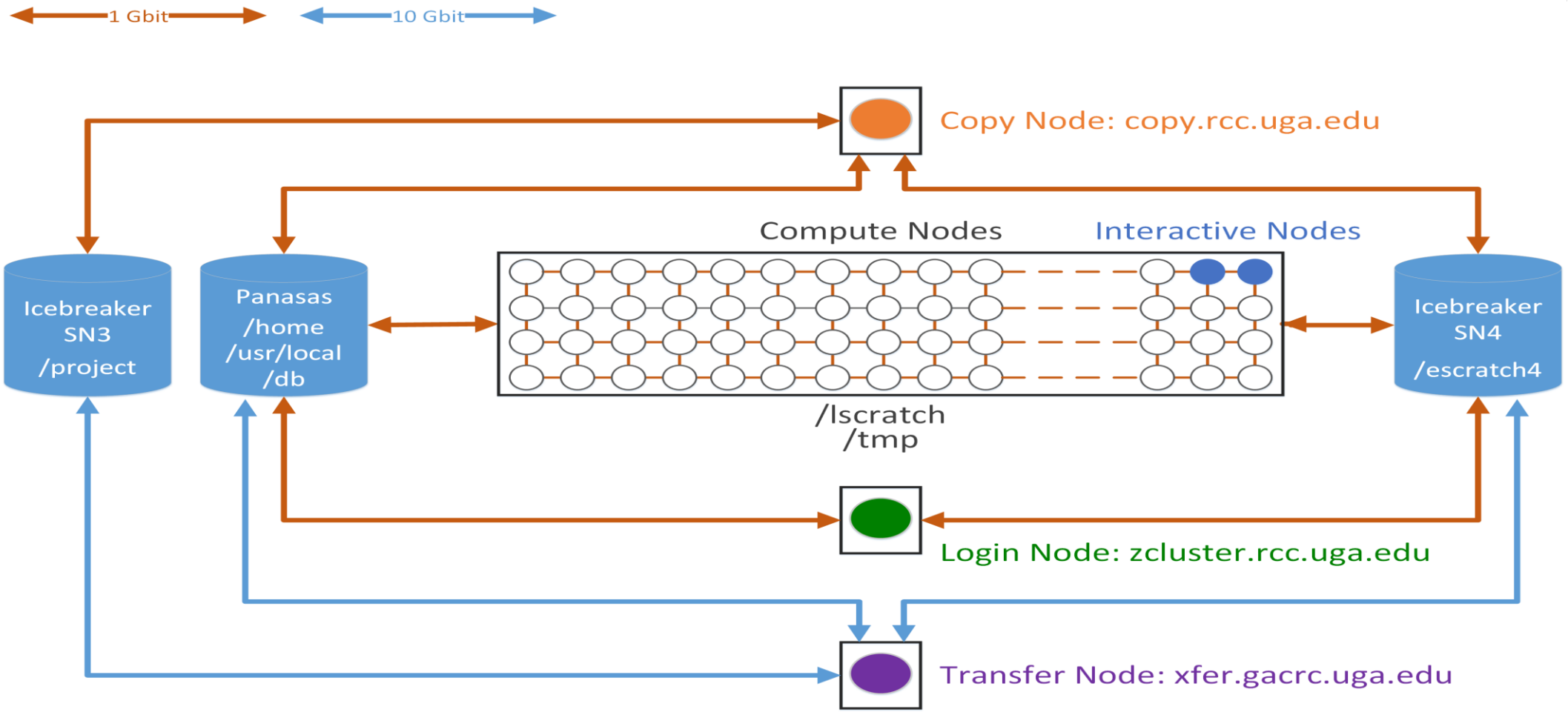
GACRC Storage Environment

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

GACRC Linux HPC Zcluster Structural Diagram



zcluster Storage Environment



zcluster Storage Environment

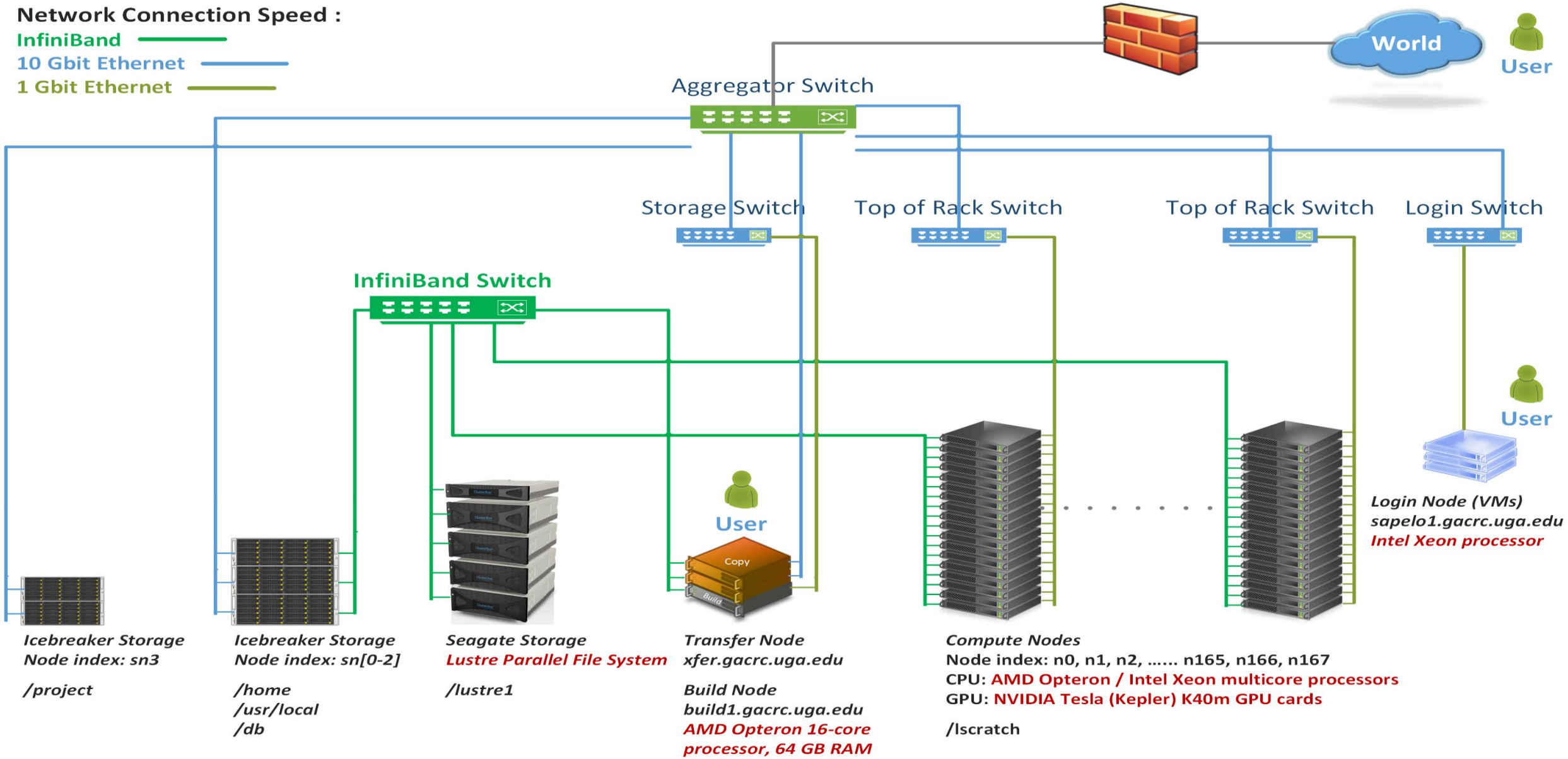
Filesystem	Role	Quota	Accessible from	Intended Use	Notes
/home/abclab/username	Home	100GB	zcluster.rcc.uga.edu (Login) Interactive nodes (Interactive) copy.rcc.uga.edu (Copy) xfer.gacrc.uga.edu (Transfer) compute nodes (Compute)	Highly static data being used frequently	Snapshots
/escratch4/username	Scratch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

Note:
1. /usr/local: Software installation directory; /db: bioinformatics database installation directory
2. `qlogin` from Login node to login to Interactive node

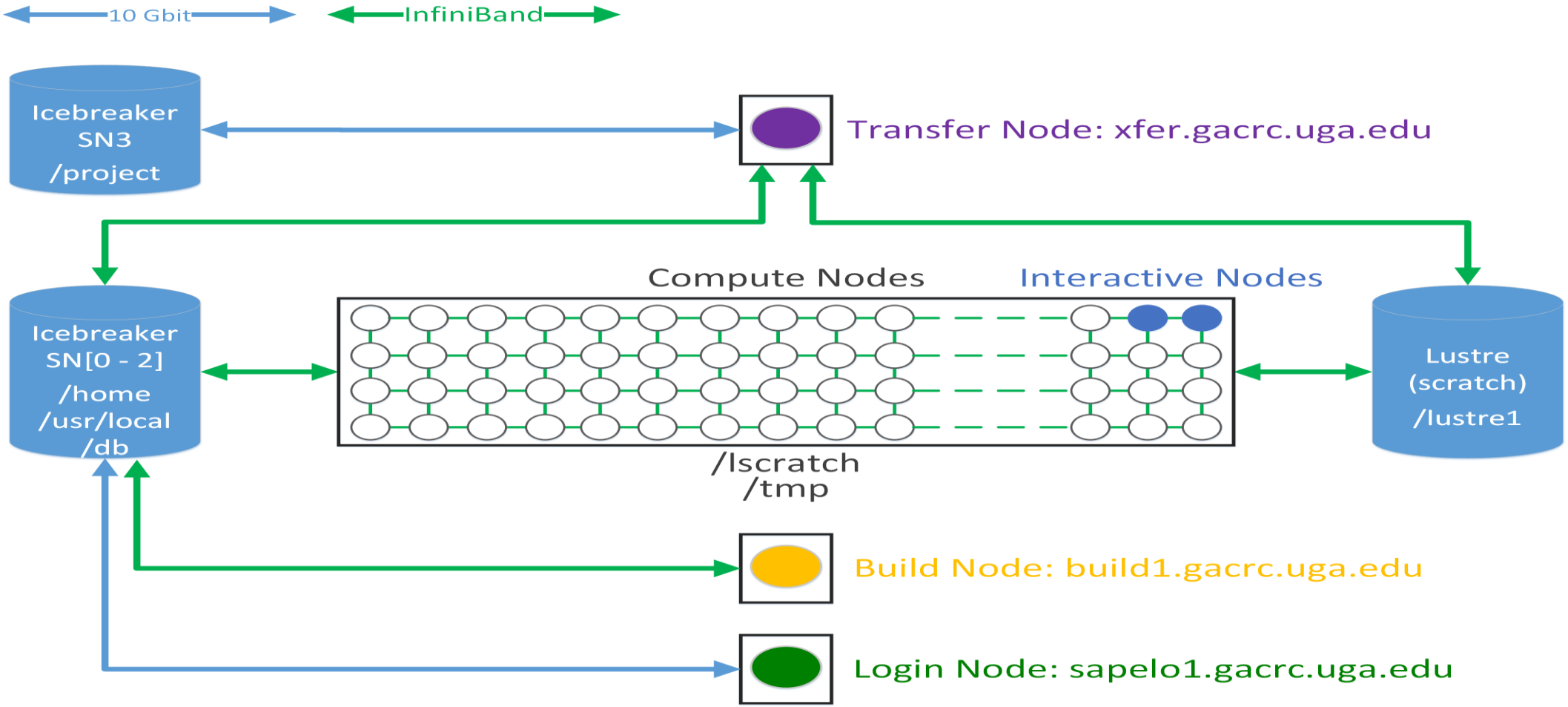
zcluster Storage Environment

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

The New GACRC Linux HPC Cluster Structural Diagram



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) xfer.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) xfer.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	xfer.gacrc.uga.edu (Transfer)	Long-term active data storage	Group sharing

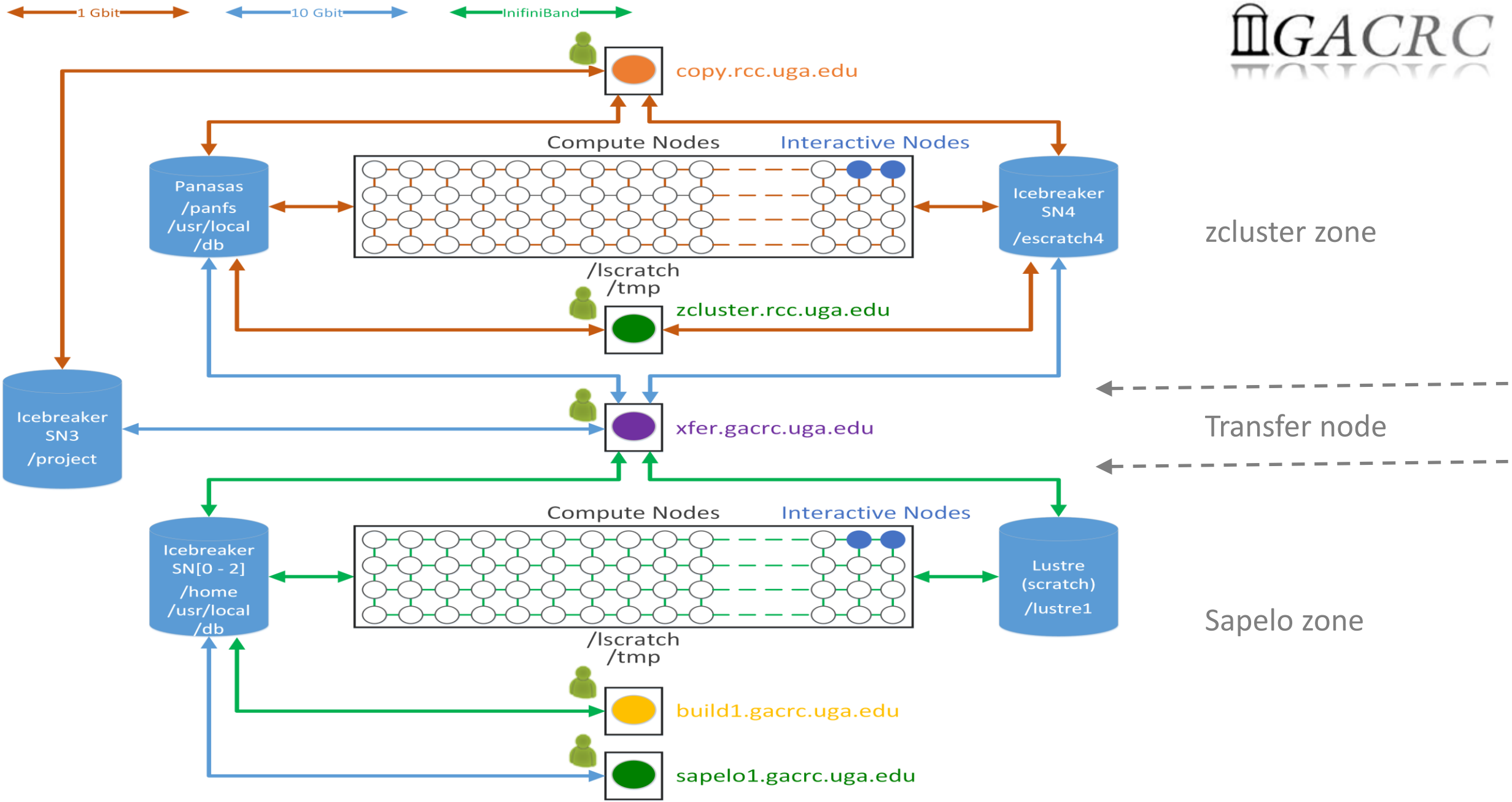
Note:
1. /usr/local/apps: Software installation directory; /db: bioinformatics database installation directory
2. `qlogin` from Login node to login to Interactive node

Sapelo Storage Environment

7 Main Functions	On/From-Node	Related Filesystem
Login Landing	Login or Transfer or Build	/home/username (Home) (Always!)
Batch Job Submitting	Login	/home/username (Home)
	Interactive	/lustre1/username (Scratch) (Suggested!) /home/username (Home)
Interactive Job Running	Interactive	/lustre1/username (Scratch) /home/username (Home)
Data Archiving , Compressing and Transferring	Transfer	/lustre1/username (Scratch) /home/username (Home)
Job Data Temporarily Storing	Compute	/lscratch/username (Local Scratch) /lustre1/username (Scratch)
Long-term Active Data Storing	Transfer	/project/abclab
Code Compilation	Build	/home/username (Home)

GACRC Storage Environment

Next Page


1 Gbit
10 Gbit
InifiniBand
copy.rcc.uga.edu
Compute Nodes
Interactive Nodes
Panasas
/panfs
/usr/local
/db
Icebreaker
SN4
/escratch4
/lscratch
/tmp
zcluster.rcc.uga.edu
zcluster zone
Icebreaker
SN3
/project
xfer.gacrc.uga.edu
Transfer node
Compute Nodes
Interactive Nodes
Icebreaker
SN[0 - 2]
/home
/usr/local
/db
Lustre
(scratch)
/lustre1
/lscratch
/tmp
build1.gacrc.uga.edu
sapelo1.gacrc.uga.edu
Sapelo zone

GACRC Storage Environment

What you should know about Transfer (xfer.gacrc.uga.edu):

- ✓ For data transferring b/w zcluster and Sapelo (`cp, mv`)
- ✓ For data transferring, compressing, packaging on Sapelo (`scp, gzip, tar etc.`)
- ✓ Home directory: `/home/username`
- ✓ File systems:
  - /home/username : Sapelo home
  - /panfs/pstor.storage/home/abclab/username : zcluster home
  - /lustre1/username : Sapelo scratch
  - /escratch4/username : zcluster scratch
  - /project/abclab : long-term active data storage
- ✓ Most file systems on Transfer are ***auto-mounted*** upon ***the first time full-path access***, e.g.,

```
cd /lustre1/username; cd /project/abclab
```

GACRC Storage Environment

What you should know about Copy (copy.rcc.uga.edu):

- ✓ For data transferring, compressing, packaging data on zcluster (`scp, gzip, tar etc.`)
- ✓ Home directory: `/home/abclab/username`
- ✓ File systems:
  - /home/abclab/username : zcluster home
  - /escratch4/username : zcluster scratch
  - /project/abclab/username : long-term active data storage

- ✓ /project file system on Copy is ***auto-mounted*** upon ***the first time full-path access***, e.g.,

```
cd /project/abclab
```

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 (username@copy.rcc.uga.edu) with *zcluster password*
  - ✓ `cd` to change directory
  - ✓ `cp` or `mv` to copy or move data

- Transfer by copy queue:
  - ✓ Create copying job submission script: `copy.sh`, e.g.:

```
#!/bin/bash
cd `pwd`
cp -r ./dataDir /project/abclab/username
```

  - ✓ Submit to copyq: `qsub -q copyq copy.sh`

Data Transferring b/w two filesystems on Sapelo

- /lustre1/username `cd` accessible on Transfer or Sapelo Interactive
- /lustre1/username NOT accessible on Sapelo Login!
- Transfer interactively on Transfer:
  - ✓ Login to Transfer (username@xfer.gacrc.uga.edu) with *UGA MyID password*
  - ✓ `cd` to change directory
  - ✓ `cp` or `mv` to copy or move data

Data Transferring b/w local and GACRC Storage

- zcluster users:
  - ✓ Use Copy (username@copy.rcc.uga.edu) with *zcluster password*
  - ✓ Linux/Mac OS X machine: `scp`, `sftp`, or *FileZilla*
  - ✓ Windows machine: *SSH file Transfer*, *FileZilla*, or *WinSCP*
- Sapelo users:
  - ✓ Use Transfer (username@xfer.gacrc.uga.edu) with *UGA MyID password*
  - ✓ 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 Transfer (username@xfer.gacrc.uga.edu) with *UGA MyID password*
  - ✓ File systems:
    - ➢ /home/username : Sapelo home
    - ➢ /panfs/pstor.storage/home/abclab/username : zcluster home
    - ➢ /lustre1/username : Sapelo scratch
    - ➢ /escratch4/username : zcluster scratch
    - ➢ /project/abclab : long-term active data storage
  - ✓ `cd` to change directory
  - ✓ `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) with *zcluster password*
- Sapelo users: Login to Transfer (xfer.gacrc.uga.edu) with *UGA MyID password*
- `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 home on zcluster and Sapelo are snapshotted
- Snapshots are like backups in that they are read-only moment-in-time captures of files and directories which can be used to restore files that may have been accidentally deleted or overwritten
- Sapelo home: once a day and maintained for 4 days
- zcluster home: once a day or every 2 days; maintained for 4 days
- Snapshots are ***eating up*** your home 100GB, if there are frequent data modifications
- Contact the GACRC staff if you need to recover data from a snapshot

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

1. From **scratch** (Sapelo /lustre1 or zcluster /escratch4), instead of from **home**, to submit your batch jobs or run interactive jobs!

From Sapelo /lustre1:

Strategy 1: on Interactive (`qlogin`) ➔ manually `cd /lustre1/username/workDir/` ➔ submit job there

Strategy 2: on Login ➔ Put `cd /lustre1/username/workDir/` in sub.sh ➔ submit sub.sh from home

From zcluster /escratch4:

Strategy 1: on Login

Strategy 2: on Interactive (`qlogin`)

➔ manually `cd /escratch4/username/workDir` ➔ submit job there

Best Practice Suggestions from GACRC

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

Please Do NOT Park Your Data on Scratch!

Otherwise, the performance of the scratch file systems will be degraded, and your and others' job will be affected!

Thank You!