01010011 Scientific
01000011 Computing
01000111 Group

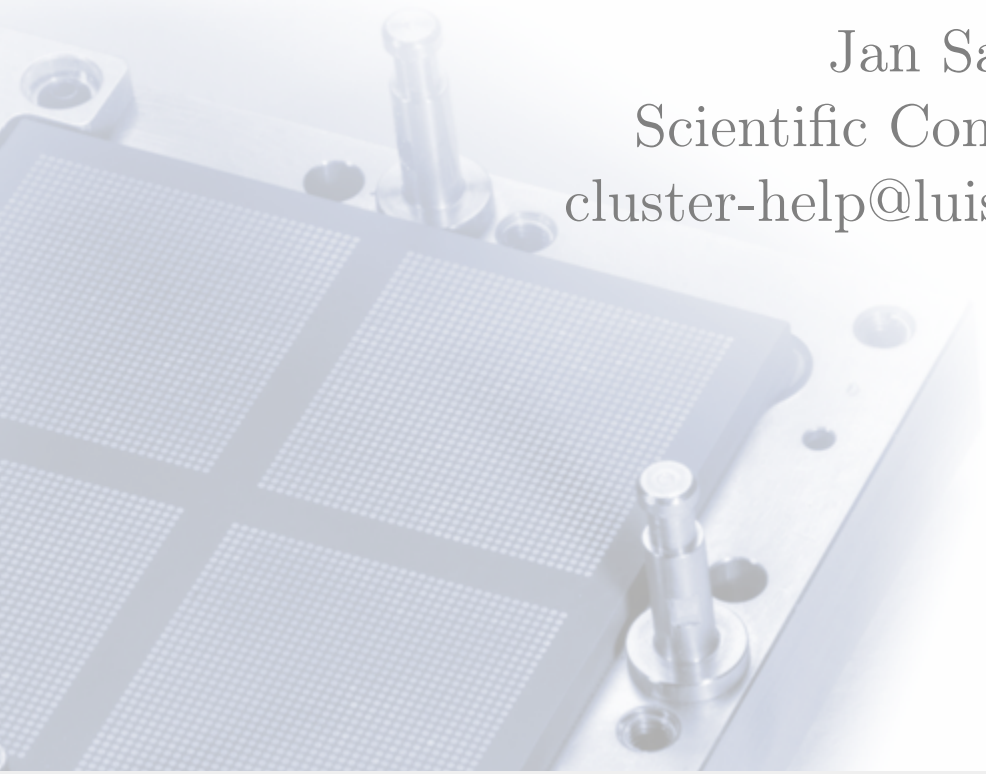
Leibniz Universität Hannover

The cluster system

Introduction

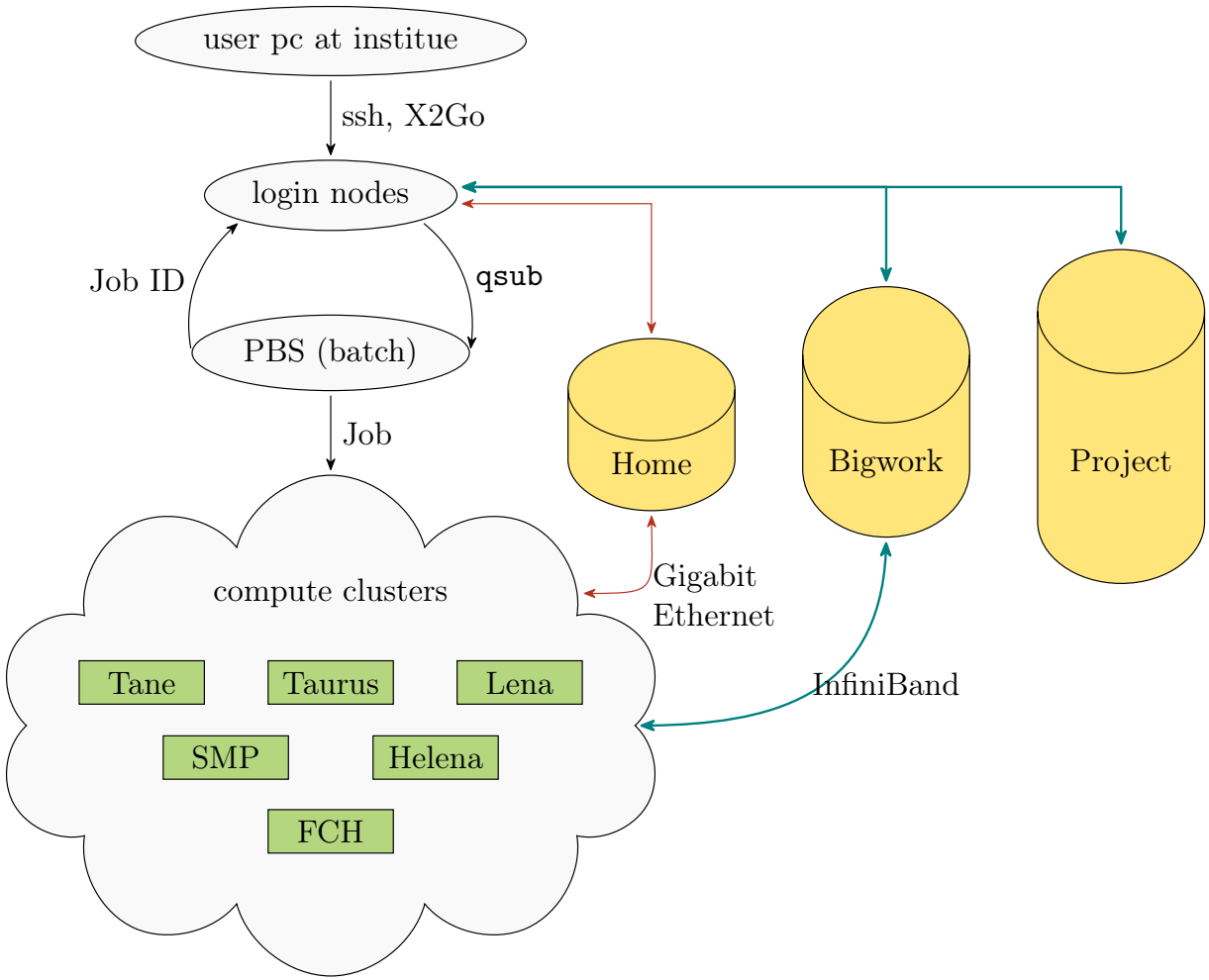
8th December 2016

Jan Saalbach
Scientific Computing Group
cluster-help@luis.uni-hannover.de



Contents

The Cluster system



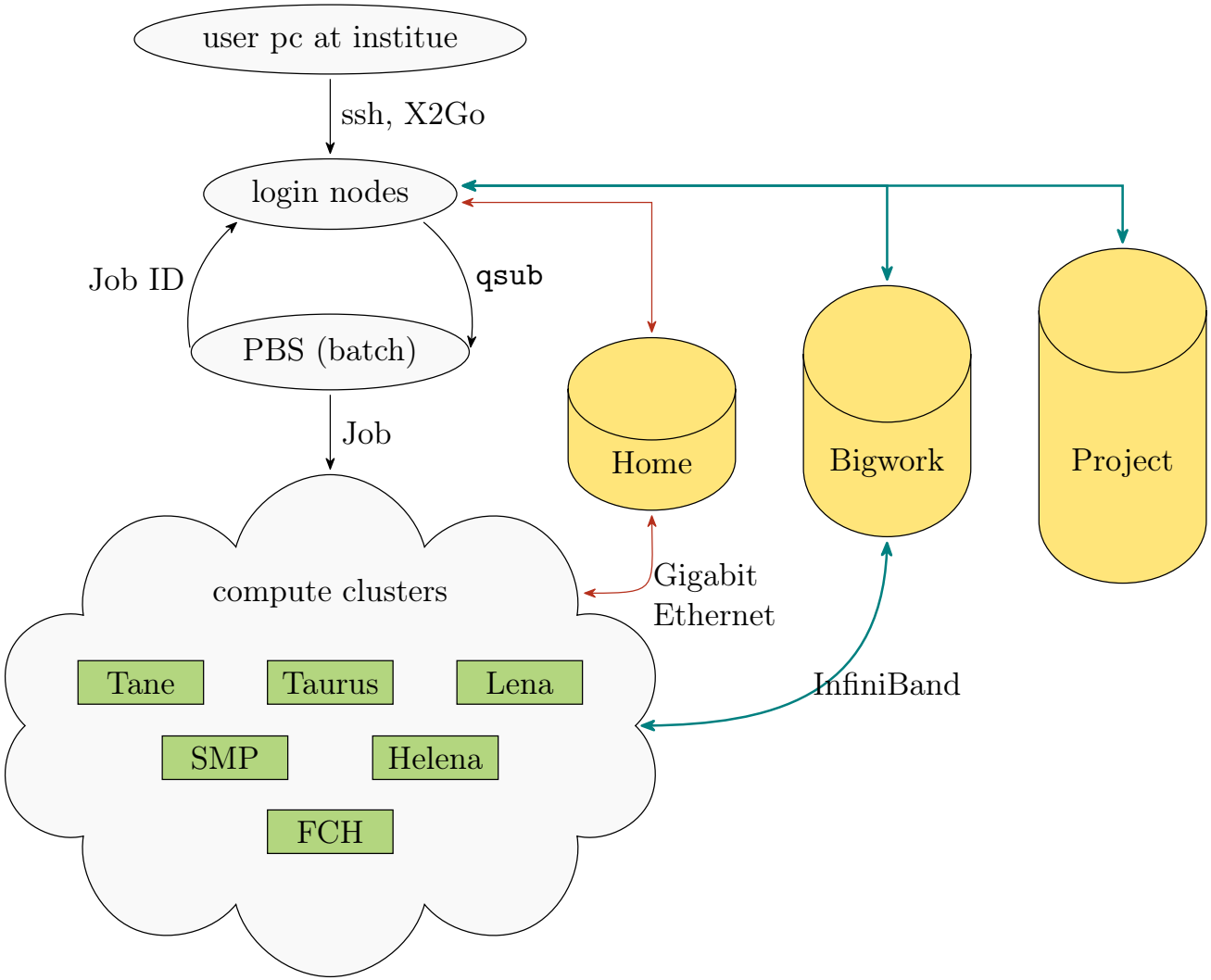
Cluster system consists of various parts serving different purposes

General information about the Cluster

- Institutes of Leibniz University Hannover only, free of charge
- Partially DFG major instrumentation »DFG Großgerät«
- Software licences are valid for academic use only
  ⇒ Research only, no industrial work
- Institutes can bring their own dedicated hardware to be a part of the cluster in Forschungscluster-Housing (FCH)
- FCH hardware is available to all cluster users over night and on weekends
- One cluster user account per person
- Please do not share accounts, create individuals ones
- See the cluster as a tool to facilitate your research, like any laboratory equipment it takes time to understand, we help you

Use the cluster the way you expect others to use it

Compute Clusters



Available computing power

- 3 Clusters for MPI jobs needing lots of CPUs
  - Lena: 80 nodes; 16 cores each @ 2.4 GHz; 64 GB RAM
  - Tane: 96 nodes; 12 cores each @ 2.9 GHz; 48 GB RAM
  - Taurus: 54 nodes; 12 cores each @ 2.66 GHz; 48 GB RAM
- 22 SMP machines for jobs needing lots of RAM
  - SMP: 4 nodes; 32 cores each @ 2.5 GHz; 256 GB RAM
  - SMP: 9 nodes; 32 cores each @ 2.0 GHz; 256 GB RAM
  - SMP: 9 nodes; 24 cores each @ 2.0 GHz; 256 GB RAM
- 3 Machines with 1 TB of memory
  - Helena: 3 nodes; 32 cores each @ 2.13 GHz; 1028 GB RAM
- Forschungscluster-Housing
  - FCH: 35 machines in various configurations

What requirements does your research have?

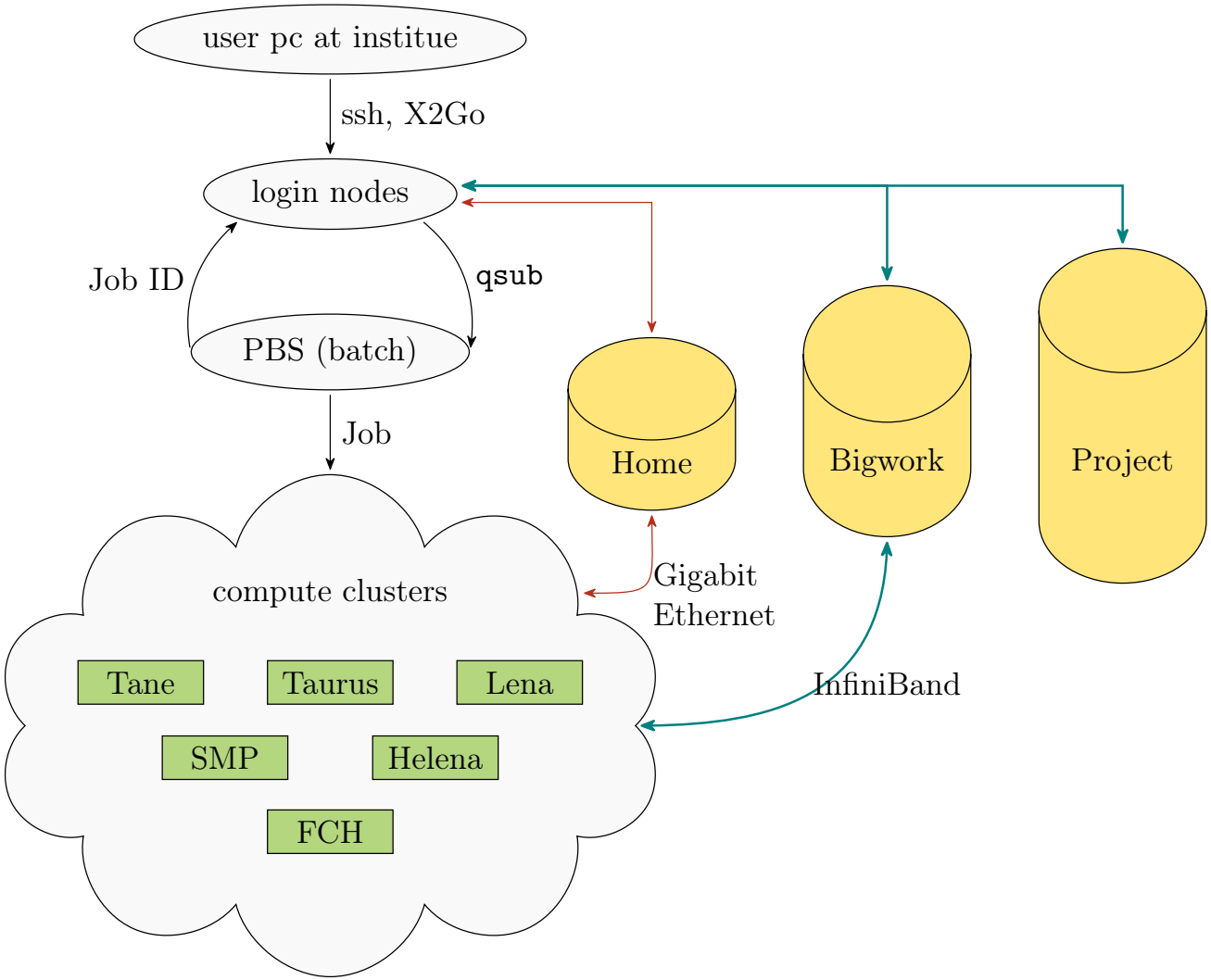
Available computing power (cont.)

- Very large projects can move up to the HLRN
  - separate application method from that of the computing centre
  - »test accounts« are available
  - http://www.hlrn.de

Requesting compute resources - Jobs

- In order to use computing resources you need to request them
- You make the request to the batch system (aka PBS, torque, maui, scheduler)
- The request is called a job which the batch system executes whenever resources are available
- These are the major resource types:
  - nodes (aka machines, server)
  - ppn: processors per node (aka cpu-cores)
  - mem: amount of physical memory (aka RAM)
  - walltime
- Jobs are submitted on the login nodes

Login Nodes


user pc at institue
ssh, X2Go
login nodes
Job ID
qsub
PBS (batch)
Job
Home
Bigwork
Project
compute clusters
Gigabit
Ethernet
InfiniBand
Tane
Taurus
Lena
SMP
Helena
FCH

Accessing the cluster system

Operating system on all cluster nodes is Scientific Linux 6, these are the recommended ways to connect:

- Windows graphical: X2Go
  - Available for free
  - `http://wiki.x2go.org/doku.php`
  - See Cluster documentation for help on installation
- File transfer: FileZilla
  - Available for free
  - Has an easy-to-use graphical user interface
  - `http://filezilla-project.org/`
- Windows console: PuTTY
  - Available for free
  - Uses SSH (the Secure Shell) for an encrypted connection
  - `http://www.putty.org/`
- Linux console ssh, scp, rsync ...

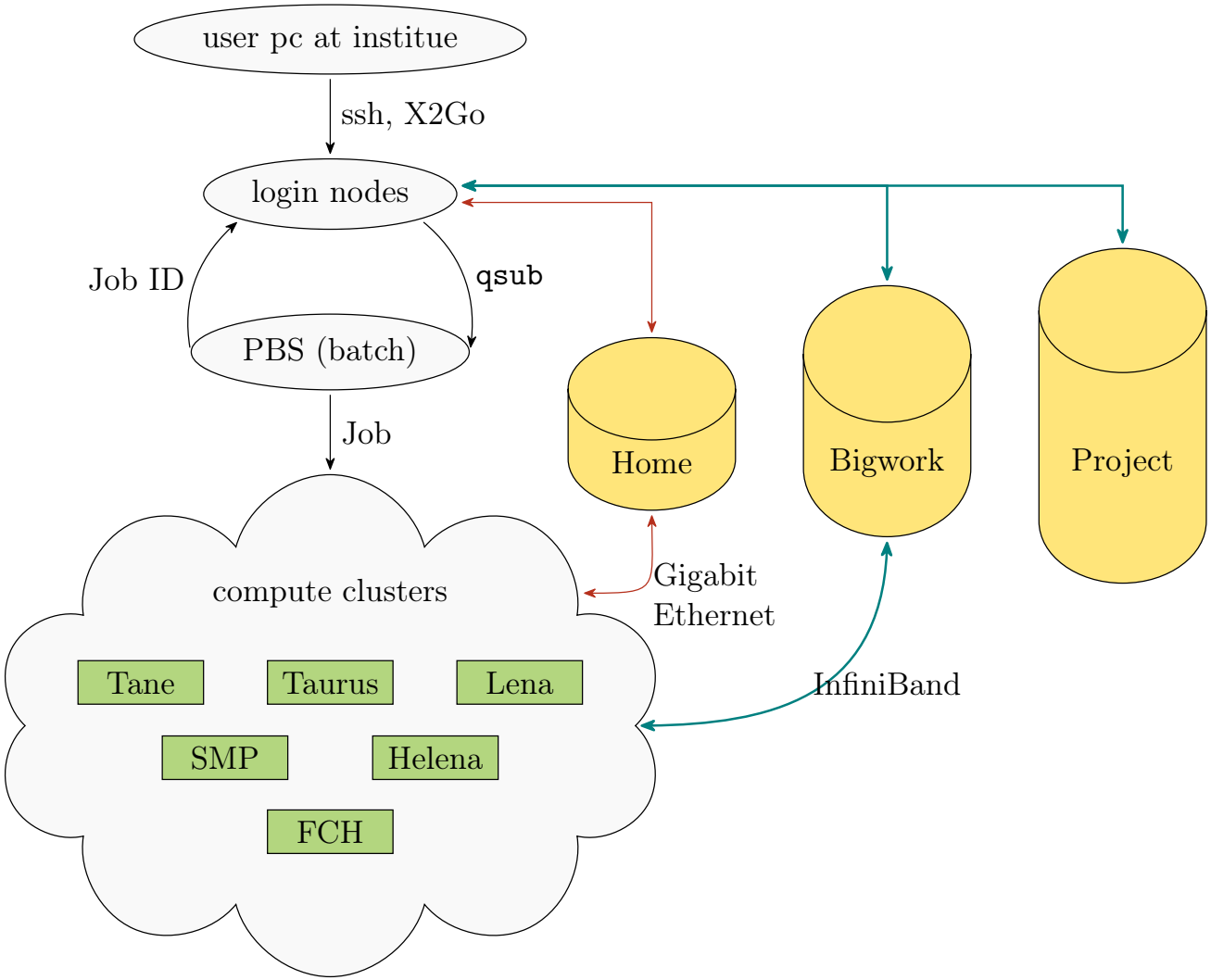
`login.cluster.uni-hannover.de`

Login nodes

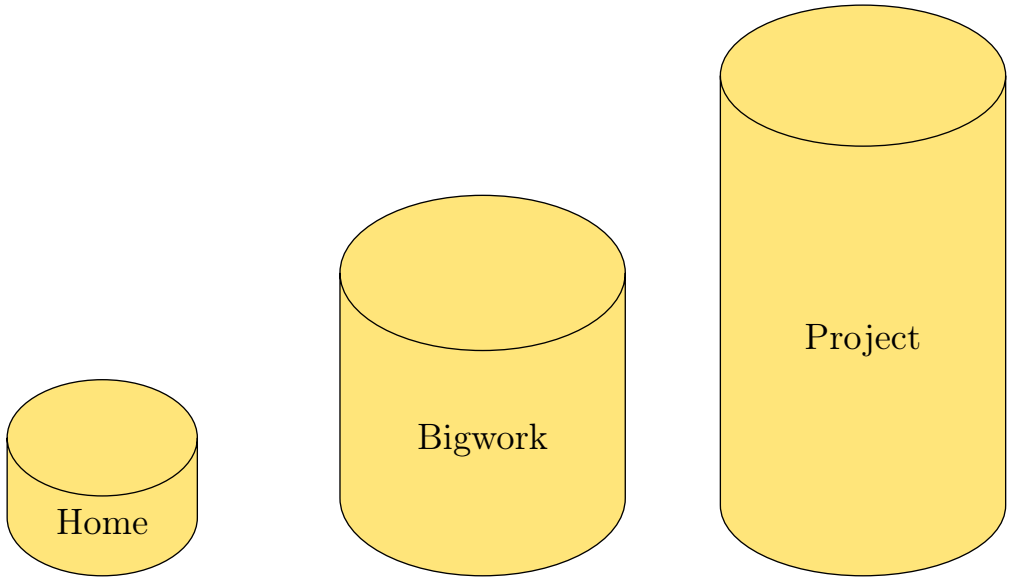
- `ssh login.cluster.uni-hannover.de`
  will route you to one of our login nodes.
- Here it is possible to:
  - Prepare computing jobs
  - Prepare batch scripts
  - Send computing jobs into the queue
  - Run small tests
  - Show the queue status
  - View simulation results
  - Copy files into the archive system
  - Copy files to your desktop computer
- All applications, compilers and tools are available
- These nodes are NOT for production simulations

Processes will be killed after a maximum run time of 30 minutes

File systems



File system specifications



$HOME	$BIGWORK	$PROJECT
◦ /home/username	◦ /bigwork/username	◦ /bigwork/groupname
◦ 10 GB soft quota	◦ 100 GB soft quota	◦ 10 TB soft quota
◦ 12 GB hard quota	◦ 1 TB hard quota	◦ 12 TB hard quota
◦ All cluster nodes	◦ All cluster nodes	◦ login & transfer nodes
◦ backup	◦ no backup	◦ no backup
◦ Gigabit Ethernet	◦ InfiniBand	◦ InfiniBand

Details of the file systems

- `$HOME`
  - Globally available home directory
  - Data are backed up daily
  - Intended for scripts, programs and small final simulation results
  - Per user quota of 10 GB
- `$BIGWORK`
  - Globally available work directory
  - Data are NOT backed up
  - Intended for large »work« files
  - Per user quota of 1 TB
  - Use it as a scratch filesystem
- `$HOME` is connected through Ethernet (slow)
- `$BIGWORK` is connected through InfiniBand (fast)

  ⇒ Do not create a link to your bigwork in your home directory, use `$BIGWORK` variable instead

Regularly backup your data to your institute's server

File systems - Exercise

```
# where are you? lost? print working directory!
pwd
# change directory to your bigwork directory
cd $BIGWORK
# what is the absolute path of your bigwork directory?
pwd
# change directory to your project directory
cd $PROJECT
# what is the absolute path of your project directory?
pwd
# change directory to your home directory
cd $HOME
# display the absolute path of your home directory
pwd
# display your home quota
fquota
# display your bigwork & project quota
lquota
# make your personal directory in your group's project
    storage
mkdir -m 0700 $PROJECT/$USER
# copy the directory mydir from bigwork to project
cp -r $BIGWORK/mydir $PROJECT/$USER
```

The module command

- Display the entire list of possible modules

```
$ module spider
```

- Show modules immediately available to load

```
$ module avail
```

- Load one or more modules

```
$ module load <modulename> <...>
```

- Unload a module

```
$ module unload <modulename>
```

- Unload all loaded modules

```
$ module purge
```

- Show already loaded modules

```
$ module list
```

- Show information about a module

```
$ module whatis <modulename>
```

The module command - Exercise

```
# find available octave modules
module spider octave

# determine how to load the selected Octave/4.0.0 module
module spider Octave/4.0.0

# load required modules
module load GCC/4.9.3-2.25 OpenMPI/1.10.2

# load module for octave
module load octave/4.0.0

# start application octave
octave
```

Working with jobs

Jobs are the framework tasks in the cluster are embedded in.

- Show all jobs

  ```
  $ showq
  ```

- Show only your jobs

  ```
  $ showq -u <username>
  ```

- Start an interactive job

  ```
  $ qsub -I <options>
  ```

- Send jobs into the queue

  ```
  $ qsub <options> <name of job script>
  ```

- Delete a job

  ```
  $ qdel <jobid>
  ```

- Show the full output for a particular job

  ```
  $ qstat -f <jobid>
  ```

Only `showq` will tell you if your job is blocked

Interactive job - Exercise

```
# start an interactive job specifying all resource
# parameters, so no defaults get used
qsub -I -X -l nodes=1:ppn=1 -l walltime=01:00:00 -l mem=2gb

# use the date command to display the current time
date

# redirect output of date command to file my-date.txt
date > my-date.txt

# exit job
exit

# display directory contents
ls -lh

# display newly created file
cat my-date.txt
```

Interactive jobs are useful for debugging, always use interactive first

Batch job - Exercise

Create a file named MyFirstBatchJob.sh

```
#!/bin/bash -login
# resource  specification
#PBS -l nodes=1:ppn=1
#PBS -l walltime=01:00:00
#PBS -l mem=2gb
# commands to execute
cd $BIGWORK
date > date.txt
# wait for 120 seconds
sleep 120
```

Submit the job script

```
qsub MyFirstBatchJob.sh
```

Check job status

```
showq -u put-your-user-name-here
```

Check files MyFirstBatchJob.sh.o* and MyFirstBatchJob.sh.e*

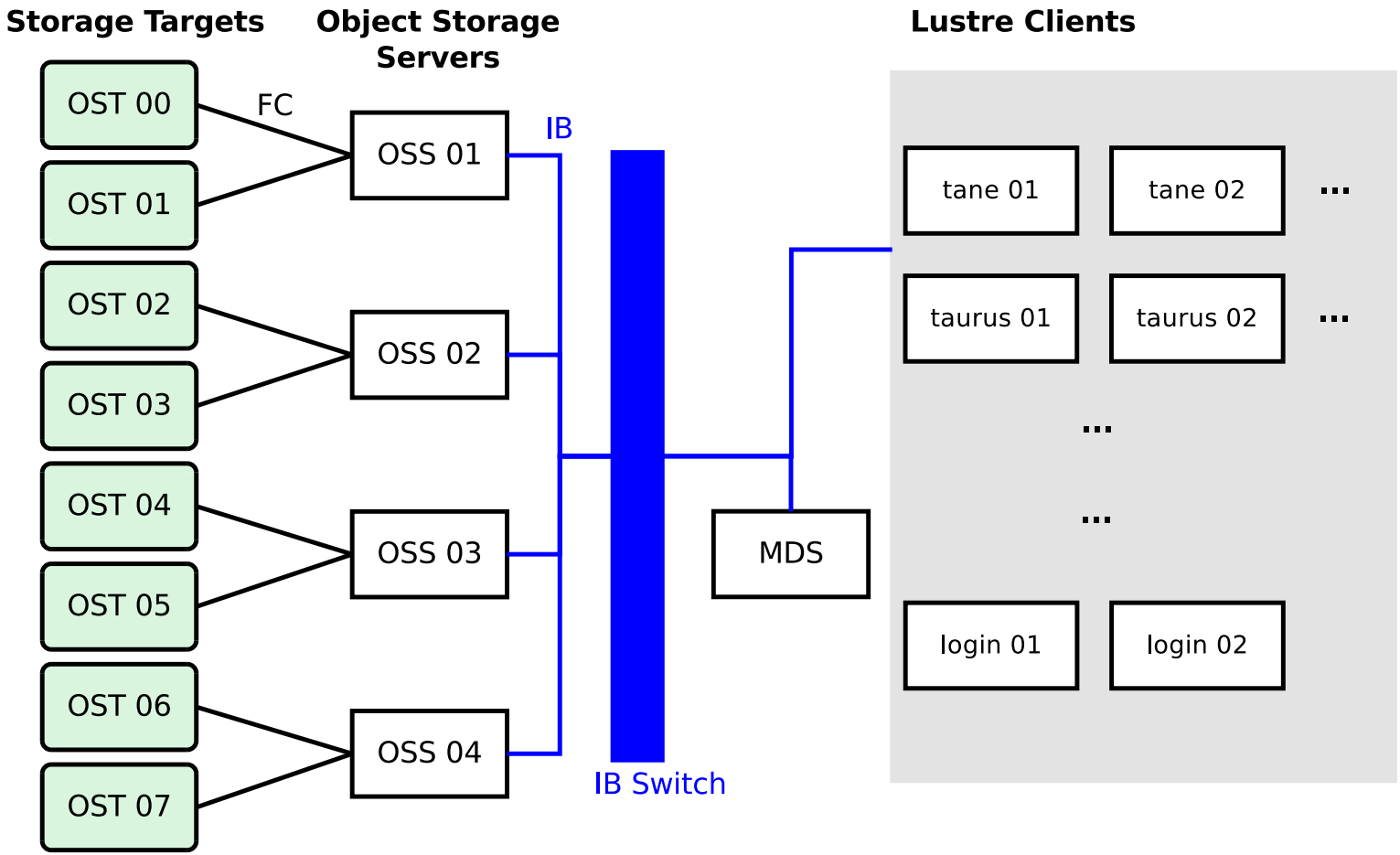
Software requiring licenses - Situation

1. You submit your job using qsub.
2. The batch system starts your job (requested resources play a role).
3. Job starts and walltime starts counting down. Your job script is executed until software that needs a license starts.
4. Software start requires license, currently none available. Depending on the software some files might get created even without a license.
5. If not configured to wait, software start fails and job ends.
6. Application start delayed, if configured to wait for license. Walltime is counting down.
7. License is available, software starts, remaining walltime does not cover the duration of a simulation run. Job ends with no walltime left even though the simulation is unfinished.

Software requiring licenses - Solution

The batch system starts jobs regardless of required licenses. Thus you need to make sure to use any available mechanism of your application in order to wait for licenses. Consequently you need to make sure, requested walltime covers any waiting periods that may occur.

Bigwork's file system Lustre



There are eight object storage targets (OSTs). In order to store large files (anything $>$ 100 MB) efficiently and fast, set your stripe count.

Setting Lustre stripe - Exercise

```
# get overall bigwork usage, note different fill levels
lfs df -h
# get current stripe settings for your bigwork
lfs getstripe $BIGWORK
# change directory to your bigwork
cd $BIGWORK
# create a directory for large files (anything over 100 MB)
mkdir LargeFiles
# get current stripe settings for that directory
lfs getstripe LargeFiles
# set stripe count to -1 (all available OSTs)
lfs setstripe --count -1 LargeFiles
# check current stripe settings for LargeFiles directory
lfs getstripe LargeFiles
# create a directory for small files
mkdir SmallFiles
# check stripe information for SmallFiles directory
lfs getstripe SmallFiles
```

Use newly created LargeFiles directory to store large files

Altering stripe

Sometimes you might not know beforehand, how large files created by your simulations will turn out. In this case you can set stripe size after a file has been created in two ways. Let us create a 100 MB file first.

```
# enter the directory for small files
cd SmallFiles
# create a 100 MB file
dd if=/dev/zero of=100mb.file bs=10M count=10
# check filesize by listing directory contents
ls -lh
# check stripe information on 100mb.file
lfs getstripe 100mb.file
# move the file into the large files directory
mv 100mb.file ../LargeFiles/
# check if stripe information of 100mb.file changed
lfs getstripe ../LargeFiles/100mb.file
# remove the file
rm ../LargeFiles/100mb.file
```

In order to change stripe, the file has to be copied (`cp`). Simply moving (`mv`) the file will not affect stripe count.

Altering stripe - Exercise

First method:

```
# from within the small files directory
cd $BIGWORK/SmallFiles
# create a 100 MB file
dd if=/dev/zero of=100mb.file bs=10M count=10
# copy file into the LargeFiles directory
cp 100mb.file ../LargeFiles/
# check stripe in the new location
lfs getstripe ../LargeFiles/100mb.file
```

Second method:

```
# create empty file with appropriate stripe count
lfs setstripe --count -1 empty.file
# check stripe information of empty file
lfs getstripe empty.file
# copy file "in place"
cp 100mb.file empty.file
# check that empty.file now has a size of 100 MB
ls -lh
# remove the original 100mb.file and work with empty.file
rm 100mb.file
```

Further reading

- Scientific Computing homepage:
  https://www.luis.uni-hannover.de/scientific_computing.html
- Cluster System Handbook:
  https://www.luis.uni-hannover.de/fileadmin/scientific_computing/anleitungen/ClusterHandbook.pdf
- Handbuch zum Clustersystem:
  https://www.luis.uni-hannover.de/fileadmin/scientific_computing/anleitungen/ClusterHandbuch.pdf
- Cluster system cheatsheet:
  https://www.luis.uni-hannover.de/fileadmin/scientific_computing/intro/clustersystem_cheatsheet.pdf
- Batch system (torque, maui), module system (Lmod), your favorite application etc. all have documentation of their own. Go check these out!

Getting help

In case an error occurs ...

1. Analyze the error message, check your program and your batch script.
2. Read the cluster documentation and search the internet.
3. Send an error report to: cluster-help@luis.uni-hannover.de
   Please provide:
   - The job ID
   - Your username
   - A short description of the problem. What did you expect to happen?

Always provide username and JobID when asking for help

Thank you for your attention

Jan Saalbach
cluster-help@luis.uni-hannover.de

Every first Monday of the month at 6:00 p.m.
Cluster User Group
http://www.luis.uni-hannover.de/cug.html