Cluster & data storage

Cluster

Connecting to the cluster

Once you have an account on the platform, you can connect via SSH to the **genossh.genouest.org** server.

genossh.genouest.org is a front-end of the cluster from which you can submit jobs with the *Sun Grid Engine* tool (SGE) job manager. You first need to connect to the front-end server via SSH from your computer.

You can connect to **genossh.genouest.org** from anywhere, but only with a properly configured SSH Key.

Connecting to genossh.genouest.org from a windows computer

On Windows, Putty can be used to load SSH keys and connect via SSH to the cluster. Have a look at this video tutorial explaining the whole procedure (creating a SSH key and then connecting to the cluster):

Connecting to genossh.genouest.org from a linux computer

You need first to generate an SSH key. To do so, launch this command on your computer:

```
ssh-keygen -t rsa -b 4096
```

The command will ask for a password: it will protect your SSH key, and you will need it everytime you will use it to connect to the cluster (depending on your configuration, a program named ssh-agent can remember this password after you entered it the first time you connect).

The ssh-keygen program will create two files in your home directory:

```
$HOME/.ssh/id_rsa
$HOME/.ssh/id_rsa.pub
```

id_rsa is your private key: keep this file secret.

id_rsa.pub is your public key. You need to open this file and copy-paste its content to http://my.genouest.org (“*Public SSH key*” form on the right side, once your are logged in).

You should then be able to connect to the cluster with this command:

```
ssh <your-login>@genossh.genouest.org
```

Data storage

Once logged, you have access to three volumes, available on all computing nodes.

Home directory

Your home directory (*/home/genouest/<your-group>/<your-login>*).

We have a total of around 100Tb of storage capacity shared between all the home directories, and each user have a quota of 100Gb. You can check your disk usage with the command “quota -s”.

A snapshot mechanism is available on this volumes, if you erased a file by mistake, your can rescue it by looking into the ~/.snapshots directory.

Project directory

A project directory (/groups/<your-group>) that you share with your team.

We have a total of around 200Tb of storage capacity shared between all these group directories.

Each project have a specific quota, and a single person in your team is responsible to grant you access to this volume.

You can check your disk usage with the command “df -h /groups/<your-group>”.

A snapshot mechanism is available on this volumes, if you erased a file by mistake, your can rescue it by looking into the ~/.snapshots directory.

High performance storage

A high performance storage space (*/omaha-beach/<your-login>*).

We have a total of around 50Tb for */omaha-beach*, and each user have a quota of 120Gb. You can check your disk usage with the command “pan_quota /omaha-beach/”.

Informations

Quotas are intentionally restrictive, if you need them to be increased, please contact support@genouest.org.

As a general rule, user should not write during the jobs in the */home* or */groups* directory, nor do heavy read operations on these volume. They are used to keep your data safe. During jobs, one should use the */omaha-beach* directory. This directory is hosted by a high performance system and designed for temporary data. It supports heavy read and write operations.

Please note that none of your data is backed up. If you would like us to backup your data for specific reasons, you can contact us and we will help you to find a solution.

We strongly advise you to anticipate your storage needs: if you plan to generate a big amount of data, please contact us **before** to check that we have the possibility to host this data. It is preferable to anticipate this when applying for grants that imply data generation and analysis.

Before generating data on the cluster, please do not forget to check the remaining available space. To do so, you may use the quota commands above, or use the **df** command for global disk usage:

```
df -h
```

Biological databanks

Some databanks are available in the */db* directory. They are automatically updated and indexed on a regular basis using BioMAJ. You can consult the list of available banks on our BioMAJ instance.

Software

To execute software, you must be logged on a cluster node, you cannot execute software directly on genossh. To log on a cluster node, you need to execute the command **qrsh**.

Pre-installed software are available in */softs/local* (see software manager for a list of installed software). To use a software, you have to load its environment. For example to load python 2.7 you can launch this command (the dot at the beggining is important):

```
.  /softs/local/env/envpython-2.7.sh
```

This will automatically configure the PATH, libraries etc… in your shell environnement. Any subsequent python command you will launch will use this 2.7 version.

If you account is configured to use *tcsh* instead of *bash* (it is the case for old GenOuest account), the command is slightly different: replace the dot by “*source*” and omit the trailing “*.sh*”

```
source /softs/local/env/envpython-2.7
```

To get a list of all environments available, just list the content of */softs/local/env/env**.

Note: DO NOT USE the python/perl/…. of the node directly, always load a specific version from /softs/local.

Conda

Since October 2016, we are experiencing the use of Conda to install software on the cluster. Conda allows you to install the software you need in your own storage volumes (/home, /groups or /omaha-beach). The software need to be available as Conda packages, for example in Bioconda.

To use Conda, first source it the usual way:

```
. /local/env/envconda.sh
```

Then the first time you use it, enable some channels this way:

```
conda config --add channels r
conda config --add channels bioconda
conda config --add channels conda-forge
```

With Conda, you can create as many environments as you want, each one containing a list of packages you need. You need to activate an environment to have access to software installed in it. You can activate only one environment at a time.

To create a new environment containing biopython, deeptools (v2.3.4), bowtie and blast, run:

```
conda create -p ~/my_env biopython deeptools=2.3.4 bowtie blast
```

To activate it:

```
source activate ~/my_env
```

To deactivate it:

```
source deactivate
```

Feel free to test this new way to install software, and to give us feedback wether you are happy or not of it.

Launching jobs on the cluster

It is **forbidden** to execute computations directly on the frontals (**genocluster2.irisa.fr** or **genossh.genouest.org**). You **MUST** first connect to a node (using qrsh) or submit a job to a node (using qsub).

When you submit a job, it is dispatched on one of the computing nodes of the cluster.

Those nodes have different characteristics (cpu, ram).We have servers from 32G up to 512G RAM on the nodes, with at 8 to 40 cores each. Launch the following command to display the list of available nodes and their characteristics and load:

```
qhost
```

You can submit a job with the “qsub” command, and you can monitor your jobs with the “qstat” command.

You can (must) select the appropriate number of resources needed for your program.
If your job needs more than 1 CPU, you can request for multiple CPU slots:

```
 qsub -pe make 8 job_script.sh
```

In this example, we request 8 free CPU on a node to execute the bash script “job_script.sh”.

If you use more than selected RAM, your process will be killed.

If your job needs 15Gb of RAM for example, you can request a node with 15Gb of RAM available:

```
 qsub -l h_vmem=15G job_script
```

For interactive operations, it is possible to use the “qrsh” command to login onto a node.

To kill a job, simply execute:

```
qdel XXX  (XXX is the job identifier)
```

A memento on usages with the cluster is available.
You can find a quick tutorial on SGE on this web site.