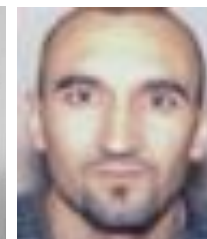
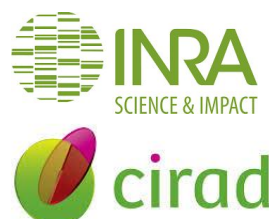


South Green
bioinformatics platform





Institut de Recherche
pour le Développement

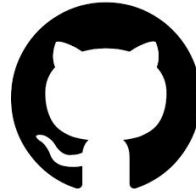
South Green
bioinformatics platform



plateau i-trop



www.southgreen.fr



<https://github.com/SouthGreenPlatform>



The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics, Current Plant Biology, 2016

Trainings 2018

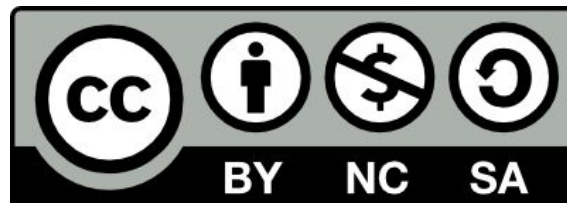


- All trainings here : <https://southgreenplatform.github.io/trainings/>
- Presentation & trainings : [HPC IRD](#)
- Work environment : [softwares to install](#)

HPC cluster Introduction

www.southgreen.fr

<https://southgreenplatform.github.io/trainings>



Objective

Knowing how to use the itrop HPC Cluster

Applications

- Knowing the architecture of the cluster
- Knowing the role of the different systems partitions
- How to use SGE (qsub, qrsh, qhost, qacct, qstat, qqdel)
- Use the modules environment
- Do some basic scripting

ARCHITECTURE

- A cluster is a logical unit composed of several servers
- Acts like a unique powerful server
- Allow to obtain a high computing performance
- A bigger storage
- A better reliability
- A high availability of the ressources

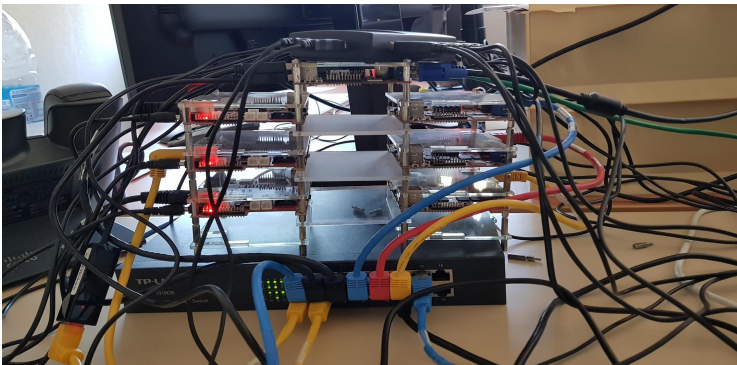
What is a cluster?

- A cluster is a logical unit composed of several servers
- Acts like a unique powerful server
- Allow to obtain a high computing performance
- A bigger storage
- A better reliability
- A high availability of the ressources



What is a cluster?

- A cluster is a logical unit composed of several servers
- Acts like a unique powerful server
- Allow to obtain a high computing performance
- A bigger storage
- A better reliability
- A high availability of the ressources



- Master node : Scheduler.
Handle the ressources and priorities of the jobs
- Computing nodes :
Ressources (CPU or RAM memory)
used by the master

COMPUTING



- Master node : Scheduler.
Handle the ressources and priorities of the jobs
- Computing nodes :
Ressources (CPU or RAM memory)
used by the master

COMPUTING



- Nas servers:
Store the users data and the analyses results

STORAGE





bioinfo-master.ird.fr
91.203.34.148

Role : launch and schedule the jobs on the computing nodes

Accessible from the Internet

Connection : `ssh login@bioinfo-master.ird.fr`



bioinfo-master.ird.fr
91.203.34.148

Role : launch and schedule the jobs on the computing nodes

Accessible from the Internet

Connection : `ssh login@bioinfo-master.ird.fr`



25 noeuds
with
bioinfo-inter.ird.fr
91.203.34.150

Role : Used by the master to execute jobs

Not accessible from the Internet

node0 à node25

Interactive node : `bioinfo-inter.ird.fr`

Connection : `ssh login@bioinfo-inter.ird.fr`



bioinfo-nas3.ird.fr
91.203.34.180



Role : Store data

Accessible from the Internet

Connection : filezilla or scp

bioinfo-nas.ird.fr
91.203.34.157

bioinfo-nas2.ird.fr
91.203.34.160

/home : Your personal folder
Quota 100Go
Hosted on : bioinfo-nas.ird.fr
Shared on all servers

/teams : project data shared between
users from the same team
Quota 200Go
Hosted on : bioinfo-nas.ird.fr
Shared on all servers

/data2 : project data shared between
several users
Quota 500Go à 1To
Hosted on: bioinfo-nas.ird.fr
Shared on all servers

/home : Your personal folder
Quota 100Go
Hosted on : bioinfo-nas.ird.fr
Shared on all servers

/data : project data shared between
several users
Quota 500Go à 1To
Hosted on : bioinfo-nas2.ird.fr
Shared on all servers

/teams : project data shared between
users from the same team
Quota 200Go
Hosted on : bioinfo-nas.ird.fr
Shared on all servers

/data3 :Données projet partagées
entre plusieurs utilisateurs
Quota 500Go à 1To
Hosted on : bioinfo-nas3.ird.fr
Shared on all servers

/data2 : project data shared between
several users
Quota 500Go à 1To
Hosted on: bioinfo-nas.ird.fr
Shared on all servers

/home : Your personal folder
Quota 100Gb
Hosted on : bioinfo-nas.ird.fr
Shared on all servers

/data : project data shared between
several users
Quota 500Gb to 1Tb
Hosted on : bioinfo-nas2.ird.fr
Shared on all servers

/teams : project data shared between
users from the same team
Quota 200Gb
Hosted on : bioinfo-nas.ird.fr
Shared on all servers

/data3 : project data shared between
several users
Quota 500Gb to 1Tb
Hosted on : bioinfo-nas3.ird.fr
Shared on all servers

/data2 : project data shared between
several users
Quota 500Gb to 1Tb
Hosted on: bioinfo-nas.ird.fr
Shared on all servers

/scratch : temporary working folder
1Tb to 5Tb
Hosted on : each node
Not shared local only
Data kept **3 weeks only**



SUN GRID ENGINE (SGE)

- SGE (SUN Grid Engine) is a linux job scheduler, able to handle from 2 to thousands of servers at the same time.
- An opensource tool
- 3 main functions :
 - Allocates ressources (CPU,RAM) to users to allow them to launch their analyses
 - Provides a frame to launch,execute et monitore the jobs on the whole allocated nodes
 - Deals with jobs in queue wait

Bioinfo.q : default queue

Nodes: node2, node8, node9, node10,
node11,node12,node13,
,node14,node15,node16,node17,
node19,node20
RAM: from 48Go to 64Go
Cores: from 12 to 20 cores

~~dynadiv.q : priority for the dynadiv
team~~

Nodes: node2, node10
RAM: 48Go
Cores: 12 cores

/scratch of 5To for node10

dynadiv2.q : priority for thomas
Couvreur

Nodes: node20
RAM: 64Go
Cores: 20 cores

smrtportal.q : priority for the smrtportal
software

Nodes: node17, node18
RAM: 64Go
Cores: 12 cores

alizon.q : priority for the samuel Alizon
team

Nodes: node8, node9, node12
RAM: 48Go
Cores: 12 cores

Bioinfo.q : default queue

Nodes: node2, node8, node9, node10,
node11,node12,node13,
,node14,node15,node16,node17,
node19,node20
RAM: from 48Go to 64Go
Cores: from 12 to 20 cores

dynadiv.q : priority for the dynadiv
team

Nodes: node2, node10
RAM: 48Go
Cores: 12 cores
/scratch of 5To for node10

dynadiv2.q : priority for thomas
Couvreur

Nodes: node20
RAM: 64Go
Cores: 20 cores

smrtportal.q : priority for the smrtportal
software

Nodes: node17, node18
RAM: 64Go
Cores: 12 cores

alizon.q : priority for the samuel Alizon
team

Nodes: node8, node9, node12
RAM: 48Go
Cores: 12 cores

r900.q : queue with **DELL nodes**

Nodes: node5, node21
RAM: 32Go
Cores: 16 cores

longjob.q : long jobs or > to10 jobs

Nodes: node0, node1, node11
RAM: 48Go
Cores: 12 cores

bigmem.q : memory need

Nodes: node3
RAM: 96Go
Cores: 12 cores

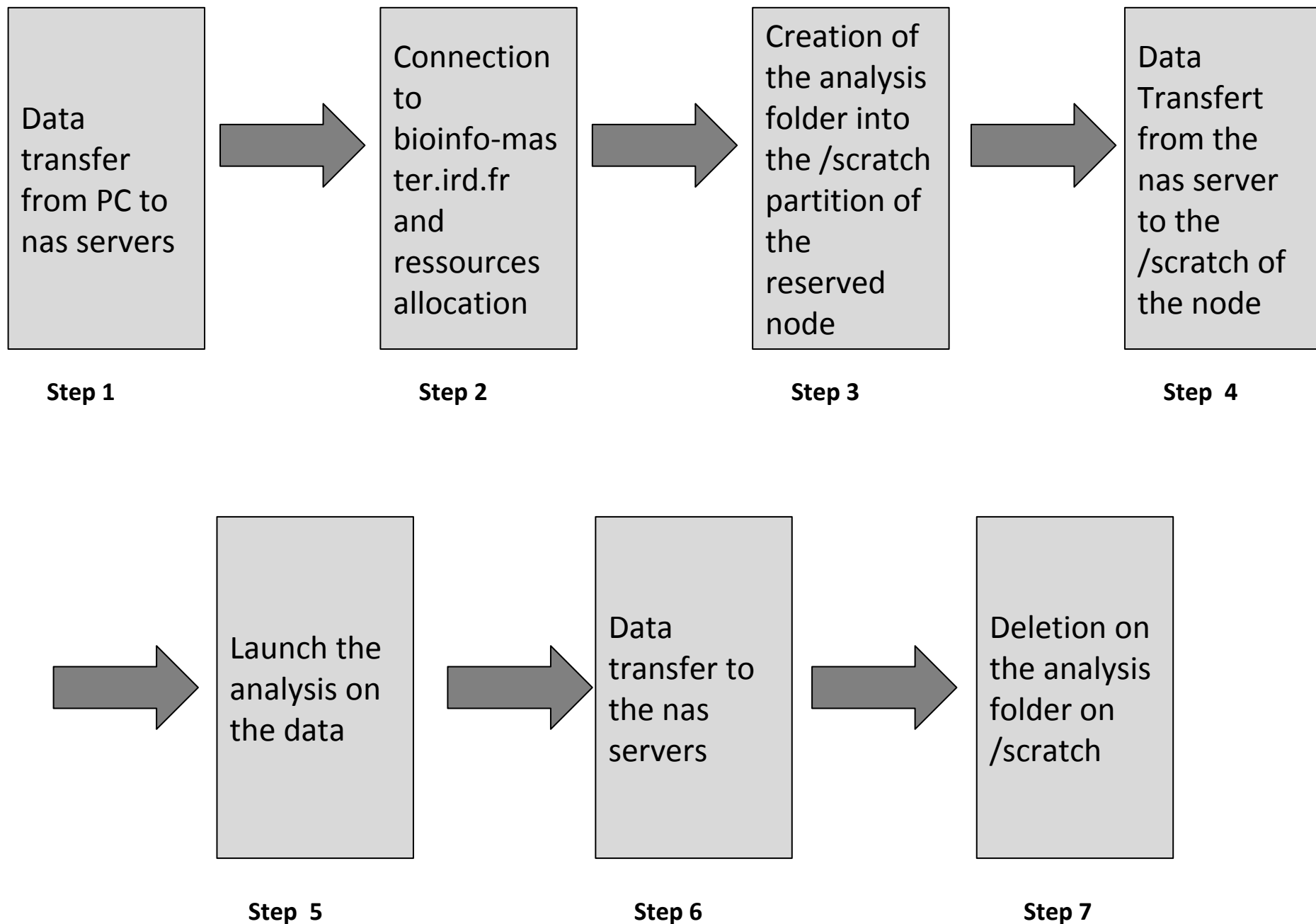
highmem.q :big need of memory

Nodes: node4 et node7
RAM: 144Go
Cores: 12 cores

1

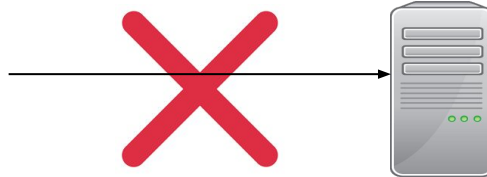


Training: Launch a blast analysis in a interactive way





Personal
Computer

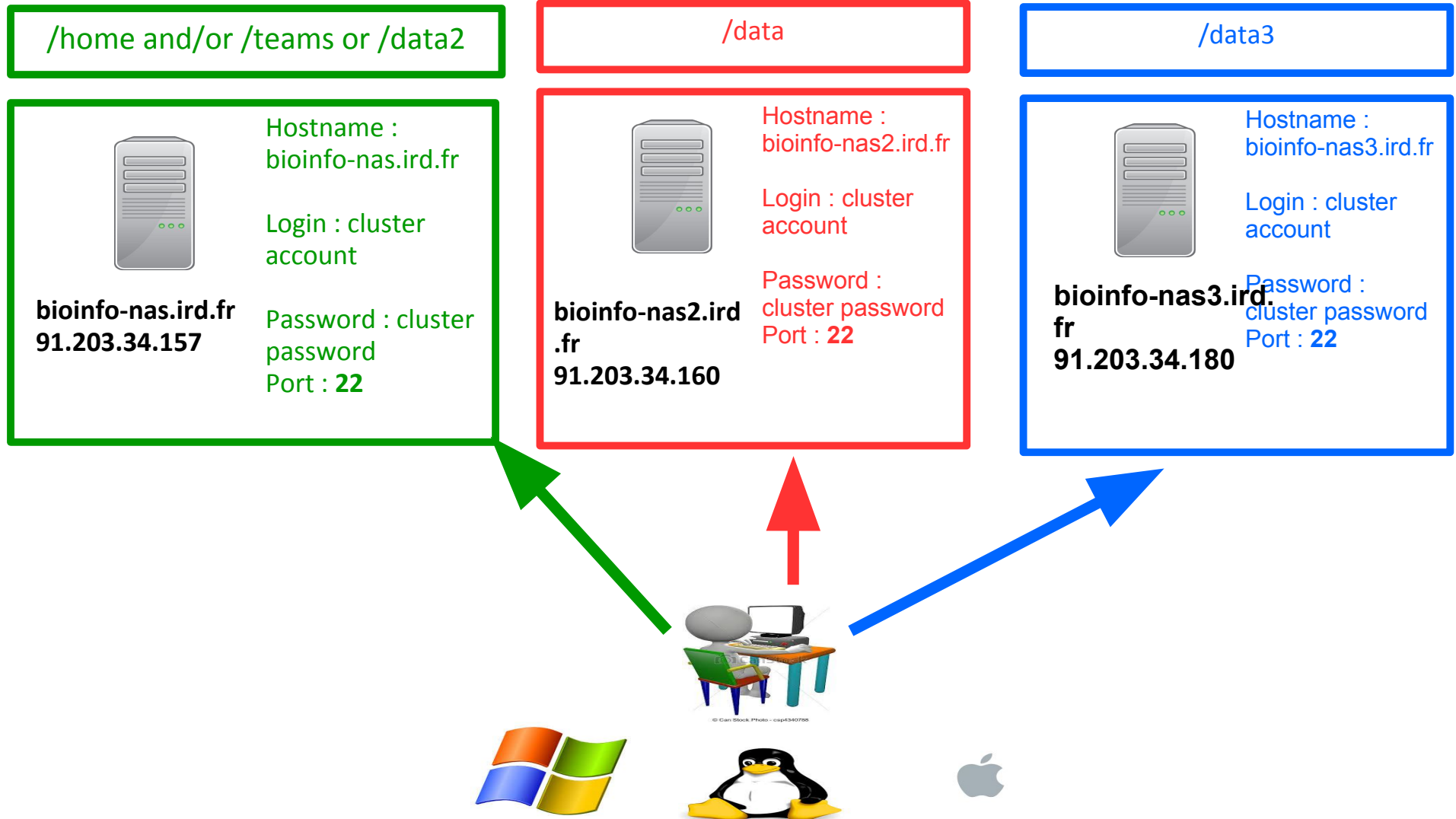


**direct transfert
through filezilla
is forbidden**



**bioinfo-master.ird.fr
91.203.34.148**

Data transfer on the cluster itrop



Open filezilla and retrieve the file « HPC_en.pdf »
Hosted in /data/projects/tp-cluster/training_2018

Open filezilla and retrieve the file « HPC_en.pdf »
Hosted in /data/projects/tp-cluster/training_2018

Enter the following parametres:

Hostname : **bioinfo-nas2.ird.fr**

Login : votre login

Password : votre login

Port :**22**

Navigate into the right window through /data/projects/tp-cluster/training_2018

Retrieve the file HPC_en.pdf with a drag-and-drop

Launch scripts to several nodes



bioinfo-master.ird.fr
91.203.34.148

Use the
qsub
command

Hostname :
bioinfo-master.ird.fr

Login : cluster
account

Password : cluster
password
Port : 22

Test your script(s)



bioinfo-inter.ird.fr
91.203.34.150

Hostname :
bioinfo-inter.ird.fr

Login : cluster
account

Password : cluster
password

Port : 22

Or use the qrun command on
bioinfo-master.ird.fr



With Putty
Use parameters above



© Can Stock Photo - csp4340788



With terminal
Use the ssh command

Connect to bioinfo-master.ird.fr via ssh

Type :

\$~ssh [login@bioinfo-master.ird.fr](https://login.bioinfo-master.ird.fr) onApple or Linux

Under windows : Download Mobaxterm to :

<https://mobaxterm.mobatek.net/download-home-edition.html>

Then connect to bioinfo-master.ird.fr

We can reserve a core of a node to launch an analysis
Through a limited time using the `qrsh` command
Type the `qstat` command and analyse the result

We can reserve a core of a node to launch an analysis
Through a limited time using the qrsh command
Type the qstat command and analyse the result

Type :
\$~qrsh
Check on wich node you are with the command
\$~ uname -a
\$~ qstat

Go into /scratch
Create a folder to host your data

Go into /scratch
Create a folder to host your data

Type the commands :
\$~cd /scratch
\$~ mkdir login (with login the name folder of your choice)

- Copy between 2 remote servers :

```
scp source destination
```

- Syntax if the source is remote :

```
scp server_name:/path/file_to_copy local_folder
```

- Syntax if the destination is remote :

```
scp /path/file_to_copy server_name:/path/remote_folder
```

Copy the folder `/data/projects/tp-cluster/training_2018/Blast` to `/scratch/login`

Copy the folder /data/projects/tp-cluster/training_2018/Blast to /scratch/login

Type the commands :

```
$~cd /scratch/login
```

```
$~ scp -r login@bioinfo-nas2.ird.fr :/data/projects/tp-cluster/training_2018/Blast  
/scratch/login
```

Go into the folder /scratch/login/Blast
List the files of the folder

Go into the folder /scratch/login/Blast
List the files of the folder

Type :
\$~cd /scratch/login/Blast
\$~ls -ali

- Allow to choose the version of software you want to use
- 2 types of softwares :
 - bioinfo : includes all the bioinformatics softwares (example BEAST)
 - system : includes all the system softwares(example JAVA)
- Overcome the environment variables

➤ 5 types of commands :

- See the available modules :

module avail

- Obtain infos on a particular module:

module whatis + module name

- Load a module :

module load + modulename

- List the loaded module :

module list

- Unload a module :

module unload + modulename

- Unload all the modules :

Module purge

Load the blast module version 2.4.0+
Use the `blastn` command to launch a blast analysis
Hat will produce the result file called `blastn.out`

Load the blast module version 2.4.0+
Use the blastn command to launch a blast analysis
That will produce the result file called blastn.out

Type :
\$~ module load bioinfo/blast/2.4.0+
\$~ blastn -db All-EST-cofea.fasta -query sequence-NMT.fasta -out blastn.out

Edit the blastn.out file with the nano tool

Edit the blastn.out file with the nano tool

Type :
\$~ nano blastn.out

Copy the file blastn.out to your home folder
Check that the file has been copied

Copy the file blastn.out to your home folder
Check that the file has been copied

Type :
\$~scp blastn.out login@bioinfo-nas.ird.fr:/home/login
\$~ls -ali /home/login

Go into the /scratch folder
Delete your working directory

Type:
\$~cd /scratch
\$~ rm -rf *login*



**TP: Launch a bwa in a
interactive way**

- Follow the steps from the last training and adapt them to this one
 - The folder to copy: /data/projects/training_2018/bwa
 - Bwa version to use: 0.7.12
 - Commands to launch:
bwa index referenceIrigin.fasta
bwa mem referenceIrigin.fasta irigin1_1.fastq irigin1_2.fastq >mapping.sam
 - Retrieve the file mapping.sam and place it in your /home/

Cf solution: [practice2](#)



**Training: Launch an analyse
via a bash script**

- Execute a bash script via sge
- We use the command:

```
$~ qsub script.sh
```

With `script.sh` : the script name

First part of the script (in green): sge execution options with the key word #

```
#!/bin/sh

##### SGE CONFIGURATION #####
# write errors in standard outputfile
#$ -j y

# Shell we want to use
#$ -S /bin/bash

# Email to follow the job
#$ -M prenom.nom@ird.fr ##### Mettre son adresse mail

# Type of messages by mail
# - (b) beginning message
# - (e) end message
# - (a) abort message
#$ -m bea

# Queue to use
#$ -q bioinfo.q

# Name of the job
#$ -N name_to_choose
#####
```

In the 2nd part of the script: the command to execute

```
path_to_dir="/data/projects/folder_to_choose";
path_to_tmp="/scratch/name_folder_to_choose-$JOB_ID"

##### Create the temporary folder on the node and load the blast module
module load bioinfo/blastn/2.4.0+
mkdir $path_to_tmp
scp -rp nas2:$path_to_dir/* $path_to_tmp # choose nas for /home, /data2 and /teams or nas2 for /data or nas3 for /data3
echo "tranfert from master -> noeud";
cd $path_to_tmp

##### Program execution
cmd="blastn -db All-EST-cofea.fasta -query sequence-NMT.fasta -num_threads $NSLOTS -out blastn1-$JOB_ID.out";
echo "executed command : $cmd";
$cmd;

##### Data transfer from node to nas
scp -rp $path_to_tmp/ nas:$path_to_dir/
echo "Transfert from node -> master";

#### Deletion of the tmp folder
rm -rf $path_to_tmp
echo "Deletion on the node";
```

- Using the Training 1 create a script to launch a blastn analysis
- Make the script launchable with

\$~ chmod 755 script.sh

- Launch the script with qsub:
\$~ qsub script.sh
- Check the running script with the command: watch qstat
[solution script blastn](#)

Use the dos2unix command when the script has been written
under Windows



- Using the Training 1 create a script to launch a bwa analysys
- Make the script launchable with

\$~ chmod 755 script.sh

- Launch the script with qsub:
\$~ qsub script.sh
- Check the running script with the command: watch qstat

[solution script bwa](#)



Use the dos2unix command when the script has been written under Windows