South Green
bioinformatics platform

HPC cluster Initiation

www.southgreen.fr

https://southgreenplatform.github.io/trainings

Bioversity International

cirad

INRA SCIENCE & IMPACT

IRD Institut de Recherche pour le Développement FRANCE

Montpellier SupAgro

i-Trop presentation

Julie ORJUELA-BOUNIOL[1], IE
Bioinformaticienne
25%

Ndomassi TANDO, IE
Administrateur systeme
100%
Animateur plateau

Christine TRANCHANT-DUBREUIL, IE
Bioinformaticienne
20%

Aurore COMTE, IE
Bioinformaticienne
20%

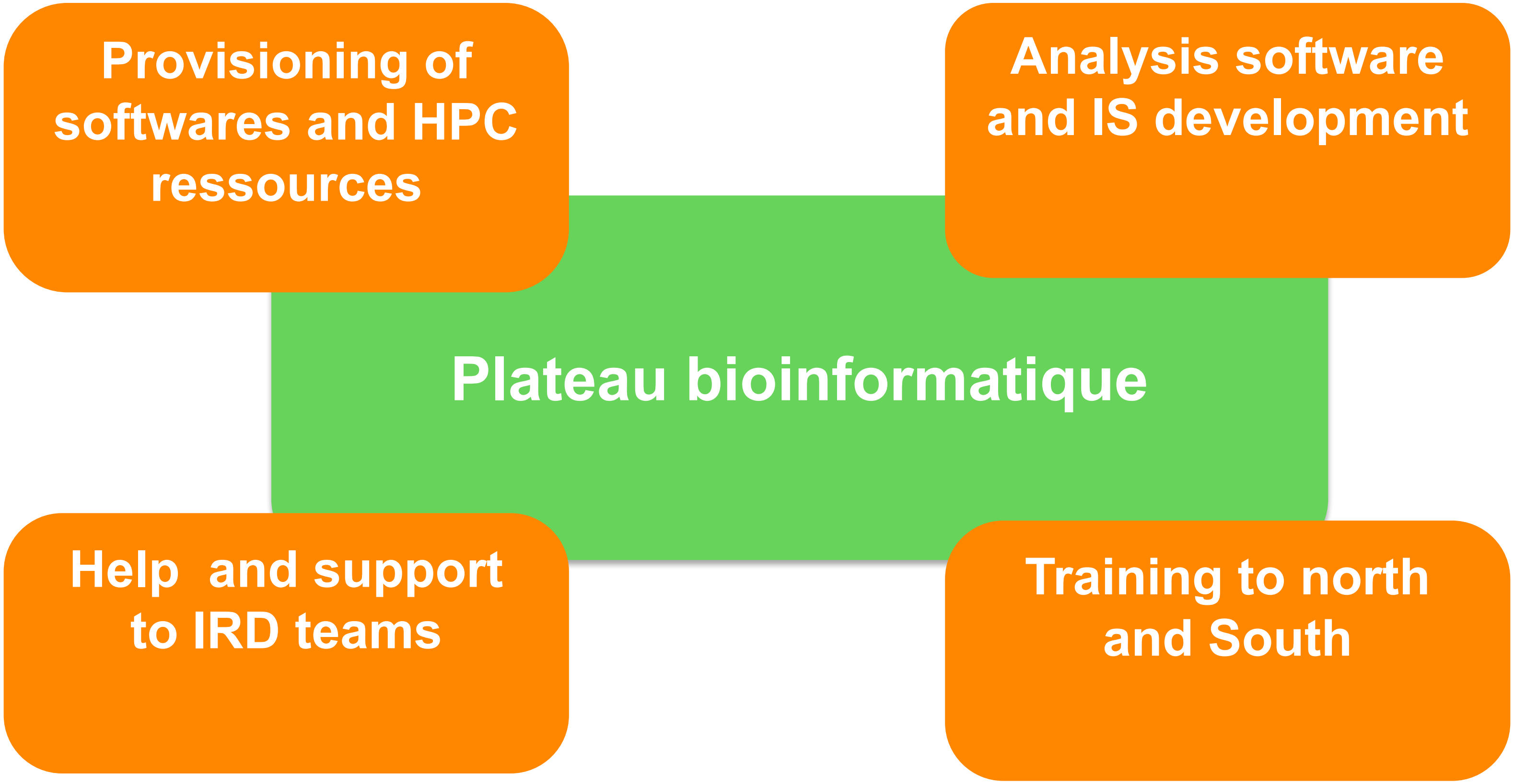
Alexis DEREEPER[2], IE
Bioinformaticien
20%

Valérie NOEL, TCS
Bioinformaticienne
25%

Bruno GRANOUILLAC[3], IE
Systèmes d'information
20%

1 Affectation multiple DIADE, IPME et BOREA 2 En disponibilité 3 Affectation multiple MIVEGEC et TRANSVIH-MI

i-Trop Presentation



Provisioning of softwares and HPC ressources

Analysis software and IS development

Plateau bioinformatique

Help and support to IRD teams

Training to north and South

Queries/incidents

- Site https://bioinfo.ird.fr
    - Accounts
    - Softwares installation
    - Projects
    - Installed softwares
- Incidents: contact bioinfo@ird.fr

ARCHITECTURE

A Cluster?

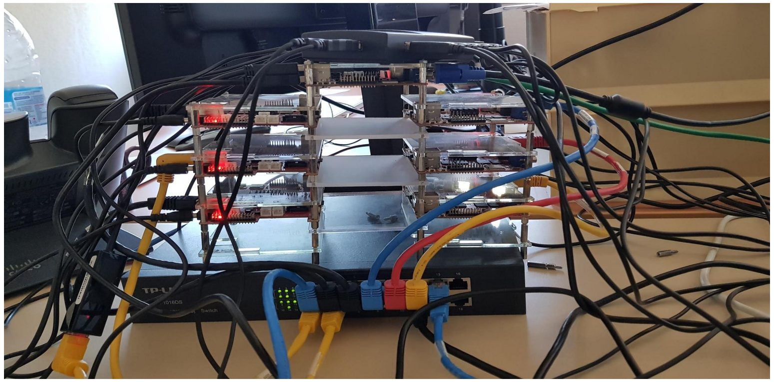
- A logical unit composed of several servers
- A powerful unique machine
- Allow to obtain high computing performance
- A bigger capacity storage
- More reliable
- A better ressources availabilty

A cluster?

- A logical unit composed of several servers
- A powerful unique machine
- Allow to obtain high computing performance
- A bigger capacity storage
- More reliable
- A better ressources availabilty

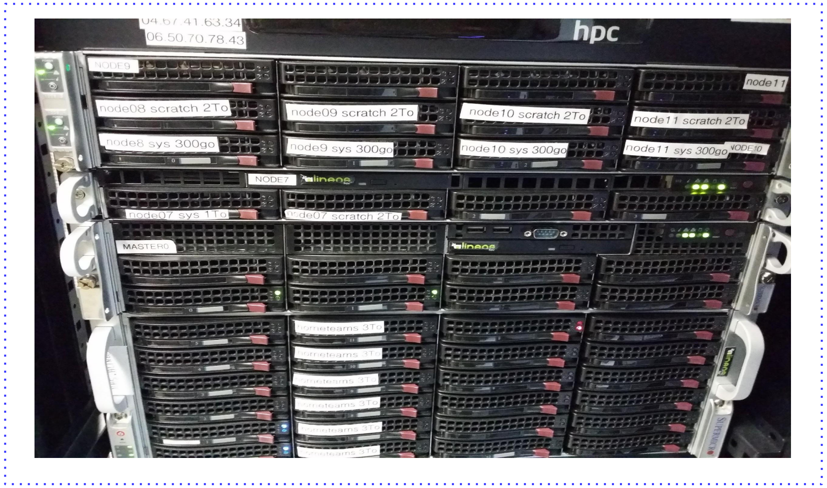
A cluster?

- A logical unit composed of several servers
- A powerful unique machine
- Allow to obtain high computing performance
- A bigger capacity storage
- More reliable
- A better ressources availabilty



Cluster components

COMPUTING



- **Master Node**
  Handle resources and jobs priorities
- **Computing nodes**
  Resources (CPU or RAM memory)

Cluster components

COMPUTING

- **Master Node**
  Handle resources and jobs priorities
- **Computing nodes**
  Resources (CPU or RAM memory)

STORAGE

- **NAS Server(s)**
  Storage

Architecture: components role

- **1 Master Node**

bioinfo-master.ird.fr

Role :

- Launch and prioritize jobs on computating nodes
- Accessible from the Internet
- Connection :

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

Architecture: components role

- **1 Master Node**

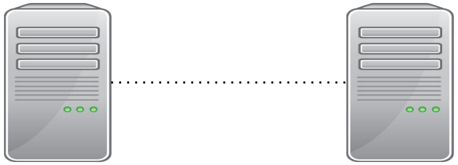
bioinfo-master.ird.fr

Role :

- Launch and prioritize jobs on computating nodes
- Accessible from the Internet
- Connection :

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

- **25 computing nodes**



nodeX
X : 1..25

Role :

- Used by the master to execute jobs
- Not accessible from the Internet
- node0 to node25
- Connection from master

```
ssh nodeX
```

Architecture: components role

1 Master Node

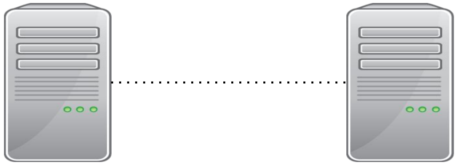
bioinfo-master.ird.fr

Role :

- Launch and prioritize jobs on computating nodes
- Accessible from the Internet
- Connection :

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

25 computing nodes



nodeX
X : 1..25

Role :

- Used by the master to execute jobs
- Not accessible from the Internet
- node0 to node25
- Connection from master

```
ssh nodeX
```

Interactif node (node6)

- Accessible from the Internet: bioinfo-inter.ird.fr
- Connection : `ssh login@bioinfo-inter.ird.fr`

Practice

Step 1: Connection, qhost

1

Go to the Practice 1 *of github*

Analyses steps of the cluster

Connection to bioinfo-master.ird.fr and resources reservation

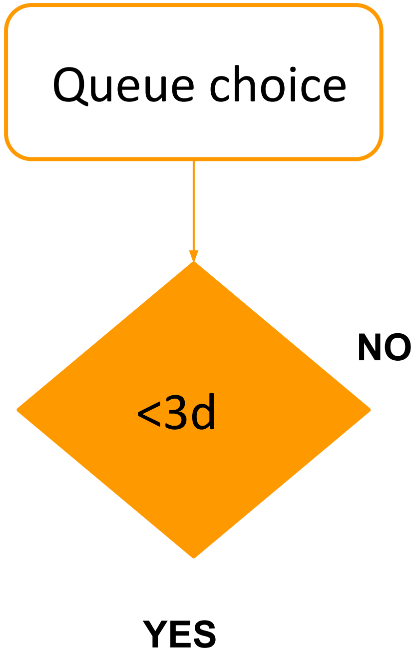
Step 1

qrsh/qlogin or qsub

Queues

Queues	Use	RAM features of the nodes	Cores features of the nodes
bioinfo.q	Short Jobs < 3days	48 to 64 GB	12 to 20 cores
longjob.q	Long Jobs > 3 days	48 GB	12 cores
bigmem.q	Jobs with extra memory needs	96 GB	12 cores
highmem.q	Jobs with big memory needs	144 GB	12 cores

How do I choose the queue?



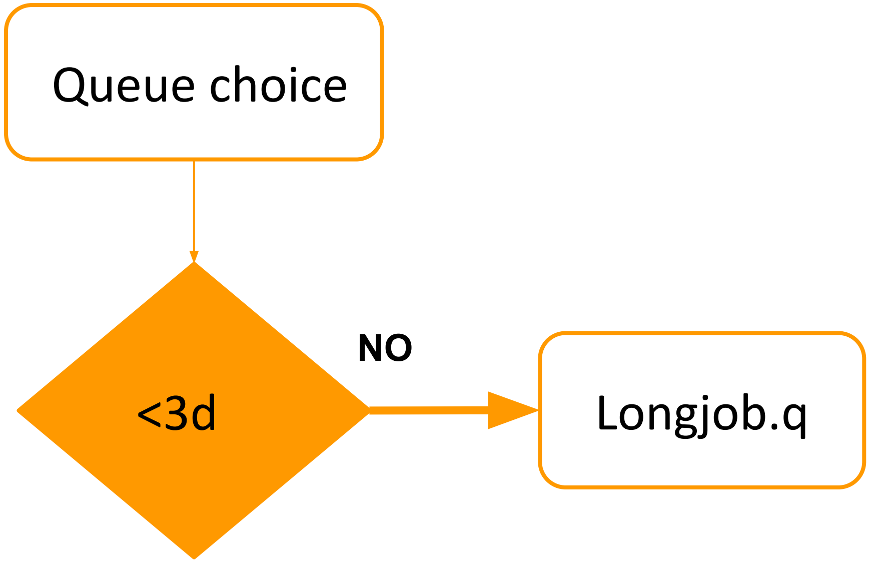
Queue choice

<3d

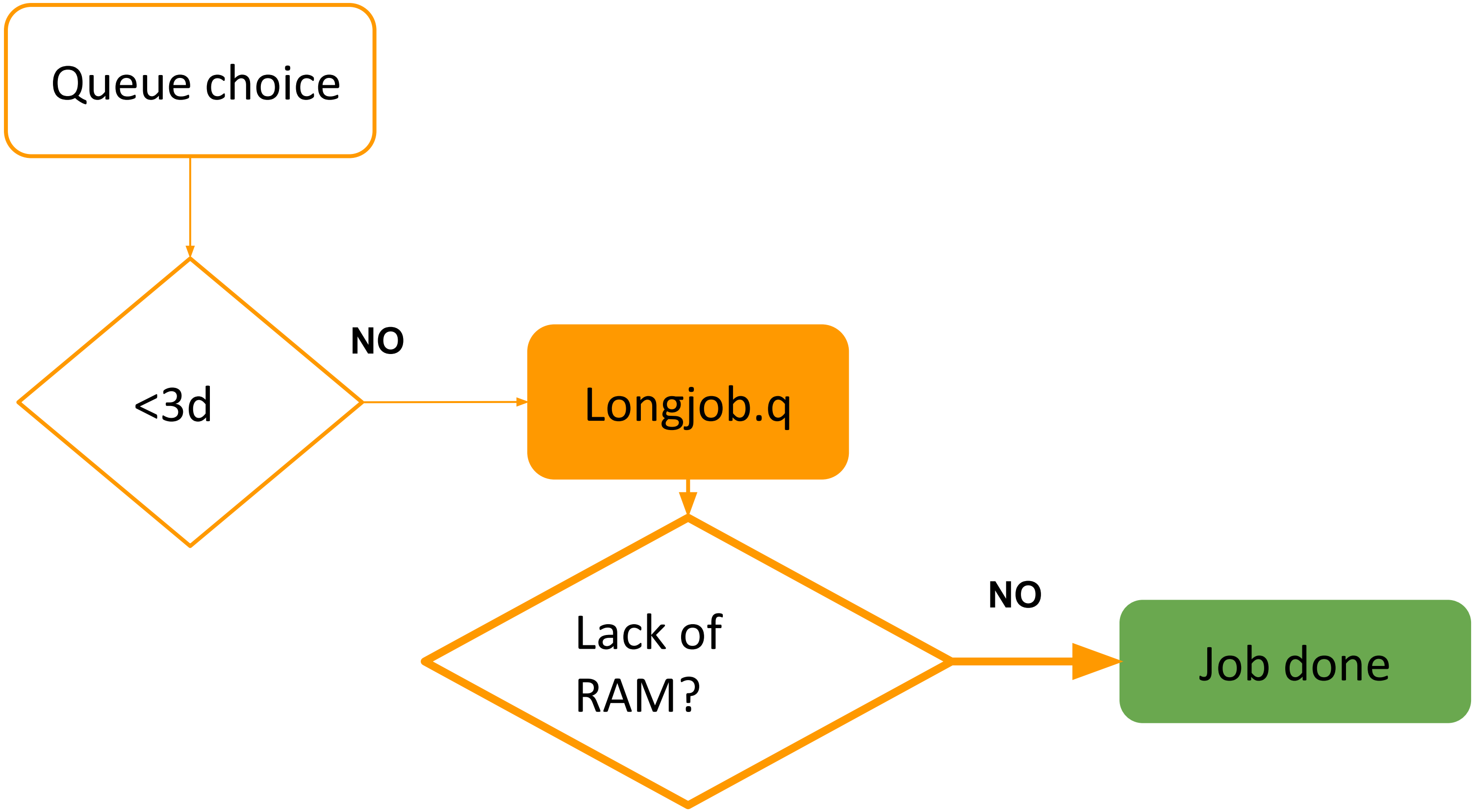
NO

YES

How do I choose the queue?



How do I choose the queue?



Queue choice

<3d

NO

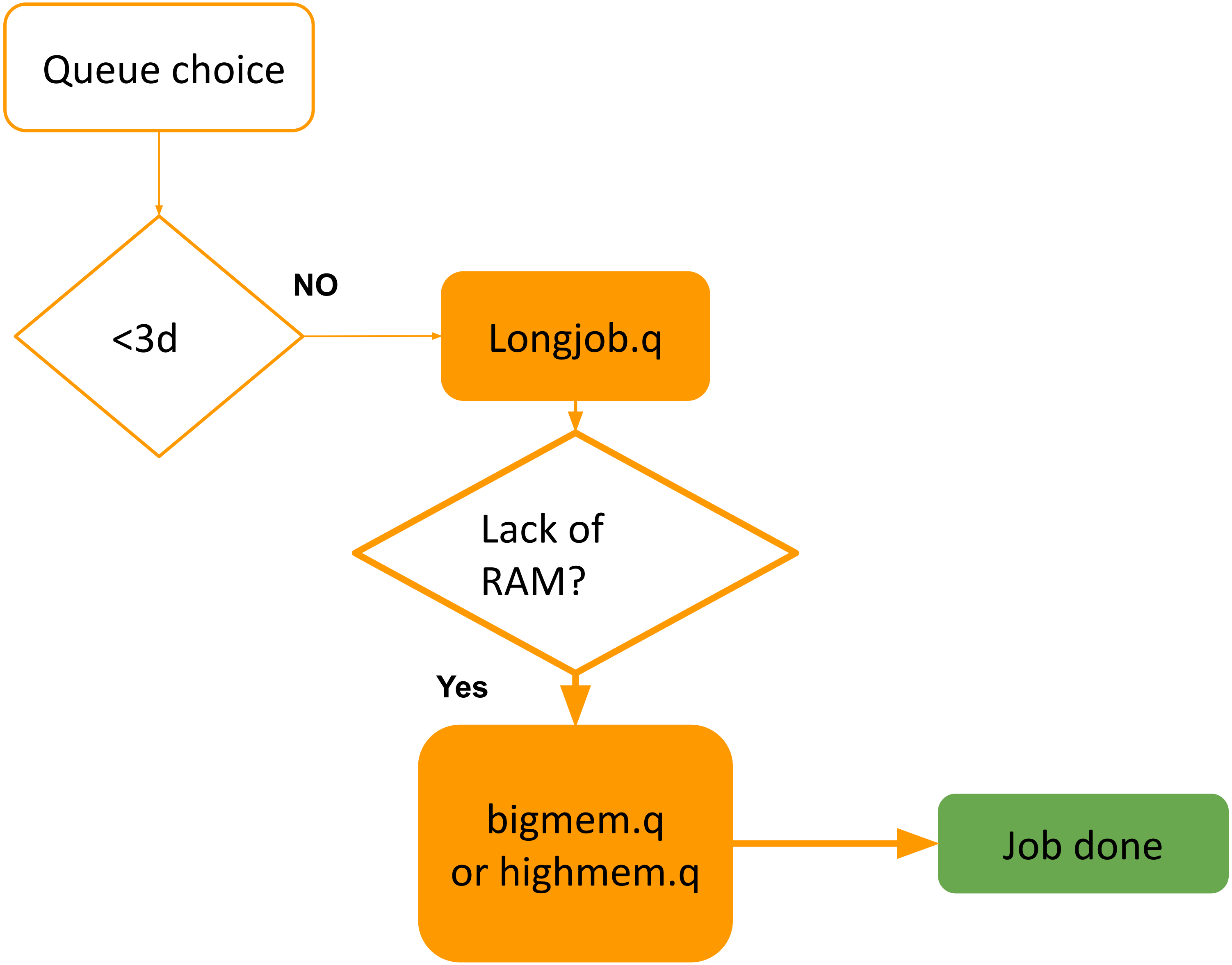
Longjob.q

Lack of RAM?

NO

Job done

How do I choose the queue?



Queue choice

<3d

NO

Longjob.q

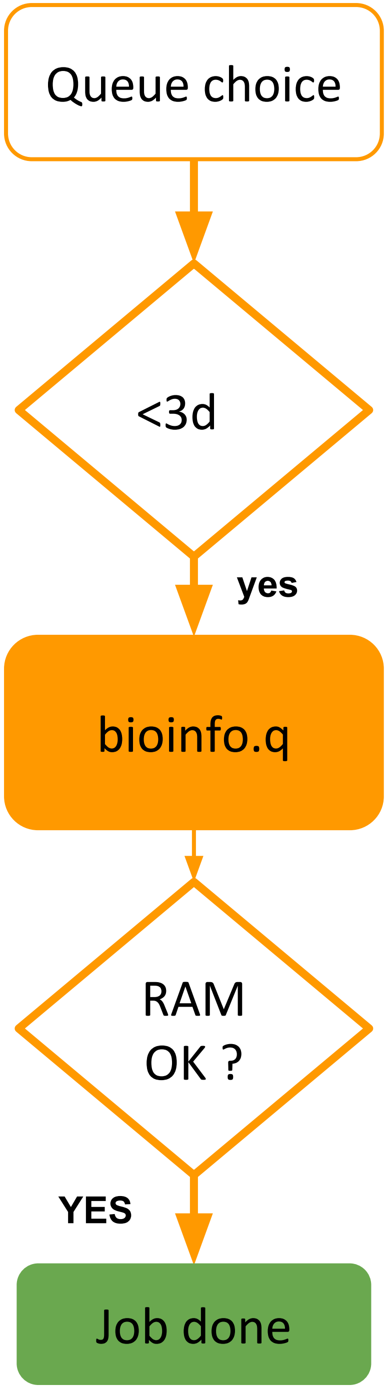
Lack of RAM?

Yes

bigmem.q or highmem.q

Job done

How do I choose the queue?



Queue choice

<3d

yes

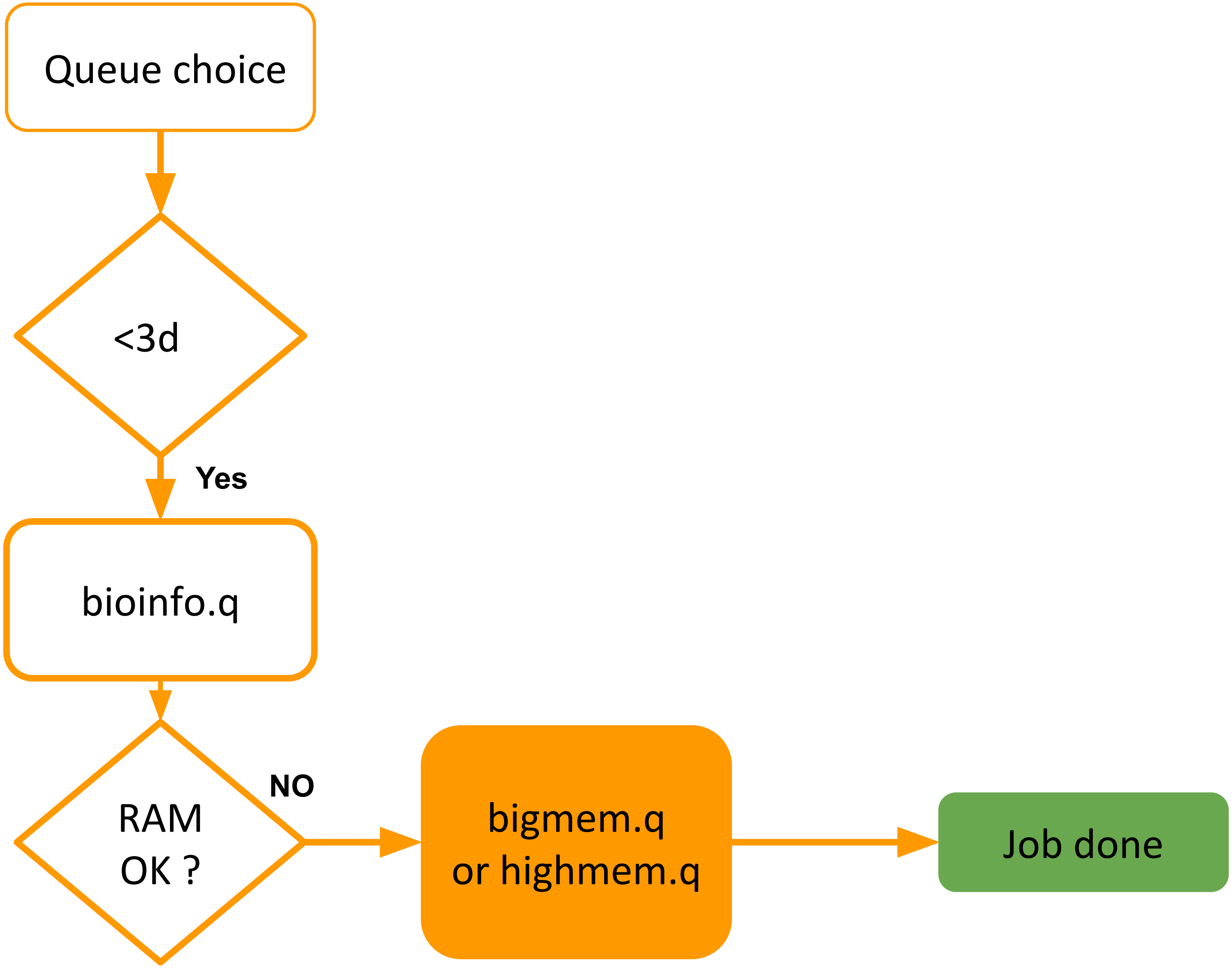
bioinfo.q

RAM OK ?

YES

Job done

How do I choose the queue?



Queue choice

<3d

Yes

bioinfo.q

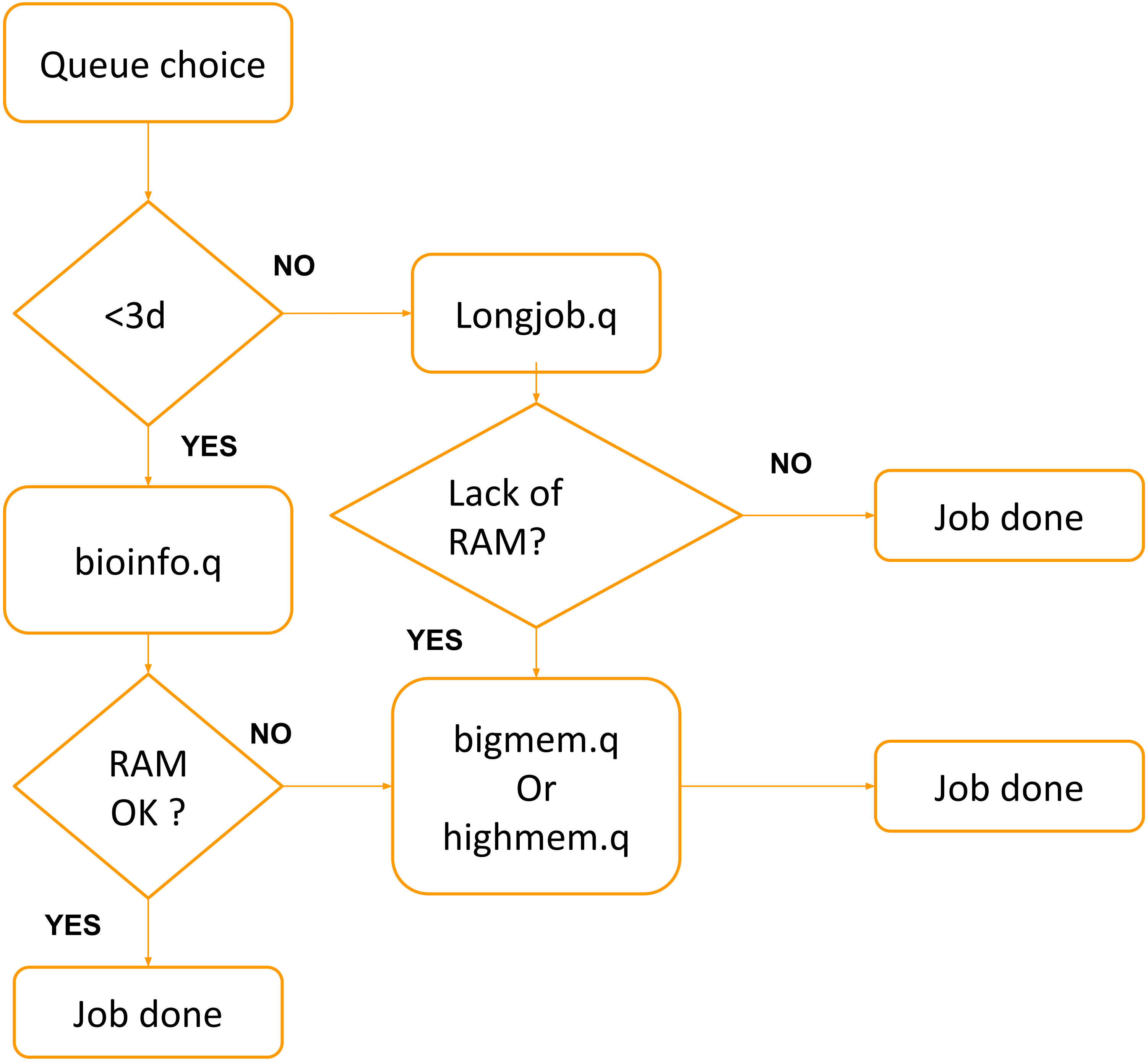
RAM OK ?

NO

bigmem.q or highmem.q

Job done

How do I choose the queue?



Queue choice

<3d

NO

Longjob.q

YES

bioinfo.q

Lack of RAM?

NO

Job done

YES

RAM OK ?

NO

bigmem.q
Or
highmem.q

Job done

YES

Job done

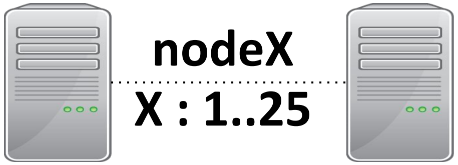
Architecture: components role

1 Master node

bioinfo-master.ird.fr

Role :
- Launch and prioritize jobs on computing nodes
- Accessible from the Internet

25 computing nodes



nodeX

X : 1..25

Role :
- Used by the master to execute jobs
- Not accessible from the Internet

3 NAS servers

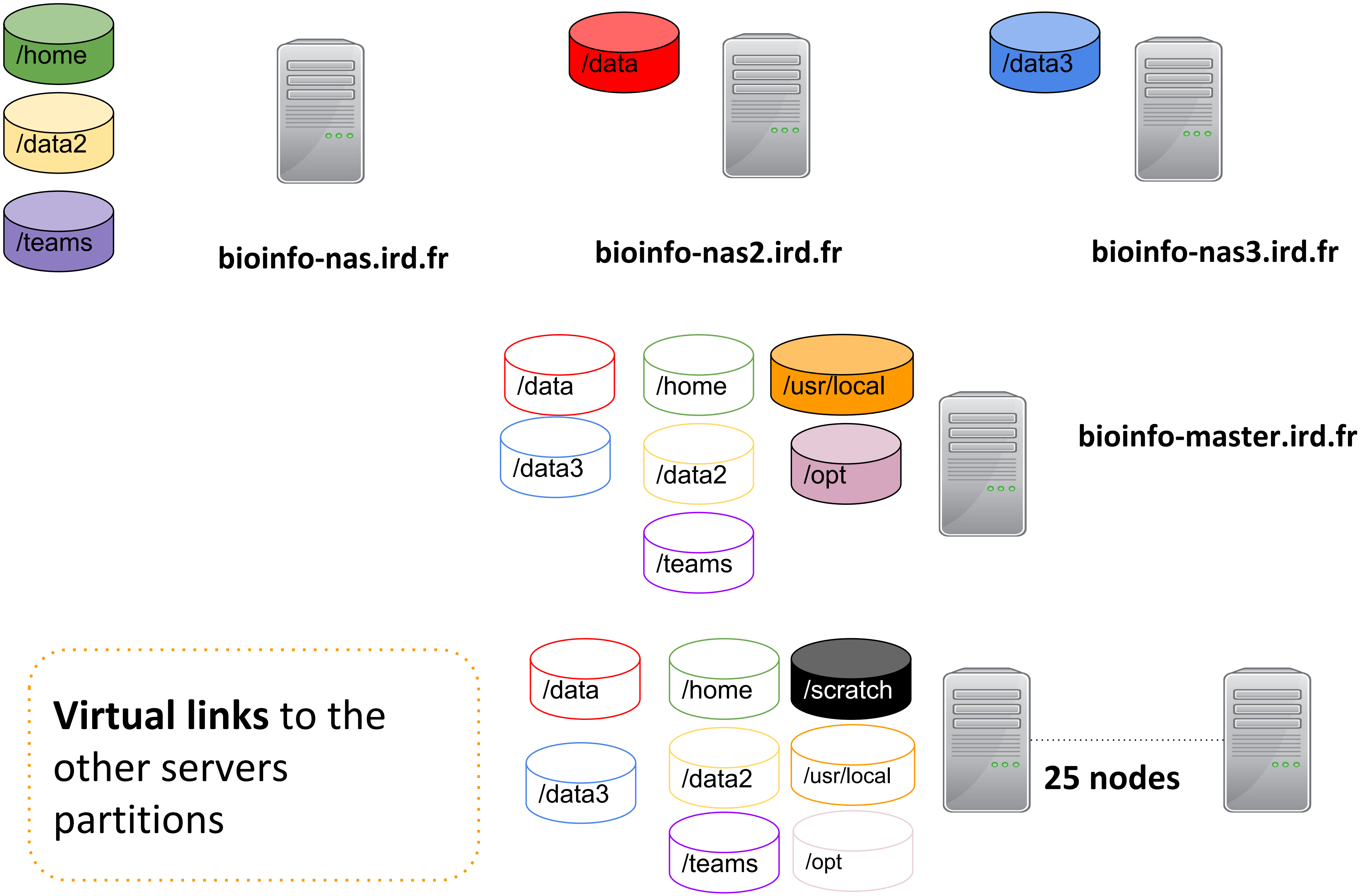
Bioinfo-nas.ird.fr
(nas)

Bioinfo-nas2.ird.fr
(nas2)

Bioinfo-nas3.ird.fr
(nas3)

Role :
- Store users data
- Accessible from the Internet
- To transfer data : *via filezilla or scp*

cluster i-Trop disk partitions



Virtual links to the other servers partitions

Analyses steps of the cluster

Connection to bioinfo-master.ird.fr and resources reservation

Step 1

Creation of the analyses folder in the /scratch of the node

Step 2
mkdir

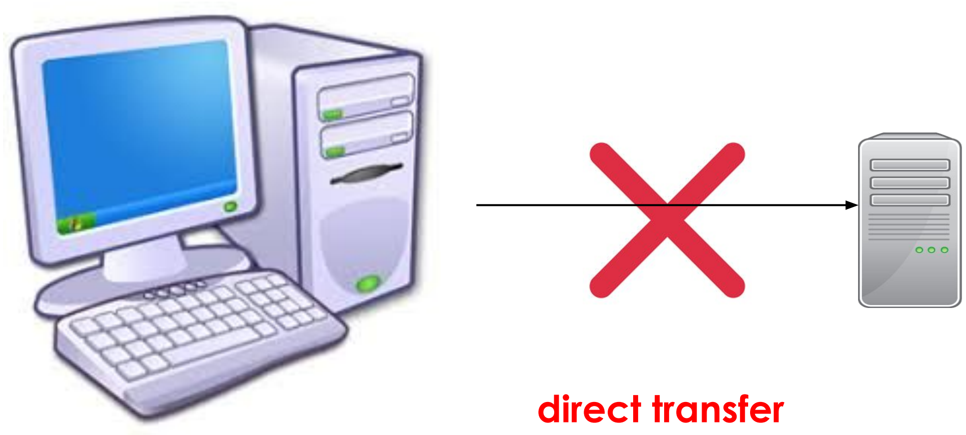
Practice

Step 2:qrsh, partition

2

Go to the Practice2 *of the github*

Data transfer on i-Trop cluster

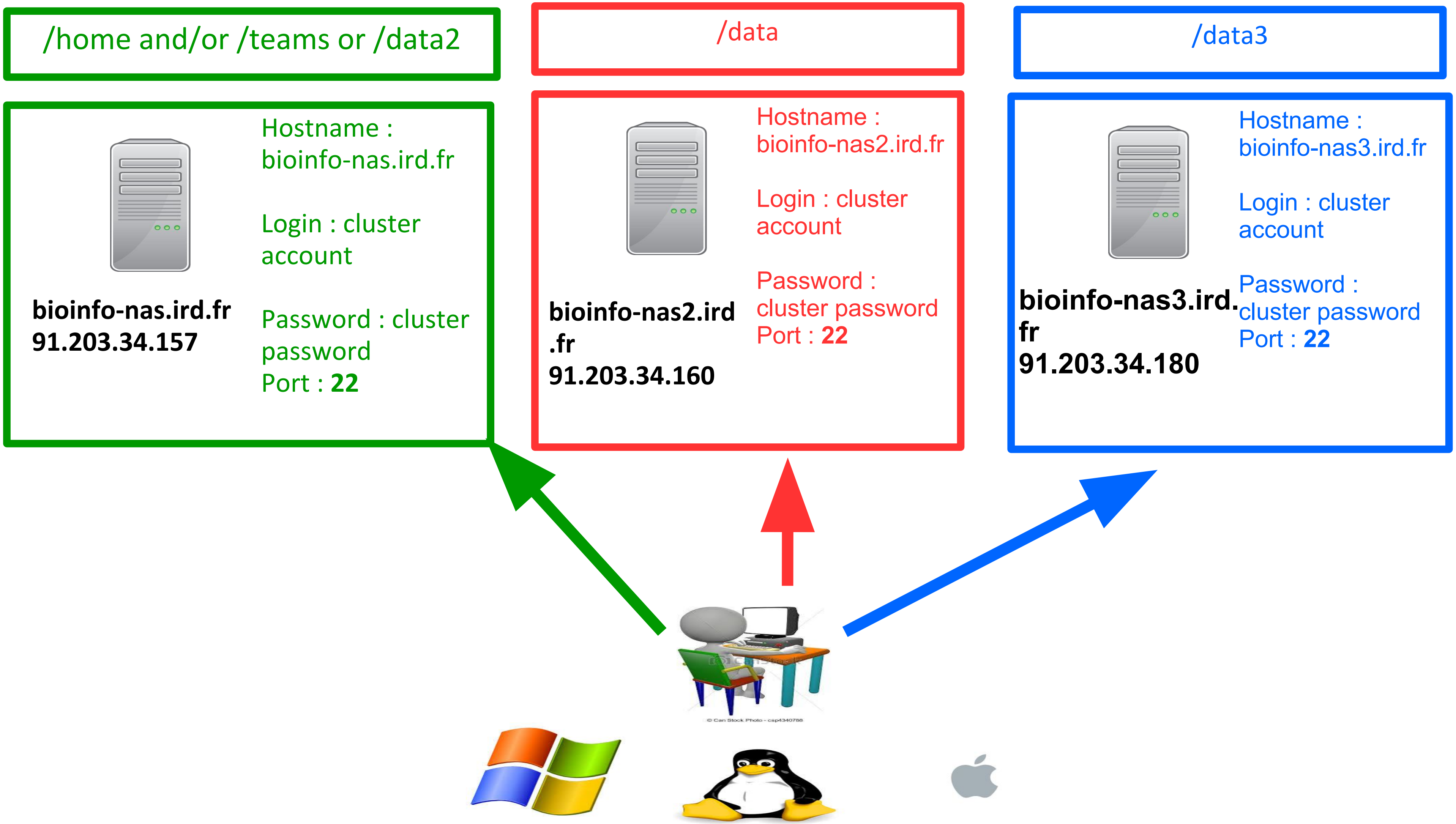


bioinfo-master.ird.fr
91.203.34.148

direct transfer
via filezilla
forbidden

PC/MAC

Data transfer on i-Trop cluster



/home and/or /teams or /data2

bioinfo-nas.ird.fr
91.203.34.157

Hostname : bioinfo-nas.ird.fr

Login : cluster account

Password : cluster password
Port : **22**

/data

bioinfo-nas2.ird.fr
91.203.34.160

Hostname : bioinfo-nas2.ird.fr

Login : cluster account

Password : cluster password
Port : **22**

/data3

bioinfo-nas3.ird.fr
91.203.34.180

Hostname : bioinfo-nas3.ird.fr

Login : cluster account

Password : cluster password
Port : **22**

© Can Stock Photo - csp4340788

Analyses steps of the cluster

Connection to bioinfo-master.ird.fr and resources reservation

Step 1

Creation of the analyses folder in the /scratch of the node

Step 2

Data transfer from your PC/MAC to the nas servers

Step 3

Copy your data from your PC/MAC to the nas serve if they are not on the cluster

Practice

Step 3: filezilla

3

Go to the Practice3 *of the github*

Copy with scp

- Copy between 2 remote servers :

```
scp  -r source destination
```

- Syntax if the source is remote :

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

- Syntax if the destination is remote :

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

Ex: `scp -r nas:/home/tando/folder/ /scratch/tando/`

Analyses steps of the cluster

Connection to bioinfo-master.ird.fr and resources reservation

Step 1

Creation of the analyses folder in the /scratch of the node

Step 2

Data transfer from your PC/MAC to the nas servers

Step 3

Data transfer from the nas servers to the /scratch of the node

Step 4

scp

Practice

Step 4: scp to nodes

4

Go to the [Practice4](#) *of the github*

Module Environment

- 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

Module Environment

➢ 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

Analyses steps of the cluster

Connection to bioinfo-master.ird.fr and resources reservation

Step 1

Creation of the analyses folder in the /scratch of the node

Step 2

Data transfer from your PC/MAC to the nas servers

Step 3

Data transfer from the nas servers to the /scratch of the node

Step 4

Load softwares with modules environment

Step 5

module

Practice

Step 5: module environment

5

Go to the Practice5 *of the github*

Analyses steps of the cluster

Connection to bioinfo-master.ird.fr and resources reservation

Step 1

Creation of the analyses folder in the /scratch of the node

Step 2

Data transfer from your PC/MAC to the nas servers

Step 3

Data transfer from the nas servers to the /scratch of the node

Step 4

Load softwares with modules environment

Step 5

Launch analyses on the data

Step 6

Launch a command from the prompt

- Load the software version to launch
- Launch the data analysis

```
$~ command <options> <arguments>
```

With *command*: the command to launch

Launch a command line job

- Execute a bash command via qsub
- Launch the command from a node
- We use:

```
$~ qsub -b y “command”
```

With *command*: the command to launch

Qsub options

Options	Description	Exemple
qsub **-N** <name>	Give a name to the job	qsub -N tando_blast
qsub **-q <**queue>	Choose a particular queue	qsub -q highmem.q
qsub **-l hostname**=<nodeX>	Choose a particular node	qsub -l hostname=node10
qsub **-pe** <ompi X>	Launch a several cores jobs	qsub -pe ompi 4
qsub **-M** <emailaddress>	Send an email	qsub -M ndomassi.tando@ird.fr
qsub **-m** <eab>	Send an email when: e: end of the job a: abort b: begin of the job	qsub -m be
qsub -cwd	Launch a job from the current working directory	qsub -cwd script.sh

Practice

Step6: launch the analysis

Go to the [Practice6]() *of the github*

Transfer your results to the nas servers

- 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
```

South Green bioinformatics platform

Analyses steps of the cluster

Connection to bioinfo-master.ird.fr and resources reservation

Step 1

Creation of the analyses folder in the /scratch of the node

Step 2

Data transfer from your PC/MAC to the nas servers

Step 3

Data transfer from the nas servers to the /scratch of the node

Step 4

Load softwares with modules environment

Step 5

Launch analyses on the data

Step 6

Results transfer on the nas servers

Step 7

scp

Practice

Step 7: Retrieve the results

7

Go to the Practice7 *of the github*

Delete results from scratchs

- Scratch= temporary spaces
- Verify that the copy is OK before
- Use rm command

```
cd /scratch
rm -rf nom_rep
```

SouthGreen bioinformatics platform

Analyses steps of the cluster

Connection to bioinfo-master.ird.fr and resources reservation

Step 1

Creation of the analyses folder in the /scratch of the node

Step 2

Data transfer from your PC/MAC to the nas servers

Step 3

Data transfer from the nas servers to the /scratch of the node

Step 4

Load softwares with modules environment

Step 5

Launch analyses on the data

Step 6

Results transfer on the nas servers

Step 7

Delete the analyses folder of the /scratch

Step 8

rm

Practice

Step8: Data deletion

8

Go to the Practice8 *of the github*

Scripts to visualize/delete données temporary data

- Scripts location: /opt/scripts/scratch-scripts
- Visualize data on scratchs: scratch_use.sh

```
sh /opt/scripts/scratch-scripts/scratch_use.sh
```

- Delete data on scratchs: clean_scratch.sh

```
sh /opt/scripts/scratch-scripts/clean_scratch.sh
```

BONUS

LAUNCH A JOB

Advantages

- Scheduler choose resources automatically
- Use up to 24 cores at the same time
- Possibility to configure this choice
- Jobs launch in background

  → possibility to turn off your PC/MAC

  → automatic results retrieving

Launch a batch job

- Execute a script via sge
- Use:

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

with script.sh : the name of the script

Qsub options

Options	Description	Exemple
qsub **-N** <name>	Give a name to the job	qsub -N tando_blast
qsub **-q <**queue>	Choose a particular queue	qsub -q highmem.q
qsub **-l hostname**=<nodeX>	Choose a particular node	qsub -l hostname=node10
qsub **-pe** <ompi X>	Launch a several cores jobs	qsub -pe ompi 4
qsub **-M** <emailaddress>	Send an email	qsub -M ndomassi.tando@ird.fr
qsub **-m** <eab>	Send an email when: e: end of the job a: abort b: begin of the job	qsub -m be
qsub -cwd	Launch a job from the current working directory	qsub -cwd script.sh

Bash scripts syntax

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

```
#!/bin/sh


############      SGE CONFIGURATION      ###################
# wirite 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 messges 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
############################################################
```

Bash scripts syntax

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-coffea.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";
```

Practice

Launch a script with sge

9

Go to the Practice9 *of the github*

Citations

If you use i-Trop Bioinformatics resources.

Thank you for citing with:

“The authors acknowledge the IRD itrop HPC (South Green Platform) at IRD montpellier

for providing HPC resources that have contributed to the research results reported within this paper.

URL: https://bioinfo.ird.fr/- http://www.southgreen.fr”

Projects

- Include a budget for bioinformatics resources in your answer to projects funding

- A need in hard drives, renewal machinesetc…

- Available quotations

- Contact bioinfo@ird.fr : help, needs definition, quotations...

Thank you for your attention !