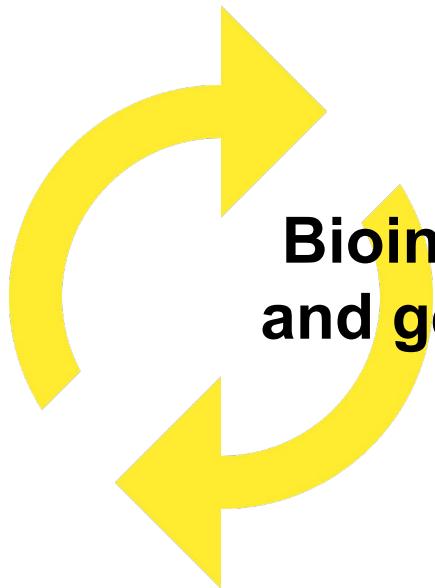




2019 Training modules





Bioinformatics platform dedicated to the genetics
and genomics of tropical and Mediterranean plants
and their pathogens

genome assembly SNP detection
phylogeny structural variation
comparative genomics transcriptome assembly differential expression
GWAS pangenomics
population genetics metagenomics
polypliody



Rice



Banana



Palm



Sorghum



Coffee



Cassava



Magnaporthe

South Green

bioinformatics platform



Larmande Pierre
Sabot François
Tando Ndomassi
**Tranchant-Dubreuil
Christine**



Comte Aurore
Dereeper Alexis



Orjuela-Bouniol Julie



Bocs Stephanie
De Lamotte Fredéric
Droc Gaetan
Dufayard Jean-François
Hamelin Chantal
Martin Guillaume
Pitollat Bertrand
Ruiz Manuel
Sarah Gautier
Summo Marilyne



Rouard Mathieu
Guignon Valentin
Catherine Breton



Mahé Frédéric
Ravel Sébastien



Sempere Guilhem



South Green bioinformatics platform

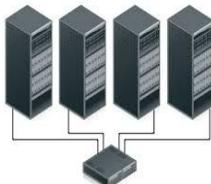
Workflow manager

TOGGLE
Toolbox for generic NGS analyses



Galaxy

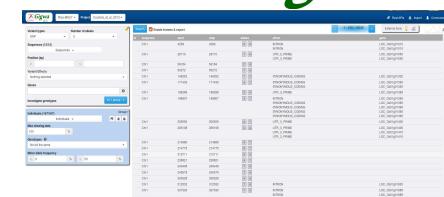
HPC and trainings....



IRD
Institut de Recherche
pour le Développement
Tropique et Méditerranéen



Genome Hubs & Information System

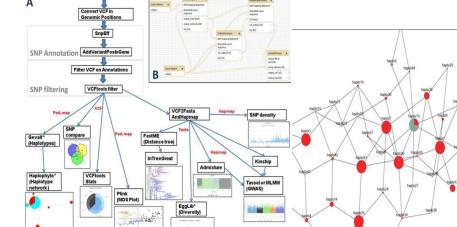


SNPs and Indels

GreenPhyl

Family Id	Family Name	Number of sequences	Status
GP000010	Cytochrome P450 superfamily	6942	green
GP000017	AP2/EREBP transcription factor family: ERF/DREB group (partial)	5142	green
GP000020	NAC transcription factor family	4574	green
GP000018	MAOS transcription factor family		
GP000019	Haem peroxidase superfamily		
GP000022	General substrate transporter superfamily		
GP000019	NPF, NRT1/PTR FAMILY		

Gene families



SNiPlay



<https://github.com/SouthGreenPlatform>

@green_bioinfo

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

18-19/03	Guide de survie à Linux - IRD
21/03	Initiation à l'utilisation du cluster CIRAD - CIRAD
22/03	Initiation à l'utilisation du cluster itrop - IRD
15-16/04	Initiation au gestionnaires de workflow SG & Gigwa - IRD
18-19/04	Guide du Jedi en Linux & bash - CIRAD
13-16/05	Python - IRD
17/05	Initiation aux analyses de données transcriptomiques - IRD
21/05	Utilisation avancée du cluster IRD - IRD
23-24/05	Initiation aux analyses de données métagénomiques - IRD
6/06	Manipulation de données et figures sous R - CIRAD
26-28/06	Assemblage et annotation de transcriptomes - IRD



2019 Training module

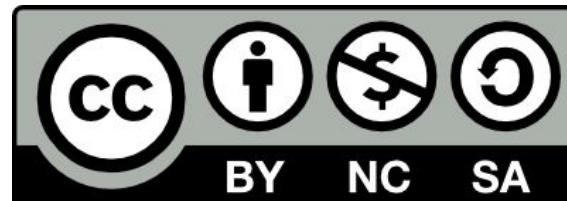
- Our trainings:
<https://southgreenplatform.github.io/trainings/>
- Topo & TPs : Initiation au cluster de calcul i-Trop
- Work Environment : Softwares to install



HPC cluster Initiation

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

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



ARCHITECTURE

- 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 resources availability

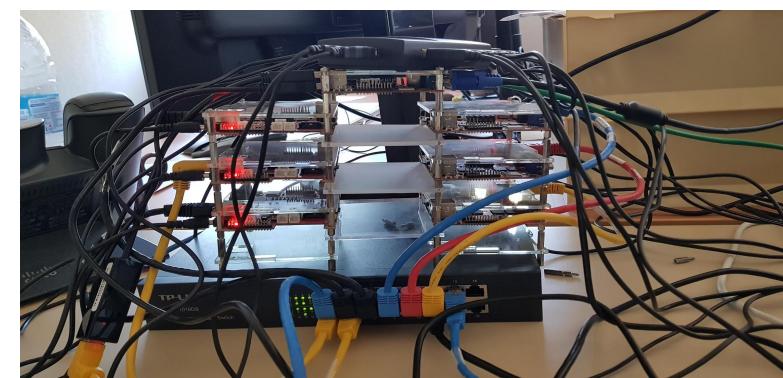
Qu'est ce qu'un cluster?

- 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 availability



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 availability



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

- **1 Master Node**



bioinfo-master.ird.fr

Role :

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

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

- **1 Master Node**



bioinfo-master.ird.fr

Role :

- Launch and prioritize jobs on computing 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`

- **1 Master Node**



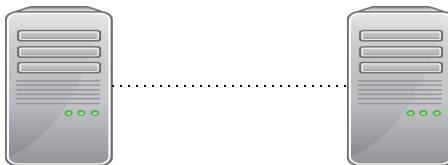
bioinfo-master.ird.fr

Role :

- Launch and prioritize jobs on computing 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`

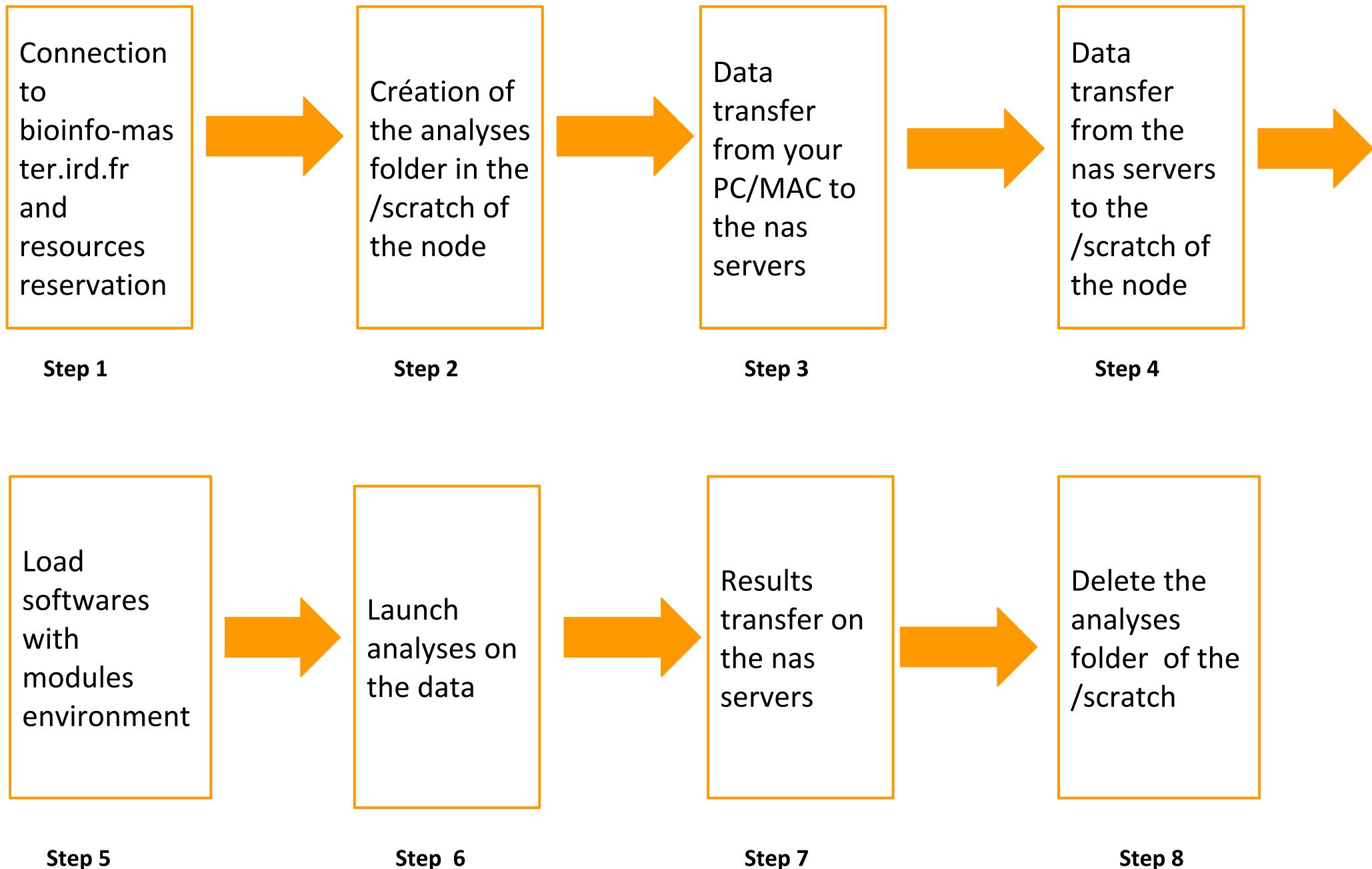


Interactif node (node6)

- Accessible from the Internet: bioinfo-inter.ird.fr
- Connection :

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

Analyses steps of the cluster



Analyses steps of the cluster

Connection
to
bioinfo-mas
ter.ird.fr
and
resources
reservation



Step 1
qrsh/qlogin
or qsub



Practice

Step 1: Connection, qhost

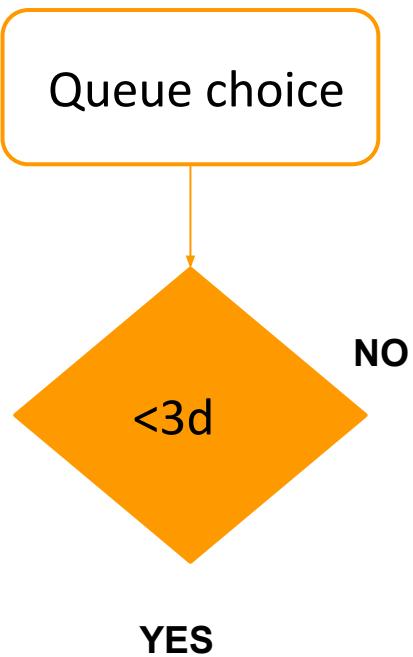
1

Go to the Practice 1 of github

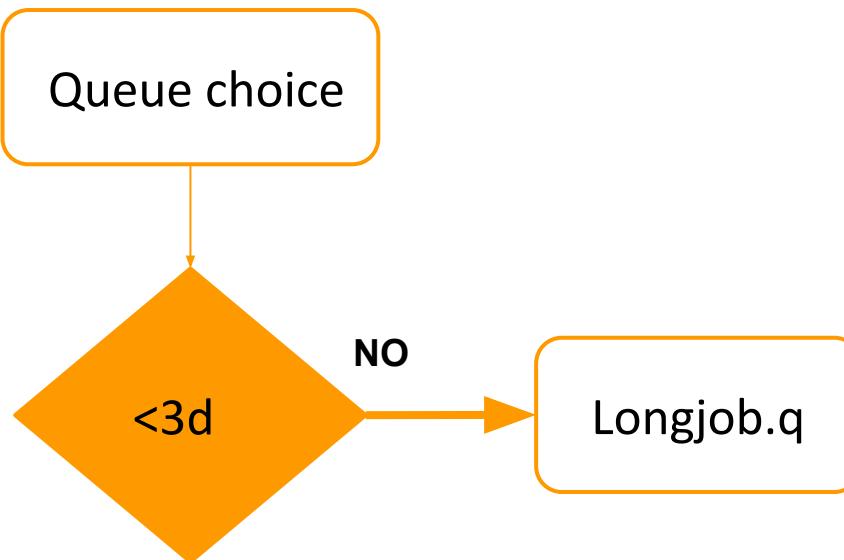
Queues

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

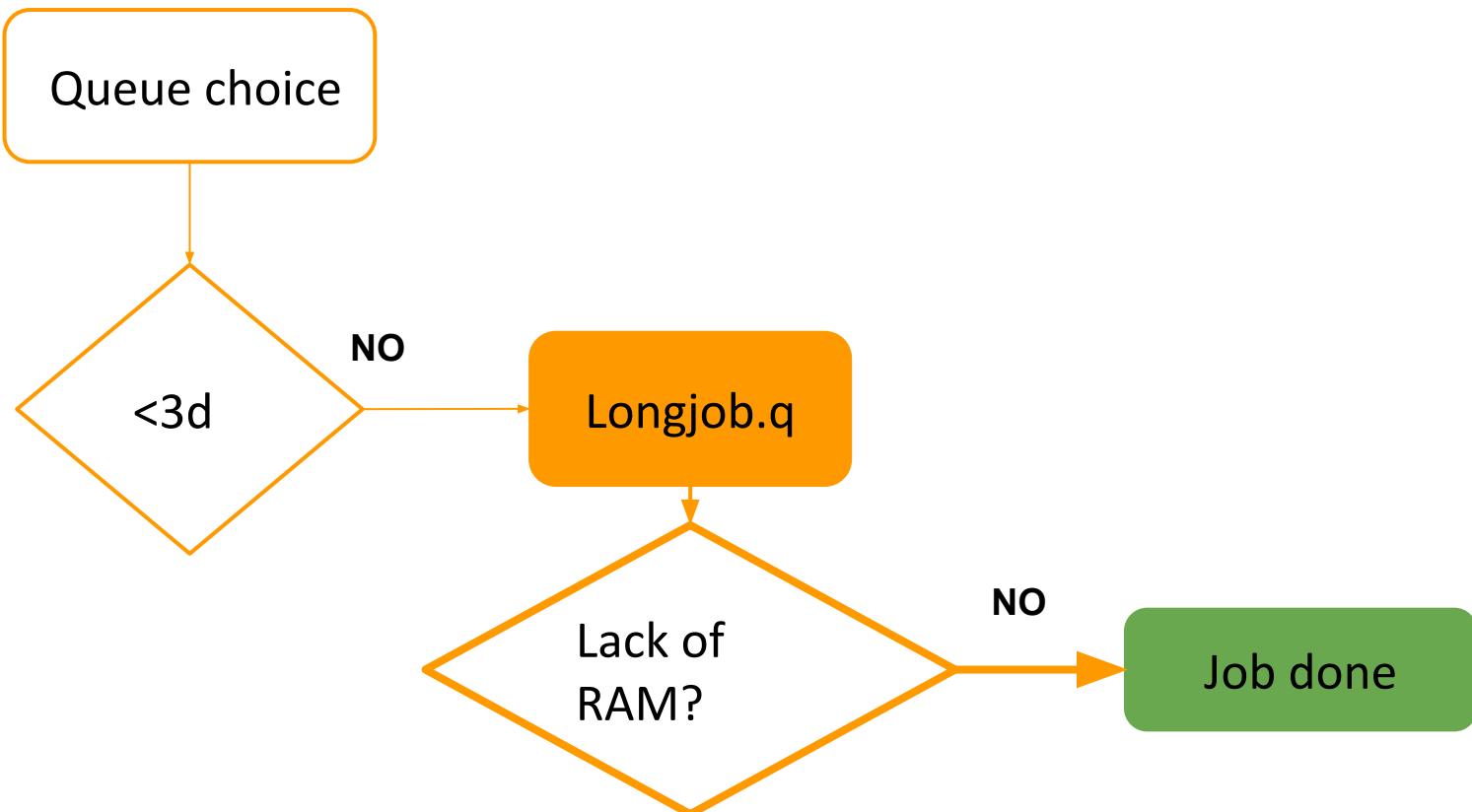
How do I choose the queue?



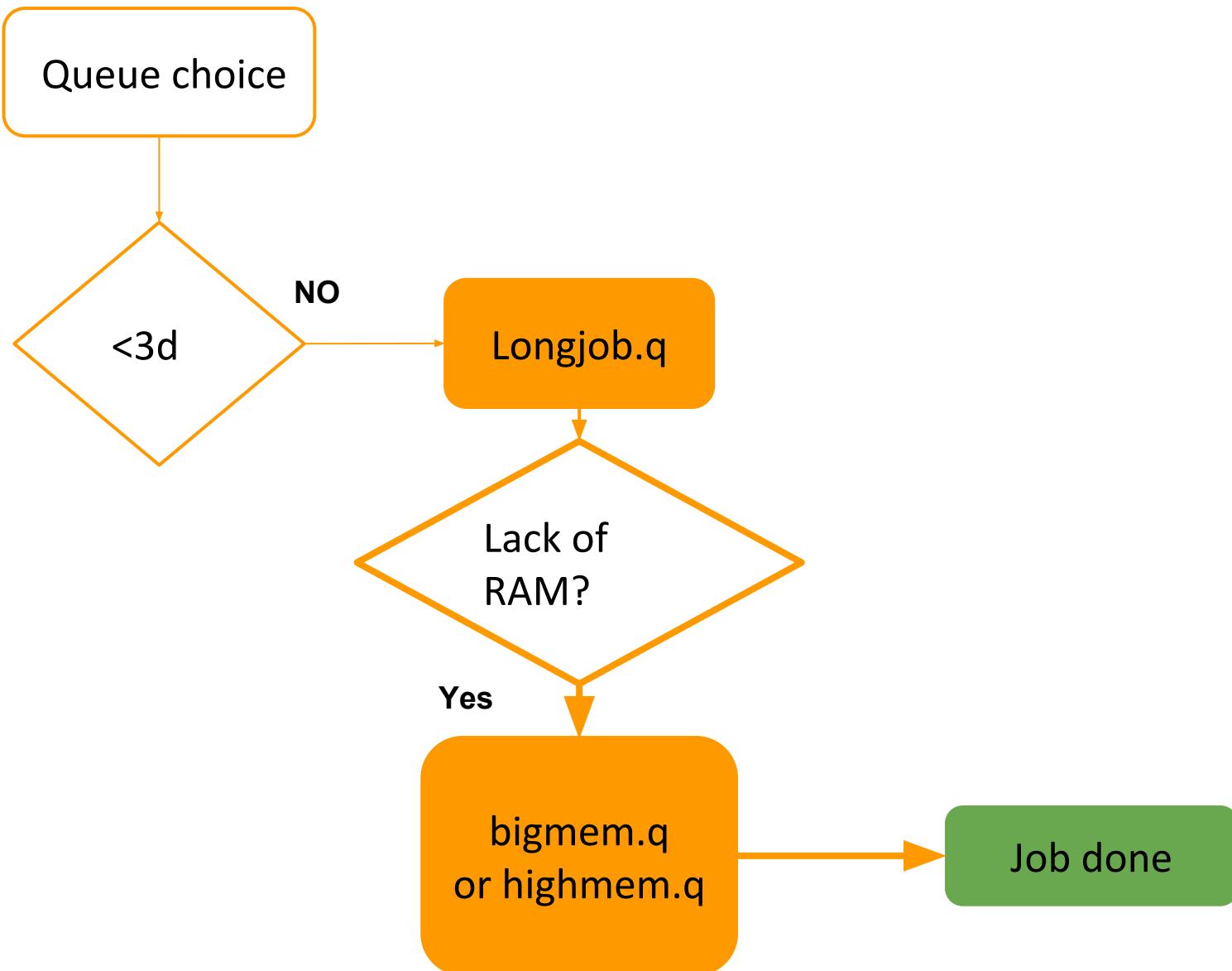
How do I choose the queue?



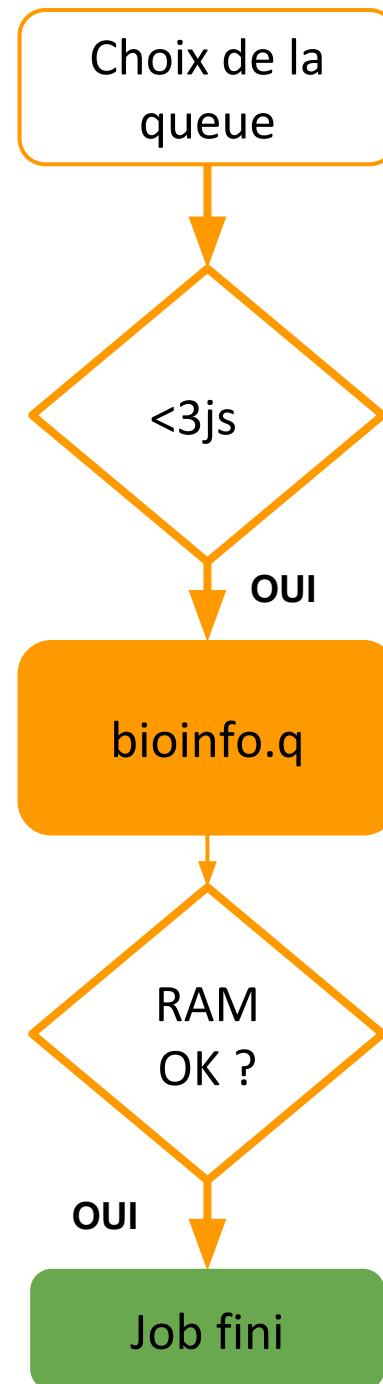
How do I choose the queue?



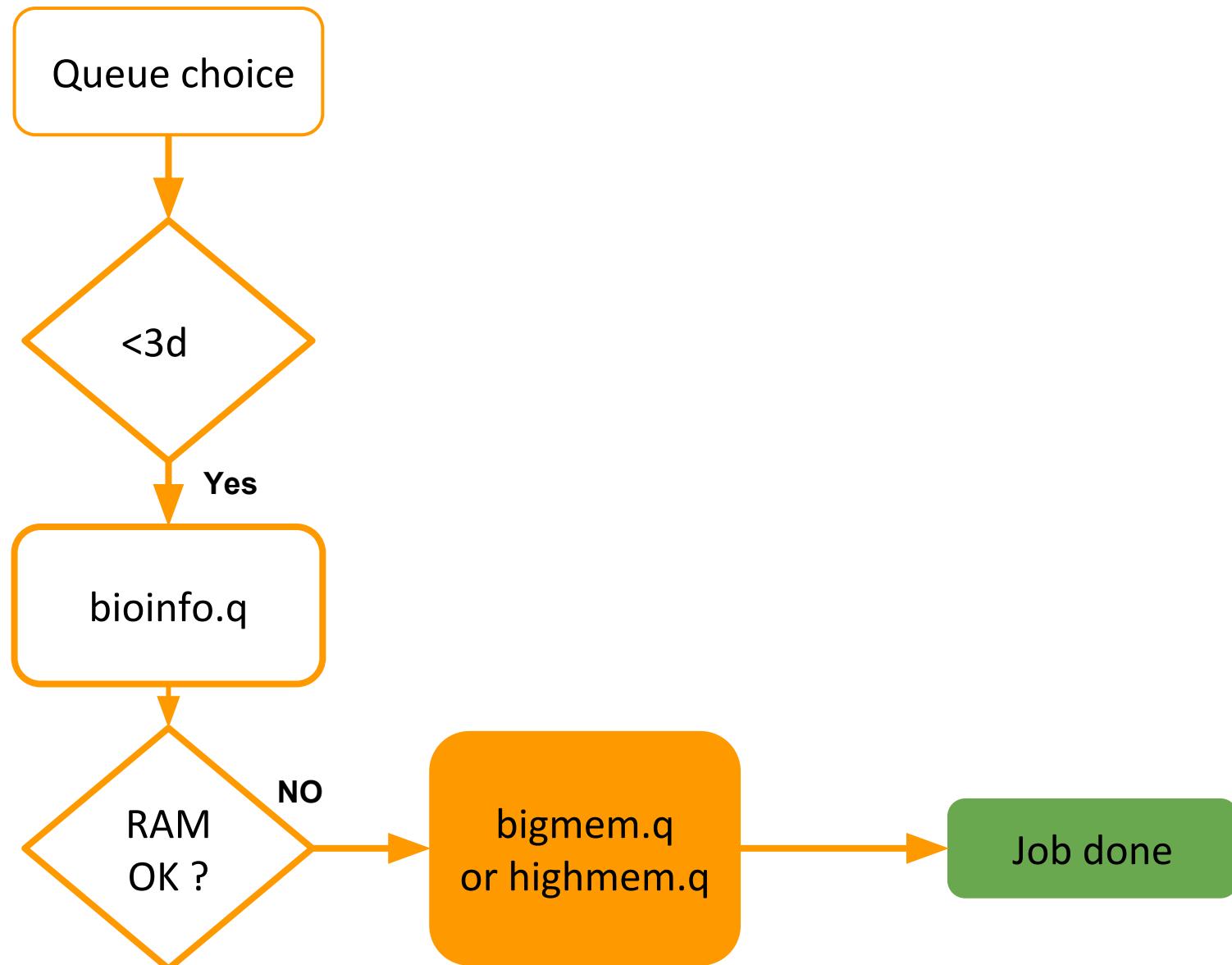
How do I choose the queue?



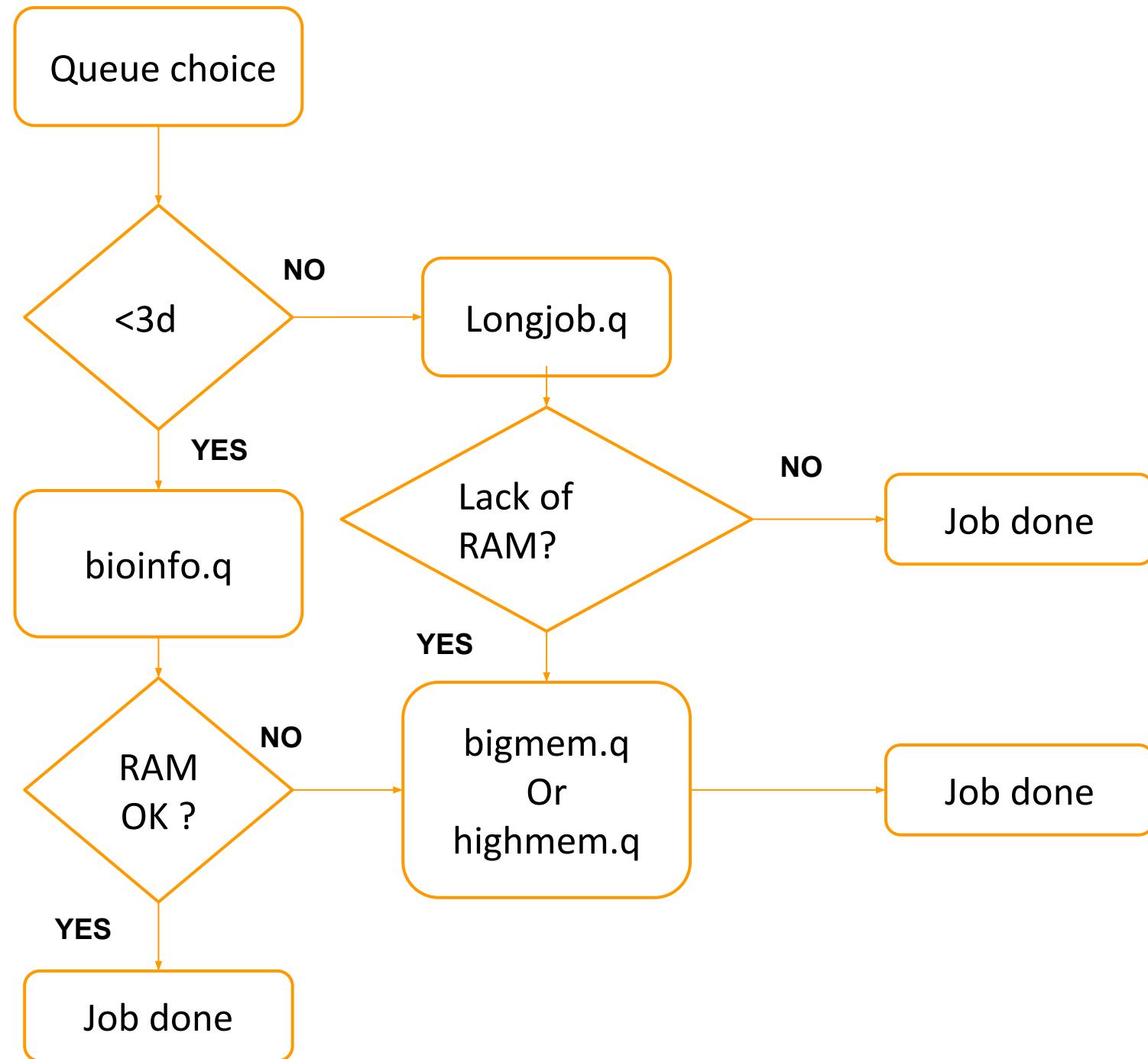
Quelle queue choisir?



How do I choose the queue?



How do I choose the queue?

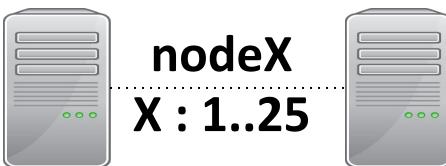


- **1 Master node**



bioinfo-master.ird.fr

- **25 computing nodes**



Role :

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

Role :

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

- **1 Master node**

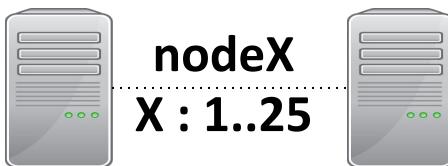


bioinfo-master.ird.fr

Role :

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

- **25 computing nodes**



Role :

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

- **3 NAS servers**



bioinfo-nas.ird.fr

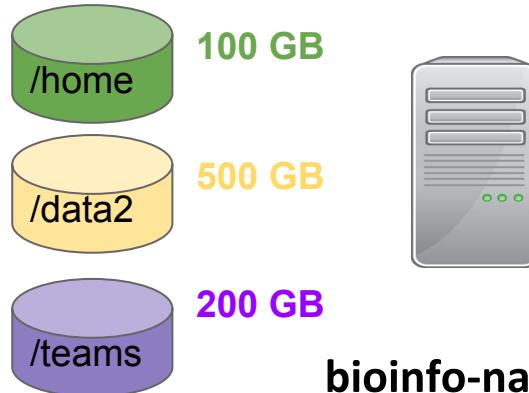
bioinfo-nas2.ird.fr

bioinfo-nas3.ird.fr

Role :

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

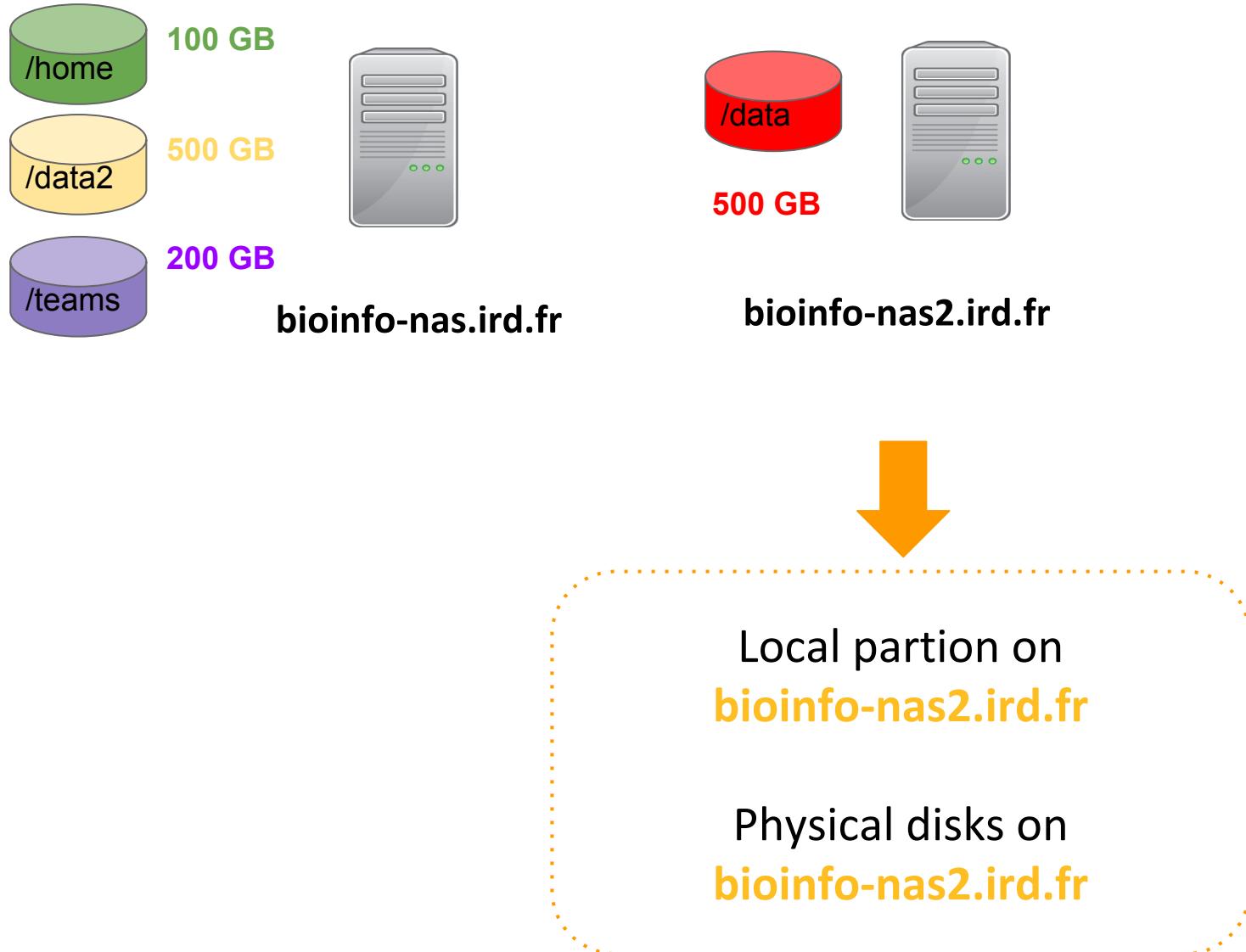
cluster i-Trop disk partitions



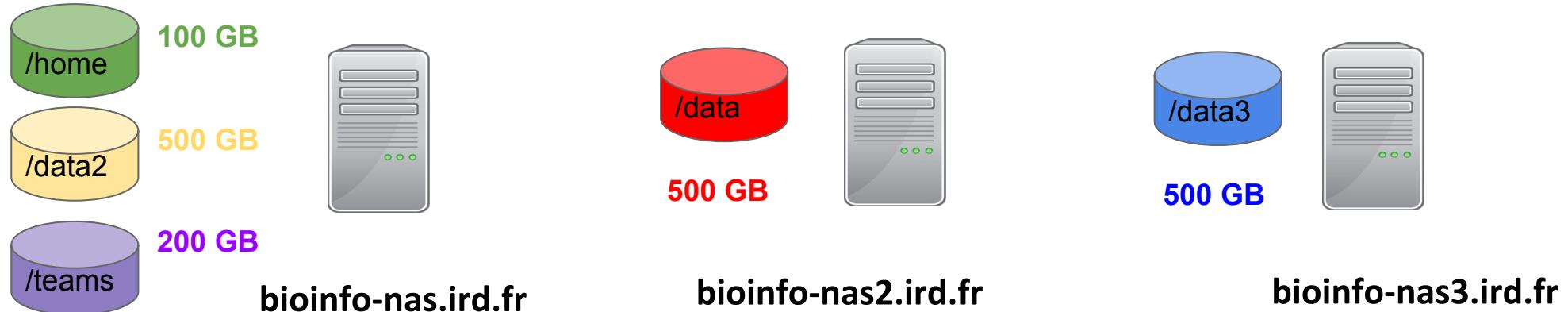
Local partition on
bioinfo-nas.ird.fr

Physical disks on
bioinfo-nas.ird.fr

cluster i-Trop disk partitions



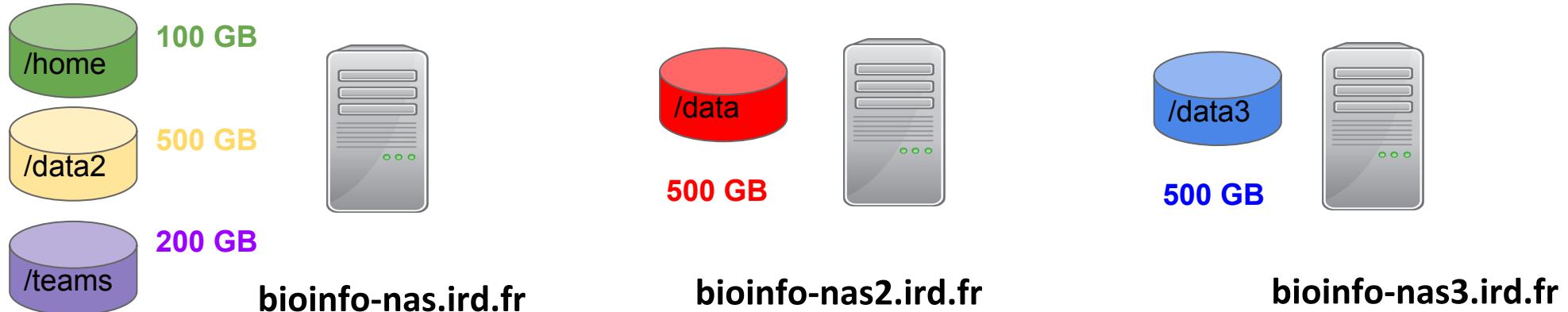
cluster i-Trop disk partitions



Local partition on
bioinfo-nas3.ird.fr

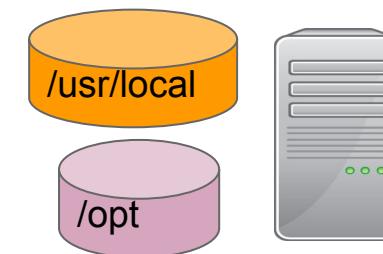
Physical disks on
bioinfo-nas3.ird.fr

cluster i-Trop disk partitions



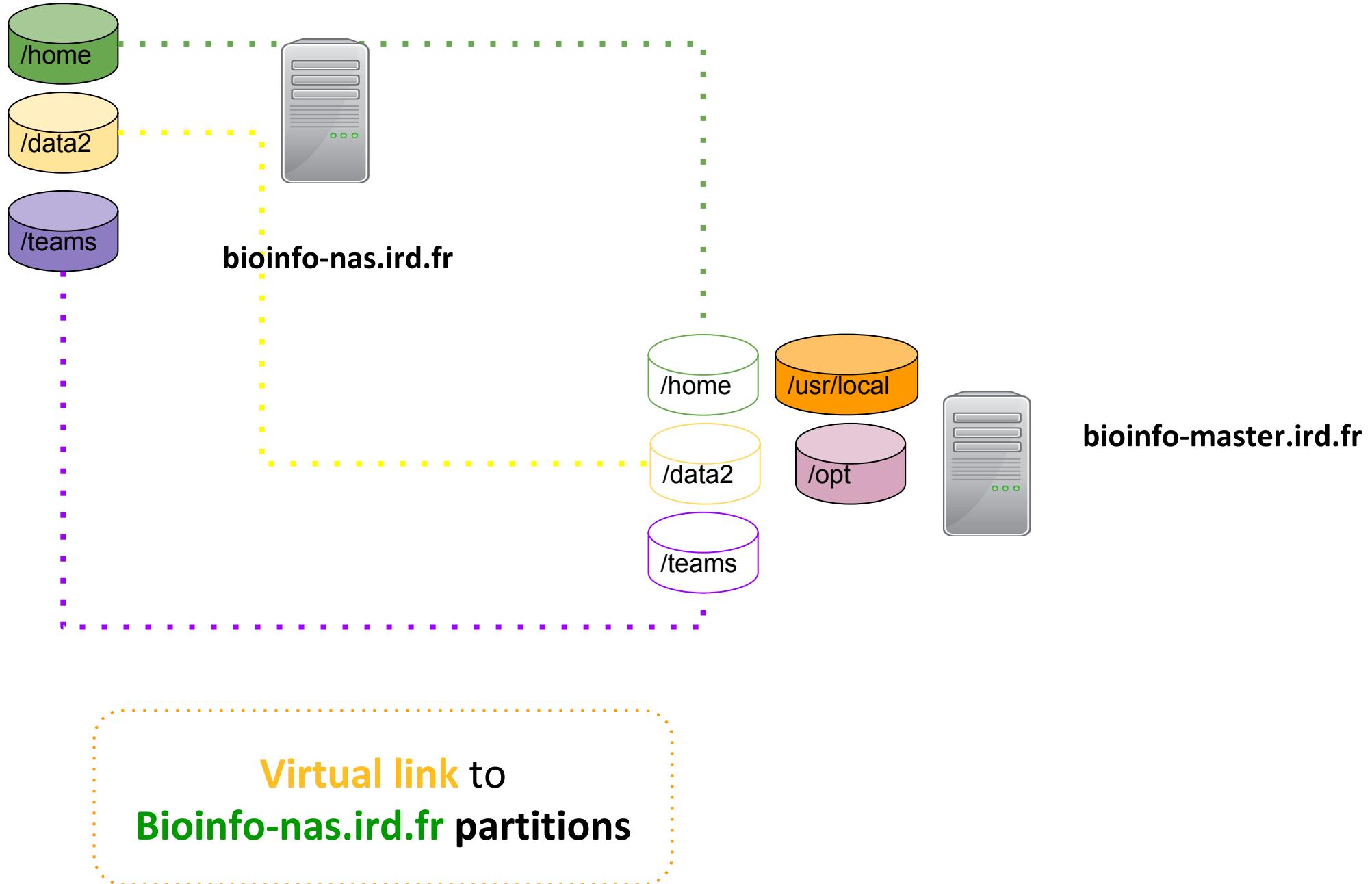
Local partitions on
bioinfo-master.ird.fr

Physical hard disk on
bioinfo-master.ird.fr

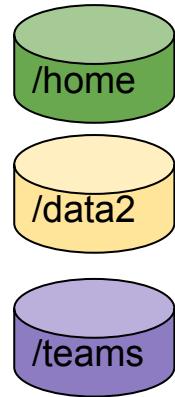


bioinfo-master.ird.fr

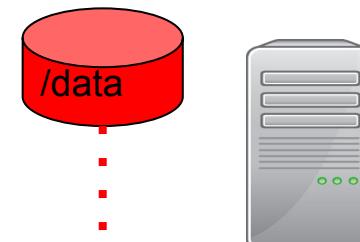
cluster i-Trop disk partitions



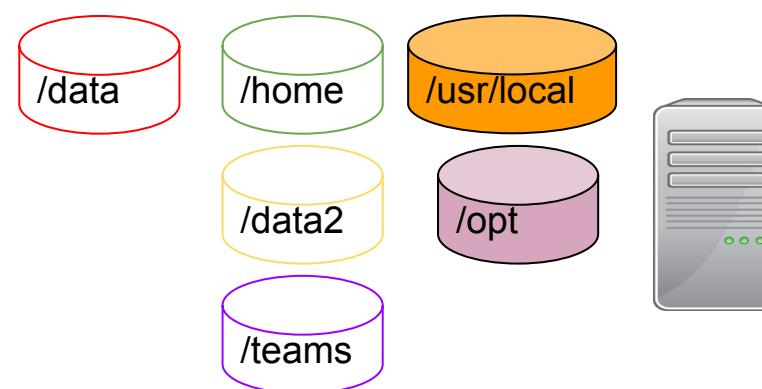
cluster i-Trop disk partitions



bioinfo-nas.ird.fr



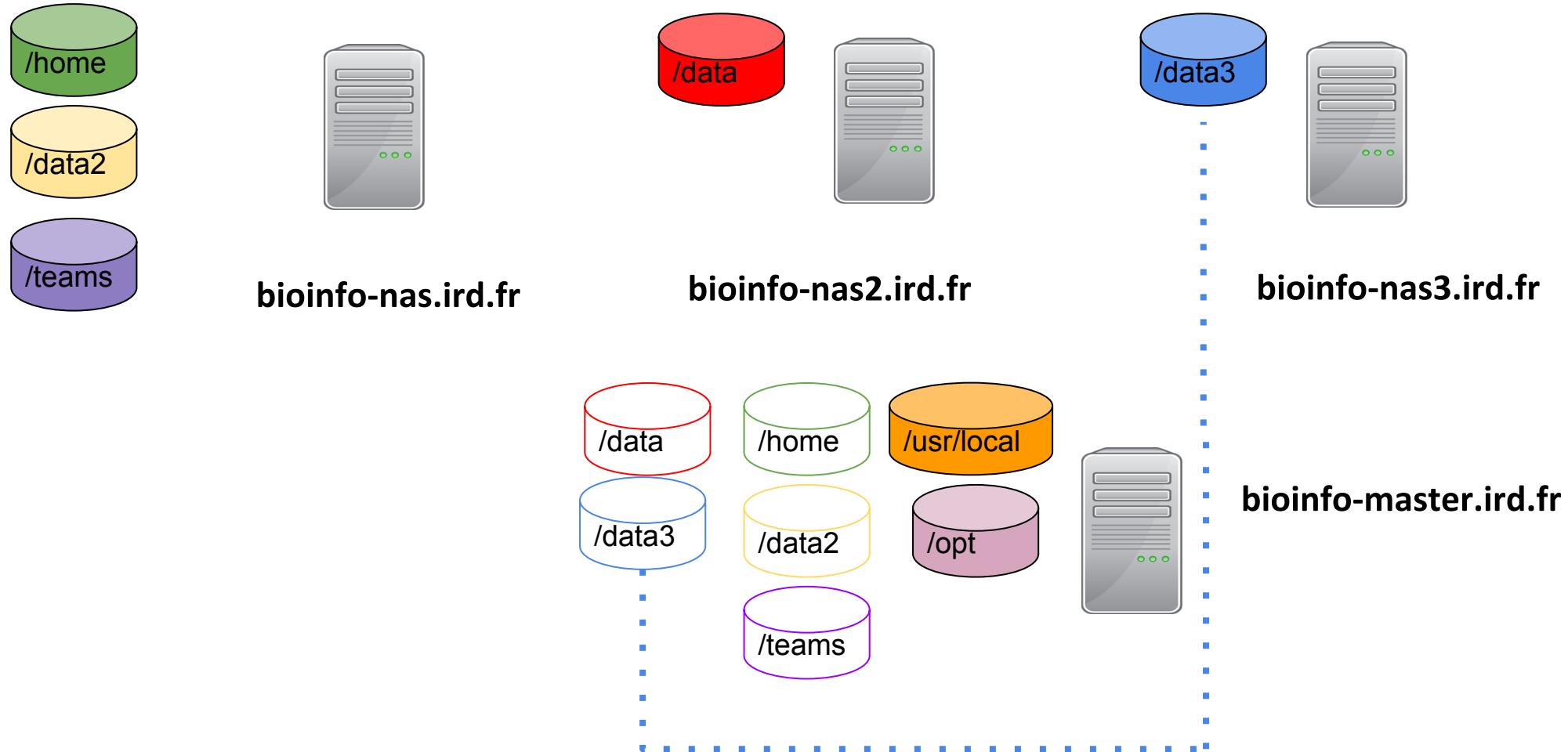
bioinfo-nas2.ird.fr



bioinfo-master.ird.fr

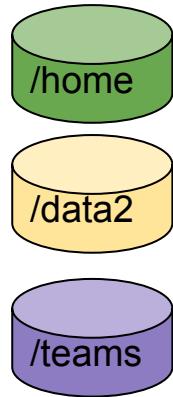
**Virtual link to the
Bioinfo-nas2.ird.fr partitions**

cluster i-Trop disk partitions

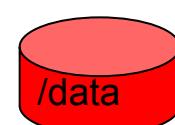


**Virtual link to the
Bioinfo-nas3.ird.fr partition**

cluster i-Trop disk partitions

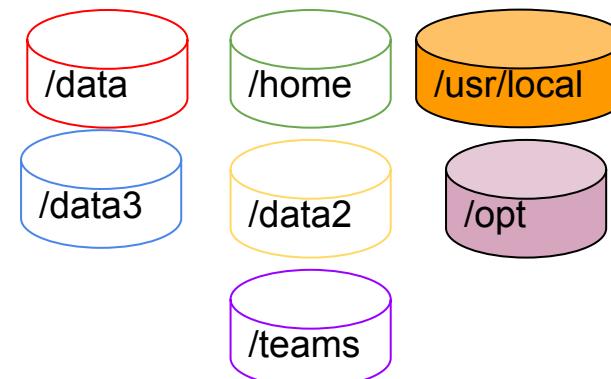


bioinfo-nas.ird.fr



bioinfo-nas2.ird.fr

bioinfo-nas3.ird.fr



bioinfo-master.ird.fr



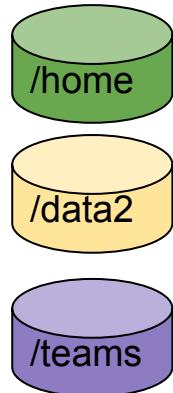
- Local partition on nodes :
temporary space
- Physical hard drives on nodes



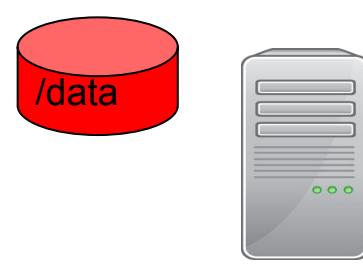
25 nodes



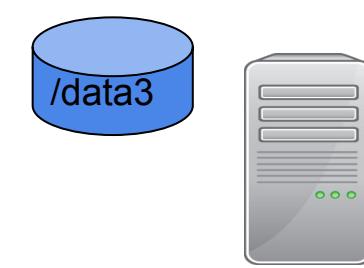
Partitions disques sur le cluster i-Trop



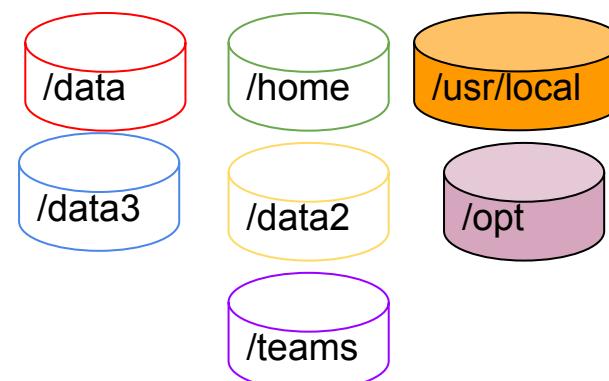
bioinfo-nas.ird.fr



bioinfo-nas2.ird.fr



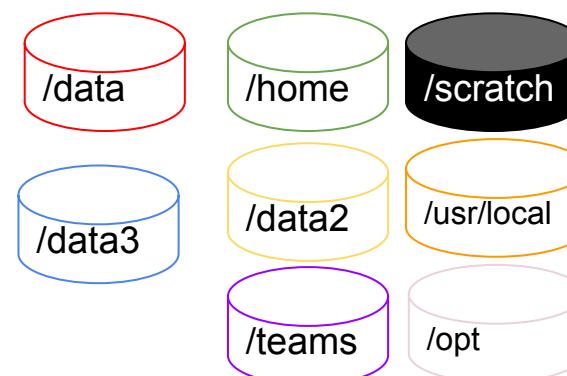
bioinfo-nas3.ird.fr



bioinfo-master.ird.fr



Virtual links to the other servers partitions



25 nodes



Analyses steps of the cluster

Connection
to
bioinfo-mas-
ter.ird.fr
and
resources
reservation



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

Step 1

Step 2
mkdir



Practice

Step 2:qrsh, partition

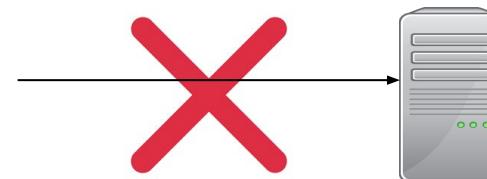
2

Go to the [Practice2](#) of the github

Data transfer on i-Trop cluster



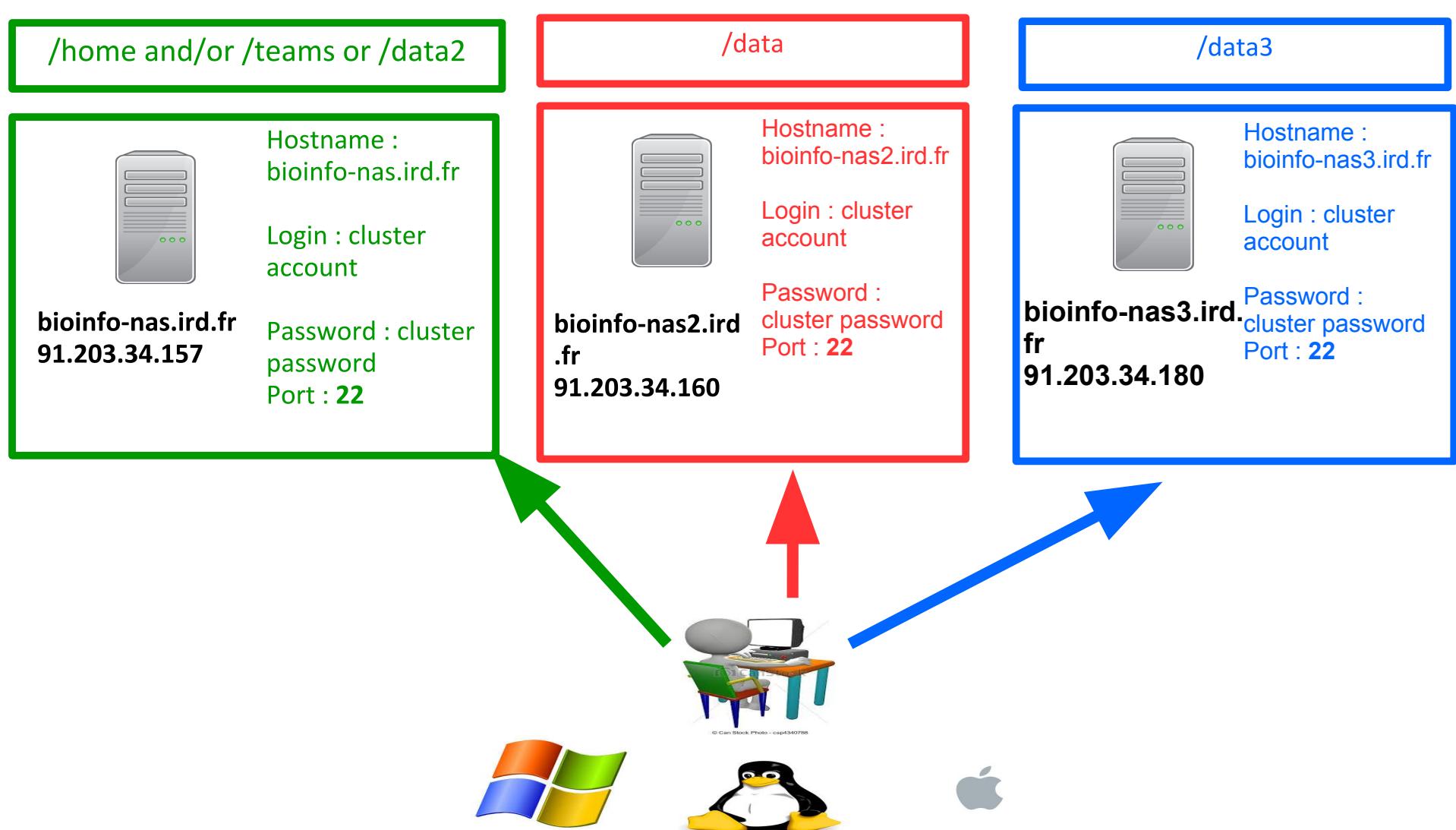
PC/MAC



direct transfer
via filezilla
forbidden

bioinfo-master.ird.fr
91.203.34.148

Data transfer on i-Trop cluster



Analyses steps of the cluster

Connection
to
bioinfo-mas
ter.ird.fr
and
resources
reservation



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



Data
transfer
from your
PC/MAC to
the nas
servers

Step 1

Step 2

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 source destination
```

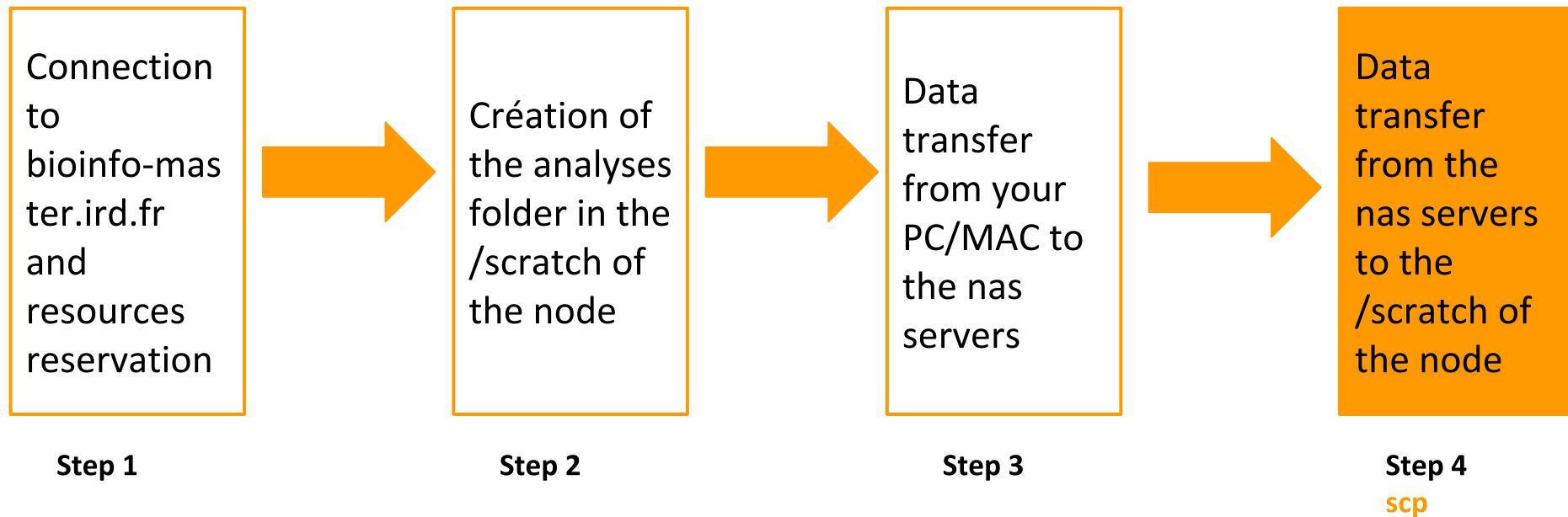
- Syntax if the source is remote :

```
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
```

Analyses steps of the cluster





Practice

Step 4: scp to nodes

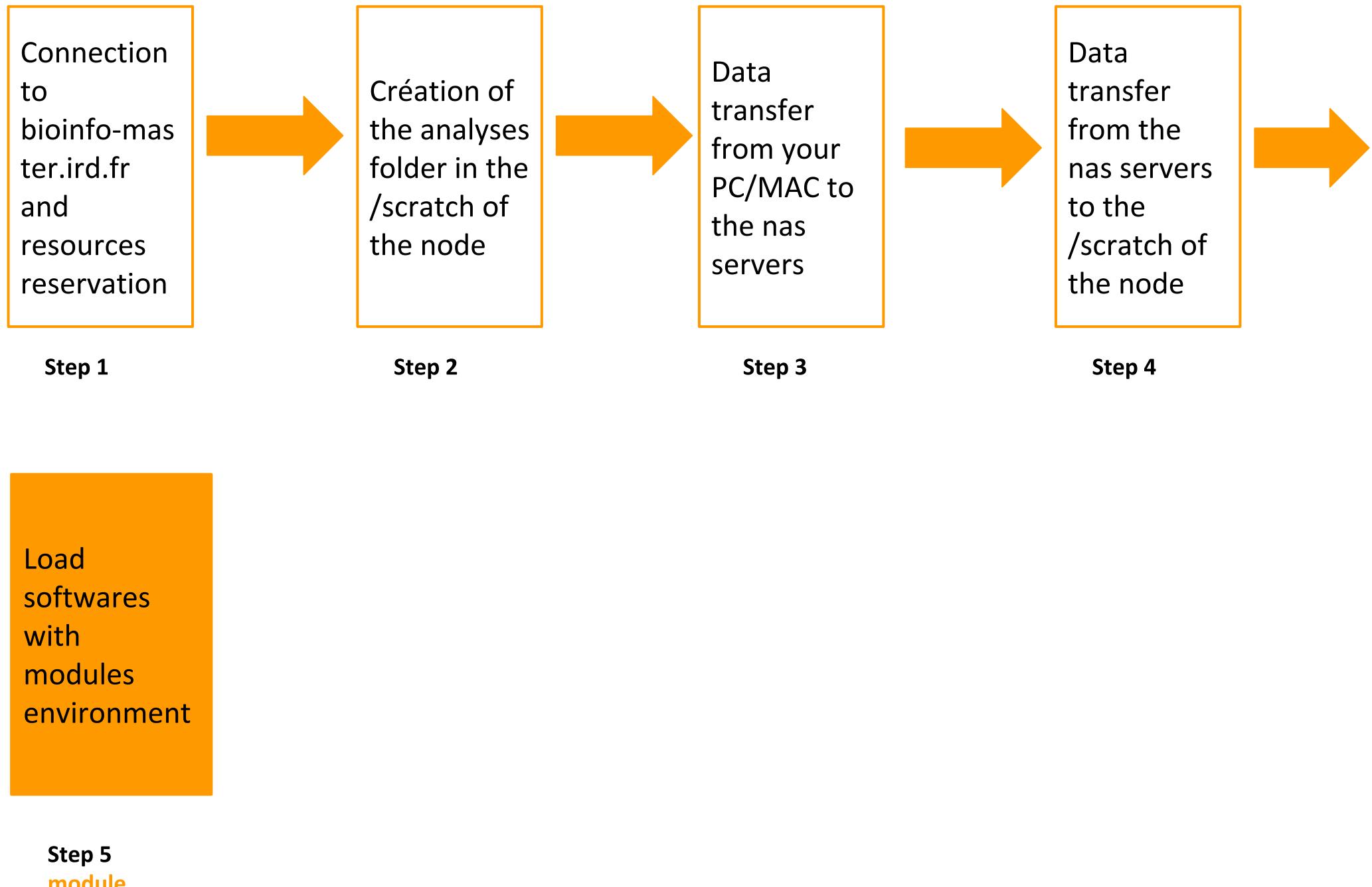
4

Go to the Practice4 of the github

- 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`

Analyses steps of the cluster





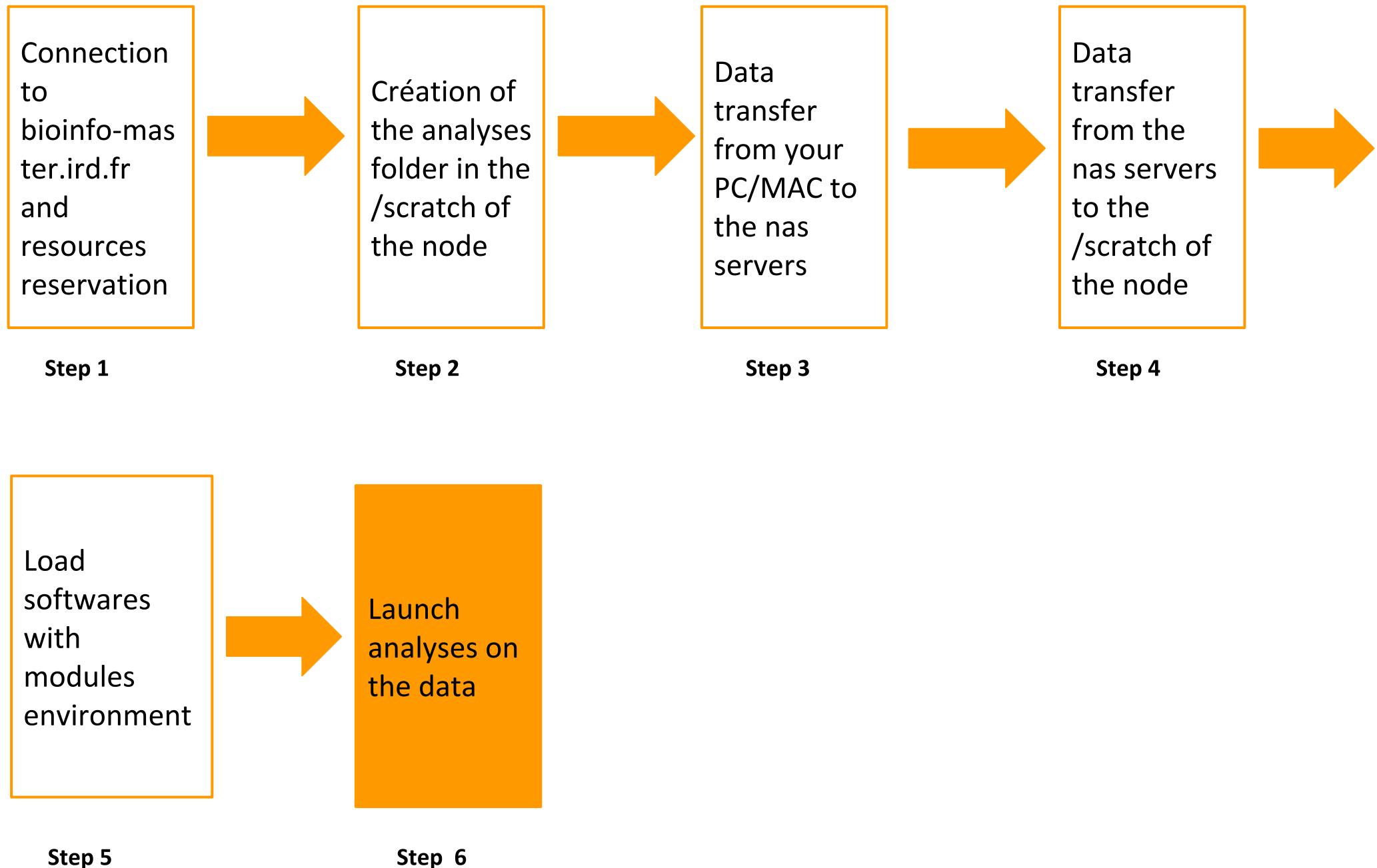
Practice

Step 5: module environment

5

Go to the [Practice5](#) of the github

Analyses steps of the cluster



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



Practice

Step6: launch the analysis

6

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

Transfer your results to the nas servers

- Copy between 2 remote servers :

```
scp source destination
```

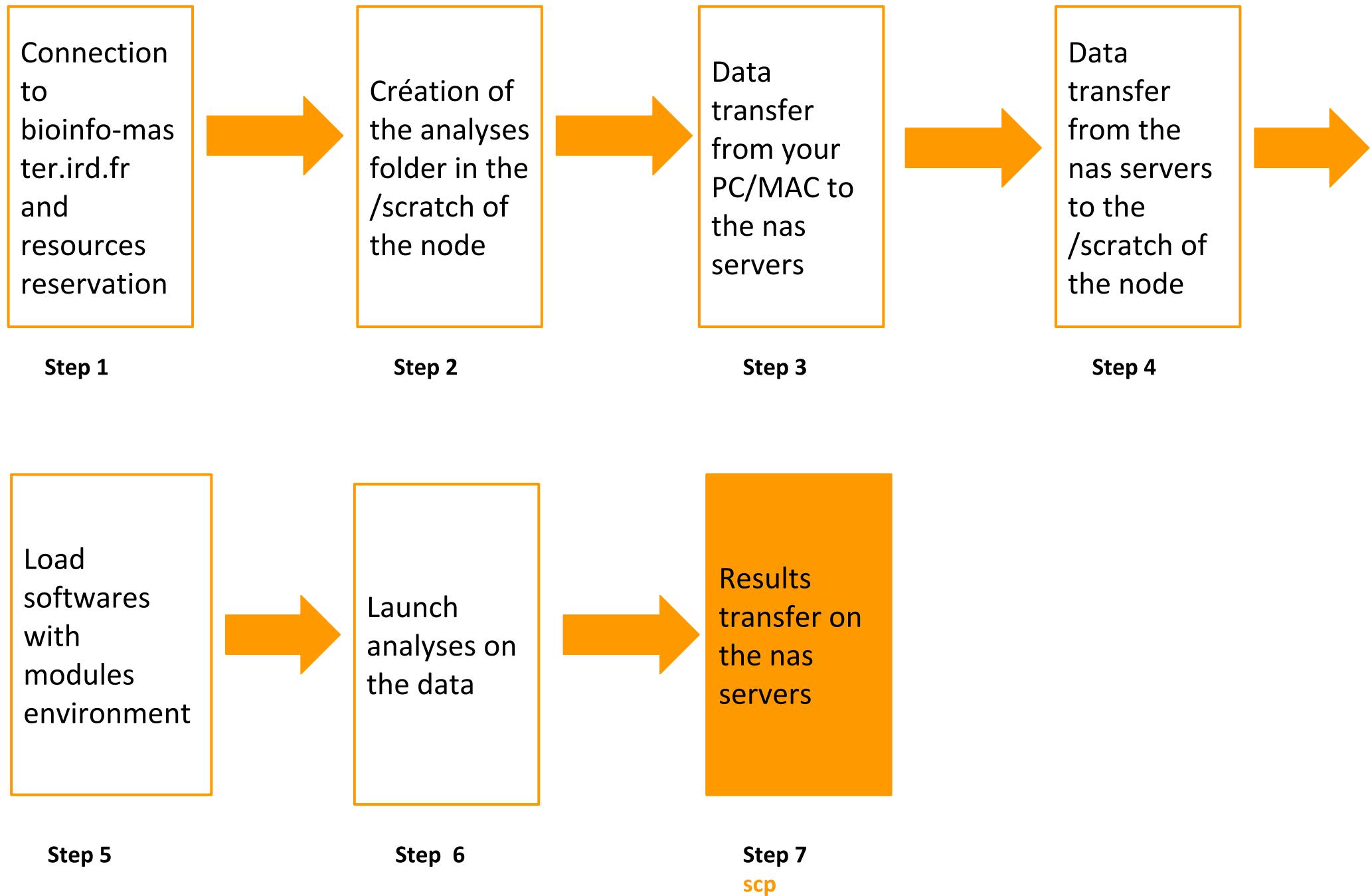
- Syntax if the source is remote :

```
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
```

Analyses steps of the cluster





Practice

Step 7: Retrieve the results

7

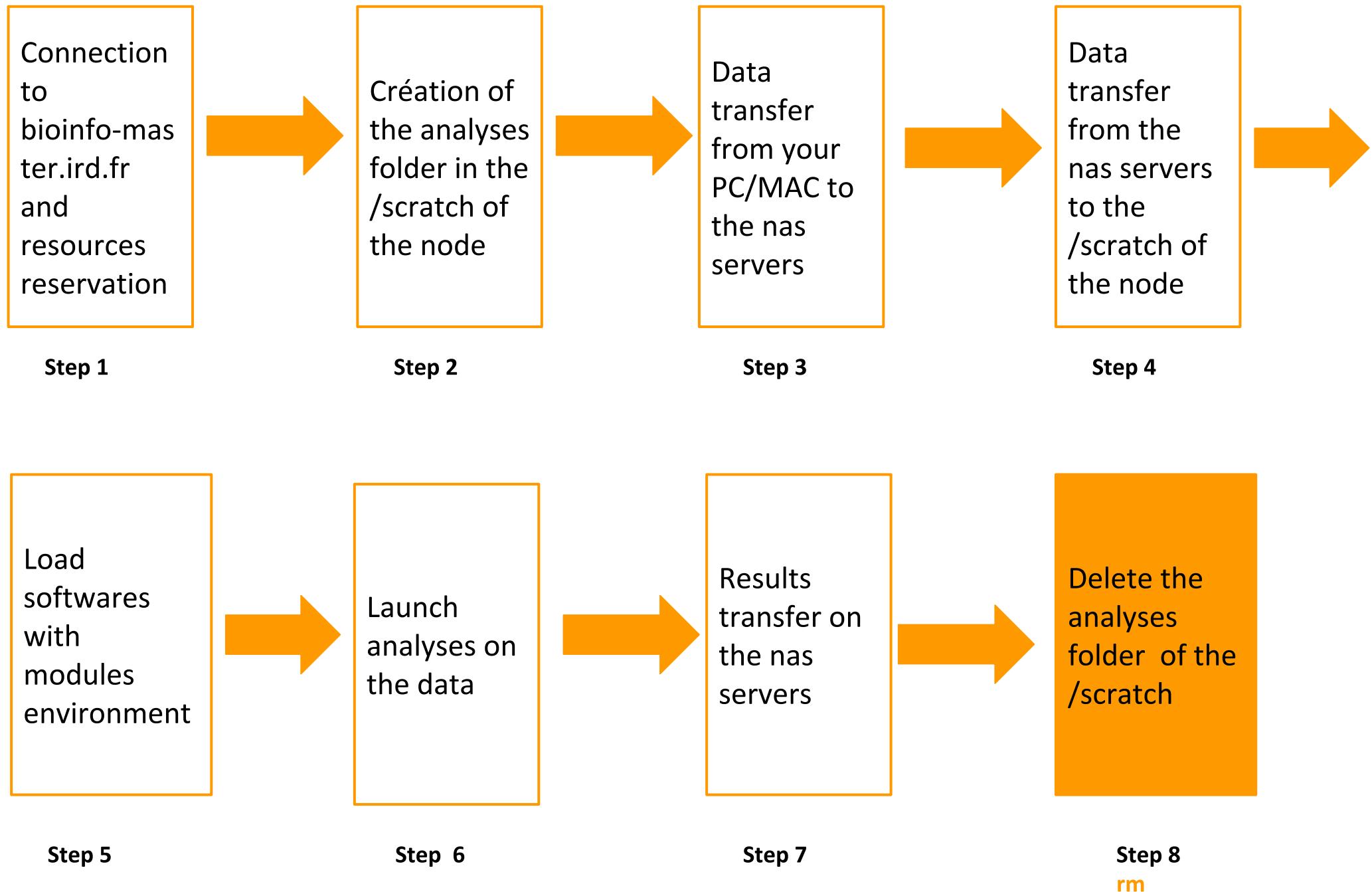
Go to the [Practice7](#) of the github

Delete results from scratchs

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

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

Analyses steps of the cluster





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
```

LAUNCH A JOB

Advantages

- Scheduler choose resources automatically
- Possibility to configure this choice
 - Jobs launch in background
 - possibility to turn off your PC/MAC
 - automatic results retrieving

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

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
#####
```

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>”

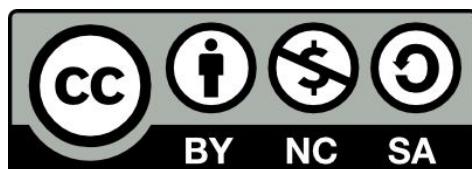
- Include a budget for bioinformatics resources in your answer to projects funding
- A need in hard drives, renewal machinesetc...
- Available quotations
- Contact bioinfo@ird.fr : help, needs definition, quotations...

Trainers

- **Christine Tranchant-Dubreuil**
- **Sebastien Ravel**
- **Alexis Dereeper**
- **Ndomassi Tando**
- **François Sabot**
- **Bruno Granouillac**
- **Valérie Noël**
- **Bertrand Pitollat**



Thank you for your attention !



Le matériel pédagogique utilisé pour ces enseignements est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions (BY-NC-SA) 4.0 International:

<http://creativecommons.org/licenses/by-nc-sa/4.0/>