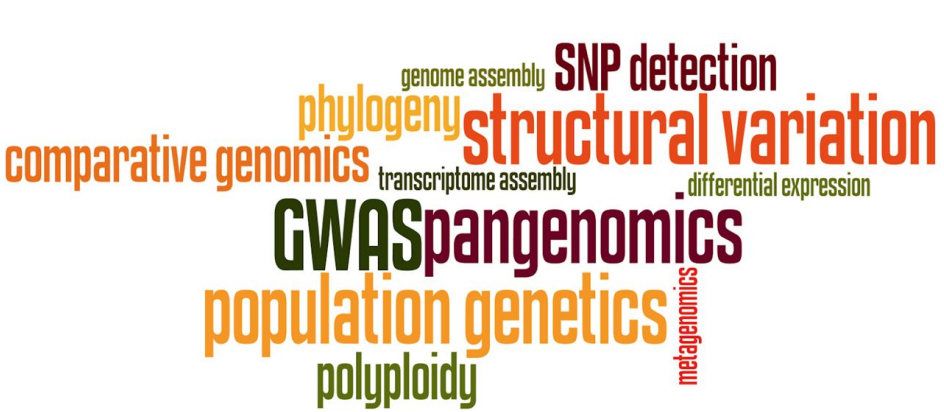
South Green bioinformatics platform

2019 Training modules

SouthGreen bioinformatics platform

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




Rice


Banana


Palm


Sorghum


Coffee


Cassava


Magnaporthe

www.southgreen.fr

SouthGreen bioinformatics platform

Charte

diade

Larmande Pierre
Sabot François
Tando Ndomassi
Tranchant-Dubreuil Christine

IPME

Comte Aurore
Dereeper Alexis

diade IPME

Orjuela-Bouniol Julie

agap

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

Bioversity International

Rouard Mathieu
Guignon Valentin
Catherine Breton

UMR BGPI

Mahé Frédéric
Ravel Sébastien

intertryp

Sempere Guilhem

Bioversity International

cirad

IRD Institut de Recherche pour le Développement FRANCE

INRA SCIENCE & IMPACT

SouthGreen bioinformatics platform

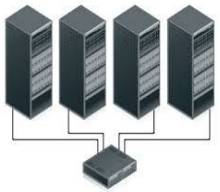
Workflow manager

TOGGLE
Toolbox for generic NGS analyses

SNAKEMAKE

Galaxy

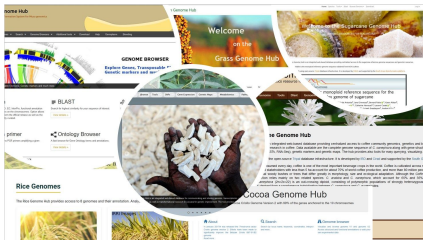
HPC and trainings….



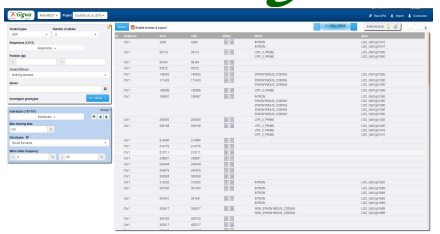
IRD Institut de Recherche pour le Développement FRANCE

cirad

Genome Hubs & Information System

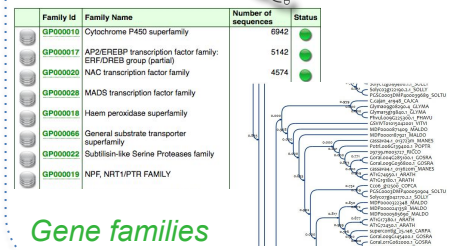


Gigwa



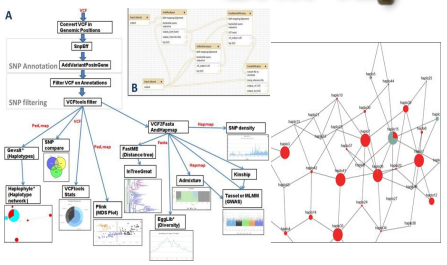
SNPs and Indels

GreenPhyl



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

SouthGreen bioinformatics platform

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

ABiMS agap ASTRE UMR BGPI Bioversity International BOREA diade IPME intertryp MIVEGEC UMR QualiSud

2019 Training module

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

HPC cluster Initiation

www.southgreen.fr

https://southgreenplatform.github.io/trainings

Objectives

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

Queries/incidents

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

- Incidents: contact bioinfo@ird.fr

ARCHITECTURE

A Cluster?

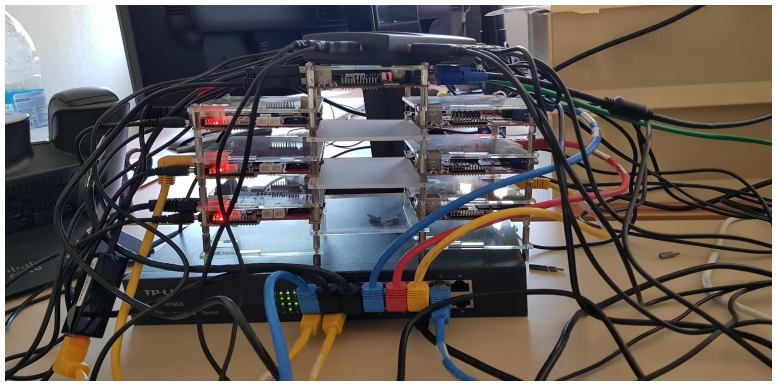
- A logical unit composed of several servers
- A powerful unique machine
- Allow to obtain high computing performance
- A bigger capacity storage
- More reliable
- A better ressources availabilty

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**

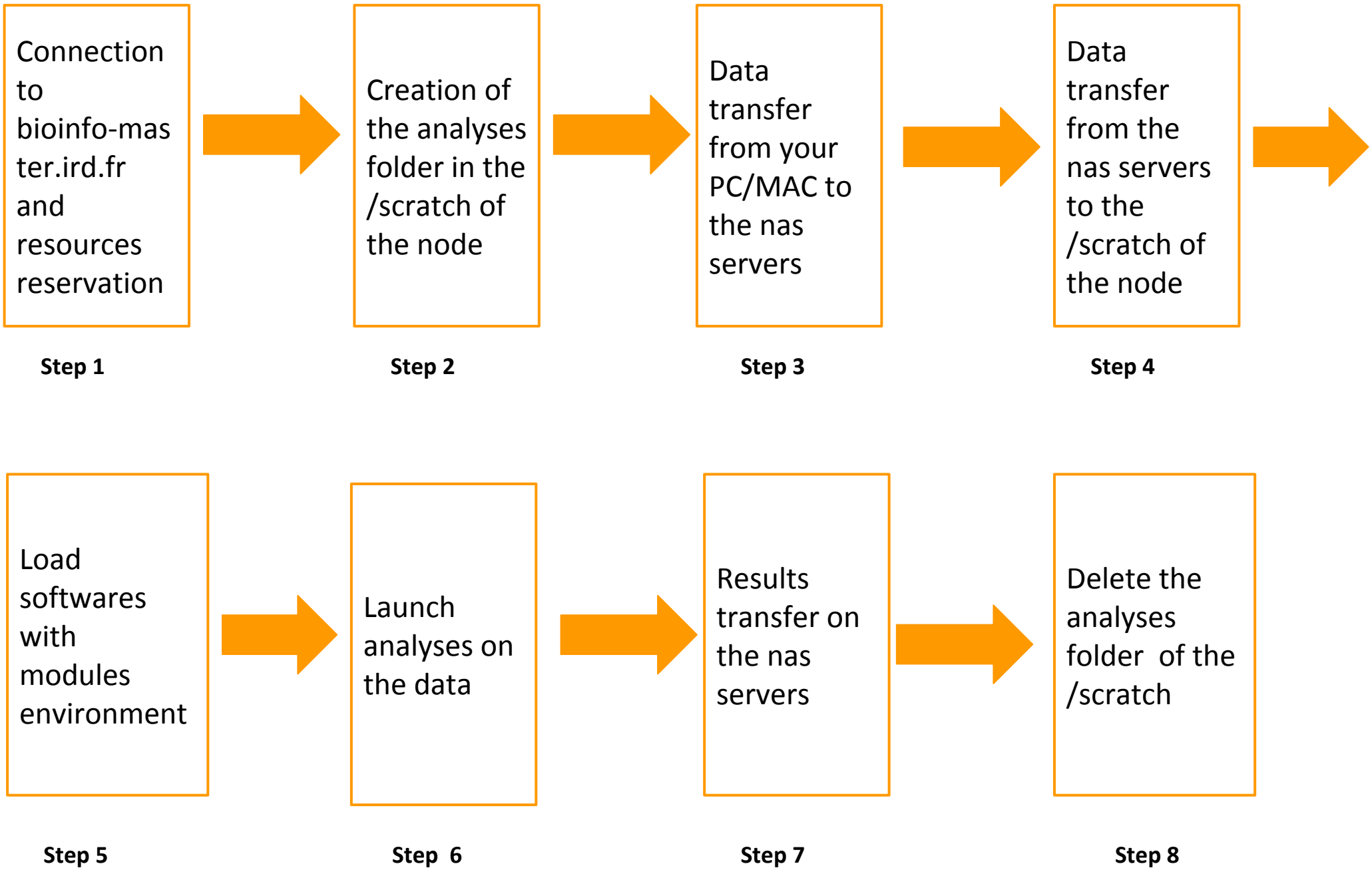
nodeX
X : 1..25

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`

Analyses steps of the cluster



Analyses steps of the cluster

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

Step 1
qrsh/qlogin or qsub

Practice

Step 1: Connection, qhost

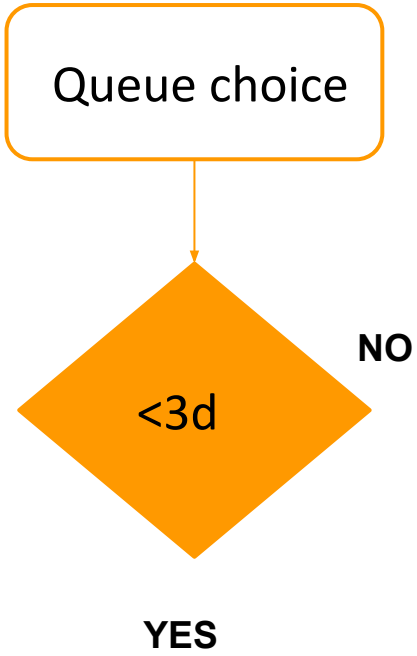
1

Go to the Practice 1 *of github*

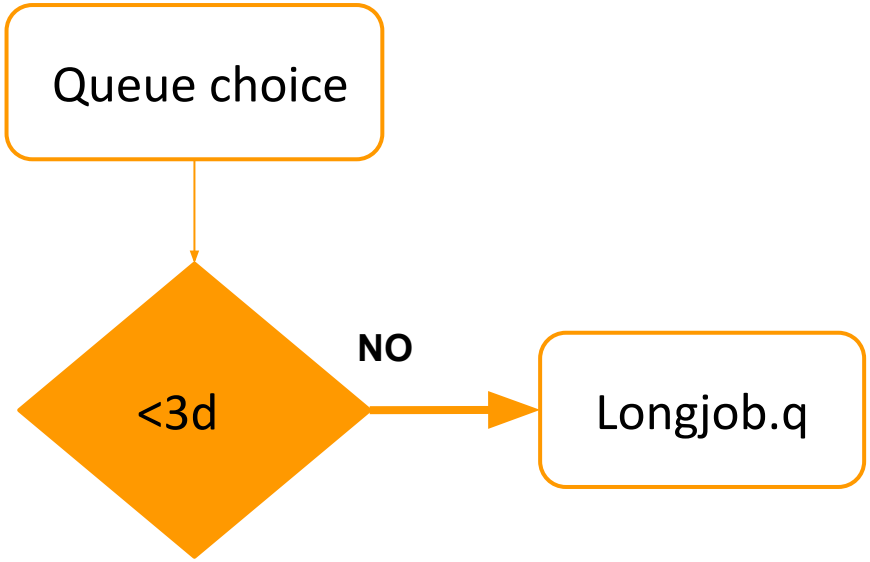
Queues

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

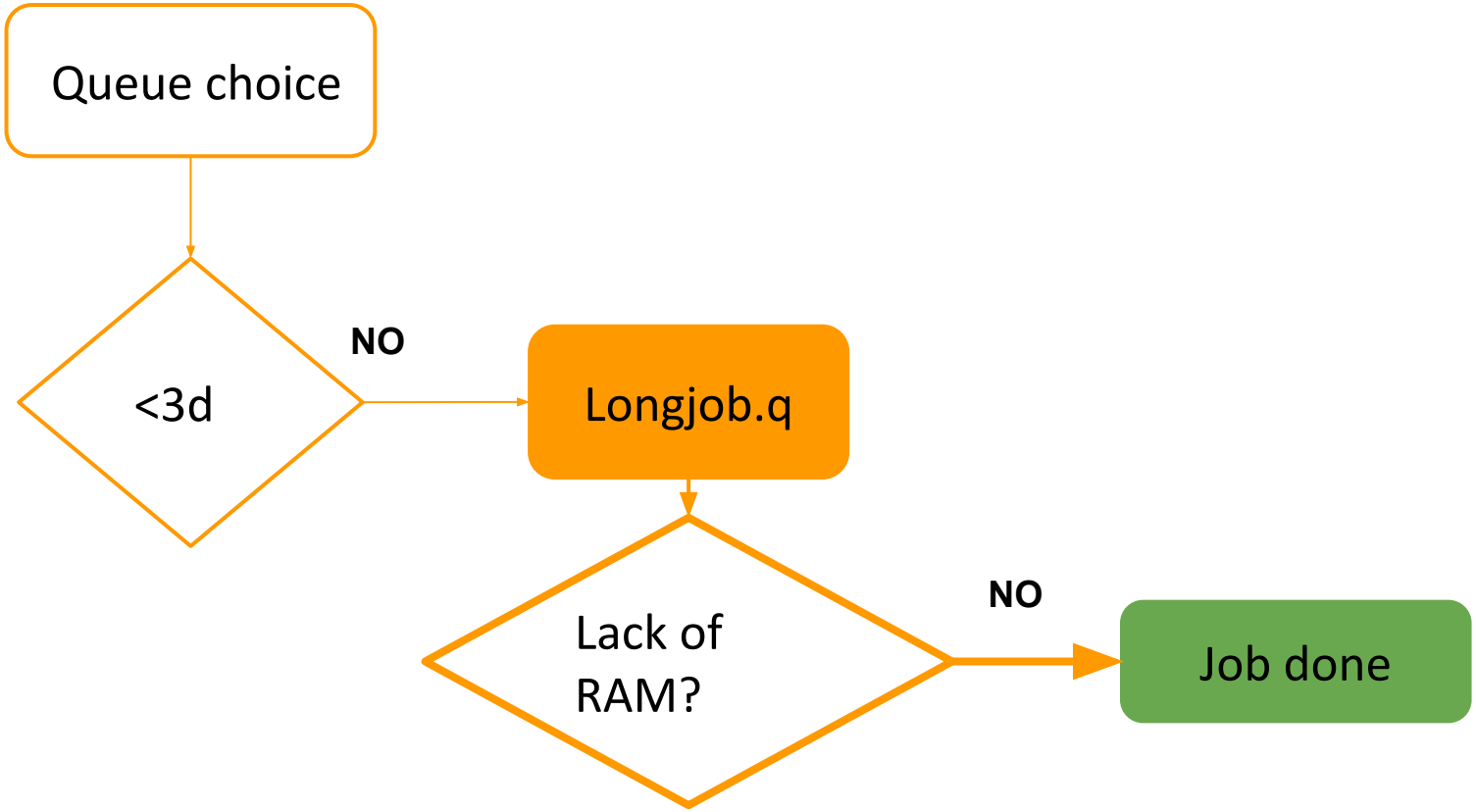
How do I choose the queue?



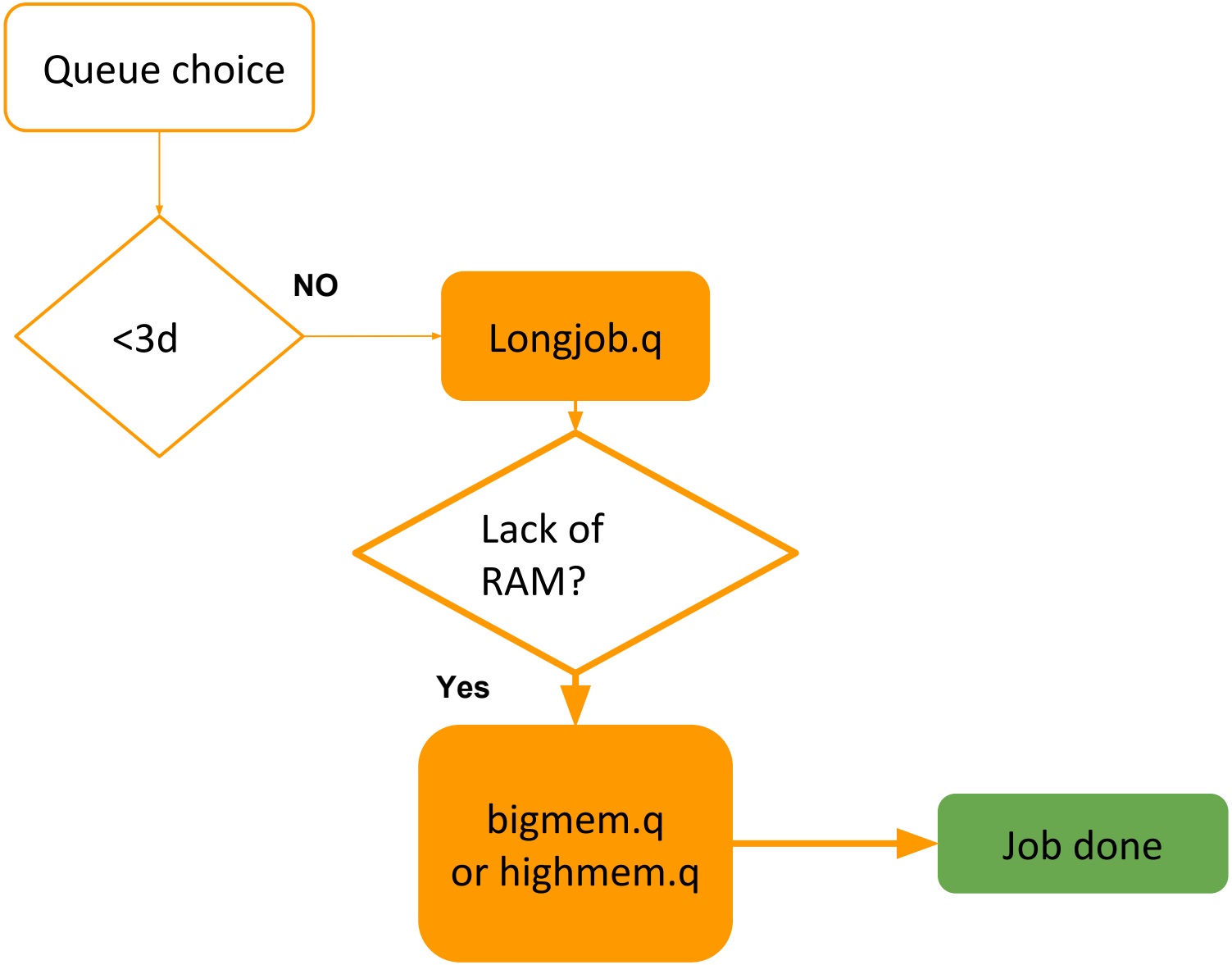
How do I choose the queue?



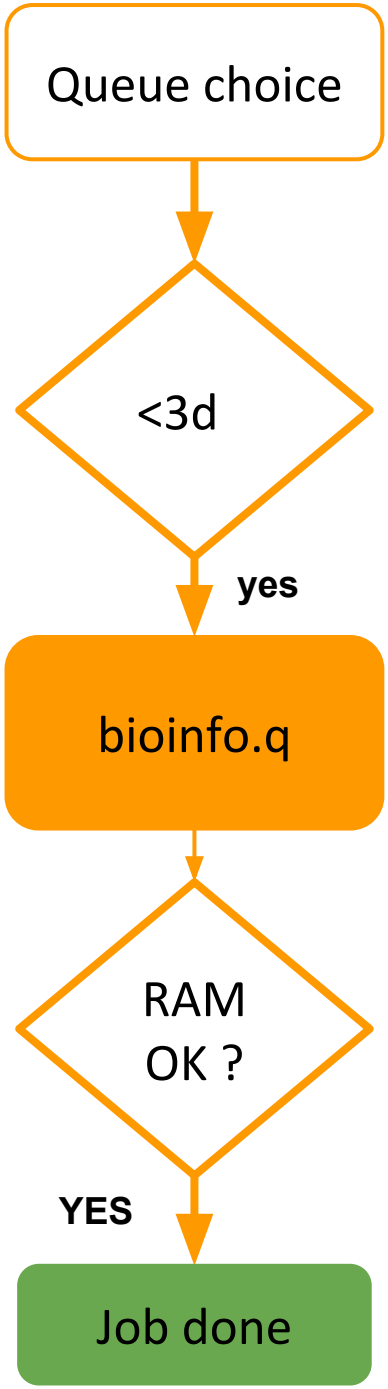
How do I choose the queue?



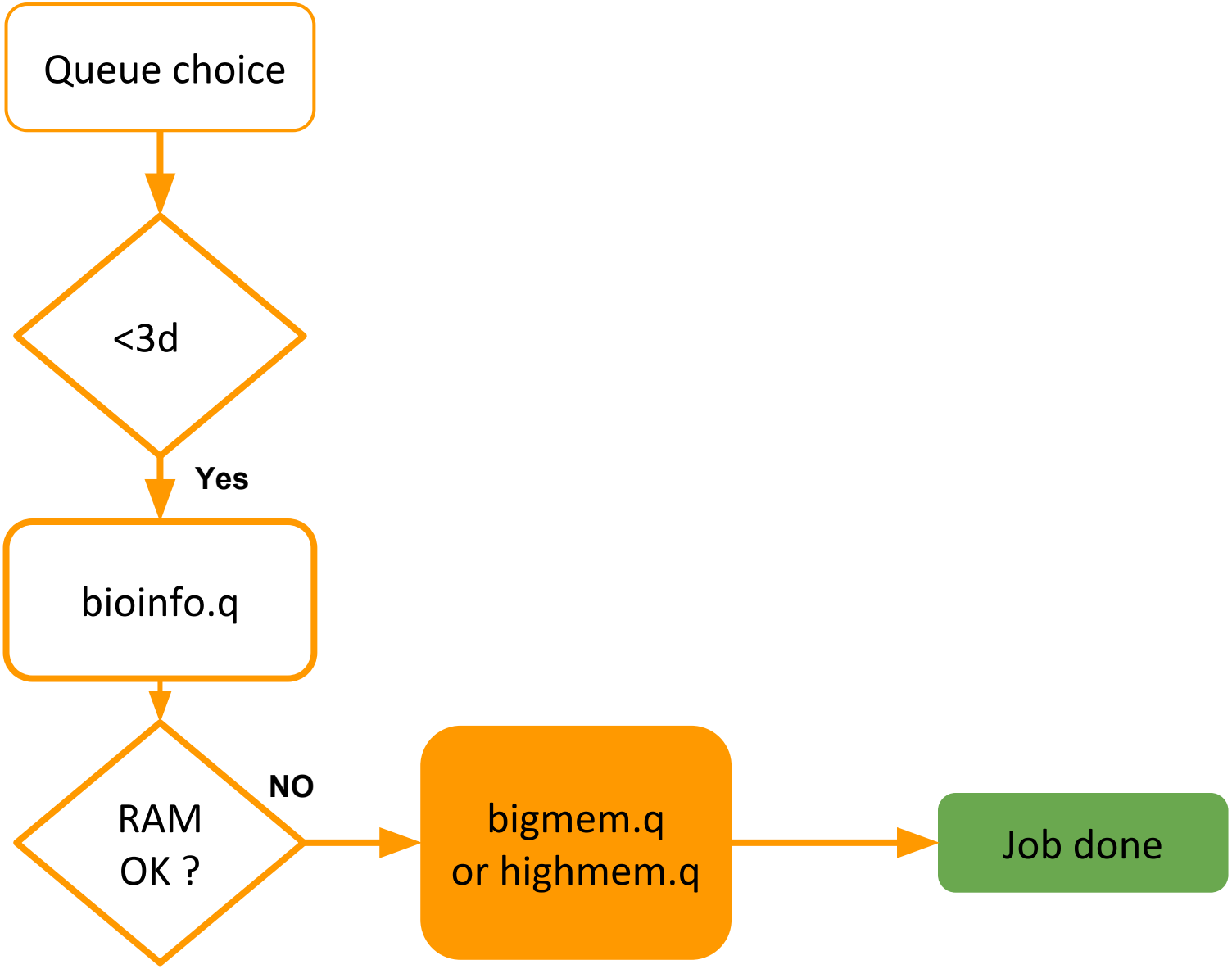
How do I choose the queue?



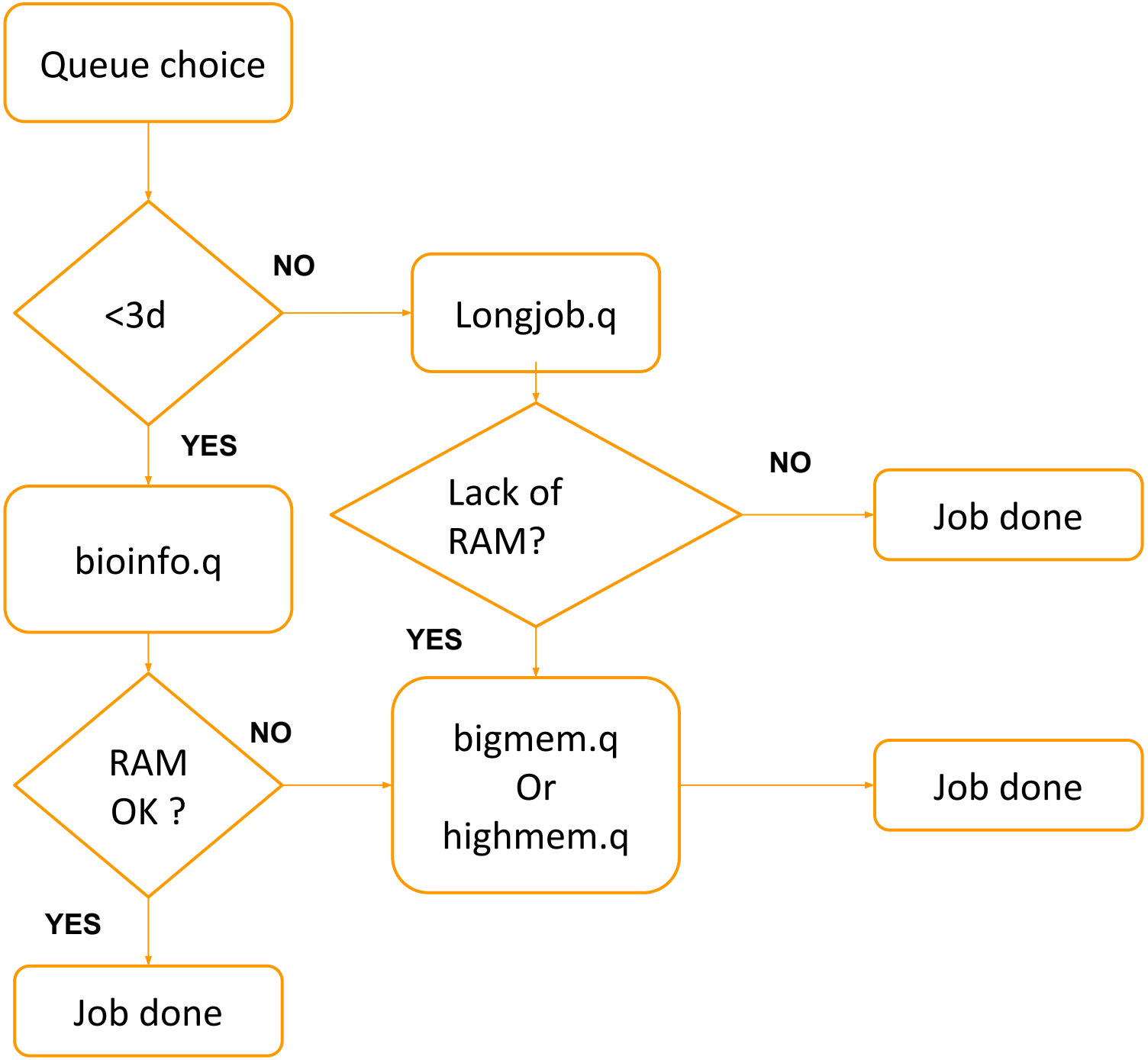
How do I choose the queue?



How do I choose the queue?



How do I choose the queue?



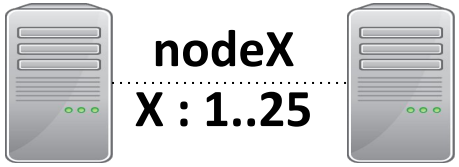
Architecture: components role

- **1 Master node**

bioinfo-master.ird.fr

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

- **25 computing nodes**



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

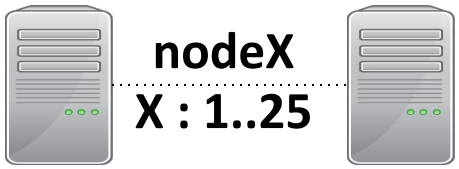
Architecture: components role

- **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**

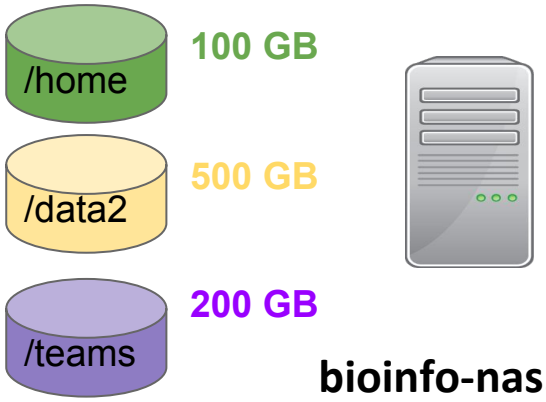
bioinfo-nas.ird.fr

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



bioinfo-nas.ird.fr

Local partition on
bioinfo-nas.ird.fr

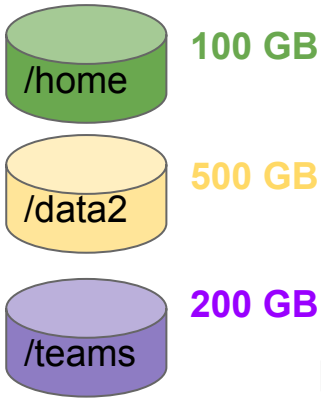
Physical disks on
bioinfo-nas.ird.fr

cluster i-Trop disk partitions

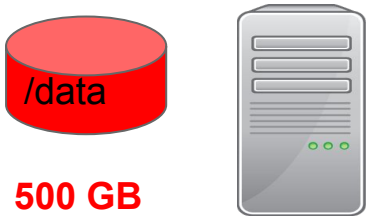
Local partion on
bioinfo-nas2.ird.fr

Physical disks on
bioinfo-nas2.ird.fr

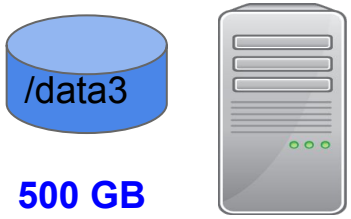
cluster i-Trop disk partitions



bioinfo-nas.ird.fr



bioinfo-nas2.ird.fr

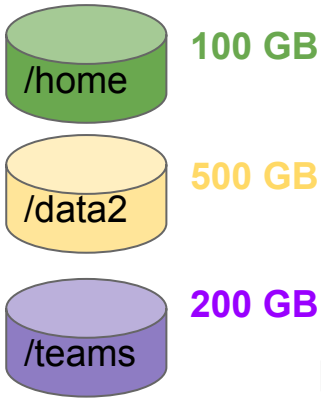


bioinfo-nas3.ird.fr

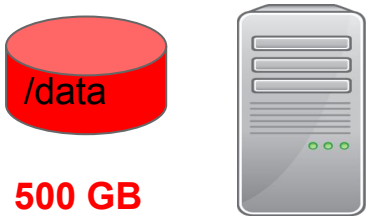
Local partition on
bioinfo-nas3.ird.fr

Physical disks on
bioinfo-nas3.ird.fr

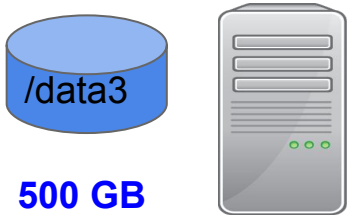
cluster i-Trop disk partitions



bioinfo-nas.ird.fr



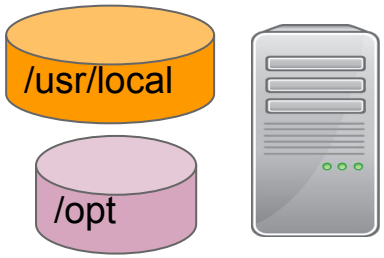
bioinfo-nas2.ird.fr



bioinfo-nas3.ird.fr

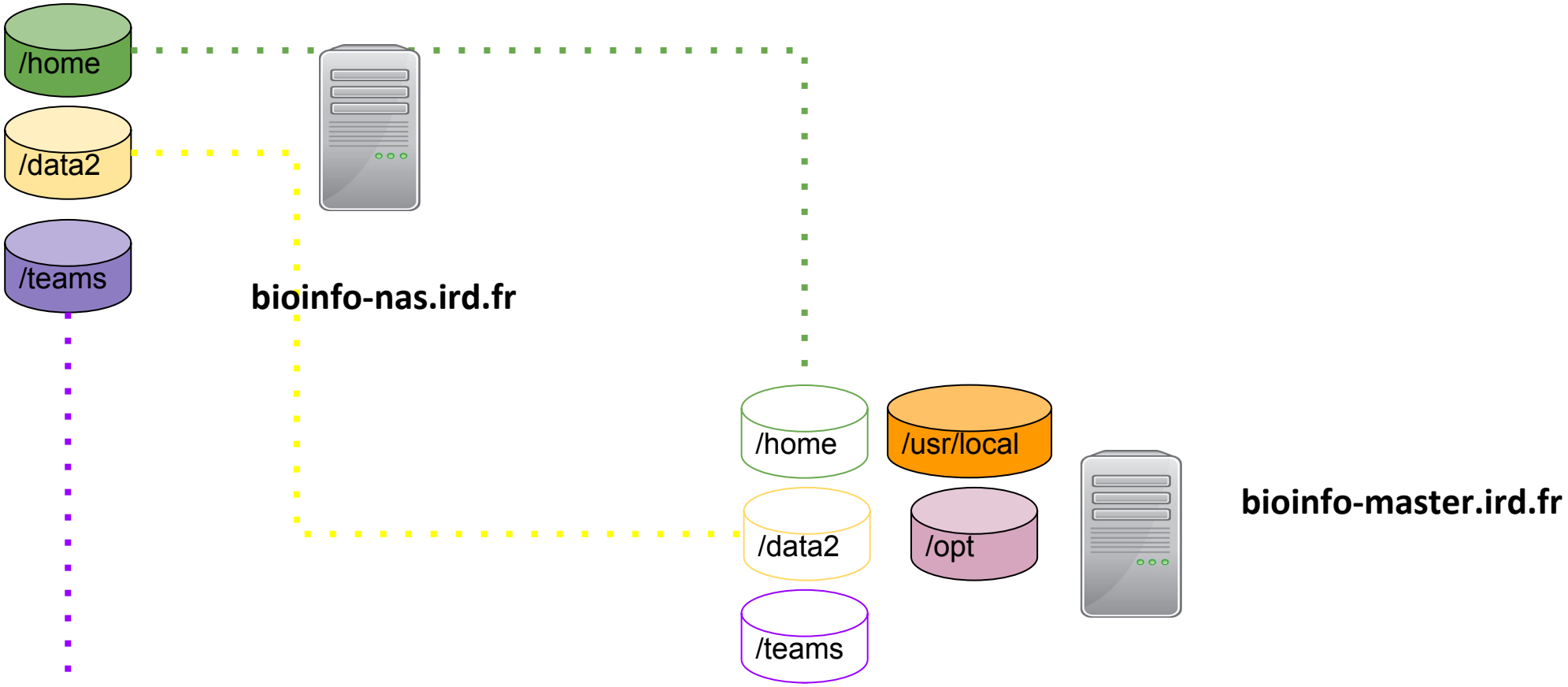
Local partitions on
bioinfo-master.ird.fr

Physical hard disk on
bioinfo-master.ird.fr



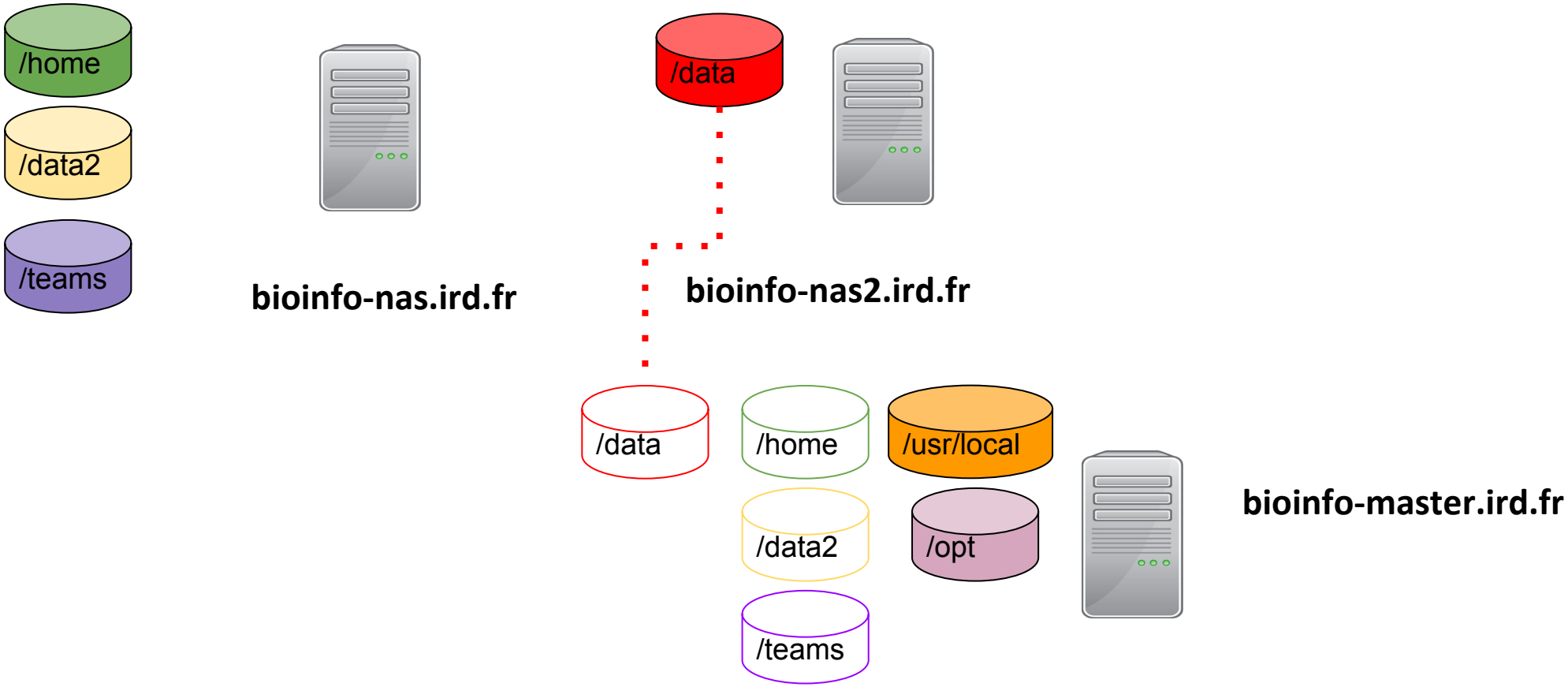
bioinfo-master.ird.fr

cluster i-Trop disk partitions



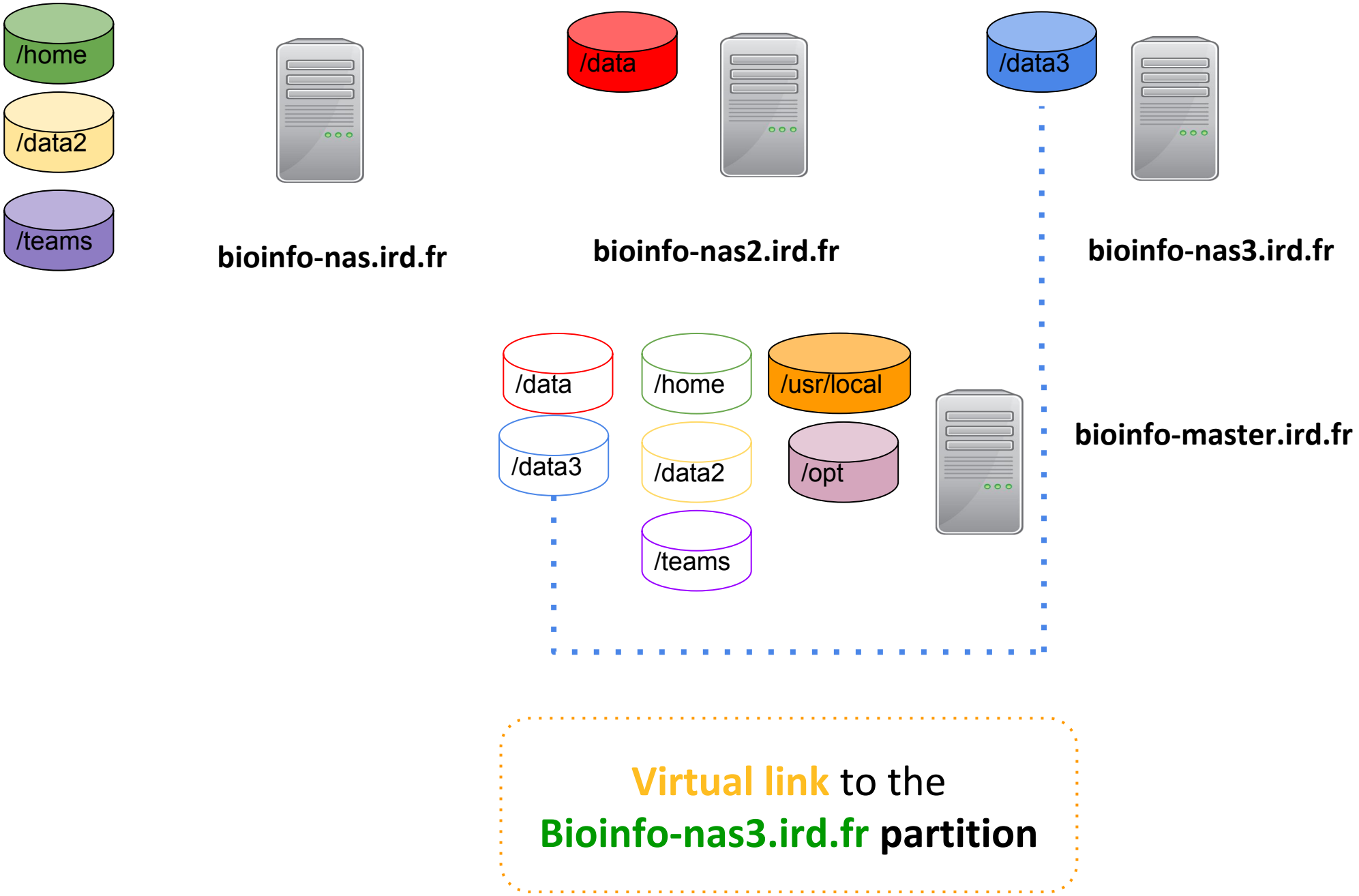
Virtual link to
Bioinfo-nas.ird.fr partitions

cluster i-Trop disk partitions

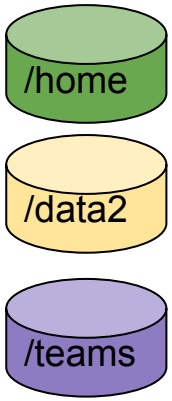


Virtual link to the
Bioinfo-nas2.ird.fr partitions

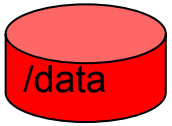
cluster i-Trop disk partitions



cluster i-Trop disk partitions

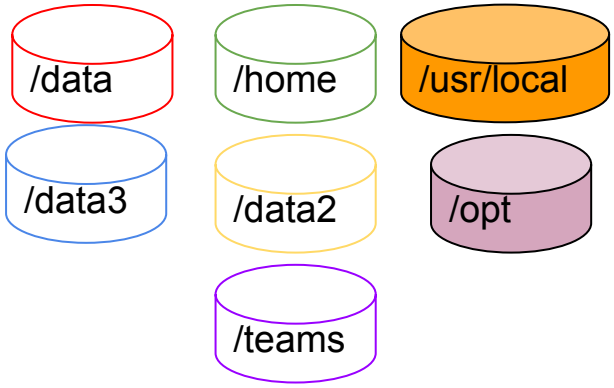


bioinfo-nas.ird.fr



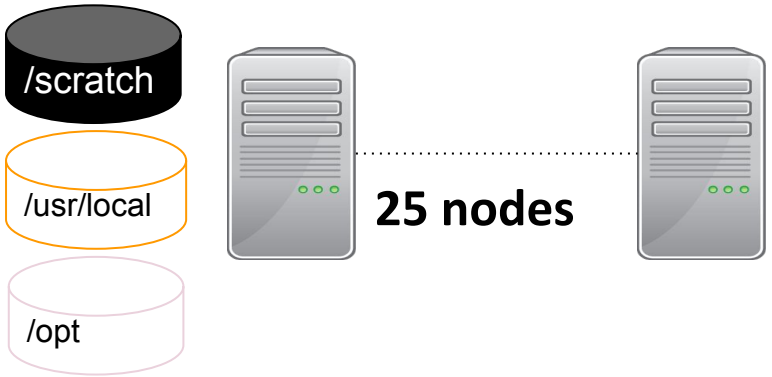
bioinfo-nas2.ird.fr

bioinfo-nas3.ird.fr

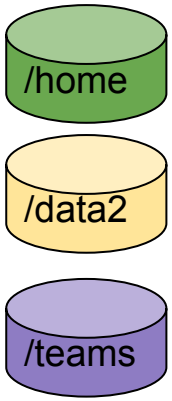


bioinfo-master.ird.fr

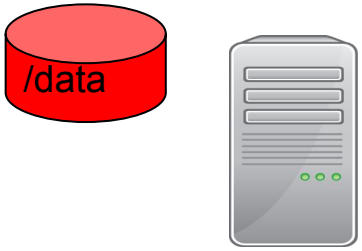
- Local partition on nodes : **temporary space**
- Physical hard drives on nodes



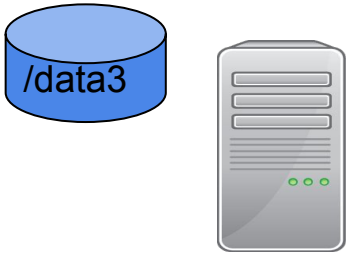
cluster i-Trop disk partitions



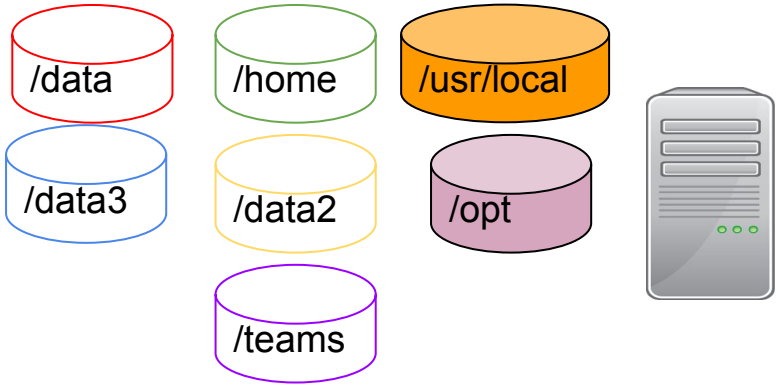
bioinfo-nas.ird.fr



bioinfo-nas2.ird.fr

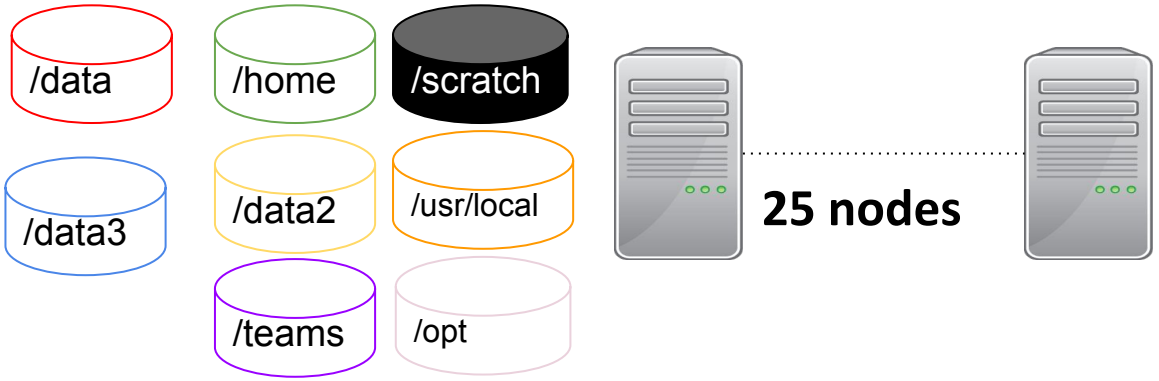


bioinfo-nas3.ird.fr



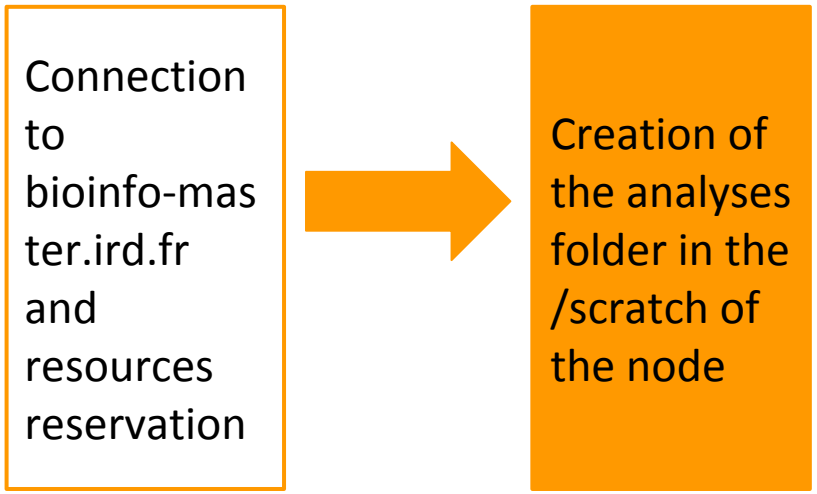
bioinfo-master.ird.fr

Virtual links to the other servers partitions



25 nodes

Analyses steps of the cluster



Step 1

Step 2
mkdir

Practice

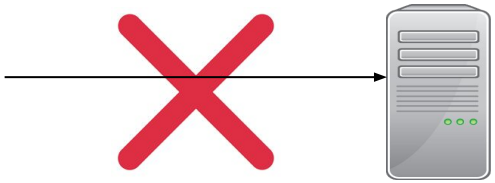
Step 2:qrsh, partition

2

Go to the Practice2 of the github

Data transfer on i-Trop cluster

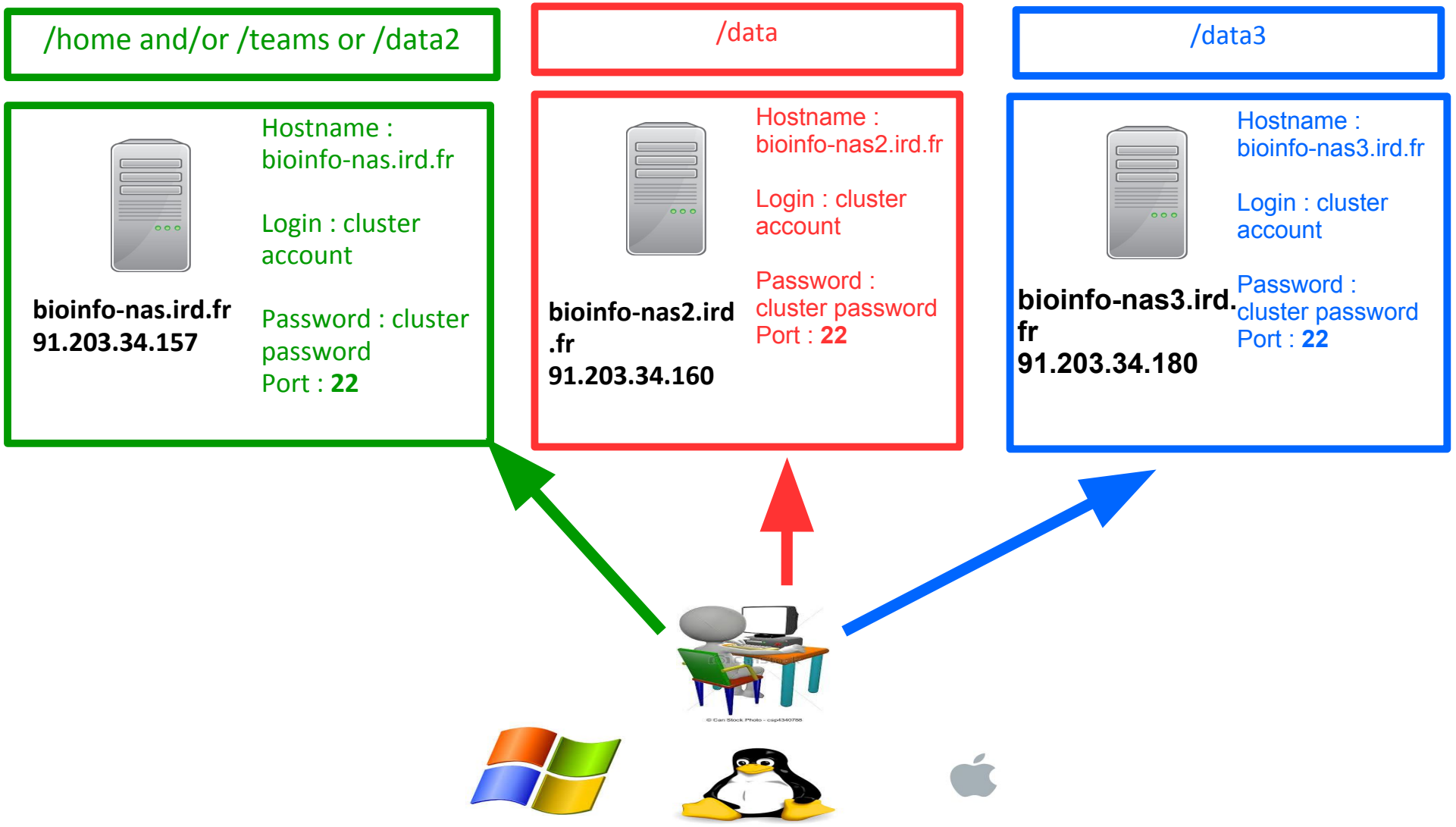
PC/MAC



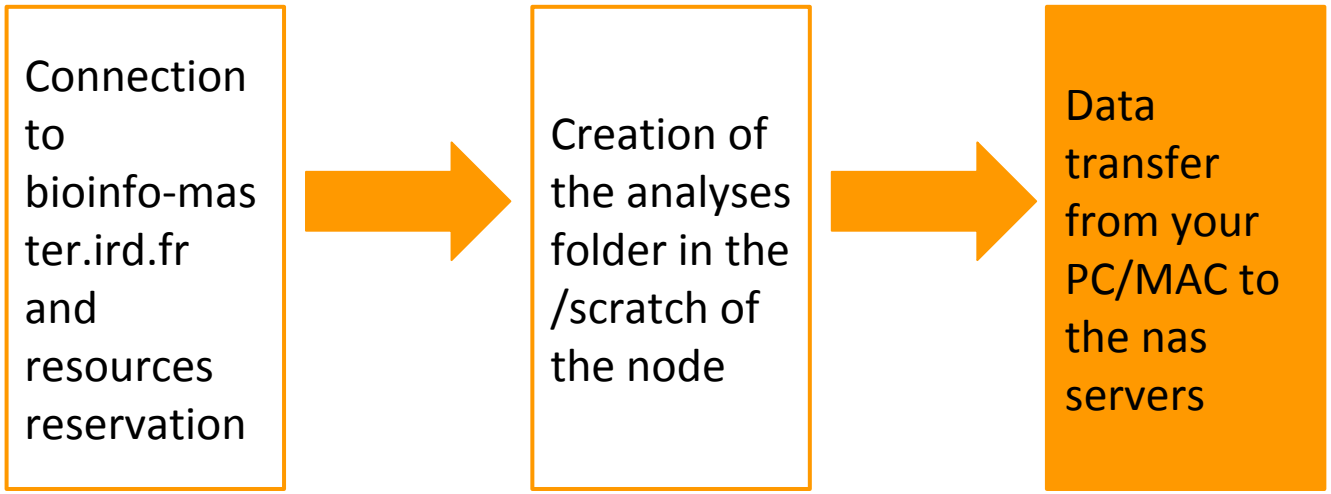
direct transfer
via filezilla
forbidden

bioinfo-master.ird.fr
91.203.34.148

Data transfer on i-Trop cluster



Analyses steps of the cluster



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

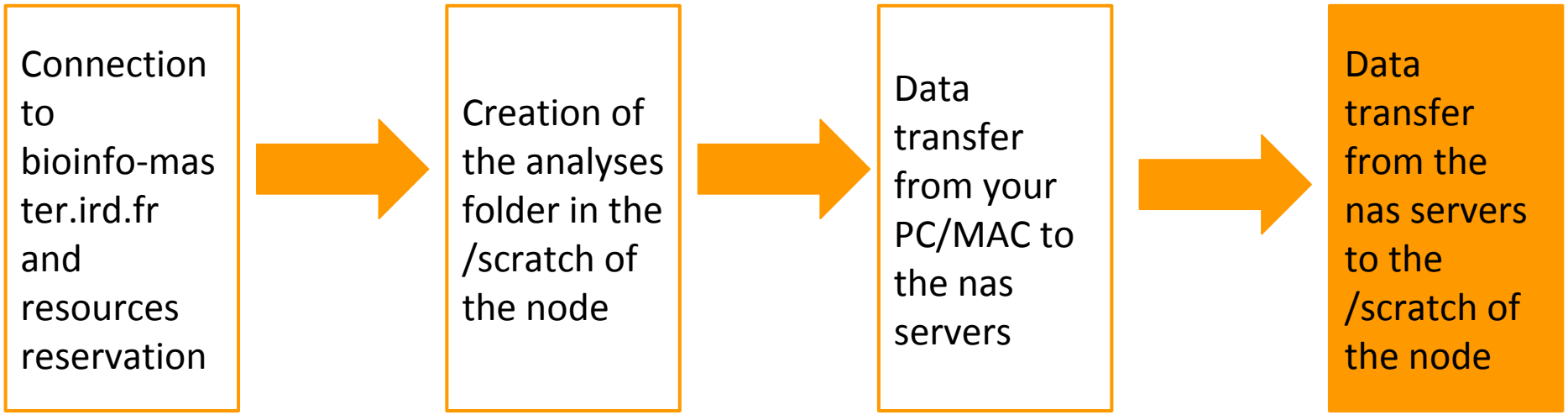
- Syntax if the source is remote :

  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



Step 1

Step 2

Step 3

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

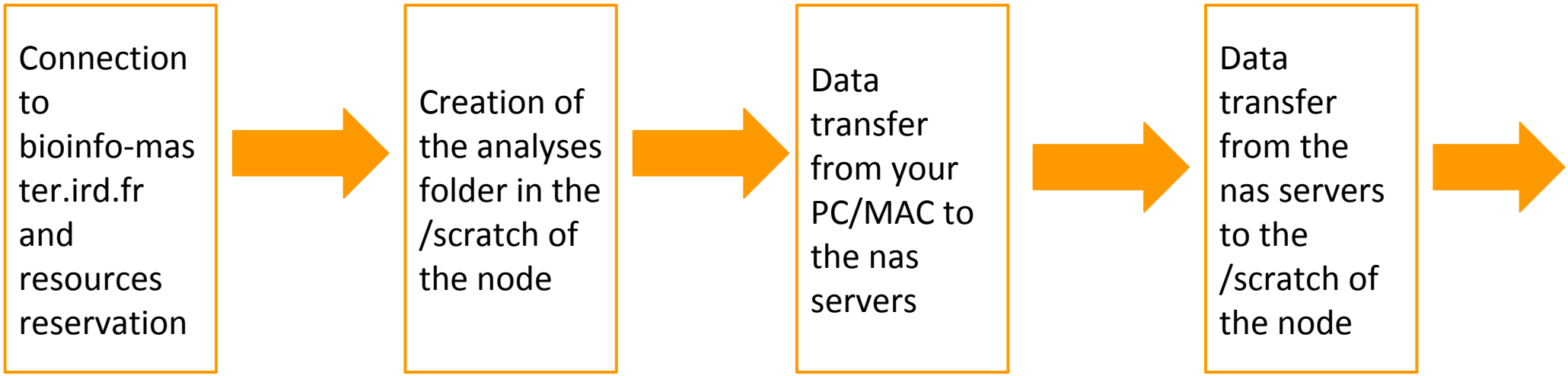
- Unload a module :

  module unload + modulename

- Unload all the modules :

  Module purge

Analyses steps of the cluster



Load softwares with modules environment

Step 5
module

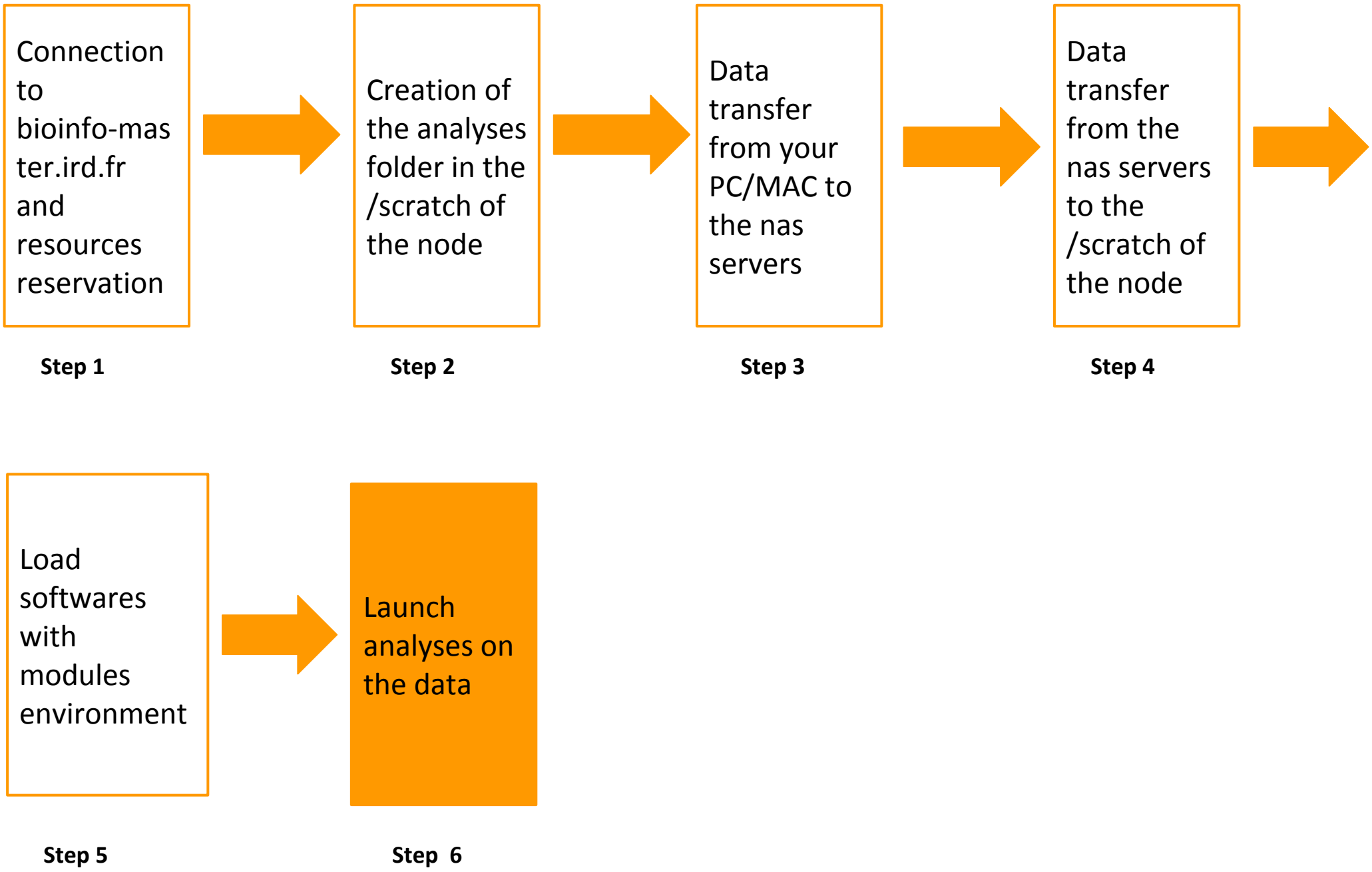
Practice

Step 5: module environment

5

Go to the Practice5 *of the github*

Analyses steps of the cluster



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

6

Go to the Practice6 *of the github*

Transfer your results to the nas servers

- Copy between 2 remote servers :

  scp source destination

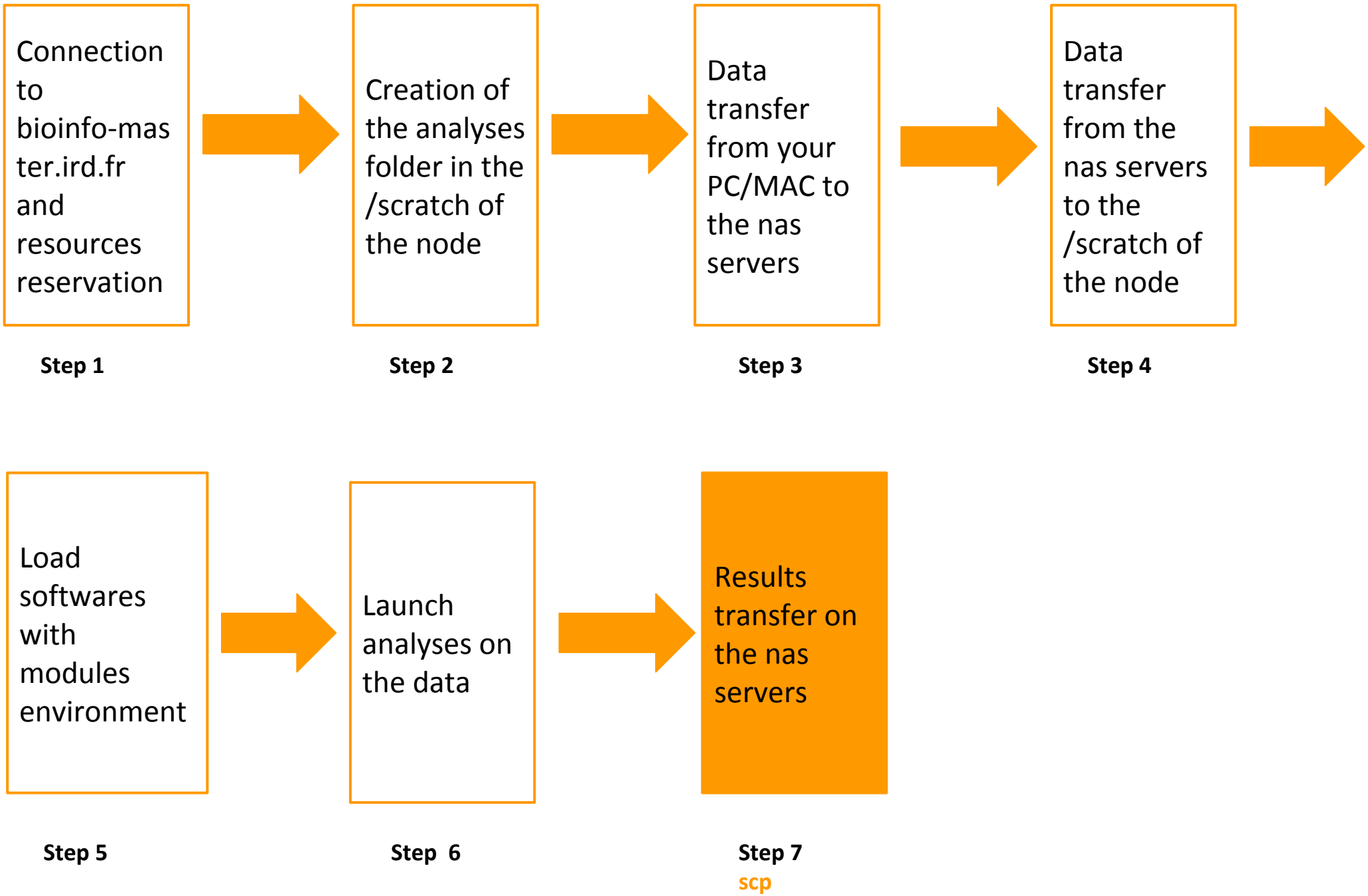
- Syntax if the source is remote :

  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

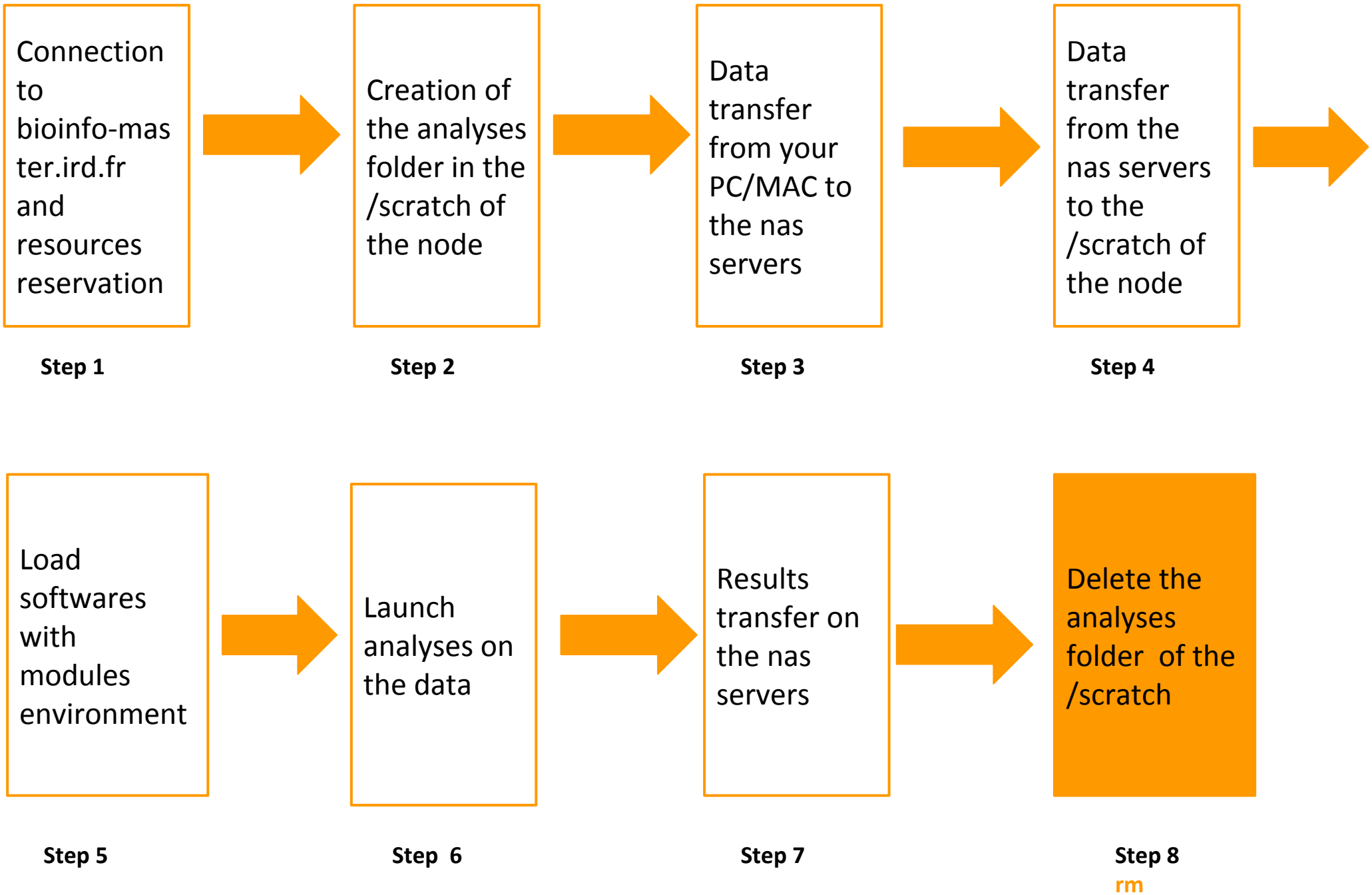
Go to the Practice7 *of the github*

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

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...

Trainers

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








Thank you for your attention !