SouthGreen bioinformatics platform

Modules de formation 2019

agap

Bioversity International

UMR BGPI Biologie et Génétique des Interactions Plante-Parasite

diade

IPME

ABiMS

ASTRE Animal - Santé Territoires - Risques - Écosystèmes

BOREA Biologie des Organismes et Écosystèmes Aquatiques

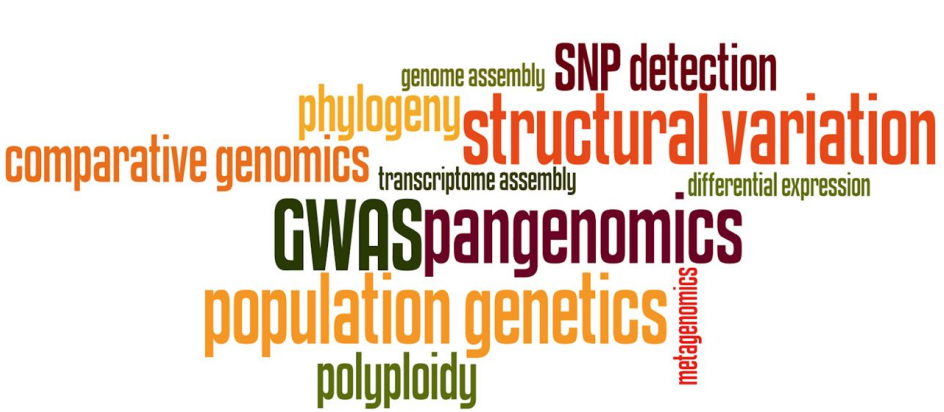
intertryp Host-vector-parasite-environment interactions in neglected tropical diseases due to trypanosomatids

MiVEGEC Maladies Infectieuses et Vecteurs Écologie, Génétique, Évolution et Contrôle

UMR QualiSud

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

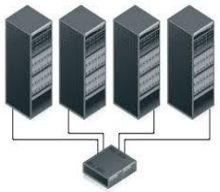
Workflow manager

TOGGLE
Toolbox for generic NGS analyses

SNAKEMAKE

Galaxy

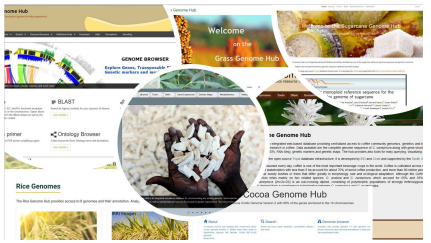
HPC and trainings….



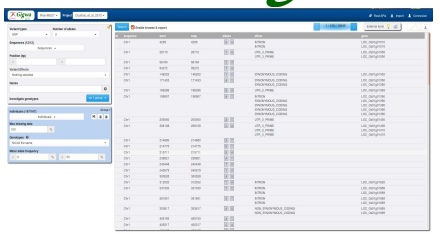
IRD Institut de Recherche pour le Développement FRANCE

cirad

Genome Hubs & Information System

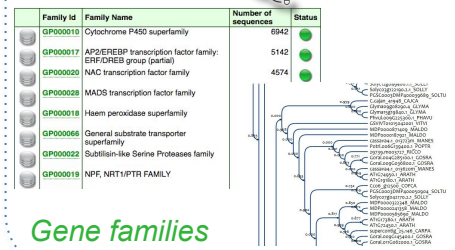


Gigwa



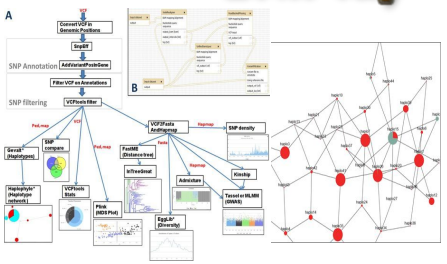
SNPs and Indels

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

Modules de formation 2019

- Toutes nos formations :
  https://southgreenplatform.github.io/trainings/
- Topo & TP : Initiation au cluster de calcul i-Trop
- Environnement de travail : Logiciels à installer
- How-tos : How-to

Initiation HPC cluster

www.southgreen.fr

https://southgreenplatform.github.io/trainings

Objectifs du module

Objectif

Acquérir les bonnes pratiques pour utiliser le cluster de calcul Itrop !

Applications

- Connaître l'architecture du cluster
- Connaître le rôle des différentes partitions
- Utiliser SGE (qusb, qrsh, qhost, qacct, qstat, qqdel)
- Utiliser les modules environment
- Faire du scripting de base

Demandes/incidents

- Site https://bioinfo.ird.fr
  - Comptes
  - Installation logiciels
  - Projets
  - Logiciels installés

- Incidents: contacter bioinfo@ird.fr

ARCHITECTURE

Qu’est ce qu’un cluster?

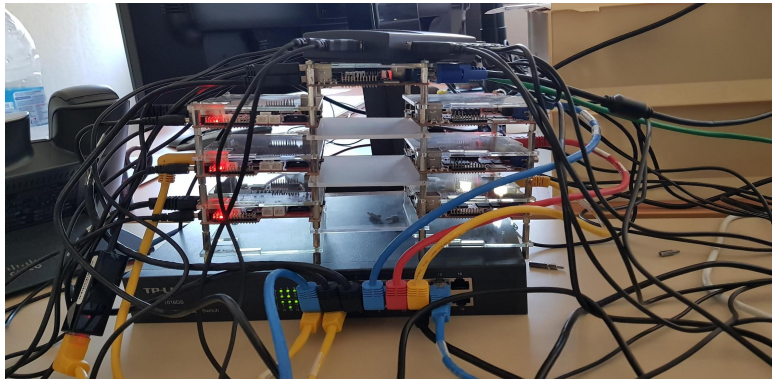
- une unité logique de plusieurs serveurs
- une unique machine puissante
- une puissance de calcul élevée
- Une plus grande capacité de stockage
- Une fiabilité supérieure
- Une plus grande disponibilité des ressources

Qu’est ce qu’un cluster?

- une unité logique de plusieurs serveurs
- une unique machine puissante
- une puissance de calcul élevée
- Une plus grande capacité de stockage
- Une fiabilité supérieure
- Une plus grande disponibilité des ressources

Qu’est ce qu’un cluster?

- une unité logique de plusieurs serveurs
- une unique machine puissante
- une puissance de calcul élevée
- Une plus grande capacité de stockage
- Une fiabilité supérieure
- Une plus grande disponibilité des ressources



Composants d'un cluster

CALCUL

- **Noeud maître**
  Gère les ressources et les priorités des jobs

- **Noeuds de calcul**
  Ressources (CPU ou mémoire RAM)

Composants d'un cluster

CALCUL

- **Noeud maître**
  Gère les ressources et les priorités des jobs
- **Noeuds de calcul**
  Ressources (CPU ou mémoire RAM)

STOCKAGE

- **Serveur(s) NAS**
  Stockage

Architecture: rôle des éléments

- **1 Noeud Maître**

bioinfo-master.ird.fr

Rôle :

- Lancer et prioriser les jobs sur les nœuds de calcul
- Accessible depuis Internet
- Connexion :

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

Architecture: rôle des éléments

- **1 Noeud Maître**

bioinfo-master.ird.fr

Rôle :

- Lancer et prioriser les jobs sur les nœuds de calcul
- Accessible depuis Internet
- Connexion :

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

- **25 Noeud de Calcul**

nodeX
X : 1..25

Rôle :

- Utilisés par le maître pour exécuter les jobs/calculs
- Pas accessibles depuis Internet
- node0 à node25
- Connexion de master

```
ssh nodeX
```

Architecture: rôle des éléments

- **1 Noeud Maître**

bioinfo-master.ird.fr
91.203.34.148

Rôle :
- Lancer et prioriser les jobs sur les nœuds de calcul
- Accessible depuis Internet
- Connexion :

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

- **25 Noeud de Calcul**

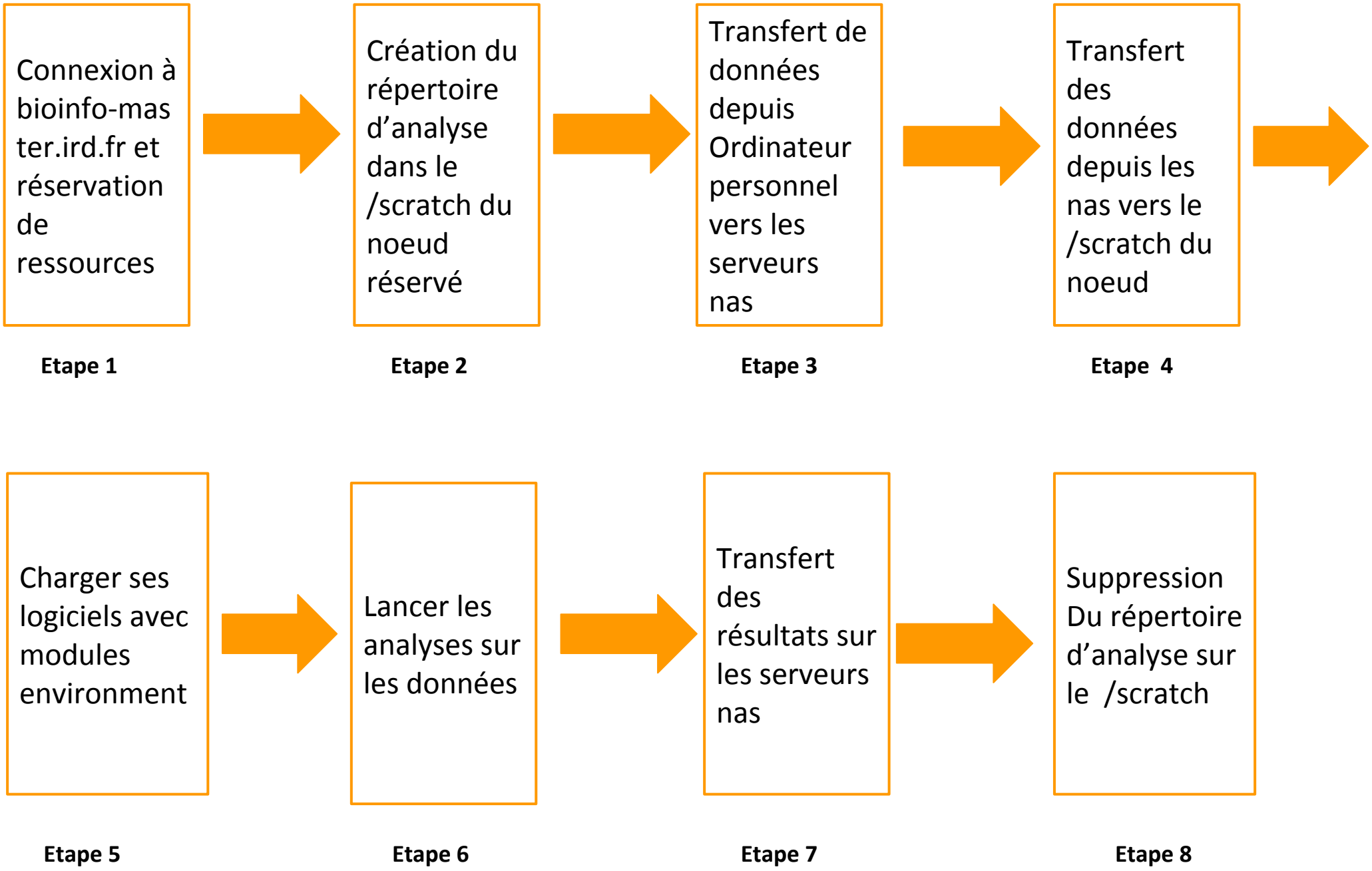
nodeX
X : 1..25

Rôle :
- Utilisés par le maître pour exécuter les jobs/calculs
- Pas accessibles depuis Internet
- node0 à node25
- Connexion de master

```
ssh nodeX
```

Noeud interactif (node6)
- Accessible de l'extérieur bioinfo-inter.ird.fr
- Connexion : `ssh login@bioinfo-inter.ird.fr`

Etapes d'une analyse sur le cluster



Etapes d’une analyse sur le cluster

Connexion à bioinfo-master.ird.fr et réservation de ressources

Etape 1
qrsh/qlogin
ou qsub

Practice

Etape 1: Connexion, qhost

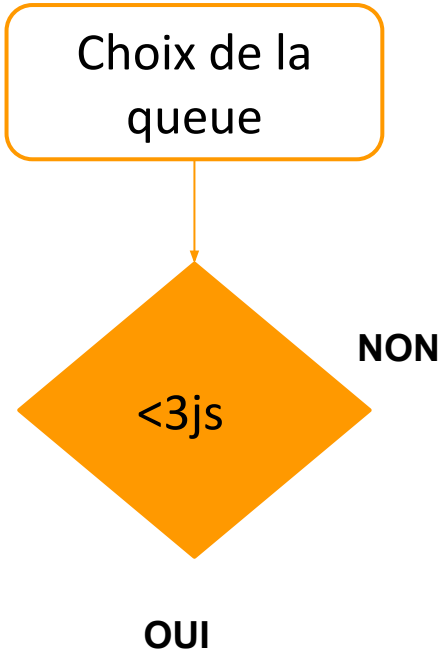
1

Aller sur le Practice 1 *du github*

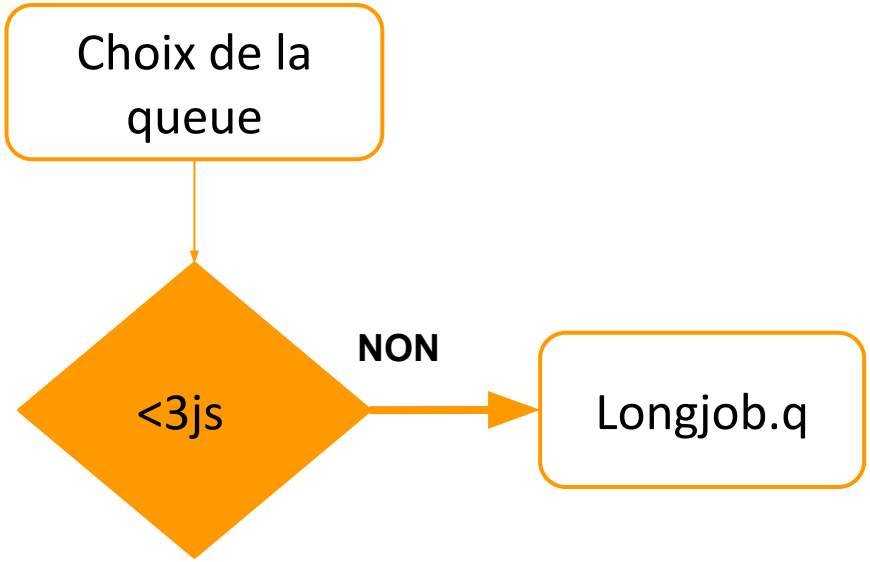
Les files d'attentes

Queues	Utilisation	Caractéristiques RAM noeuds	Caractéristiques coeurs noeuds
bioinfo.q	Jobs courts < 3jours	48 à 64 Go	12 à 20 coeurs
longjob.q	Jobs longs > 3 jours	48 Go	12 coeurs
bigmem.q	Jobs avec besoin de plus de mémoire	96 Go	12 coeurs
highmem.q	Jobs avec besoin de beaucoup de mémoire	144 Go	12 coeurs

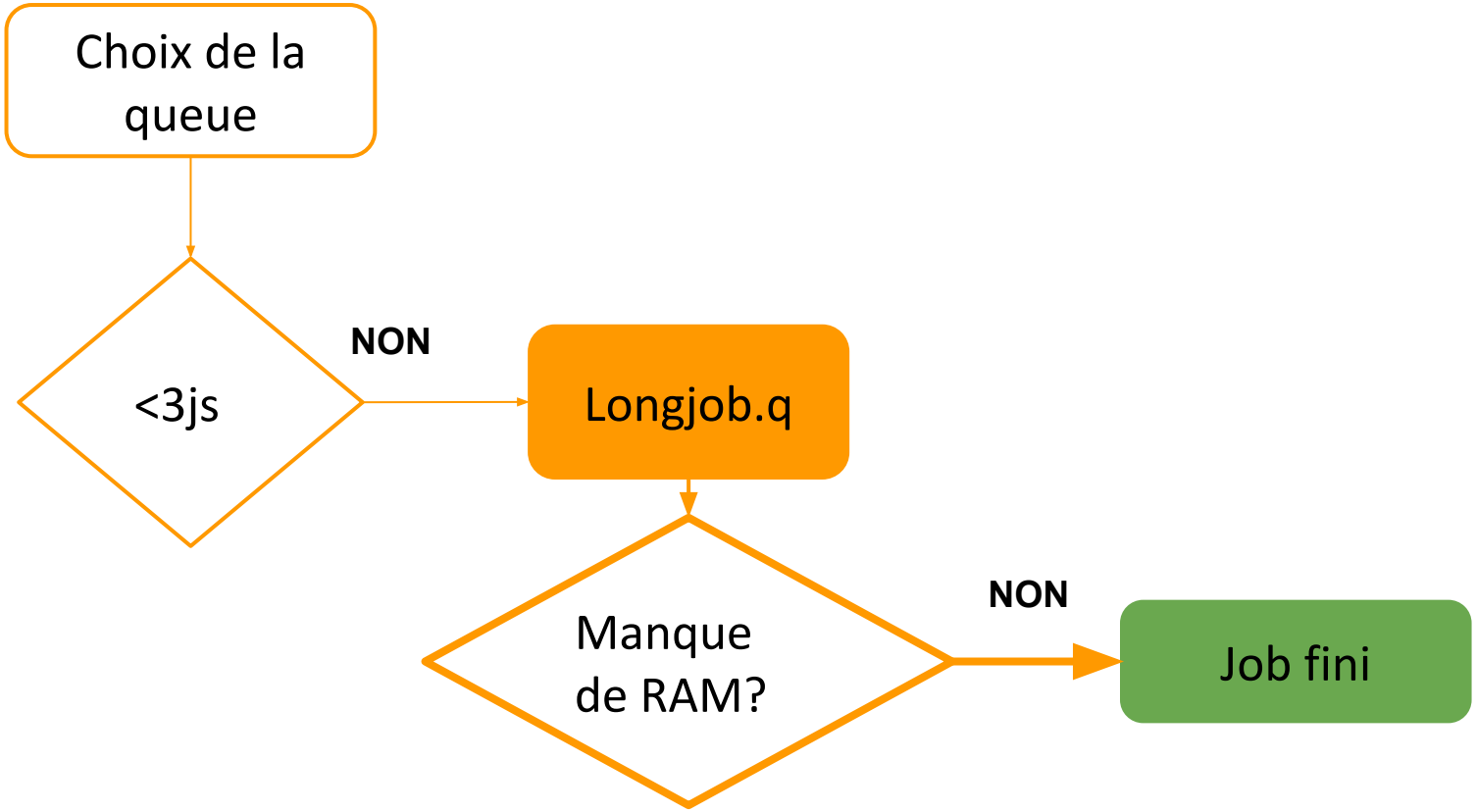
Quelle queue choisir?



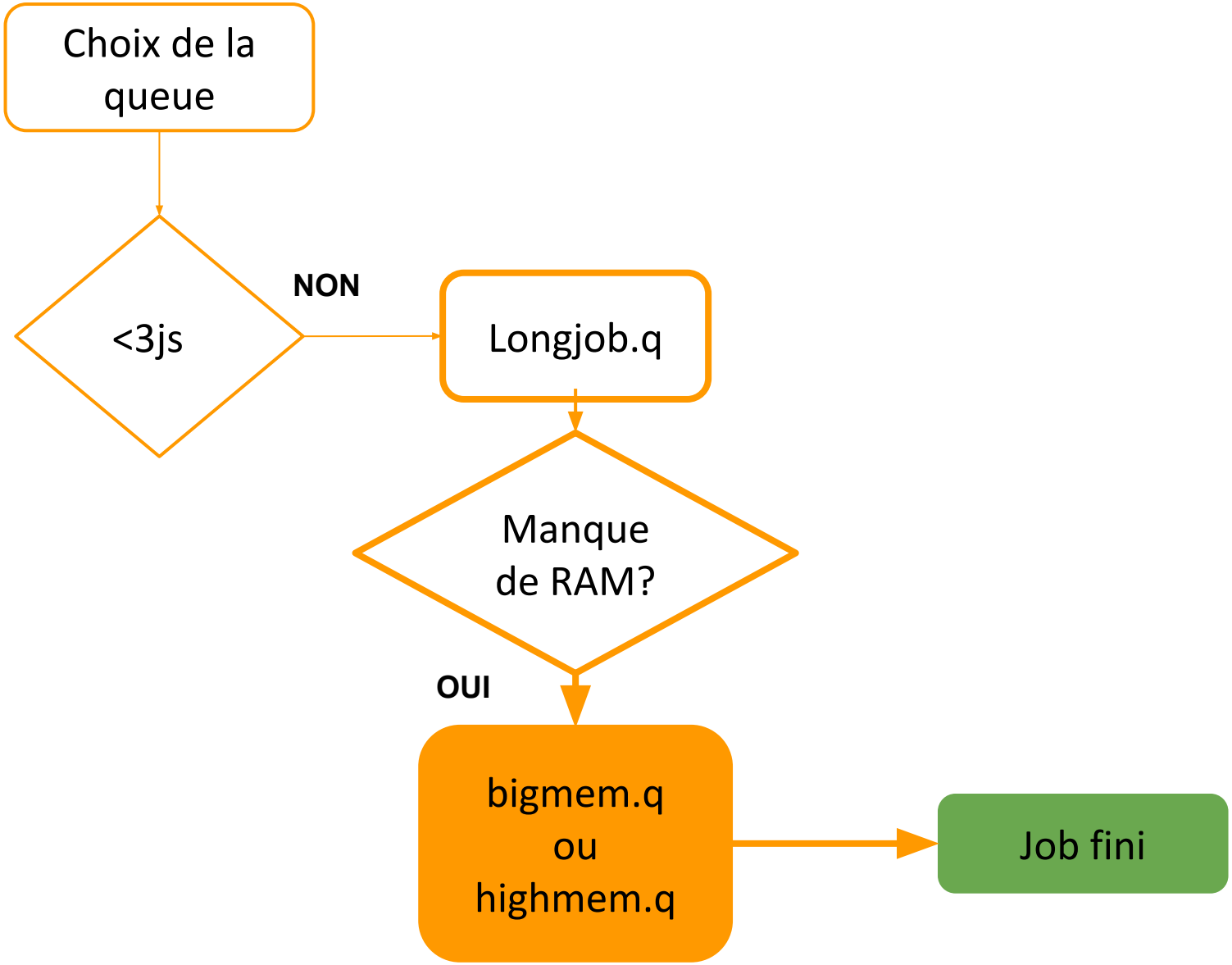
Quelle queue choisir?



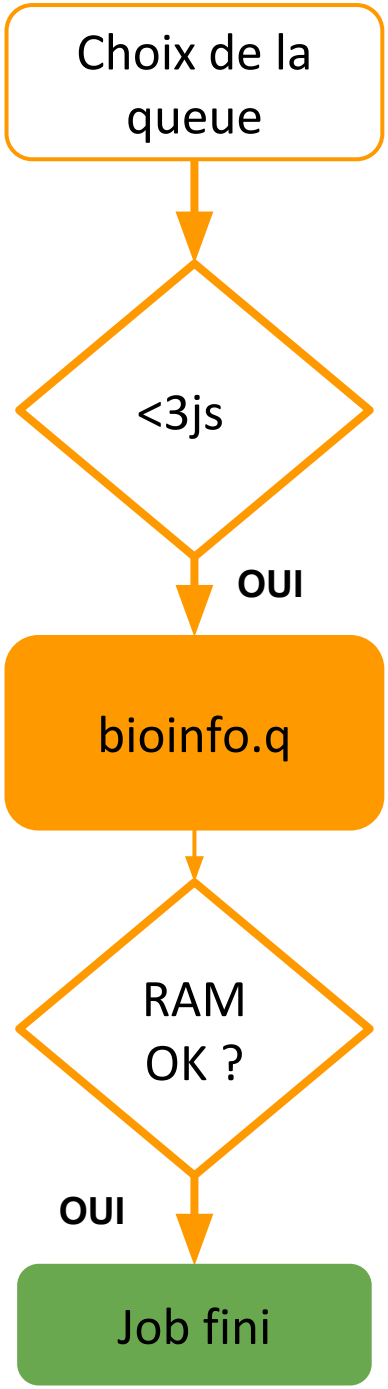
Quelle queue choisir?



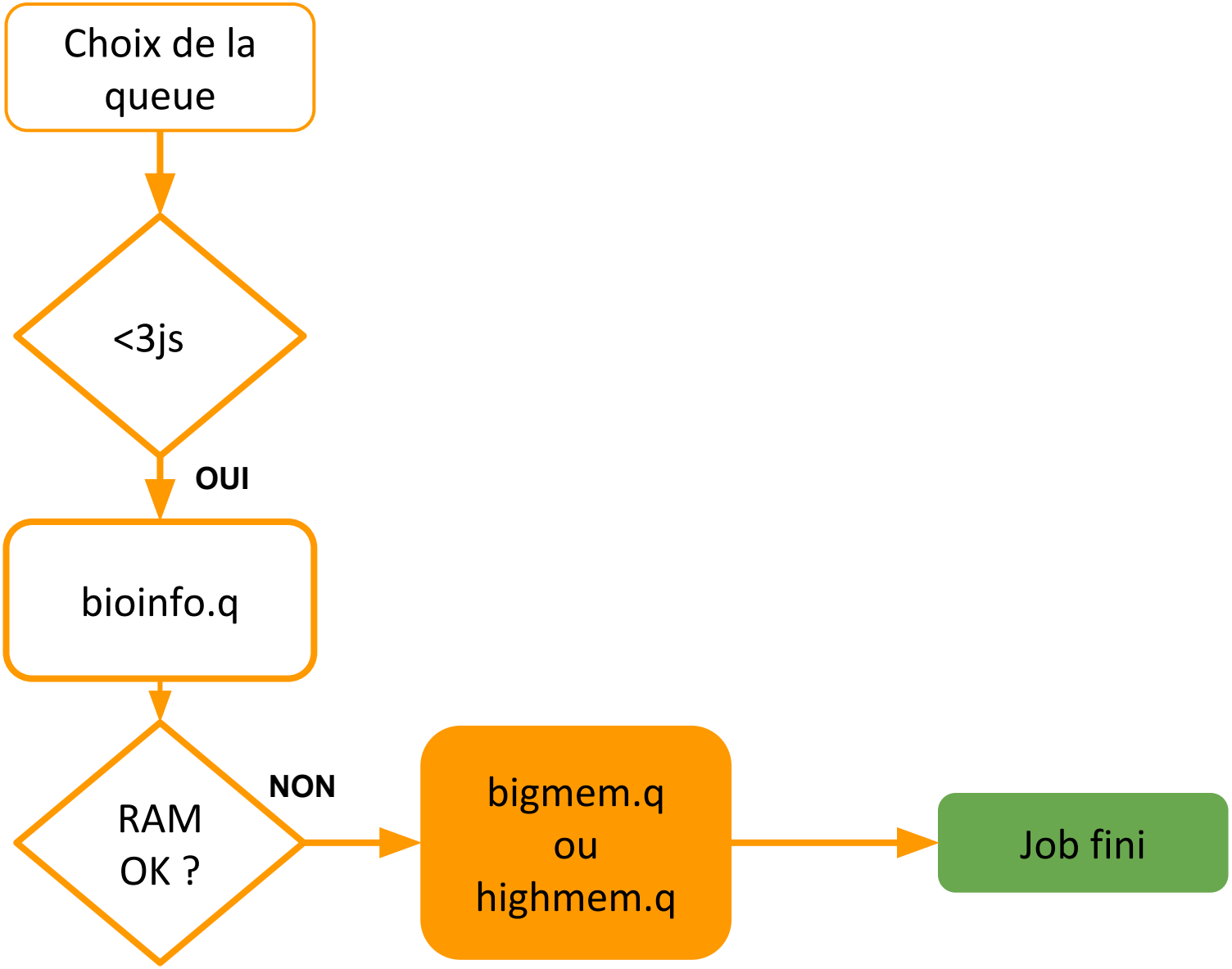
Quelle queue choisir?



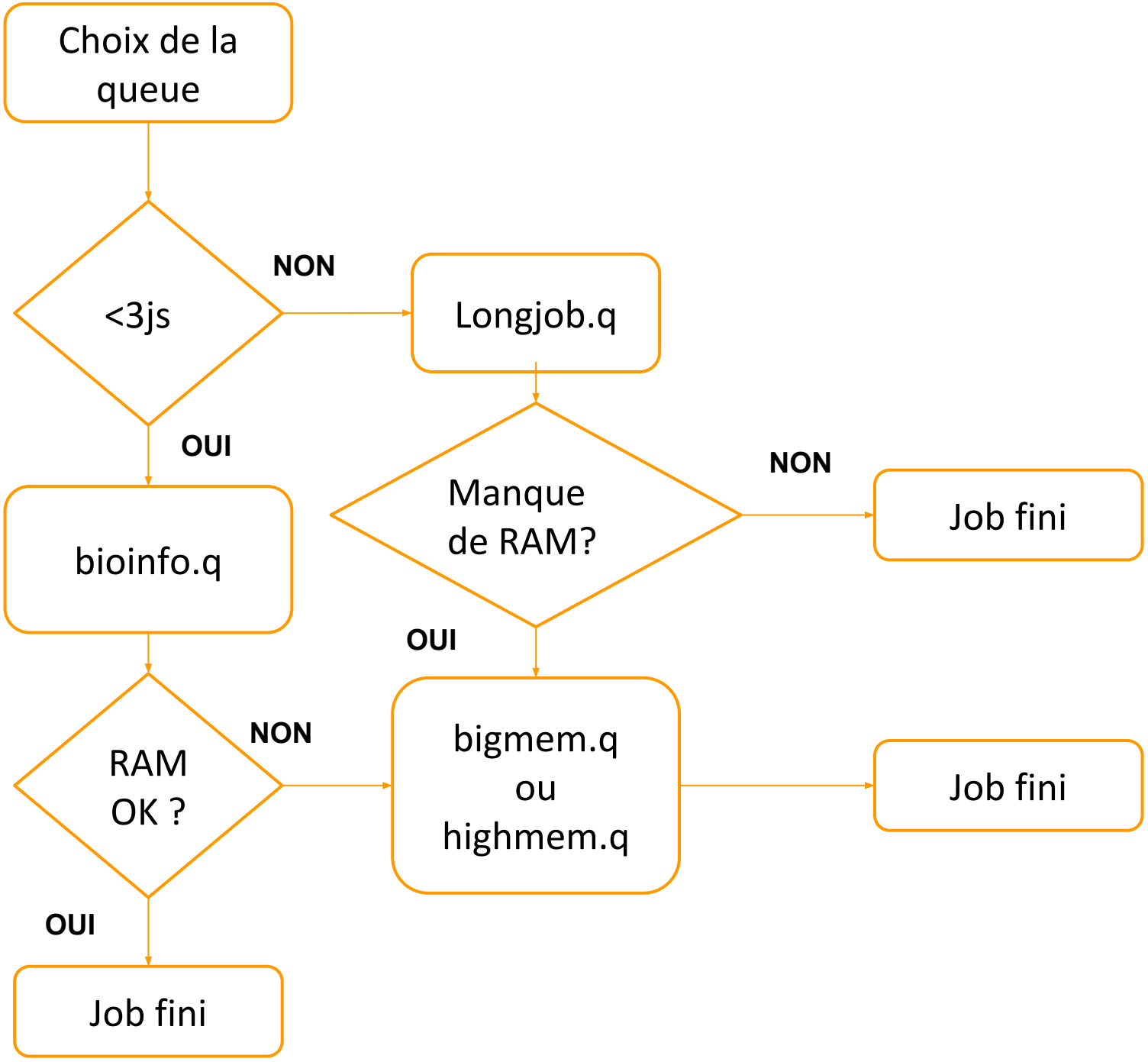
Quelle queue choisir?



Quelle queue choisir?



Quelle queue choisir?



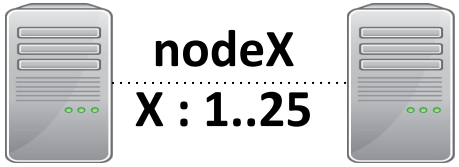
Architecture: rôle des éléments

- **1 Noeud Maître**

bioinfo-master.ird.fr

Rôle :
- Lancer et prioriser les jobs sur les nœuds de calcul
- Accessible depuis Internet

- **25 Noeud de Calcul**



Rôle :
- Utilisés par le maître pour exécuter les jobs/calculs
- Pas accessibles depuis Internet

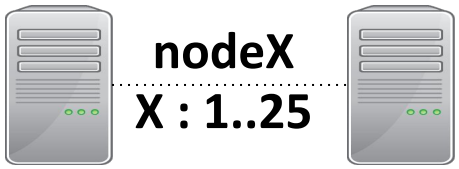
Architecture: rôle des éléments

- **1 Noeud Maître**

bioinfo-master.ird.fr
91.203.34.148

Rôle :
- Lancer et prioriser les jobs sur les nœuds de calcul
- Accessible depuis Internet

- **25 Noeud de Calcul**



nodeX
X : 1..25

Rôle :
- Utilisés par le maître pour exécuter les jobs/calculs
- Pas accessibles depuis Internet

- **3 serveurs NAS**

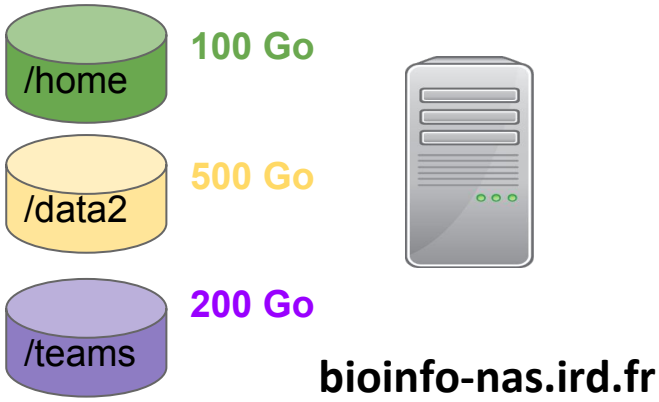
bioinfo-nas.ird.fr

bioinfo-nas2.ird.fr

bioinfo-nas3.ird.fr

Rôle :
- Stocker les données utilisateurs
- Accessibles depuis Internet
- Pour transférer les données : *via filezilla ou scp*

Partitions disques sur le cluster i-Trop



Partition locale sur
bioinfo-nas.ird.fr

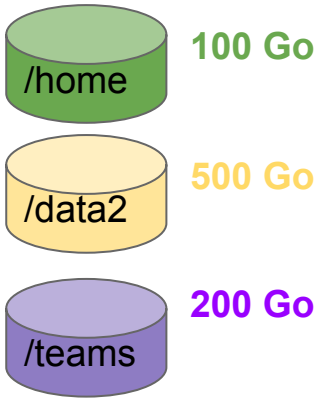
Disques durs physiques sur
bioinfo-nas.ird.fr

Partitions disques sur le cluster i-Trop

Partition locale sur
bioinfo-nas2.ird.fr

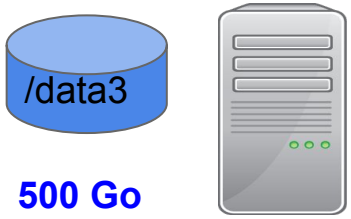
Disques durs physiques sur
bioinfo-nas2.ird.fr

Partitions disques sur le cluster i-Trop



bioinfo-nas.ird.fr

bioinfo-nas2.ird.fr

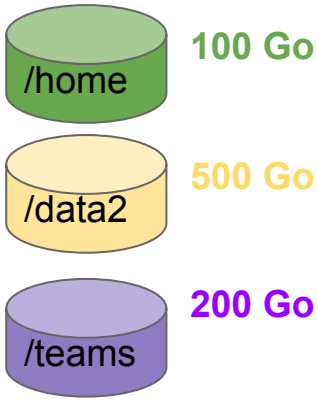


bioinfo-nas3.ird.fr

Partition locale sur
bioinfo-nas3.ird.fr

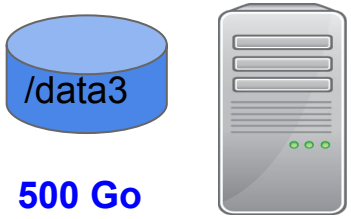
Disques durs physiques sur
bioinfo-nas3.ird.fr

Partitions disques sur le cluster i-Trop



bioinfo-nas.ird.fr

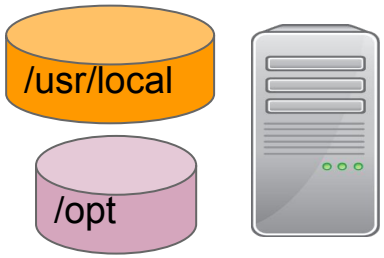
bioinfo-nas2.ird.fr



bioinfo-nas3.ird.fr

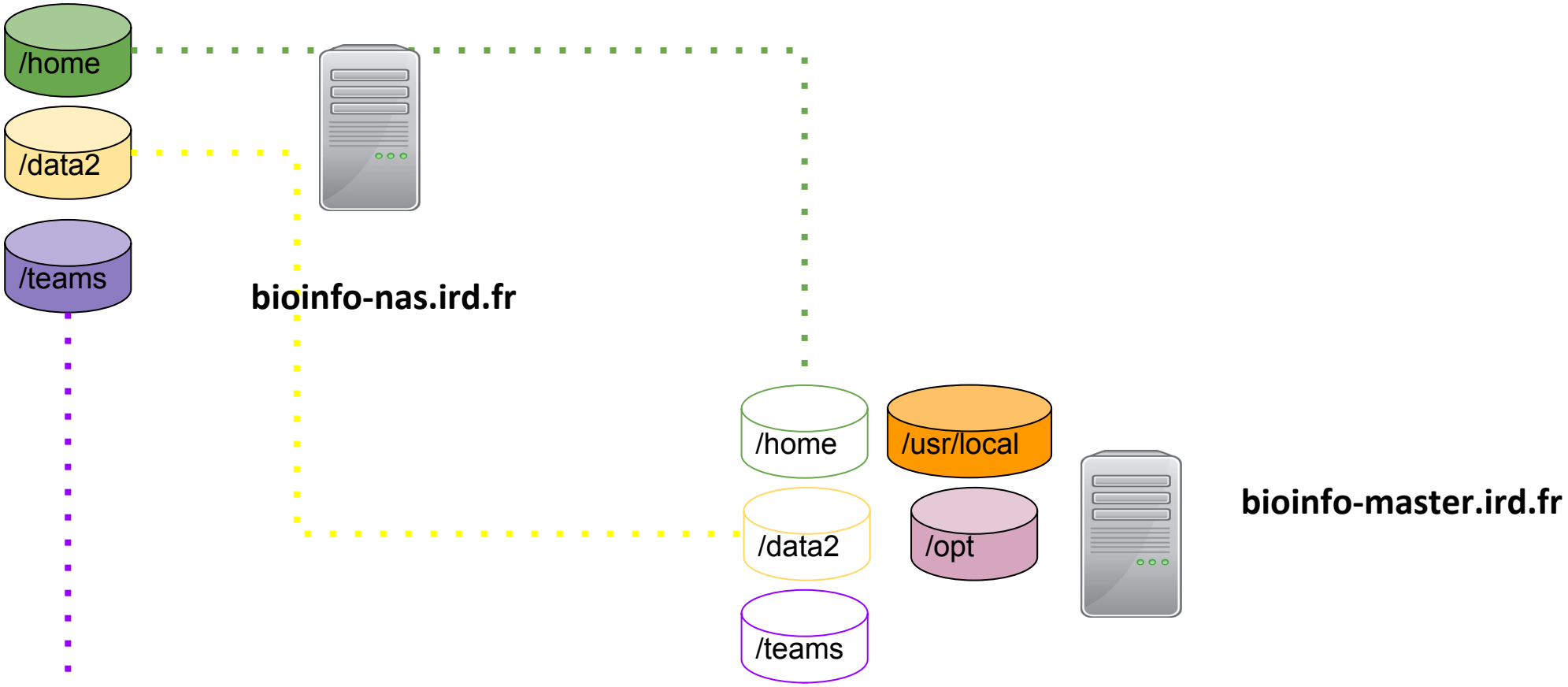
Partitions locales sur
bioinfo-master.ird.fr

Disques durs physiques sur
bioinfo-master.ird.fr



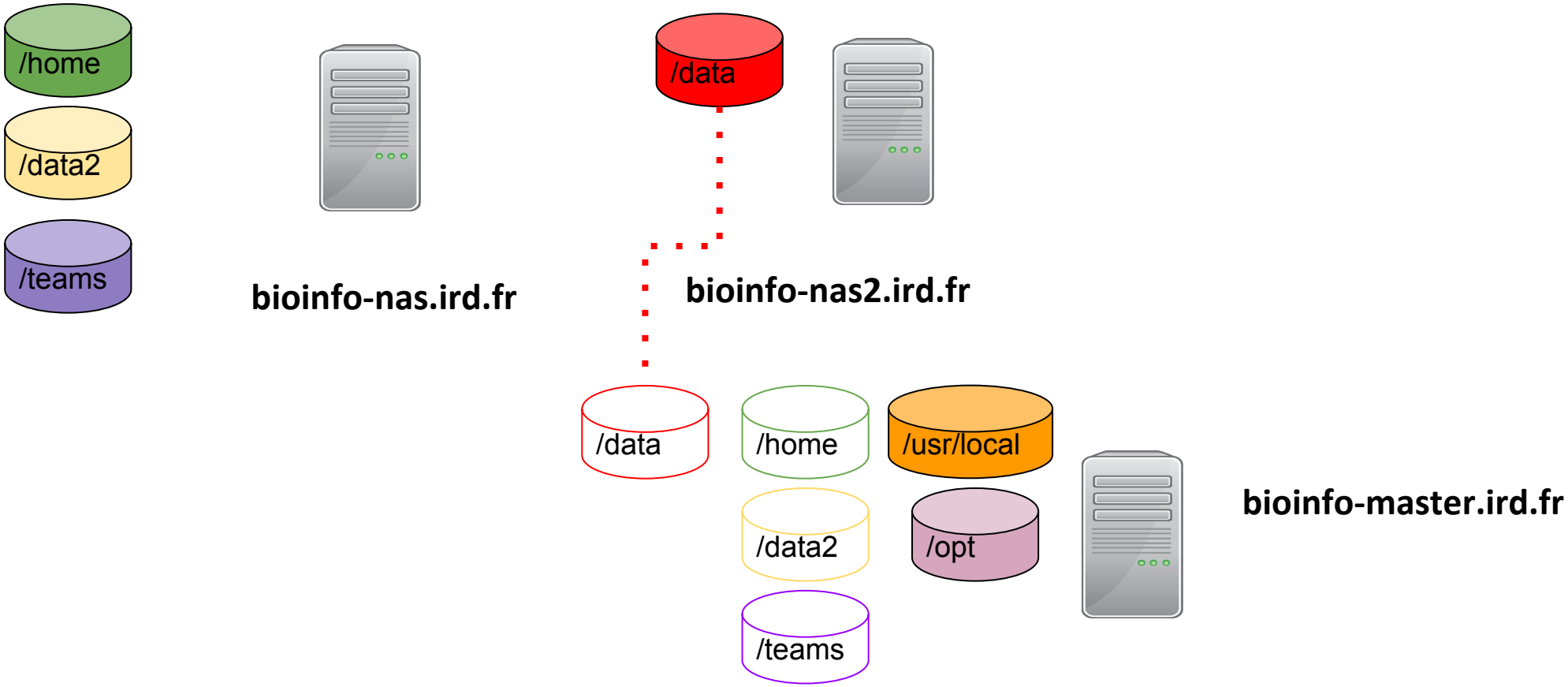
bioinfo-master.ird.fr

Partitions disques sur le cluster i-Trop



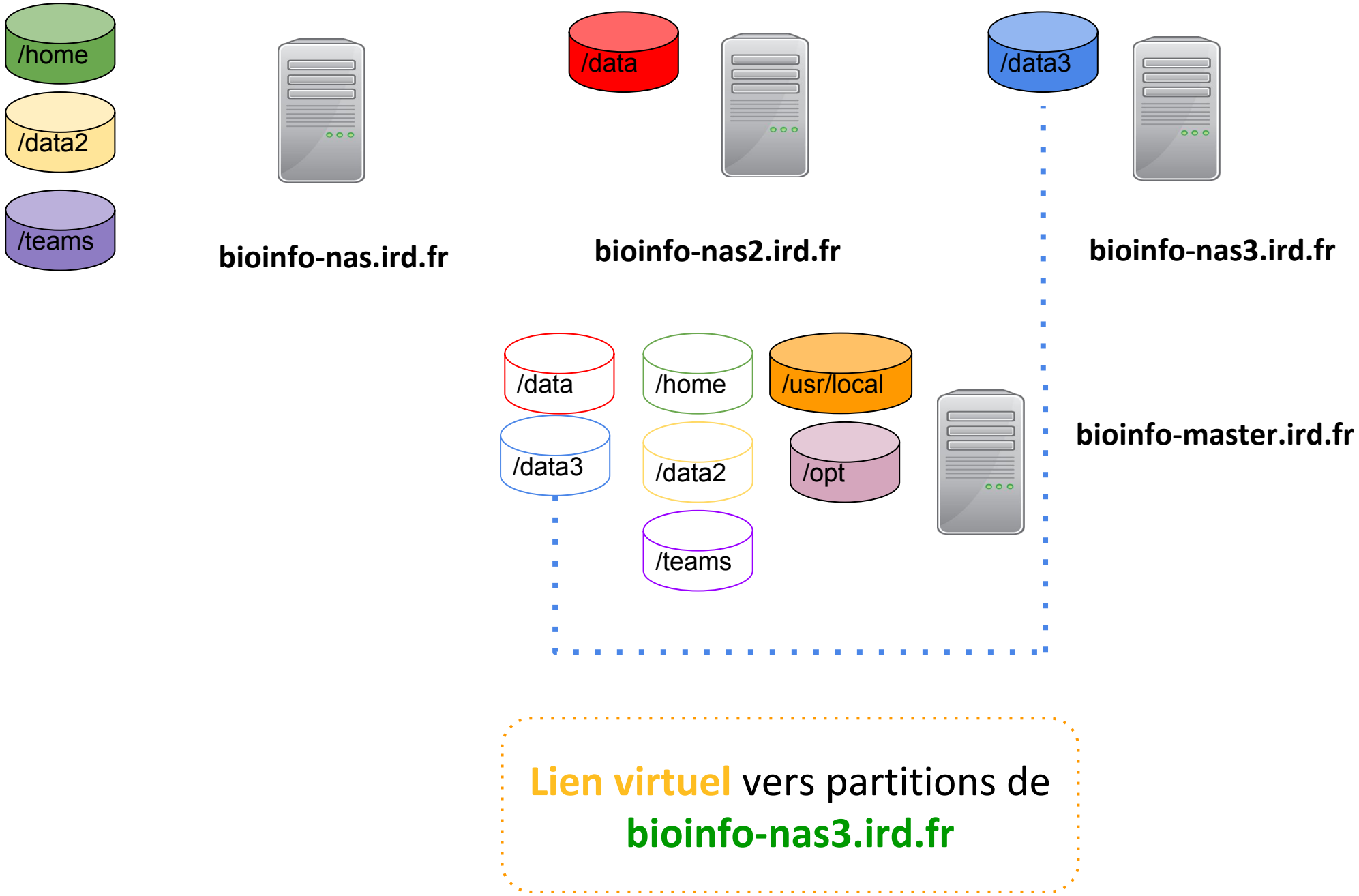
Lien virtuel vers partitions de **bioinfo-nas.ird.fr**

Partitions disques sur le cluster i-Trop

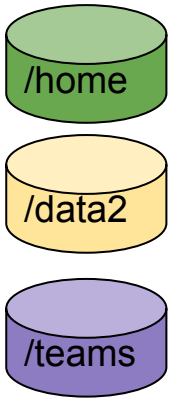


Lien virtuel vers partitions de
bioinfo-nas2.ird.fr

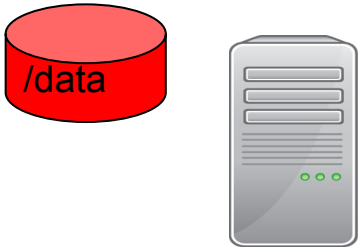
Partitions disques sur le cluster i-Trop



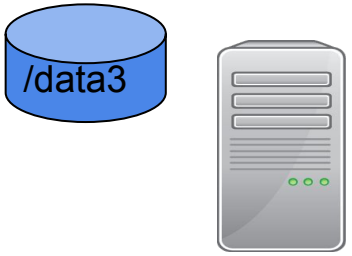
Partitions disques sur le cluster i-Trop



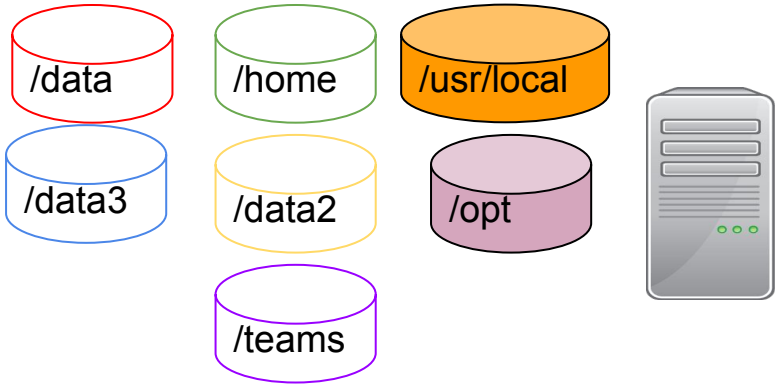
bioinfo-nas.ird.fr



bioinfo-nas2.ird.fr



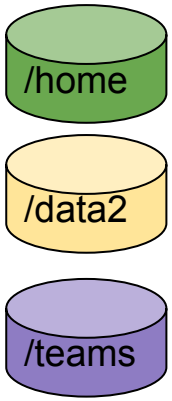
bioinfo-nas3.ird.fr



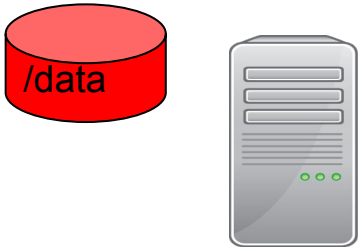
bioinfo-master.ird.fr

- Partition locale sur les noeuds : **espace temporaire**
- Disques durs physiques sur les noeuds

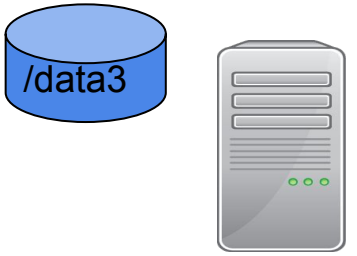
Partitions disques sur le cluster i-Trop



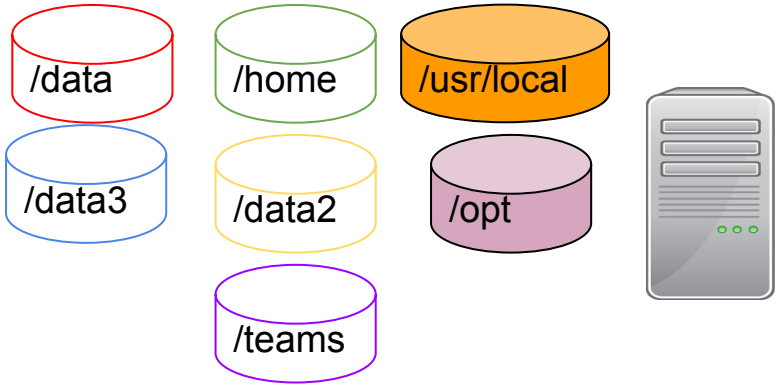
bioinfo-nas.ird.fr



bioinfo-nas2.ird.fr

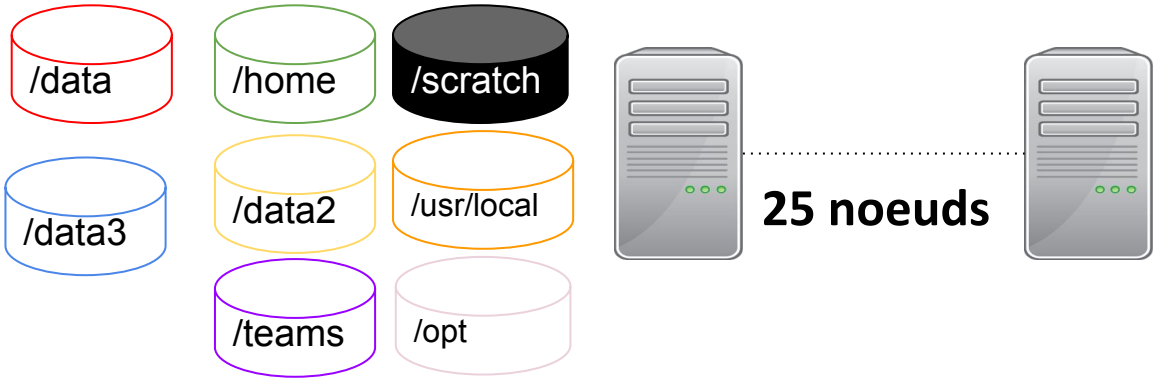


bioinfo-nas3.ird.fr



bioinfo-master.ird.fr

Liens virtuels vers les partitions des autres machines



25 noeuds

Etapes d'une analyse sur le cluster

Connexion à bioinfo-master.ird.fr et réservation de ressources

Création du répertoire d'analyse /scratch dans le noeud réservé

Etape 1

Etape 2
mkdir

Practice

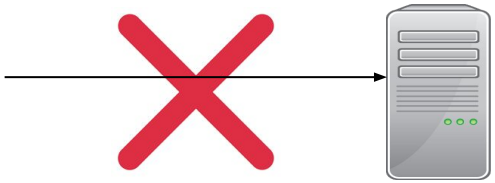
Etape 2:qrsh, partition

2

Aller sur le Practice2 *du github*

Transferts de données sur le cluster itrop

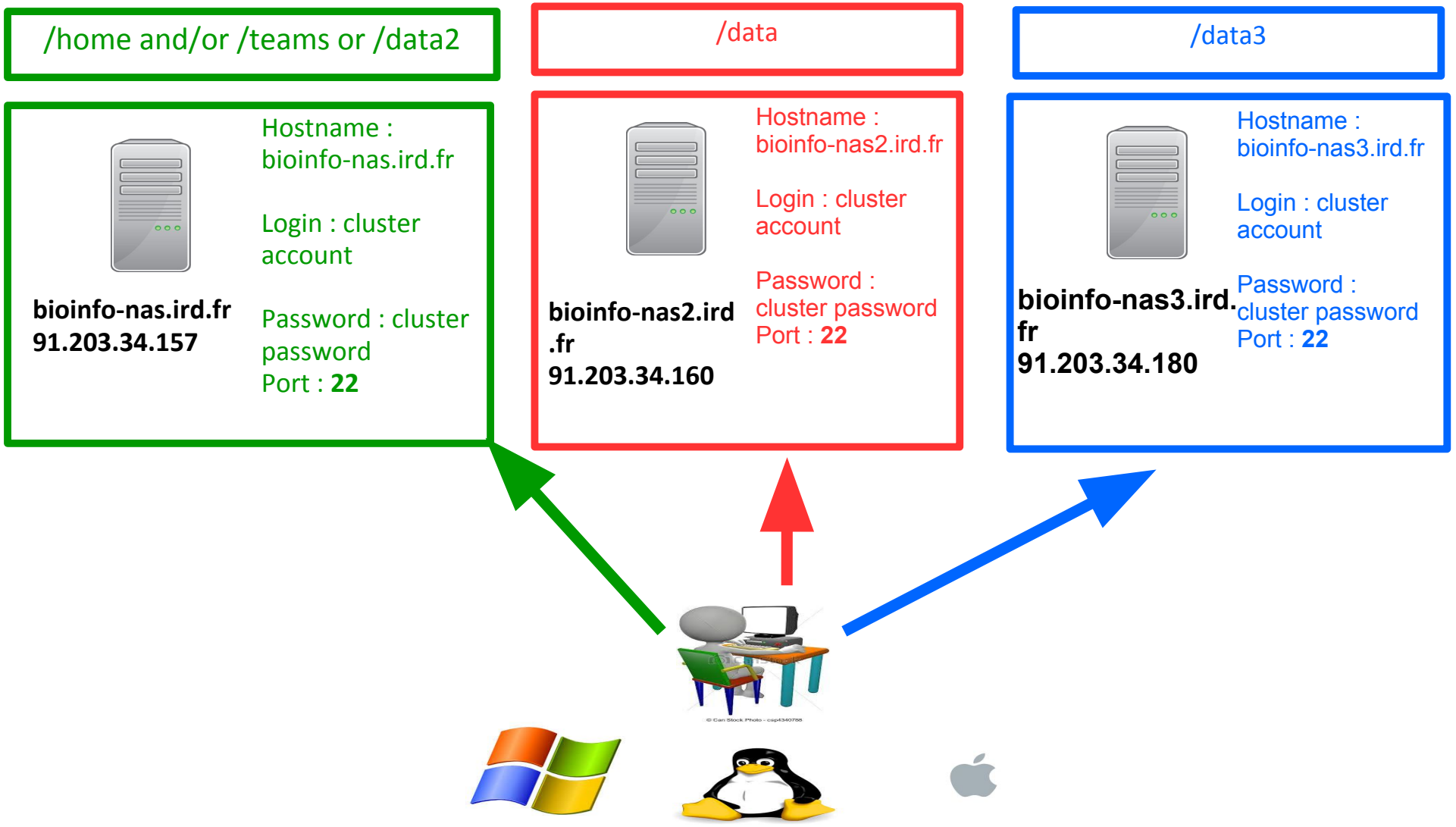
Ordinateur personnel



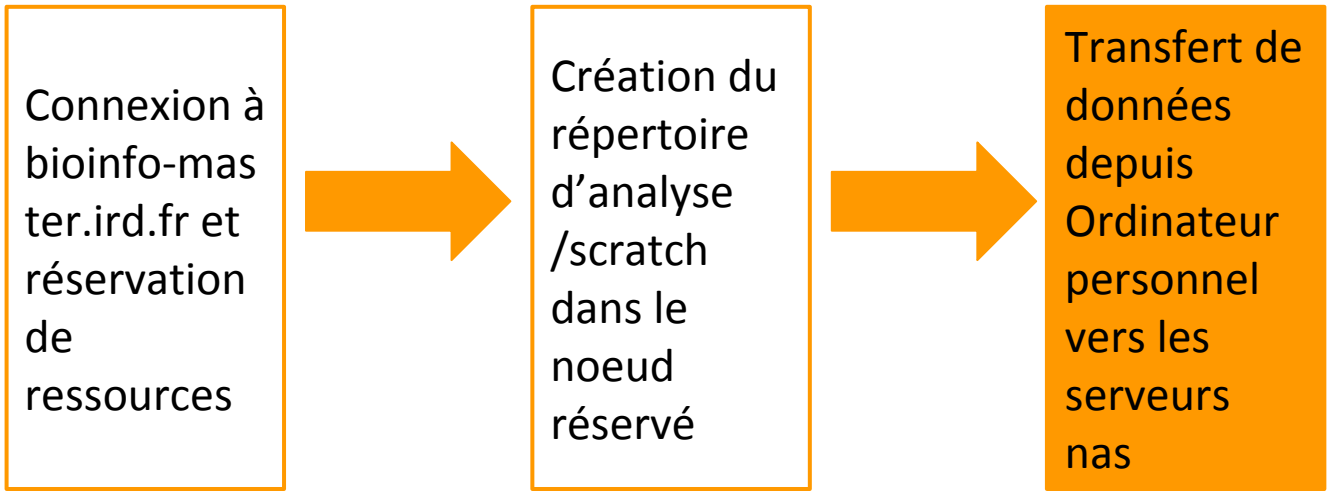
Transfert direct via filezilla interdit

bioinfo-master.ird.fr
91.203.34.148

Transferts de données sur le cluster itrop



Etapes d'une analyse sur le cluster



Etape 1

Etape 2

Etape 3
filezilla

Copier les données depuis son ordinateur personnel vers les serveurs nas si les données à analyser ne sont pas sur le cluster

Practice

Etape3: filezilla

3

Aller sur le Practice3 *du github*

La copie avec scp

- Copie entre 2 serveurs distants :

  scp source destination

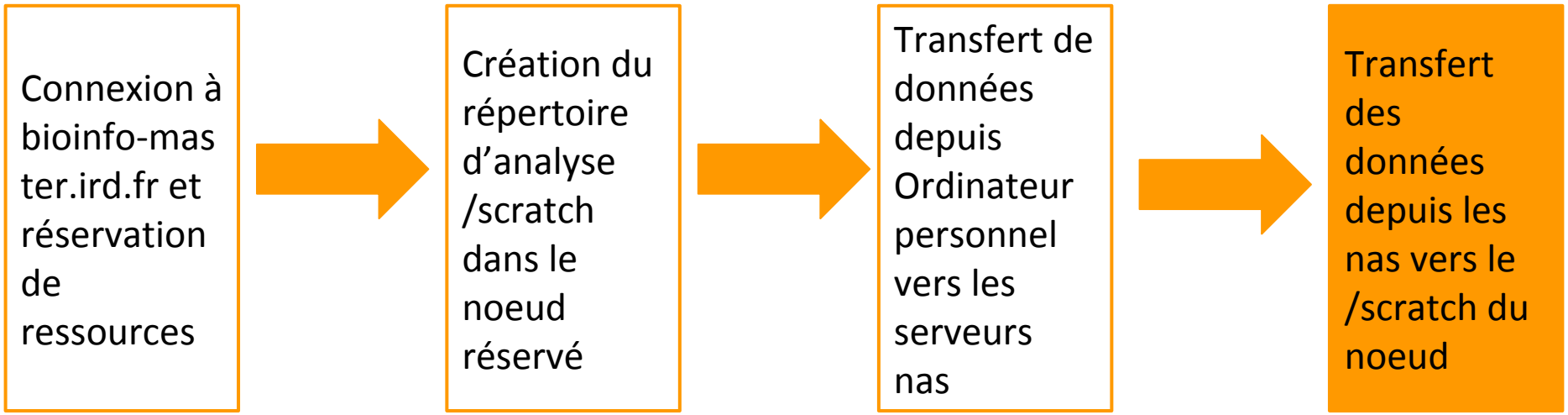
- Syntaxe si la source est distante :

  scp nom_serveur:/chemin/fichier_a_copier répertoire_local

- Syntaxe si la destination est distante :

scp /chemin/fichier_a_copier nomserveur:/chemin/répertoire_distant

Etapes d'une analyse sur le cluster



Etape 1

Etape 2

Etape 3

Etape 4

scp

Practice

Etape4: scp vers noeuds

4

Aller sur le Practice4 *du github*

Module Environment

- Permet de choisir la version du logiciel que l'on veut utiliser
- 2 types de logiciels :

  bioinfo : désigne les logiciels de bioinformatique (exemple BEAST)

  system : désigne tous les logiciels systèmes(exemple JAVA)

- Surpassent les variables d'environnement

Module Environment

➢ 5 types de commandes :

- Voir les modules disponibles :

  module avail

- Obtenir une info sur un module en particulier :

  module whatis + module name

- Charger un module :

  module load + modulename

- Lister les modules chargés :

  module list

- Décharger un module :

  module unload + modulename

- Décharger tous les modules :

  Module purge

Etapes d'une analyse sur le cluster

Connexion à bioinfo-master.ird.fr et réservation de ressources

Etape 1

Création du répertoire d'analyse /scratch dans le noeud réservé

Etape 2

Transfert de données depuis Ordinateur personnel vers les serveurs nas

Etape 3

Transfert des données depuis les nas vers le /scratch du noeud

Etape 4

Charger ses logiciels avec modules environment

Etape 5

module

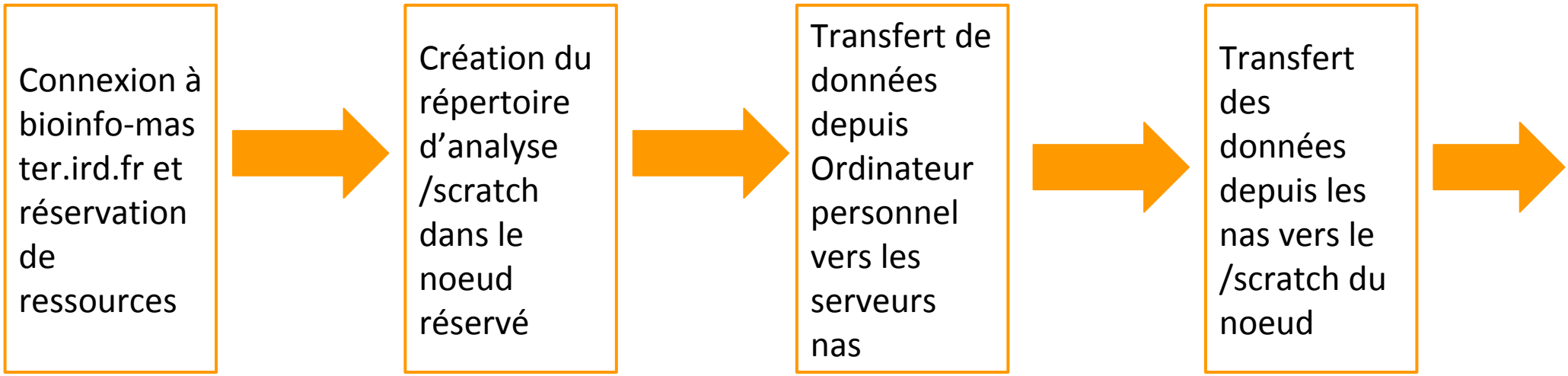
Practice

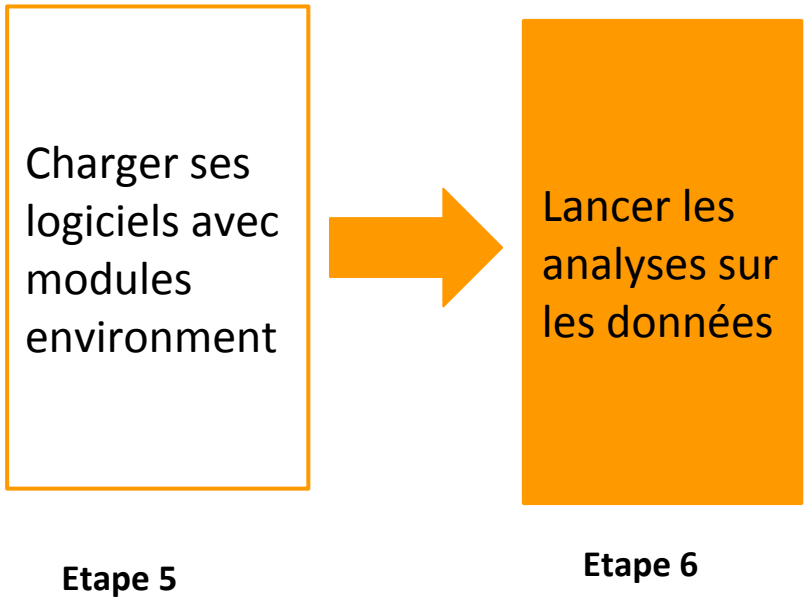
Etape5: module environment

5

Aller sur le Practice5 *du github*

Etapes d'une analyse sur le cluster





Lancer une commande depuis le prompt

- Charger la version du logiciel à lancer
- Lancer l'analyse des données

```
$~ commande <options> <arguments>
```

Avec *commande*: la commande à lancer

Lancer un job en ligne de commande

- Exécuter une commande bash via qsub
- Lance la commande sur un noeud
- On utilise la commande:

$~ **qsub** *-b y "commande"*

Avec *commande*: la commande à lancer

Options de la commande qsub

Options	Description	Exemple
qsub **-N** <name>	Donner un nom au job	qsub -N tando_blast
qsub **-q <**queue>	Choisir une queue en particulier	qsub -q highmem.q
qsub **-l hostname**=<nodeX>	Choisir un noeud en particulier	qsub -l hostname=node10
qsub **-pe** <ompi X>	Lancer un job avec plusieurs coeurs	qsub -pe ompi 4
qsub **-M** <emailaddress>	Envoyer un mail	qsub -M ndomassi.tando@ird.fr
qsub **-m** <eab>	Envoyer un mail quand: e: fin du job a: abandon b: début du job	qsub -m be
qsub -cwd	Lancer un job depuis le répertoire courant	qsub -cwd script.sh

Practice

Etape6: lancer l'analyse

6

Aller sur le Practice6 *du github*

Le transfert des résultats vers les nas

- Copie entre 2 serveurs distants :

  scp source destination

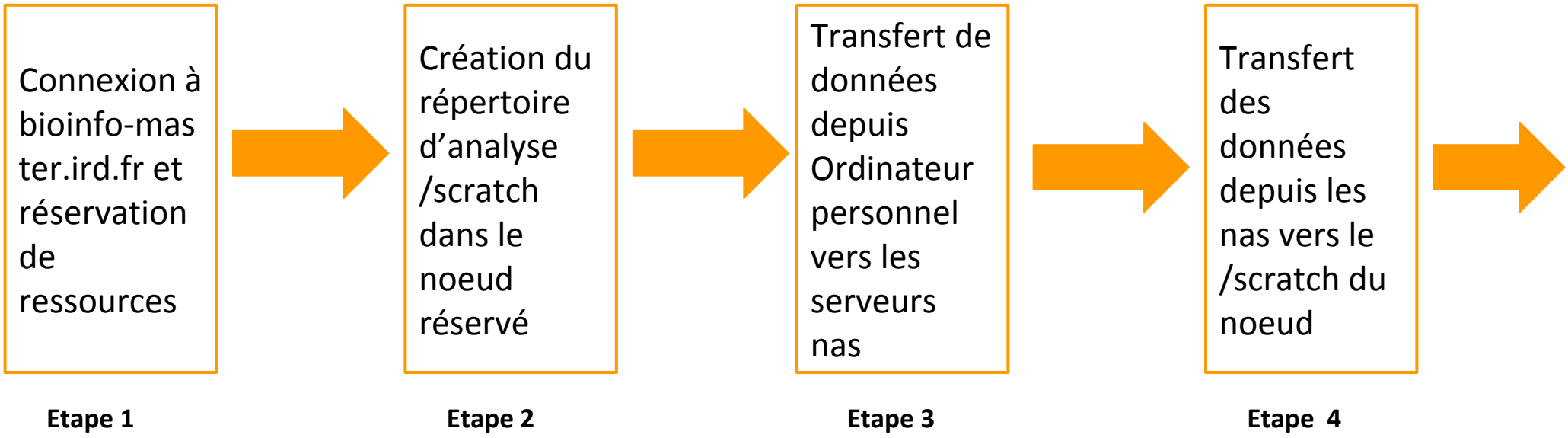
- Syntaxe si la source est distante :

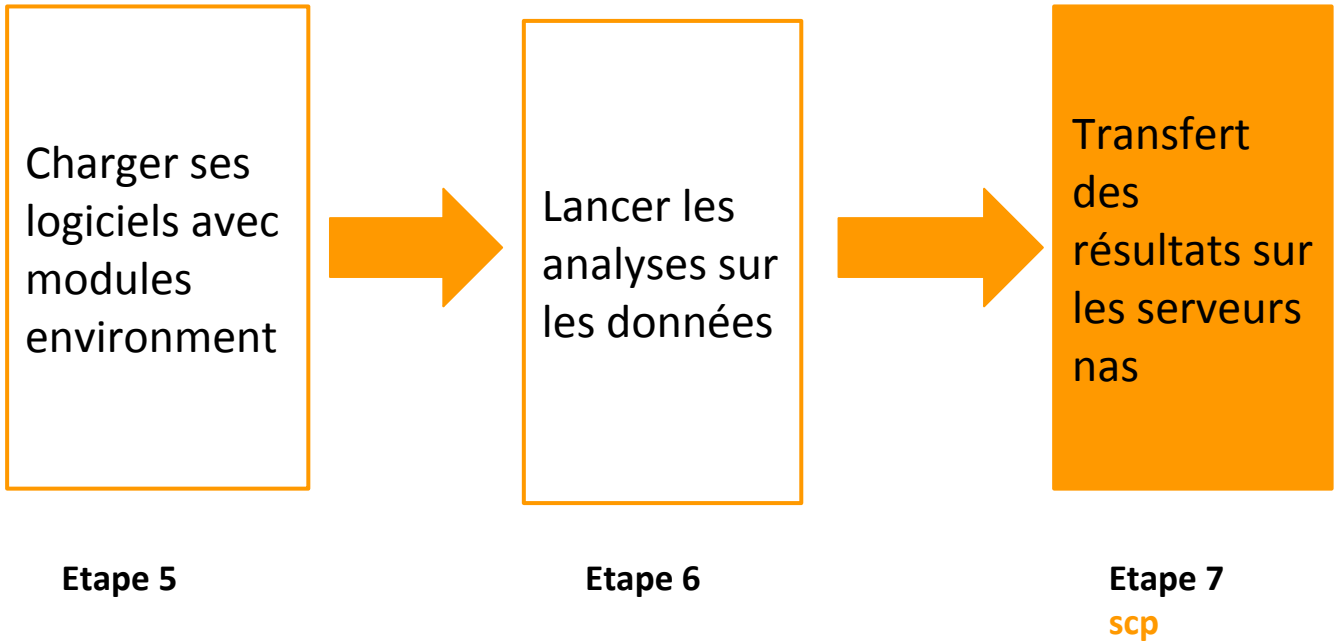
  scp nom_serveur:/chemin/fichier_a_copier répertoire_local

- Syntaxe si la destination est distante :

scp /chemin/fichier_a_copier nomserveur:/chemin/répertoire_distant

Etapes d'une analyse sur le cluster





Practice

Etape7: Récupérer les résultats

7

Aller sur le Practice7 *du github*

Supprimer les résultats des scratchs

- Scratch= espaces temporaires
- Vérifier la copie des résultats avant
- Utiliser la commande rm

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

Etapes d'une analyse sur le cluster

Connexion à bioinfo-master.ird.fr et réservation de ressources

Etape 1

Création du répertoire d'analyse /scratch dans le noeud réservé

Etape 2

Transfert de données depuis Ordinateur personnel vers les serveurs nas

Etape 3

Transfert des données depuis les nas vers le /scratch du noeud

Etape 4

Charger ses logiciels avec modules environment

Etape 5

Lancer les analyses sur les données

Etape 6

Transfert des résultats sur les serveurs nas

Etape 7

Suppression Du répertoire d'analyse sur le /scratch

Etape 8

rm

Practice

Etape8: suppression des données

8

Aller sur le Practice8 *du github*

Scripts pour visualiser/supprimer données temporaires

- Emplacement des scripts: /opt/scripts/scratch-scripts
- Visualiser ses données sur les scratchs: scratch_use.sh

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

- Supprimer ses données sur les scratchs: clean_scratch.sh

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

LANCER UN JOB

Avantages

- Le scheduler choisit les ressources automatiquement
- Possibilité de paramétrer ce choix
- Jobs lancés en arrière plan

→ possibilité d'éteindre son ordinateur

→ récupération des résultats automatique

Lancer un job en mode batch

- C'est le fait d'exécuter un script bash via sge
- On utilise la commande:

$~ **qsub** *script.sh*

Avec script.sh : le nom du script

Options de la commande qsub

Options	Description	Exemple
qsub **-N** <name>	Donner un nom au job	qsub -N tando_blast
qsub **-q <**queue>	Choisir une queue en particulier	qsub -q highmem.q
qsub **-l hostname**=<nodeX>	Choisir un noeud en particulier	qsub -l hostname=node10
qsub **-pe** <ompi X>	Lancer avec plusieurs coeurs	qsub -pe ompi 4
qsub **-M** <emailaddress>	Envoyer un mail	qsub -M ndomassi.tando@ird.fr
qsub **-m** <eab>	Envoyer un mail quand: e: fin du job a: abandon b: début du job	qsub -m be
qsub -cwd	Lancer un job depuis le répertoire courant	qsub -cwd script.sh

Syntaxe des scripts bash

Dans la première partie du script on renseigne les options d'exécution de sge avec le mot clé #$ (partie en vert)

```
#!/bin/sh


#############     SGE CONFIGURATION     ####################
# Ecrit les erreur dans le fichier de sortie standard
#$ -j y

# Shell que l'on veut utiliser
#$ -S /bin/bash



# Email pour suivre l'execution
#$ -M prenom.nom@ird.fr         ######### Mettre son adresse mail

# Type de message que l'on reçoit par mail
#    -  (b) un message au demarrage
#    -  (e) a la fin
#    -  (a)  en cas d'abandon
#$ -m bea

# Queue que l'on veut utiliser
#$ -q formation.q

# Nom du job
#$ -N Nom_a_choisir
#################################################################
```

Syntaxe des scripts bash

Dans la 2e partie du script on renseigne les actions à effectuer

```
path_to_dir="/data/projects/rep_a_choisir";
path_to_tmp="/scratch/nom_rep_a_choisir-$JOB_ID"

###### Creation du repertoire temporaire sur noeud et chargement du module blast
module load bioinfo/blastn/2.4.0+
mkdir $path_to_tmp
scp -rp nas2:$path_to_dir/* $path_to_tmp # choisir nas pour/home, /data2 et /teams ou nas2 pour /data ou nas3 pour /data3
echo "tranfert donnees master -> noeud";
cd $path_to_tmp

###### Execution du programme
cmd="blastn -db All-EST-coffea.fasta -query  sequence-NMT.fasta -num_threads $NSLOTS  -out blastn1-$JOB_ID.out";
echo "Commande executee : $cmd";
$cmd;

##### Transfert des données du noeud vers master
scp -rp $path_to_tmp/ nas:$path_to_dir/
echo "Transfert donnees node -> master";

#### Suppression du repertoire tmp noeud
rm -rf $path_to_tmp
echo "Suppression des donnees sur le noeud";
```

Practice

Lancer un script avec qsub

9

Aller sur le Practice9 *du github*

Citations

Si vous utilisez les ressources du plateau i-Trop.

Merci de nous citer avec:

“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”

Projets

- Pensez à inclure un budget ressource de calcul dans vos réponses à projets

- Besoin en disques dur, renouvellement de machines etc…

- Devis disponibles

- Contactez bioinfo@ird.fr : aide, définition de besoins, devis...

Formateurs

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








Merci pour votre attention !