Computing resources: IFB cloud and PSMN

Description, Usefulness, Guidelines

Laurent Gilquin (IGFL) & Carine Rey (CIRI)

November 14th 2023

Table of contents

IFB cloud (Biosphere)

Introduction

The IFB: French Institute of Bioinformatics

"It is the National Bioinformatics Infrastructure that provides support, deploys services, organizes training and carries out innovative developments for the life sciences communities."

The IFB cloud: This is one of the services provided by the IFB

"IFB is providing cloud services to analyze life science data. These services rely on **a federation of clouds** - **Biosphere** - built on interconnected IT infrastructures of some IFB's platforms."

Documentation: https://www.france-bioinformatique.fr/en/ifb-cloud/

1. Create an account and/or log in on **Biosphere** web page https://biosphere.france-bioinformatique.fr/
2. Deploy a VM based on a template (an **appliance**) and a configuration (a **flavor**)
3. Connect to your VM according the appliance
   - with a web interface (HTTPS)
   - with the SSH command line interface through a terminal window
   - with a remote graphical desktop (X2Go)
4. Upload your data, scripts, ...
5. Analyse your data (1hour -> 1 day -> 1 week -> ...)
6. Download results
7. Shutdown the VM (!)

Complete Documentation : https://ifb-elixirfr.github.io/biosphere/

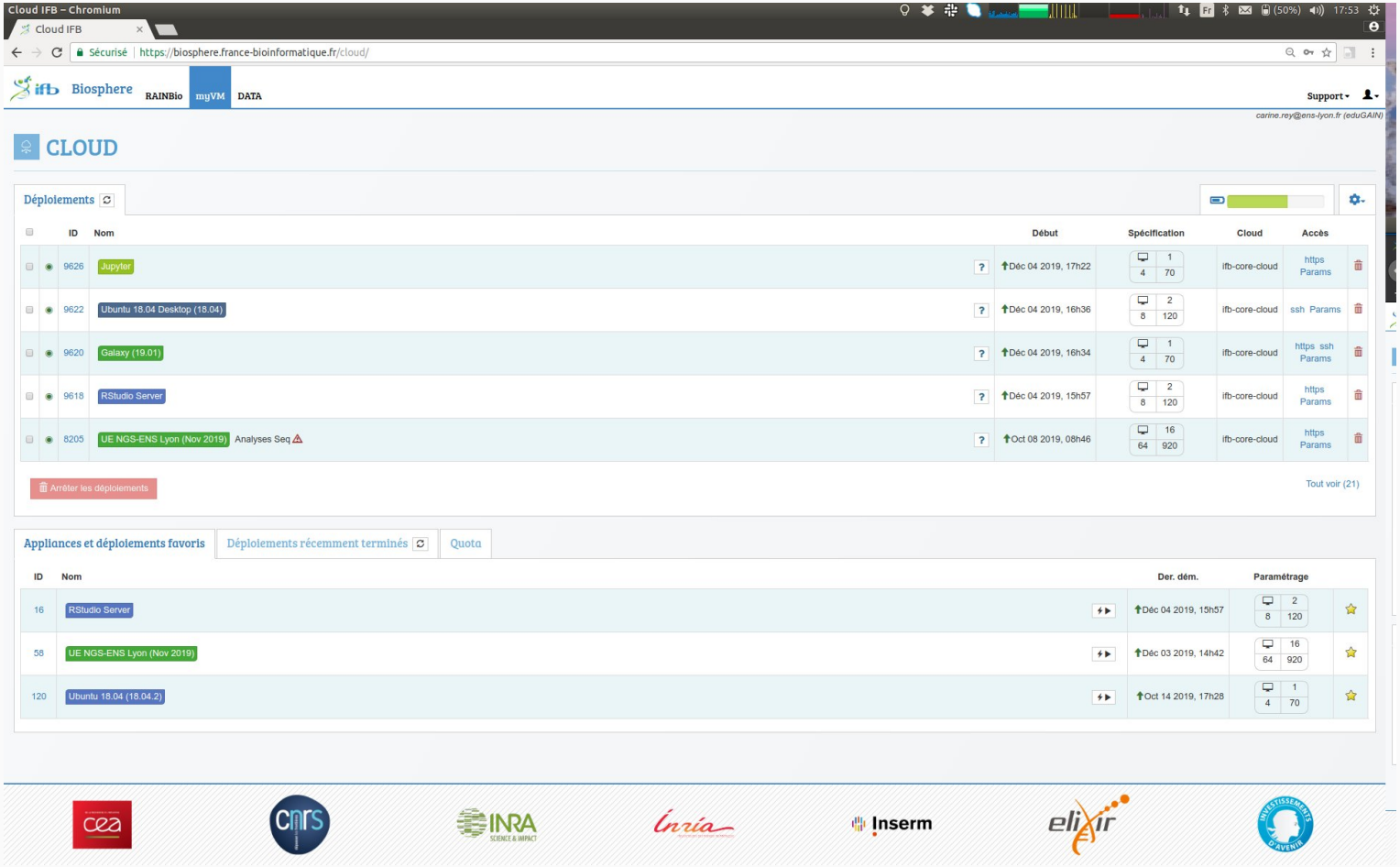
IFB cloud (Biosphere)

In practice

- Go to https://biosphere.france-bioinformatique.fr/
- At the top right, click on **login**.
- Click again on **login**
- Click on **accept conditions**
- Select **ENS de Lyon|CNRS|Inserm|...** according your employer
- Identify yourself with your institutional **login and password**
- Fill in **your name, first name, city and postal code**, leaving all other information as default and accept.
- You'll get a new page. In the top right-hand corner, click **on the little man icon**, then on group.
- Click on **Join a group** in the tabs at top left.
- Search for the group corresponding to your laboratory *LBMC/IGFL/CIRI/RDP/...* and click on the **+** button to apply.
- And you're all set! Now , your application need to be validated by the IFB

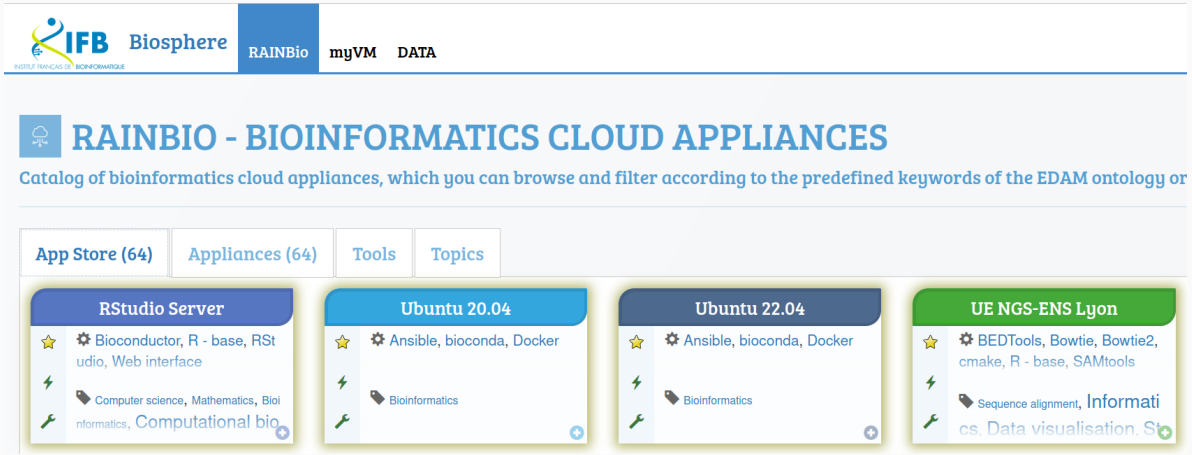
1. 2 | The dashboard:

- Once, your account is approved you can access to your dashboard:



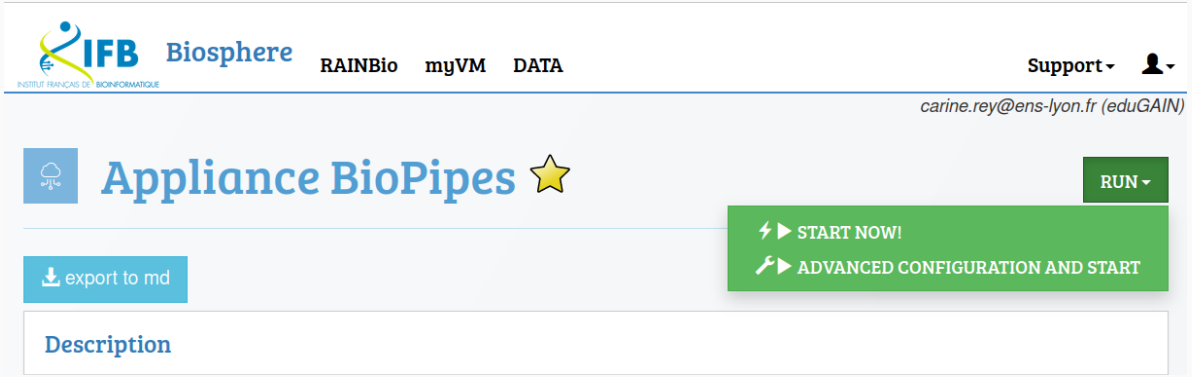
Direct link: https://biosphere.france-bioinformatique.fr/cloud/

1) Click on **RainBio** in the head bar to select an appliance according your need and click on it

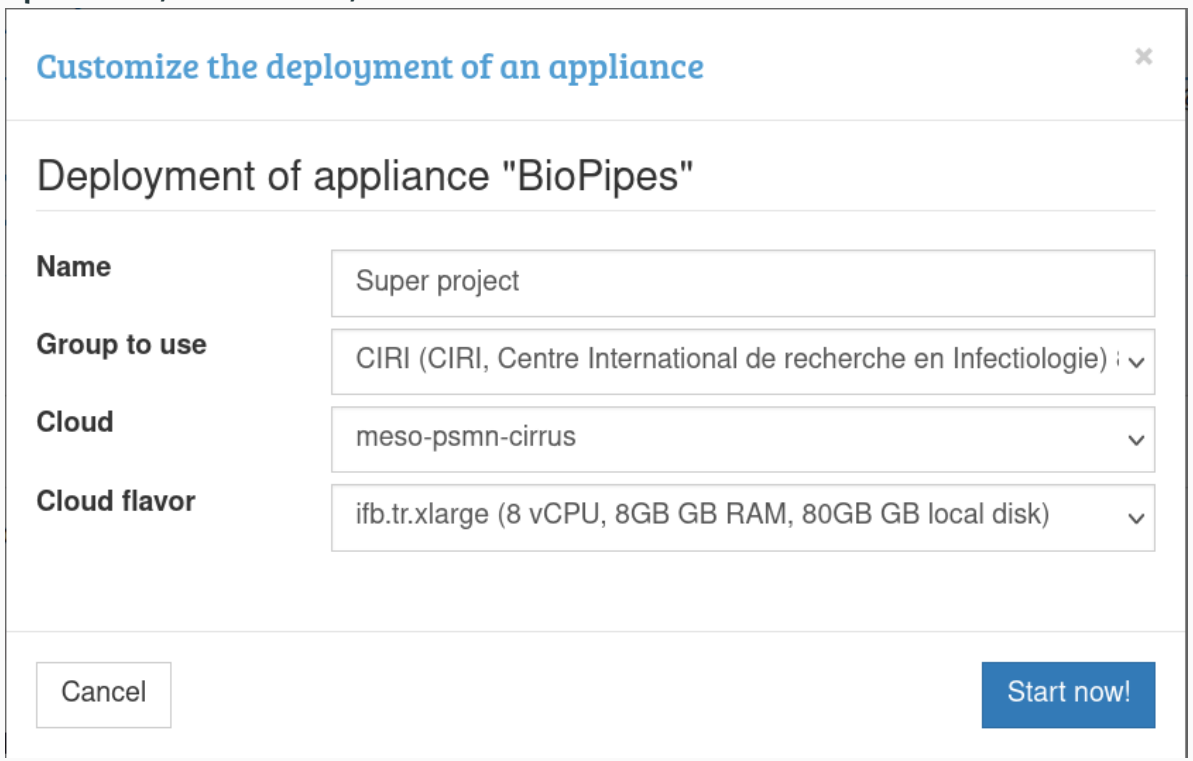


2) Check the description and click on **Run** in the top right corner

3) Chose **Advanced configuration and Start**

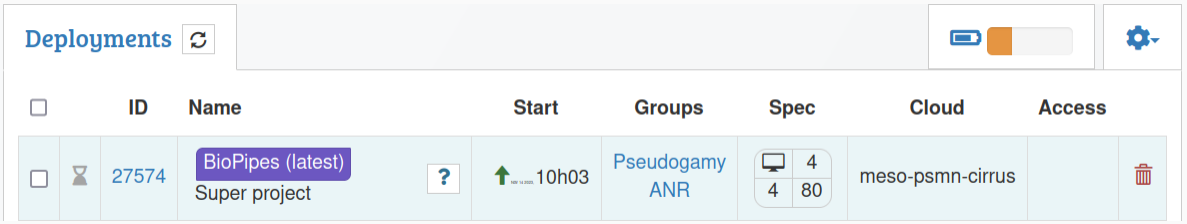


4) Configure your VM by specifying: a **name**; a **group** from which take quotas; a **cloud**; a **flavor**

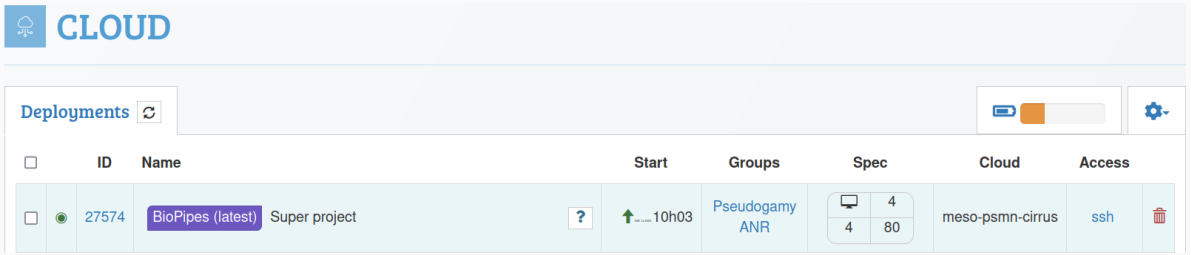


5) Click on **Start Now!**, you will be redirected to your dashboard.

6) You can see on your dashboard your new VM in deployement (grey hourglass)



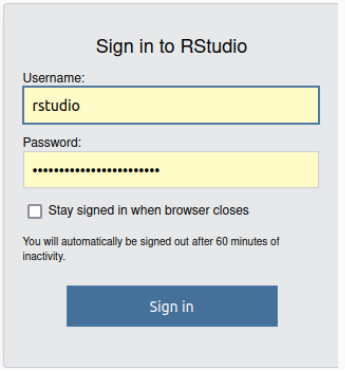
7) Wait for the complete deployement (15-20min; green light) and get the connexion parameters on the right



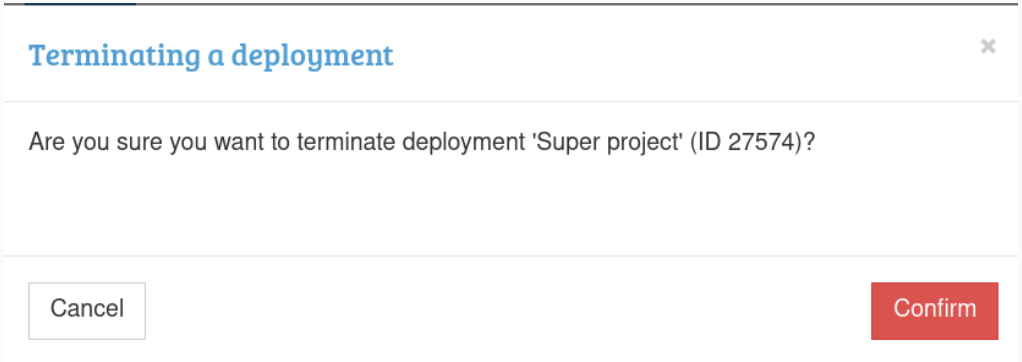
8) Connexion parameters will differ according your appliance

9) Sometime the deployement may go wrong (red light), try again.
10) Using the connexion parameters, you can log in on your VM

```
(base) admincarine@bibs-3571-cr:~$ ssh ubuntu@192.33.153.182
Welcome to Ubuntu 22.04.3 LTS (GNU/Linux 5.15.0-69-generic x86_64)
```



11) When you have finished with your VM, you will click on **the trash**.

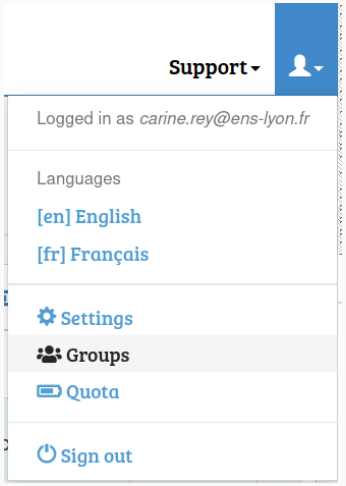


! Important

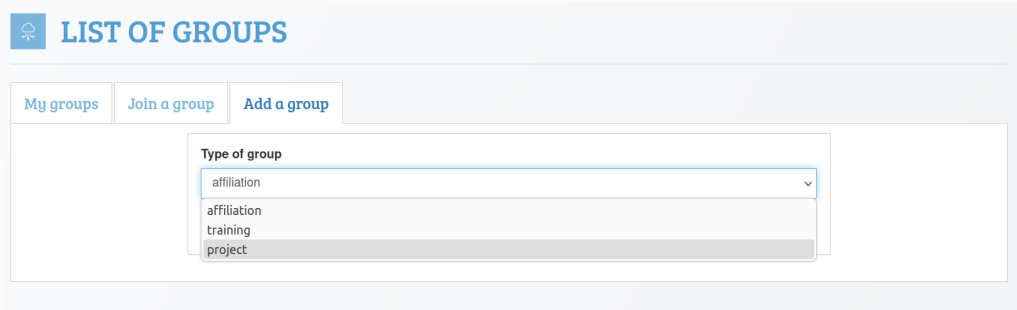
This action is irreversible, all the data on the VM will be destroyed.

- They are 2 types of quotas:
  - your personnal quota (small: by default 30.000 vCPU.h)
  - quotas for the groups you belong to (may be big)
- Quotas are counted in **vCPU.h**, which means that a VM with 8 CPUs will cost 8 vCPU.h each hour.
- If you use group quotas, the quota total is shared between all the people who belong to the group.

- You can get an extention of quotas by creating your group
- Click on the **little man icon** on the top right corner and **Group**



- Click on **Add a group** and select **project**

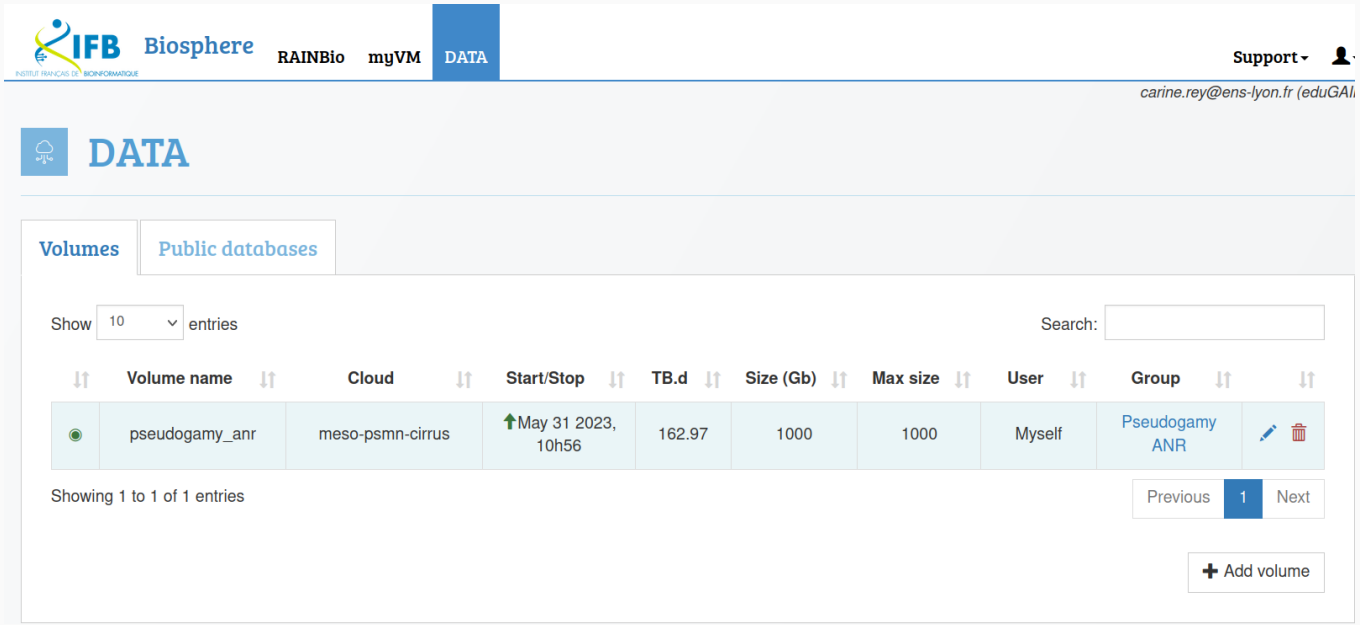


- Fill the **form** and wait for its acceptance (15 days)

Tip

Using a group is the only way to have persistent and shared storage mounted on each VM (of the group).
(For the moment only for meso-psmn-cirrus)

- You can view your persistent storage in the **Data** tab in the head bar
- Warning: A volume is attached to a cloud

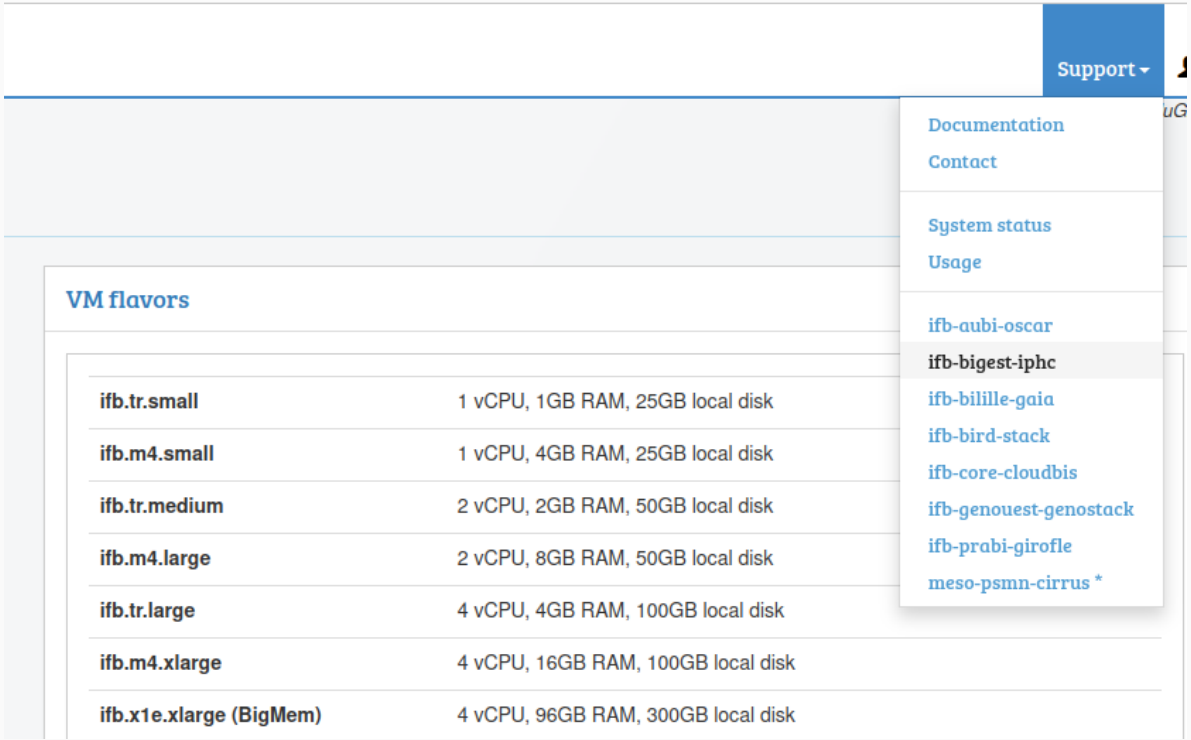


Platform	Location	Calcul (#CPU HT*)	Storage (#TB)	RAM (#GB)
IFB Core	Lyon (CC-IN2P3)	3 616	180	20 116
AuBI	Clermont-Ferrand (Université)	384	12	1 536
GenOuest	Rennes (IRISA)	600	350	2 600
PRABI	Lyon (Université)	536	144	3 300
BiRD	Nantes (Université)	860	150	2 500
BIGEst	Strasbourg (IPHC)	1024	500	4 000
BILILLE	Lille (Université)	192	0	768
CBP-PSMN	Lyon (ENS-Lyon)	3 408	24	7 552
Total federation of clouds		**10 620**	**1 360**	**43 372**

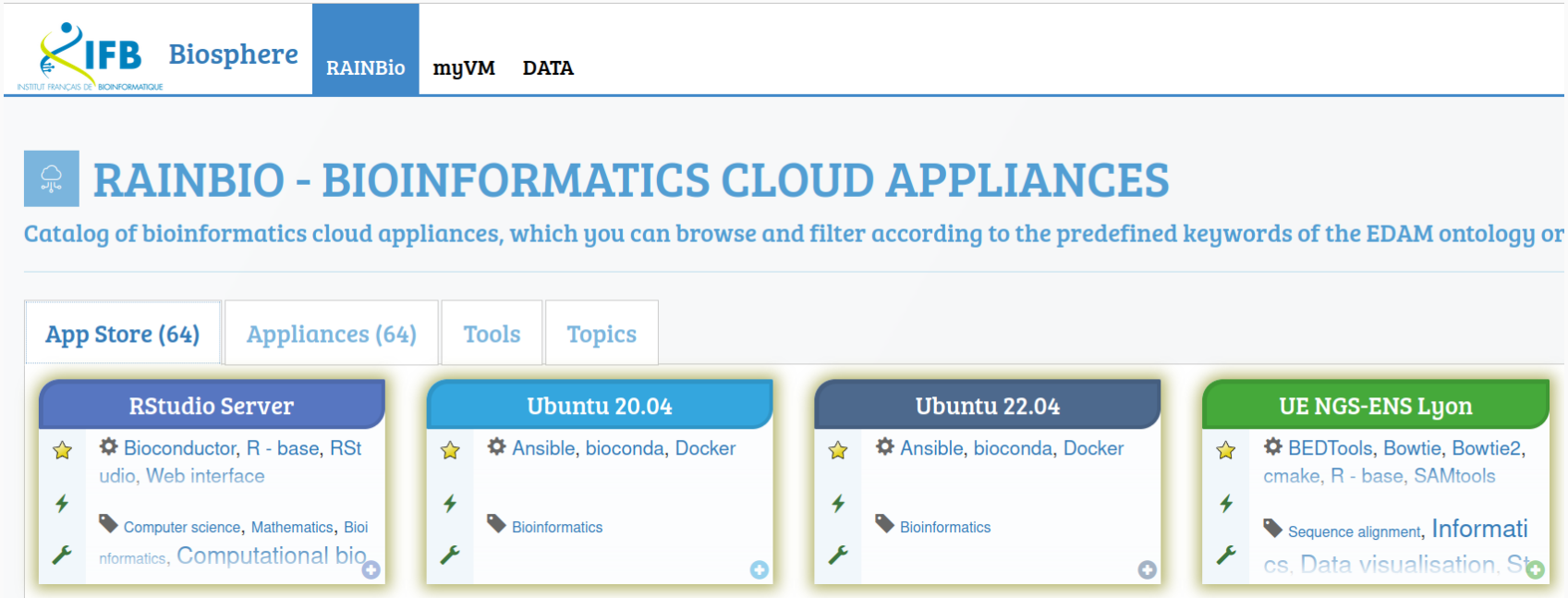
* CPU HT = Hyperthreading CPU

- You can use any of them (If one of them is undergoing maintenance, you can use another)
- Each of them will provide different VM sizes called **flavors**

- You can view the different **flavors** available by clicking on "Support" and choosing a cloud.
- If you require a specific configuration (flavor), do not hesitate to contact the support.



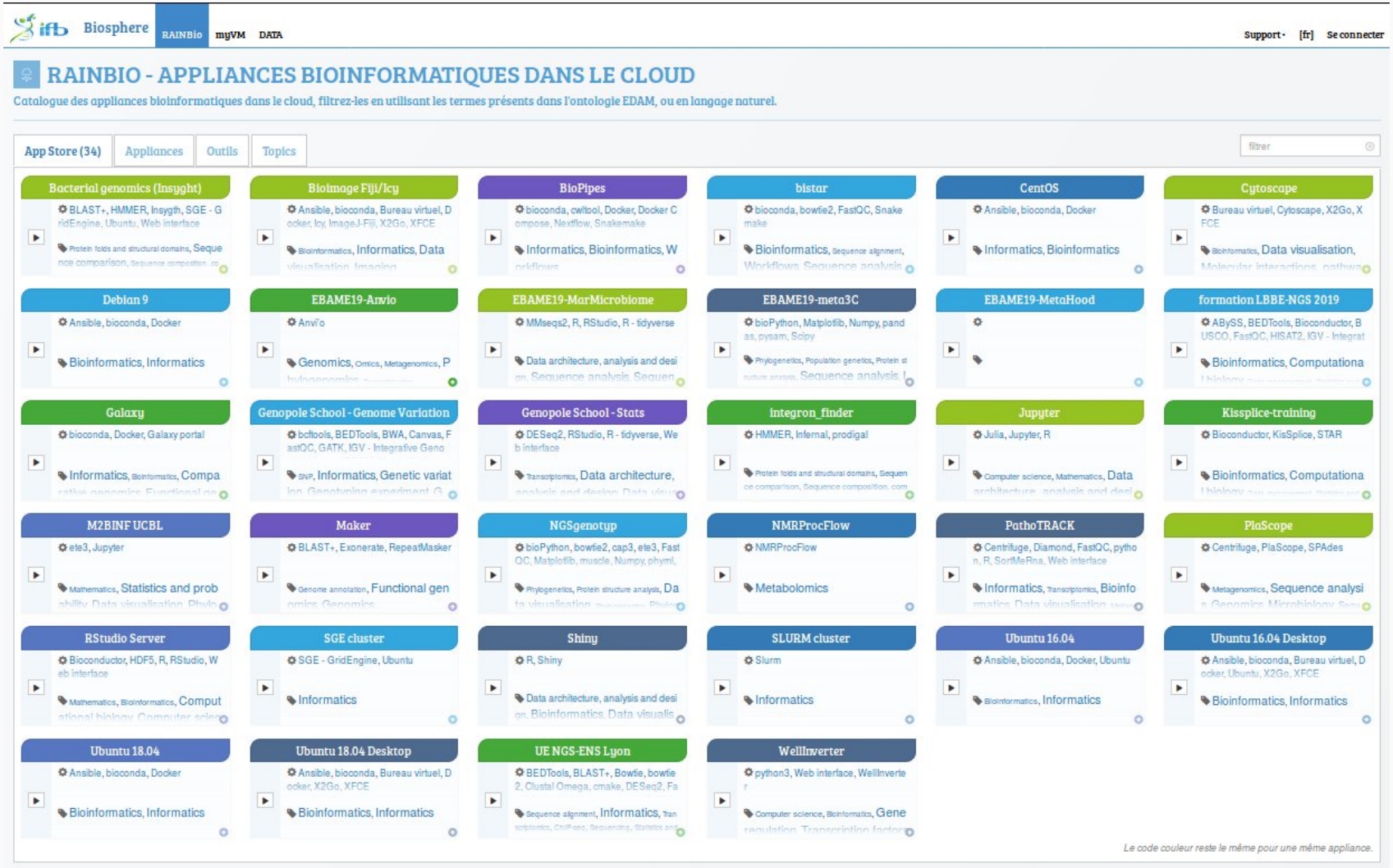
- You can use each appliance in all VM flavors / cloud
- If you require a specific appliance, you can built it, do not hesitate to contact the support. (Require advanced skills)



IFB cloud (Biosphere)

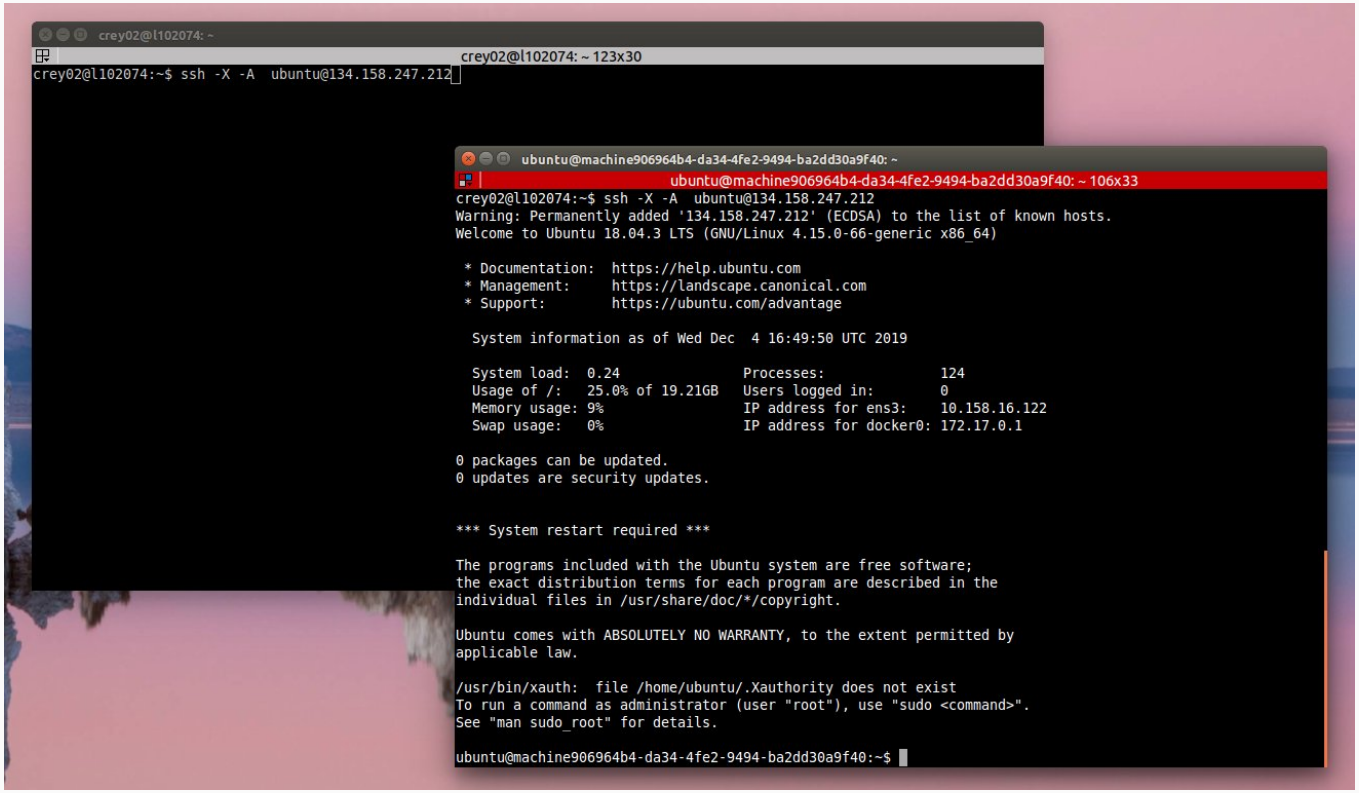
The RainBio catalogue

1. 3 | The different appliances available : the RAINBio catalogue

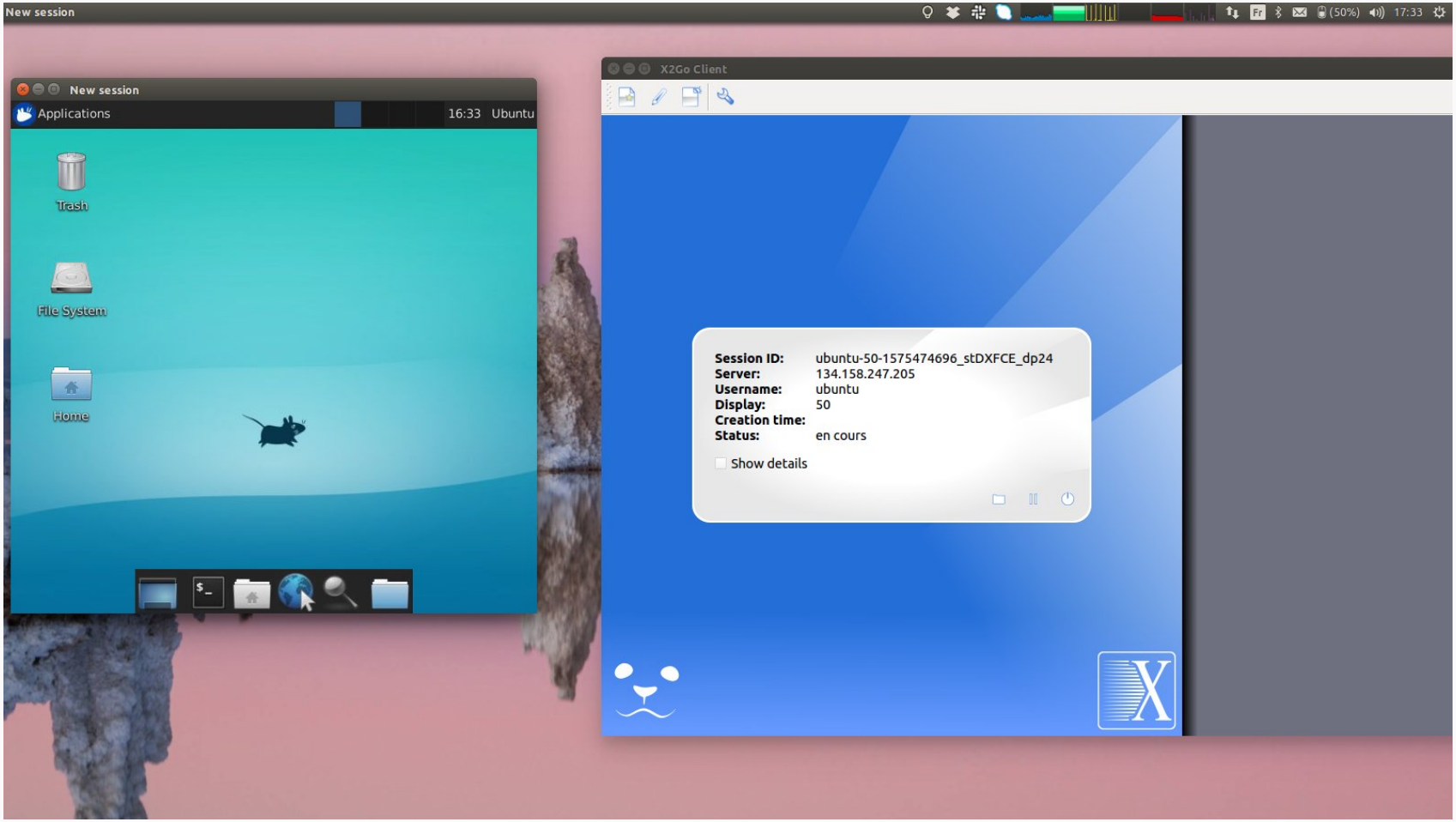


https://biosphere.france-bioinformatique.fr/catalogue/

- Filter by tools available by clicking on the ”Tools” tab

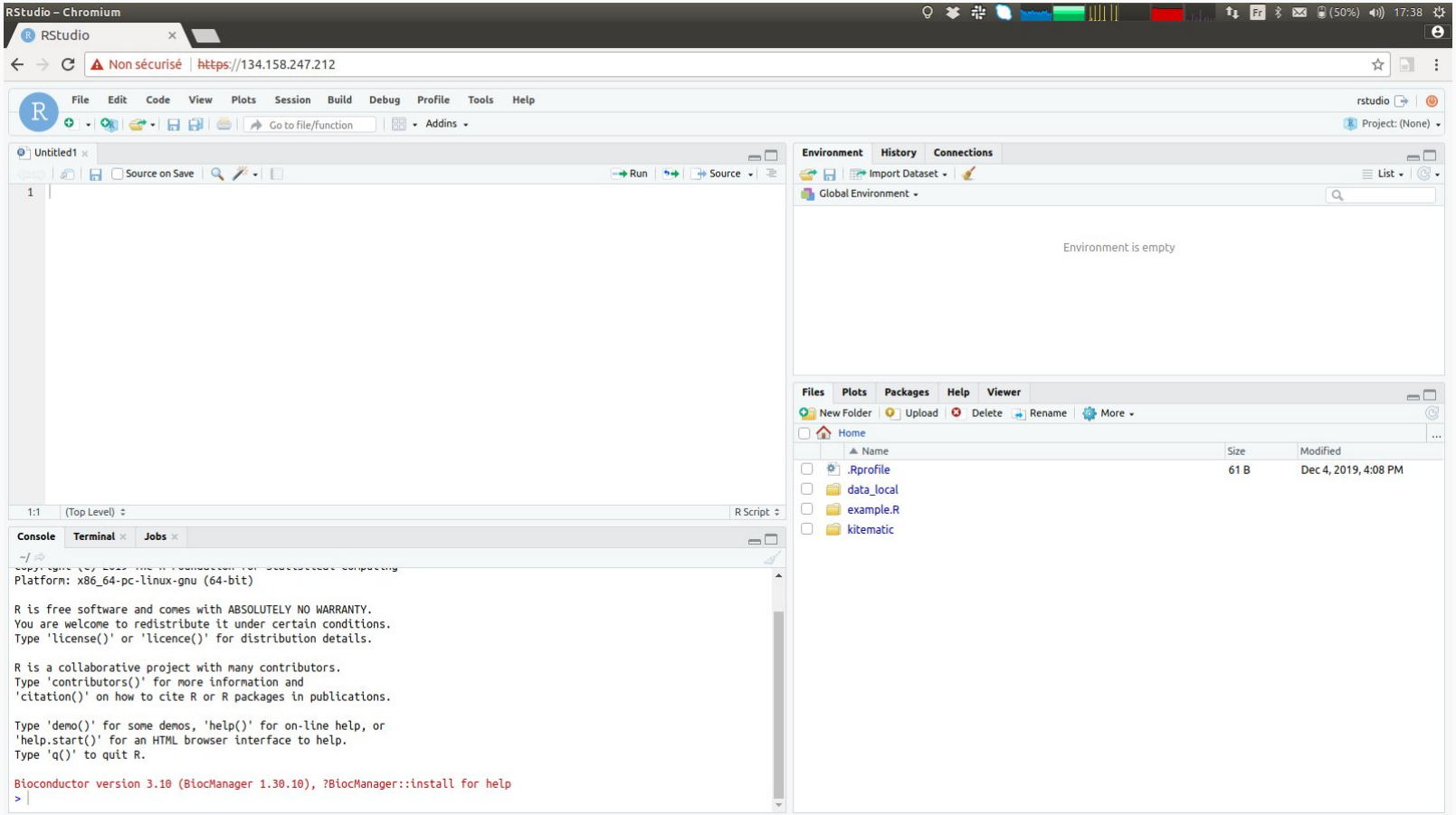


- Need ssh keys configuration **before** deployement : https://ifb-elixirfr.github.io/biosphere/vm_connect
  ”See: Annex - Configure your SSH parameters”



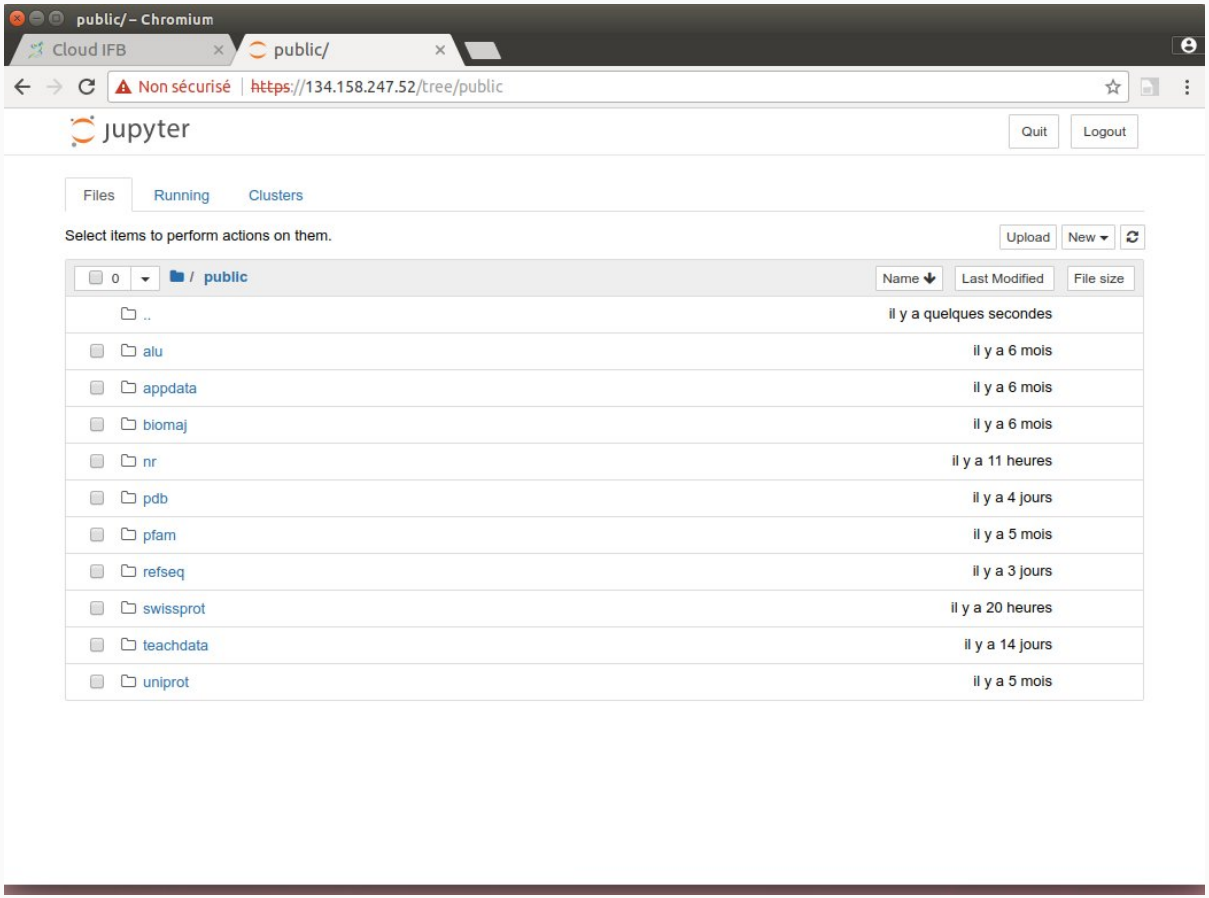
- Need ssh keys and install X2GO (whatever your operating system, Windows, Mac, Linux): https://ifb-elixirfr.github.io/biosphere/vm_connect "See: Connecting with X2Go to a VM"

- Many tools already pre-installed (check the catalogue)
  - For example: BioPipes appliance is a Biosphere-commons app providing usual bioinformatics pipeline tools (nextflow, snakemak, cwltool)
- Easy installation of new tools
  - Conda
  - Docker



Access via your browser (no ssh keys required) :

- you must accept the safety warning
- sometime, you must also modify your firewall rules



Access via your browser (no ssh keys required) :

- you must accept the safety warning
- sometime, you must also modify your firewall rules

Documentation: https://ifb-elixirfr.github.io/biosphere/

Useful links:

- Sign in
  https://ifb-elixirfr.github.io/biosphere/signin
- Deploy your VM
  https://ifb-elixirfr.github.io/biosphere/vm_deploy
- Connecting to your VM
  https://ifb-elixirfr.github.io/biosphere/vm_connect
- Transferring data with your VM
  https://ifb-elixirfr.github.io/biosphere/data

IFB cloud (Biosphere)

Conclusion

- Access to large-scale computing resources and/or specific ready to use environments
- Testing/learning the Linux environment without fear of breaking your machine
- Administrator rights on your machines
- No queue, VMs are directly available (For the moment)
- Direct access to public reference databases (RefSeq, nr, Uniprot, PDB, ...)
- Very useful for training
- No backup of your data, so be sure to keep copies locally
- An often unknown resource that has changed the way I do bioinfo

PSMN

Introduction

PSMN: Scientific Pole of Numerical Modelling

"The High Performance Computing (HPC) facility of ENS Lyon that hosts a large computational power and provide a fast and flexible access to it."

Organization chart

PSMN	CBP
Ralf EVERAERS	
Loïs TAULELLE Cerasela Iliana CALUGARU Micaël CALVAS	Emmanuel QUEMENER Micaël CALVAS

New since 2023: PSMN + CBP $\rightarrow$ CBPsmn

- administrative change
- platform & tools identical (for now)

PSMN users

- ENS Lyon biology labs: LBMC, RDP, CIRI, IGFL
- other fields: Chemistry, Geology, Physics, Informatics, Astrophysics

Useful links

→ PSMN user documentation

→ News feed: server issues, upgrades, maintenance, and so on

→ contact forms: signaling an issue, request a software update/installation, . . .

Resources

→ LBMC guide for PSMN/CBP

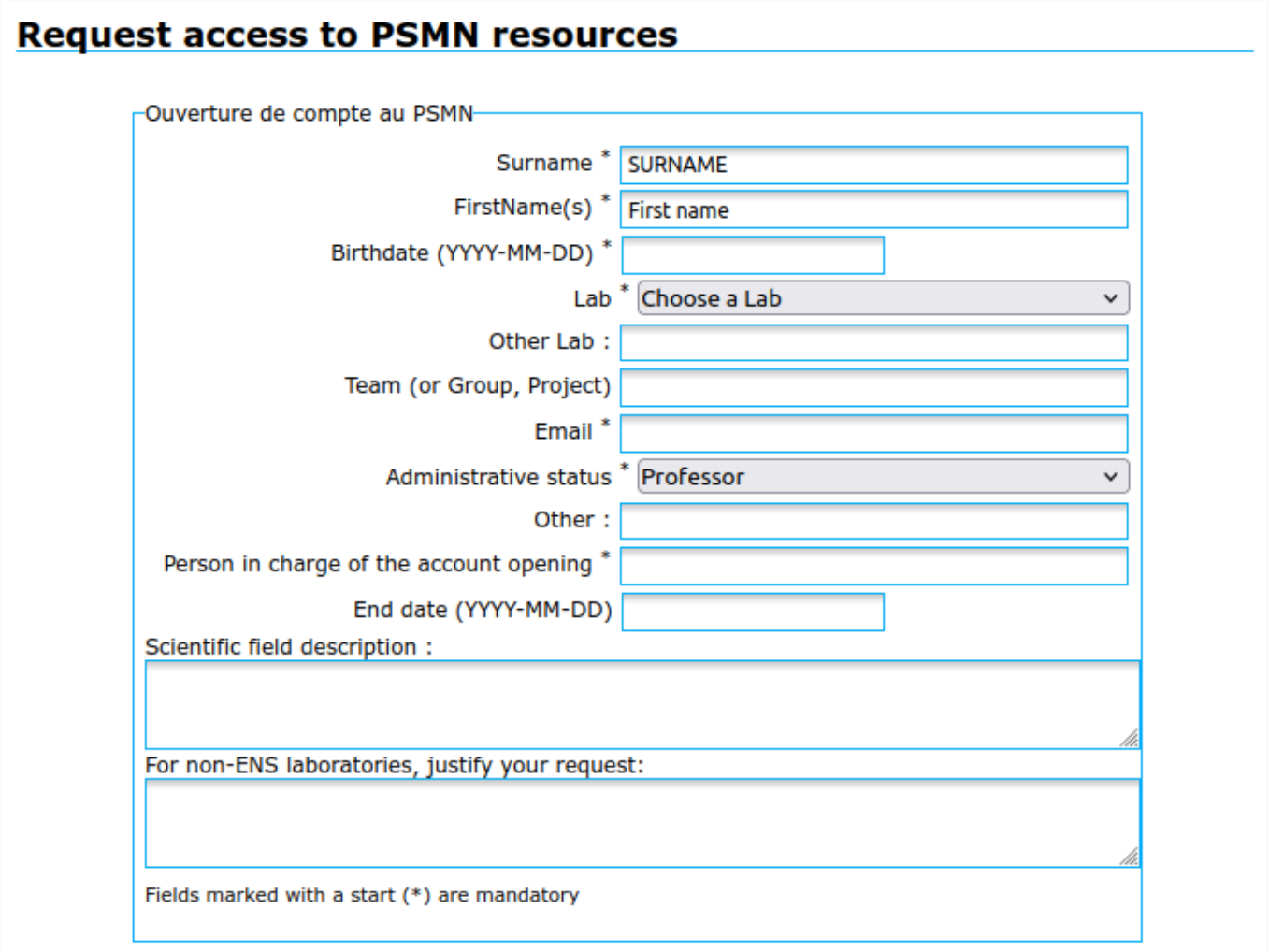
→ IGFL Practical introduction for beginners

Common use cases

- Process large volume of data (10^2GB $\sim$ 10TB)
  examples: sequence (huge FASTQ) or image analysis (3D stack $\times$ time)
- Run many tasks
  example: performing the same task with $\neq$ input file
- Long-running time task (several days)
- Exploit parallelism (single-node) or distributed computing (multi-nodes)
- Share files:
  (i) databases and genomes
  (ii) collaborative project requiring several TB

Prerequisites

1) Request an account through the form



Request access to PSMN resources

Ouverture de compte au PSMN

Surname * SURNAME

FirstName(s) * First name

Birthdate (YYYY-MM-DD) *

Lab * Choose a Lab

Other Lab :

Team (or Group, Project)

Email *

Administrative status * Professor

Other :

Person in charge of the account opening *

End date (YYYY-MM-DD)

Scientific field description :

For non-ENS laboratories, justify your request:

Fields marked with a start (*) are mandatory

Prerequisites

1) Request an account through the form

2) An SSH client:

 : native

: native

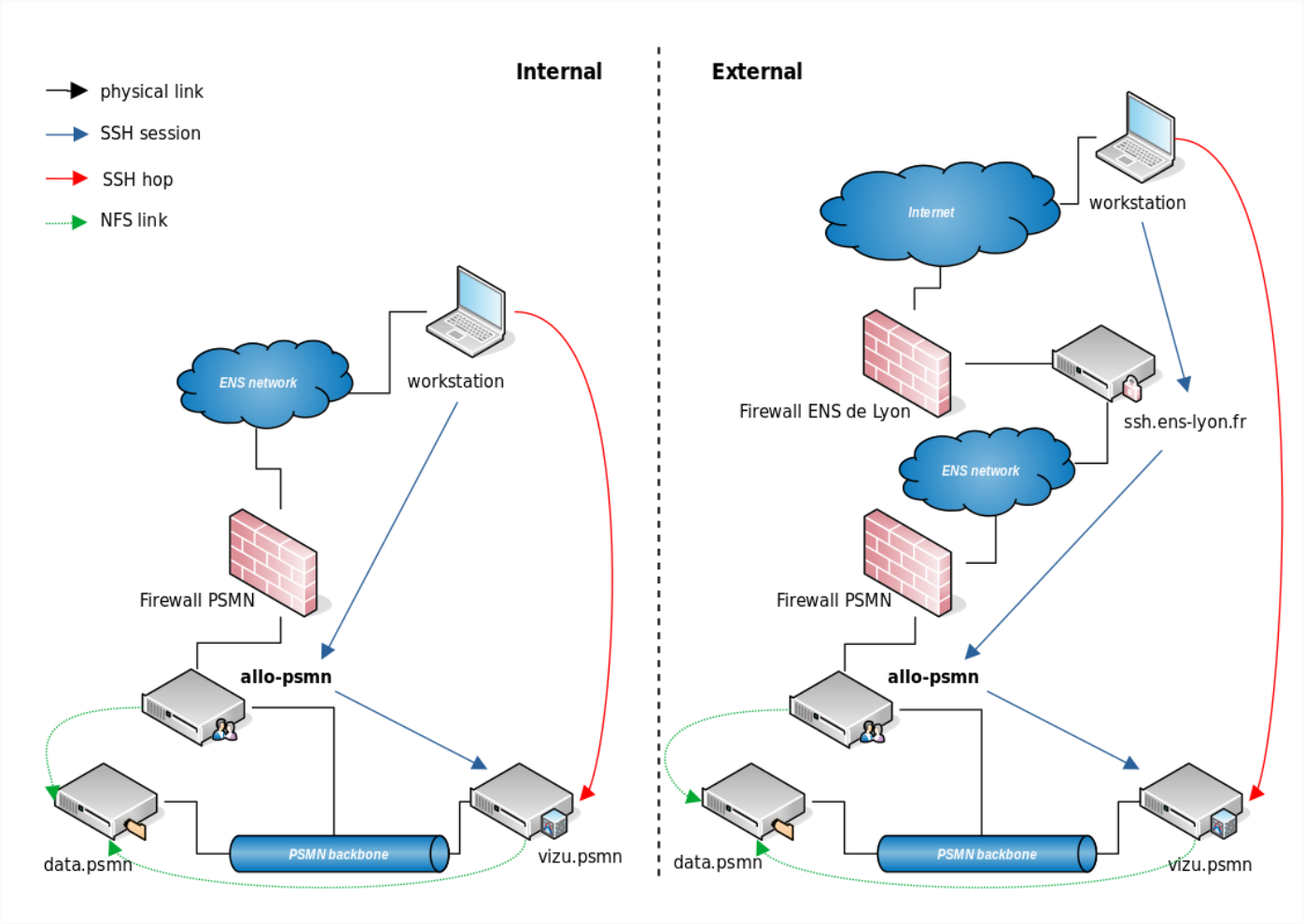
 : OpenSSH (native since Win10 v.1803), Putty

3) Knowledge of basic terminologies

4) Basics of shell scripting (CAN UNIX training or tutorial)

$\Rightarrow$ PSMN mandatory training ($\approx$ 1h)

Connection servers



- internal gataway $\Leftrightarrow$ the ENS network
- external gataway $\Leftrightarrow$ outside the ENS network

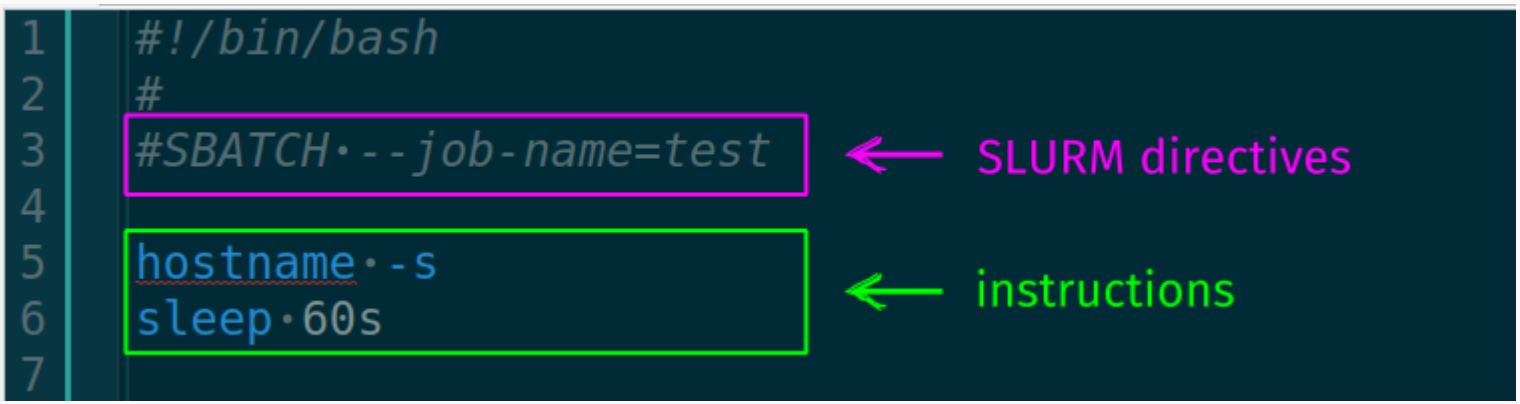
Job submission

Job

A **shell** script that contains a series of directives and instructions to be run

On a **shared** cluster, you submit a job to a **scheduler**, here **SLURM**, that will decide if and where the job can be run

sbatch myjob.sh



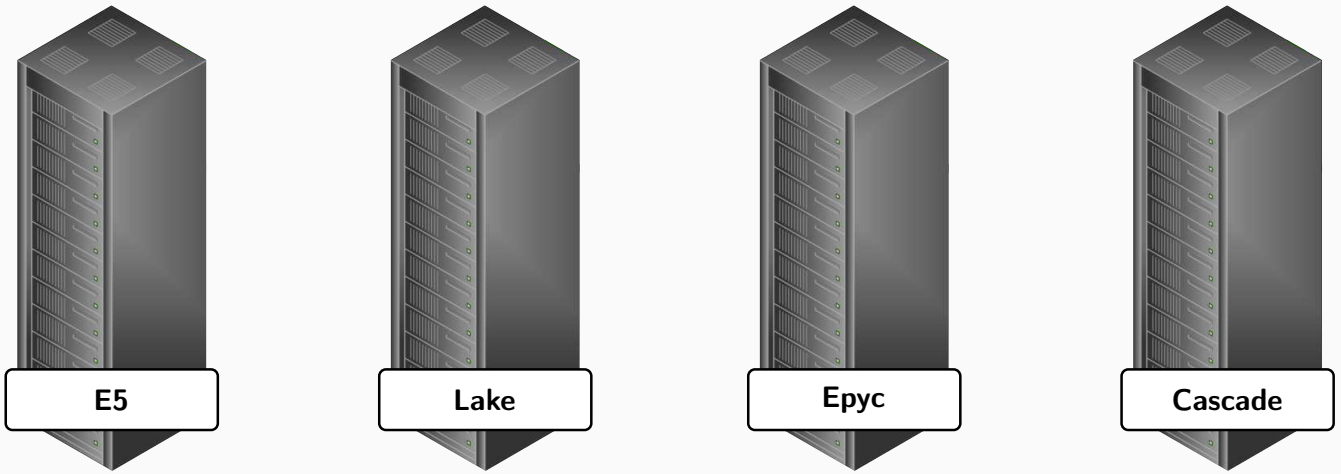
PSMN

Overview

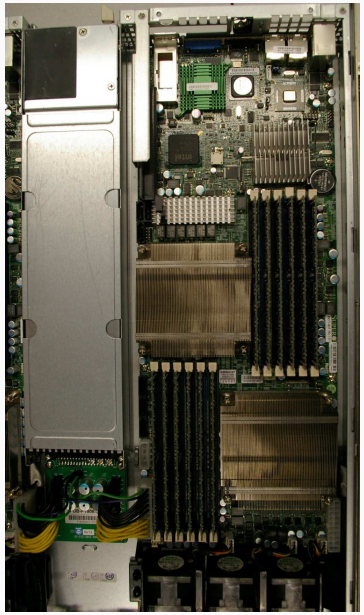
The PSMN is composed of 4 **clusters**

Cluster

A set of computers linked to each other so that their **resources** can be **pooled** together

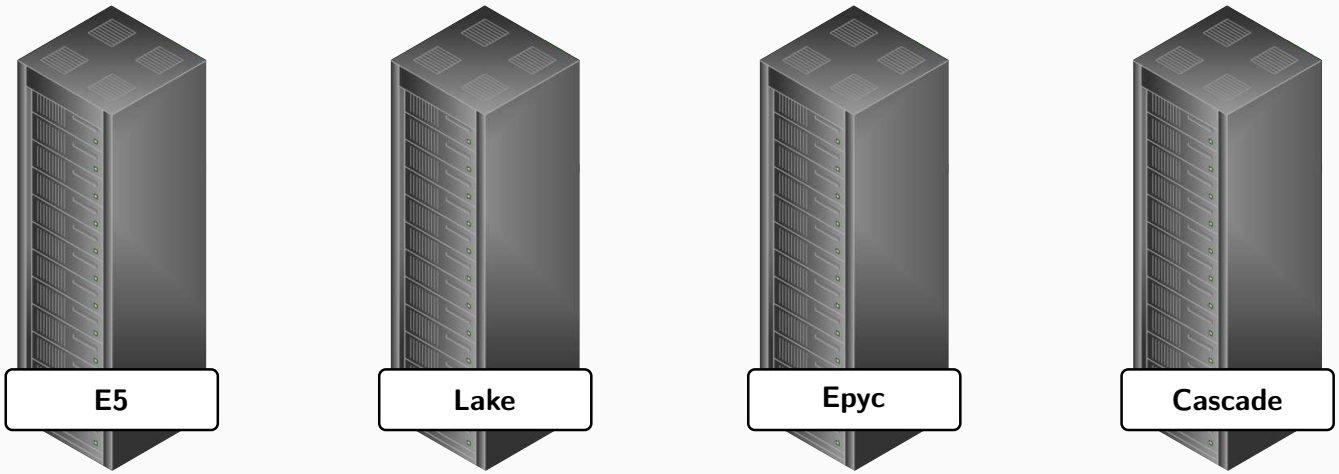


Each cluster contains several **compute nodes**



Compute node

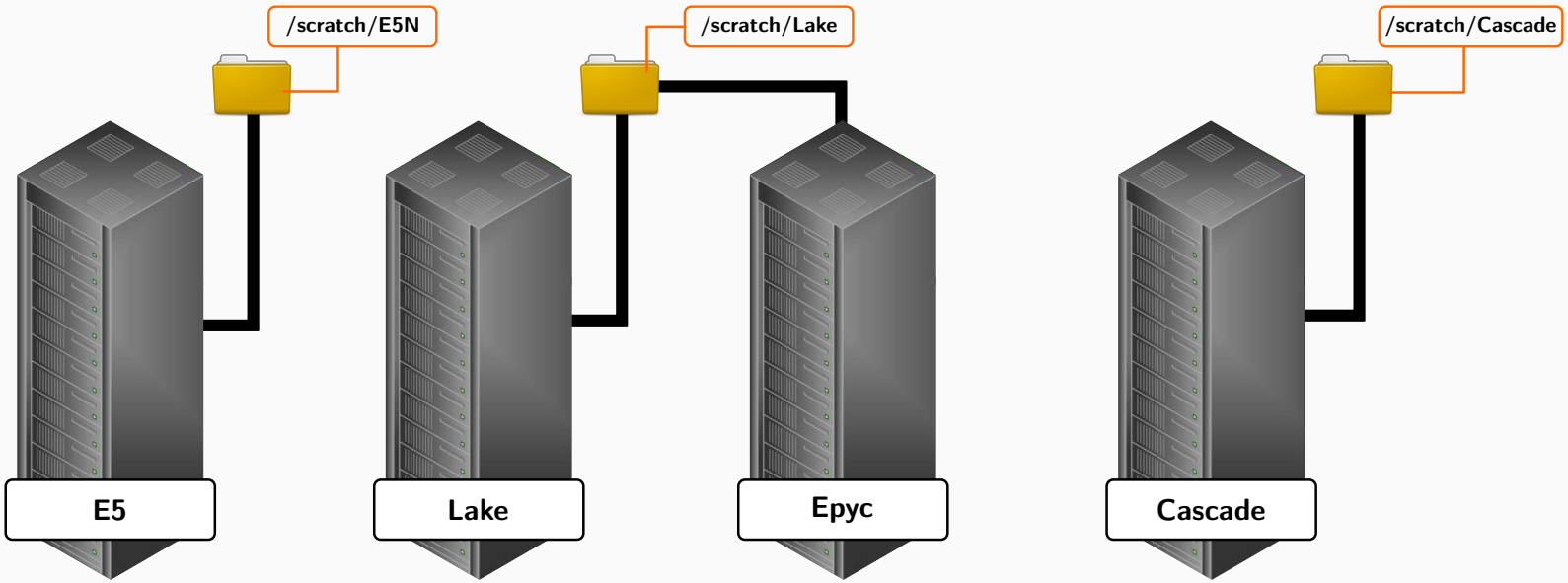
A unit of a cluster acting as a provider of **computing resources** (CPU, RAM) for **jobs** to run on



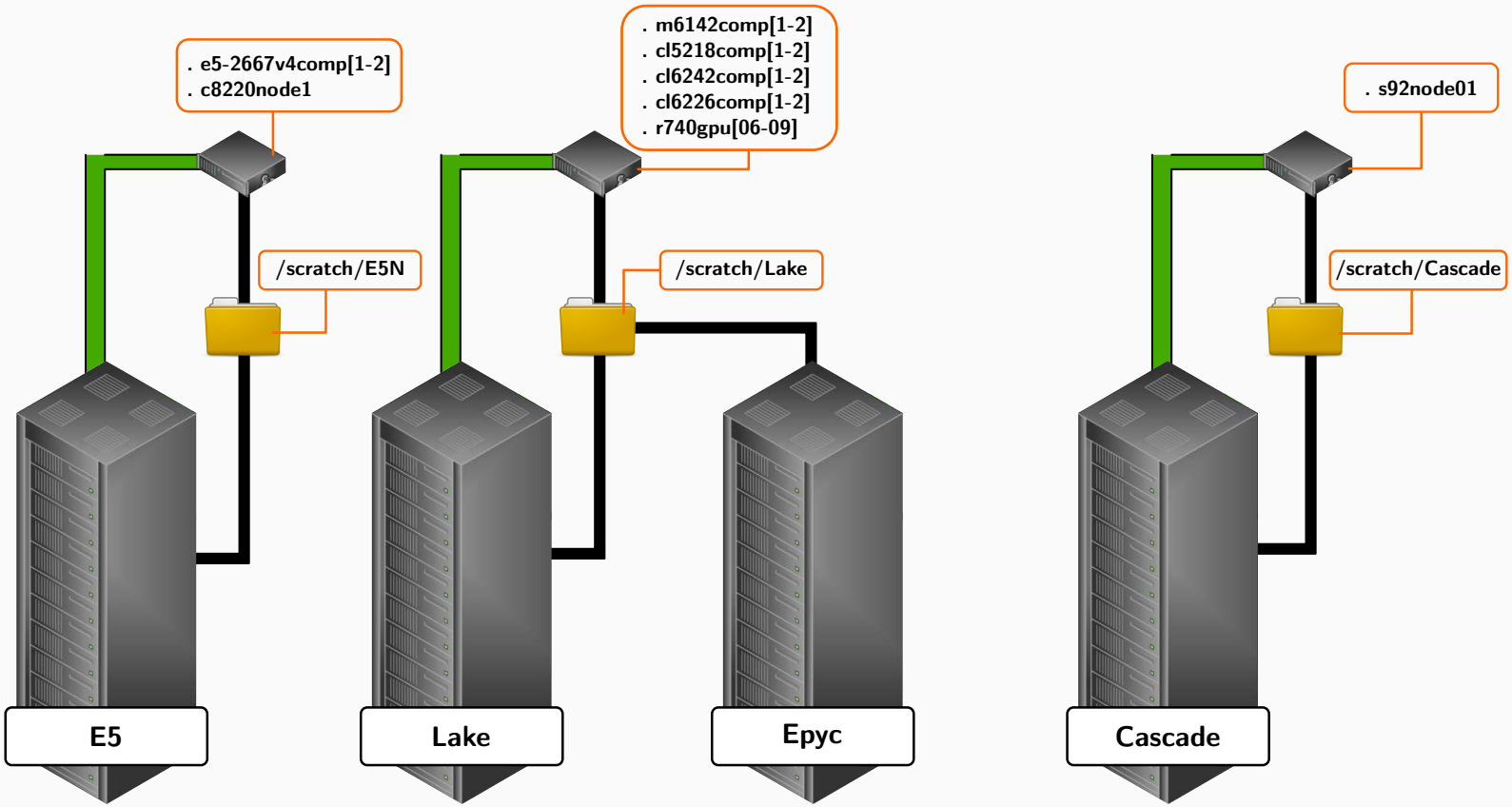
Each cluster possesses a **scratch** volume to store data **during** a job

! Important

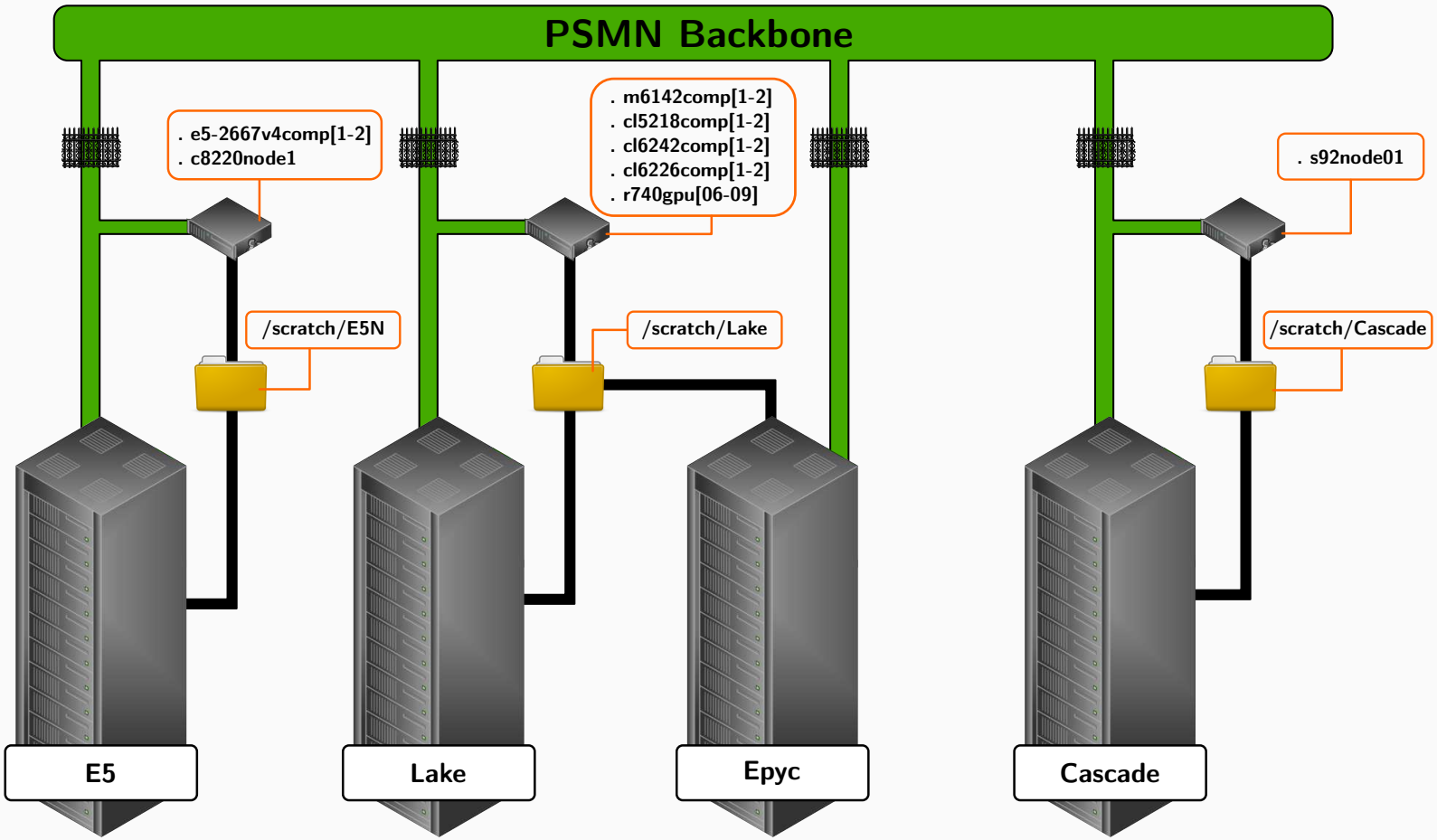
Scratch data are often erased to make space



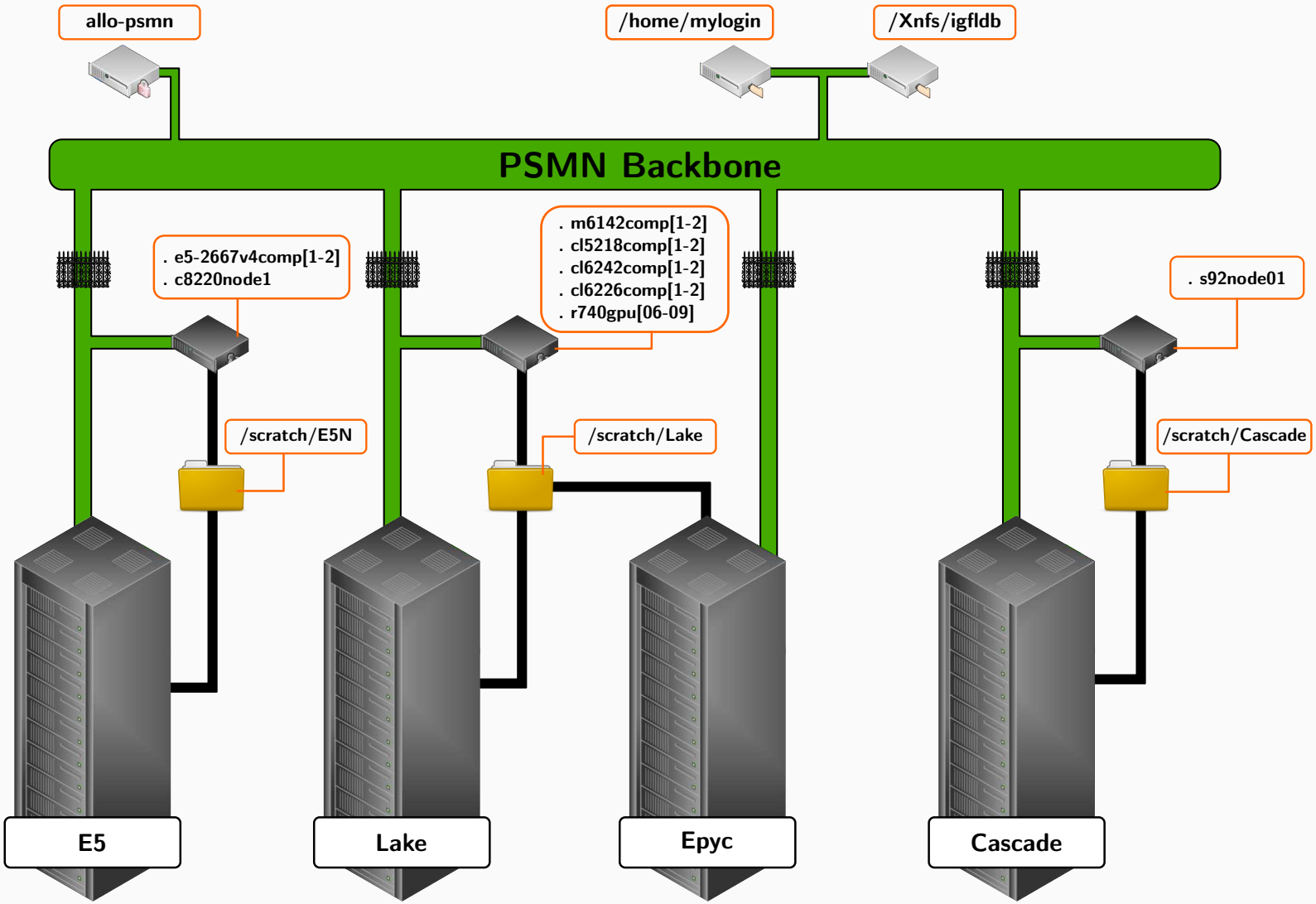
Each cluster (except Epyc) contains one or multiple **login nodes**



login nodes are primarily used to access clusters via SSH



Each user has access to two spaces: its **home** and the lab **Xnfs** volume



PSMN

Guidelines

Login nodes

Front-end servers on which you can:

- edit and transfer files,
- compile and build software,
- test scripts or software on small datasets,
- submit and monitor jobs.

! Important

Do not run **heavy** calculations directly on the login nodes

⚠ allo-psmn

allo-psmn is a connexion server and should only be used to transfer files or access one of the login nodes

Compute nodes

- sole purpose is to run submitted jobs
- you can only connect to it if you have a job running on it

> **Tip**
>
> You can display information about the compute nodes of a cluster

PARTITION	CORES	CPUS	MEMORY(MB)	NODES	NODELIST
E5*	8	16	128872(MB)	4	c8220node[31-34]
E5*	8	16	128828(MB)	24	c6320node[1-24]
E5*	8	16	257852(MB)	24	c6320node[101-124]

"Cluster **E5** contain 4 nodes (c8220node31, c8220node32, c8220node33, c8220node34) that each possesses 8 cores, 16 CPUs and 128GB of RAM."

Job scheduler (SLURM)

Multiple purposes:

- match and allocate compute nodes with user requests
- provide a framework to execute and monitor requests
- manage and order the queue of pending requests

SLURM directives

Placed **at the head** of the job script to specify computing resources:

- resources allocation (CPU, RAM)
- estimated run time
- constrains (cluster, node, GPU, . . .)

Two ways to submit a job:

(i) through a script: **sbatch** (ii) interactively: **srun**

Available spaces

- your home folder: **/home/{username}**
- your lab Xnfs folder: **/Xnfs/{lab}db**
- cluster scratch: **/scratch/{cluster}/{username}**
- folder for Nextflow job: **/Xnfs/abc** (maintained by LBMC)

Project space

It is possible to get a temporary volume on **/Xnfs** for project requiring a decent amount of space ($\approx$ 10TB)

Remark

Discuss within your lab on how to **properly** organize your lab space (avoid duplicata, keep it tidy, . . .)

SSH-based tools

- **SCP**: command line tool limited to file transfer
- **SFTP**: CLI or GUI for file transfer, access and management
- **rsync**: command line tool to synchronize a set of directories
- **SSHFS**: mount your PSMN home on your workstation

Modules

PSMN provides pre-installed software through **Lua Environment Modules**.

3-steps process:

- initialize the environment
- choose a list of modules (one per cluster)
- load one or several modules

J	E5	Lake	Epyc	Cascade
JasPer/2.0.33-GCCcore-11.2.0	✓	✓		✓
Java/11.0.2	✓	✓		✓
jbigkit/2.1-GCCcore-11.2.0	✓	✓		✓
Jellyfish/2.3.0-GCC-11.2.0	✓	✓		✓
jemalloc/5.2.1-GCCcore-11.2.0	✓	✓		✓
Julia/1.7.1-linux-x86_64	✓	✓		✓

Local install

You can download and compile an external program/software on your PSMN home.

Requirements:

- the program must be available for Debian 11
- the libraries must be already installed on PSMN

> ⚠ **Warning**
>
> You have no admin rights on PSMN ⇒ no **sudo** commands

Alternatively, you can use containers: Singularity, Charliecloud (not Docker)

PSMN

Utilities

Shellcheck

Command line tool to check for syntax errors and sometimes bugs:

shellcheck myjob.sh

Warning

- **shellcheck** cannot detect incorrect paths
  ⇒ check them at the login node or through an interactive job
- **shellcheck** cannot detect syntax errors in SLURM directives
  ⇒ see Caution block

A web version is available here: https://www.shellcheck.net

Text editor

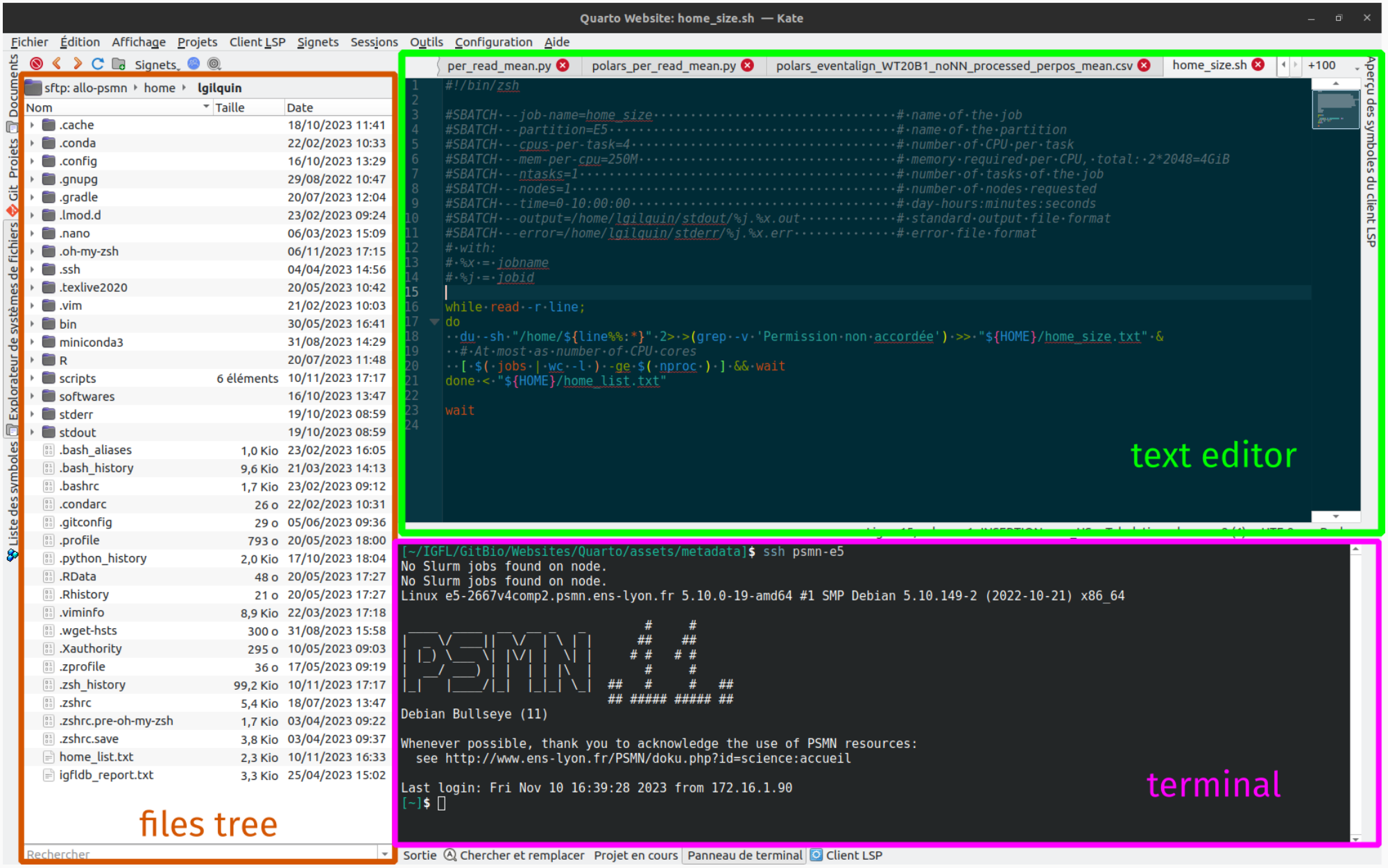
Available on PSMN: Nano, Vim, Emacs

Limitations

- juggling with terminal and editor
- keep track of the file tree
- no langage services: code completion, error and warning, and so on

$\Rightarrow$ use an IDE that support SSH connexion to remote server:
VSCodium, NeoVIM(+plugins), KDevelop, . . .

2. 4 | Utilities: SSH-based IDE



Syntax highlighting

Highlights source code in different colors and fonts according to the category of terms
⇒ improves readability and finding error

Example: zsh with Oh-My-Zsh plugin

```
[~]$ df -h --total /Xnfs/igfldb
Sys. de fichiers       Taille Utilisé Dispo Uti% Monté sur
r730data7:/data/igfldb    40T     34T  6,1T  85% /Xnfs/igfldb
total                     40T     34T  6,1T  85% -
[~]$ fdisk -l
```

Conclusion

- Two complementary tools to use in function of your computational skills and your needs:
  - IFB cloud to start with bioinformatic or for specific analyses
  - PSMN for advanced users and high computing analyses
- Both tools require users to manage their environment and data
- Do not hesitate to get help from your bioinformatics hub

⚠ They are mutualized tools

Take care to respect the other users by:

- managing your storage
- tidying shared spaces
- shutting down useless VM / removing intermediate files

Acknowledgments

“The Institut Français de Bioinformatique (IFB) is funded by the Programme d’Investissements d’Avenir (PIA), grant Agence Nationale de la Recherche, number ANR-11-INBS-0013.”

“We gratefully acknowledge support from the PSMN (Pôle Scientifique de Modélisation Numérique) of the ENS de Lyon for the computing resources.”