Virtual machine on cloud for Galaxy training

F. Samson, S. Dérozier, V. Loux, V. Martin, C. Blanchet, C. Gauthey

Agenda

- Galaxy portal of the Migale platform
- Introduction of Migale platform's trainings around Galaxy
- Problems observed during the trainings
- A solution: training on the cloud

Galaxy portal on the Migale platform

Migale Platform

- Assignments of Migale platform:
  - Develop an IT infrastructure for genomics
  - Disseminate knowledge in bioinformatics
  - Design and development of bioinformatics software and workflows
- Galaxy group:
  - F. Samson: Galaxy installation
  - S. Dérozier: Galaxy administration, software integration, databanks, … and Galaxy support
  - V. Loux: software integration and Galaxy support
  - V. Martin: software and databanks installation on Migale server

Galaxy instances

- Two Galaxy instances:
  - Production: http://migale.jouy.inra.fr/galaxy
  - Development
- 3 web / 2 handlers / 1 manager
- PostgreSQL database on a dedicated server
- Upload local files (Curie) + Galaxy home user
- Development and formation instance on a R900/24 cores 196GoRAM (CentOS 5), with dedicated queue 16 cores of a 500 cores cluster (CentOS 5)
- Production instance on a Virtual Machine R710/16 cores 50Go RAM (CentOS 6)

Migale Galaxy informations

- 95 users
- around **30 added tools** (in-house or from the tool-shed):

Rmes	RmesFormat	SurfG+	Prodigal	Sed
FASTA Stats	Blastall	HMMScan	RiboPicker	
SortmeRNA	NCBI Tools	Sickle	Prinseq	Quast
FASTQc	Velvetg	Velveth	VelvetOptimiser	

Galaxy Cluster Jobs (2012)

2012		
month	*number of jobs*	*average time*
Septembre	179	00:01:43
Octobre	60	00:00:11
Novembre	203	01:51:25
Décembre	29	01:58:26

Galaxy Cluster Jobs (2013)

2013		
month	*Number of jobs*	*Average time*
Janvier	43	00:22:34
Février	17	00:03:04
Mars	267	00:08:36
Avril	566	00:44:39
Mai	1.192	00:16:38
Juin	1.060	00:34:52
Juillet	177	00:41:01
Août	17	00:03:18
Septembre	237	00:59:30
Octobre	552	01:44:03
Novembre	219	02:14:22

Introduction of Migale platform's trainings using Galaxy

« Using galaxy » training session

- Duration: 1 day (two sessions in Jouy, one in Montpellier in 2013)
- Theory: 20% / Practice: 80%
- 33 trained persons
- Objectives :
  - treatment of files,
  - execution of tools,
  - creation and sharing of workflows, history and pages.

« NGS primary analysis » with Galaxy

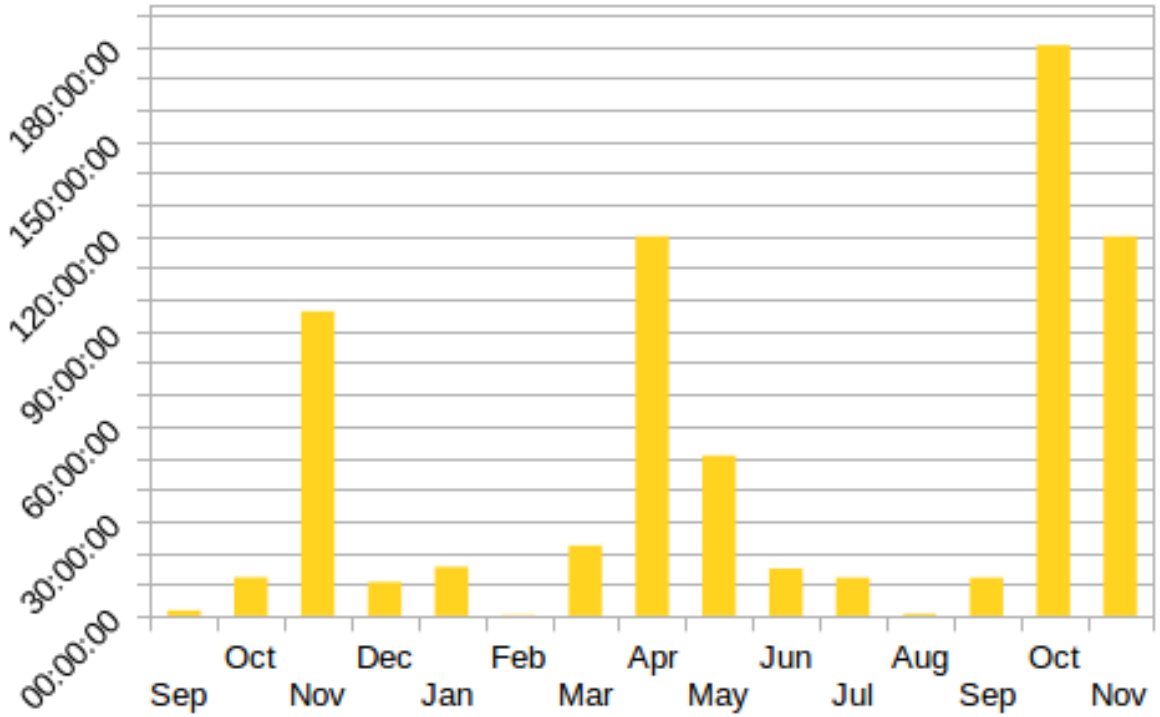
- Duration: 1 day (two sessions in Jouy, one in Montpellier in 2013)
- Theory: 40% / Practice: 60%
- 32 trained persons
- Objectives: discover the concepts and the bioinformatics methods used for the primary analysis of NGS data. Application on the tools of mapping and assembly.
- Contribution of galaxy:
  - practice without the need to know command line,
  - time economy (about 1h)

Problems observed during the training sessions

Problems

- 12 users launch at the same time very expensive jobs
- Web interface highly stressed
- Need to dedicate resources (nodes) and avoid to saturate the production instance (development portal)
- Some problems with the synchronization of instance due to OS version and local python version
- Use of small datasets adapted to the node resource

Galaxy jobs maximum duration



A solution: training on the cloud

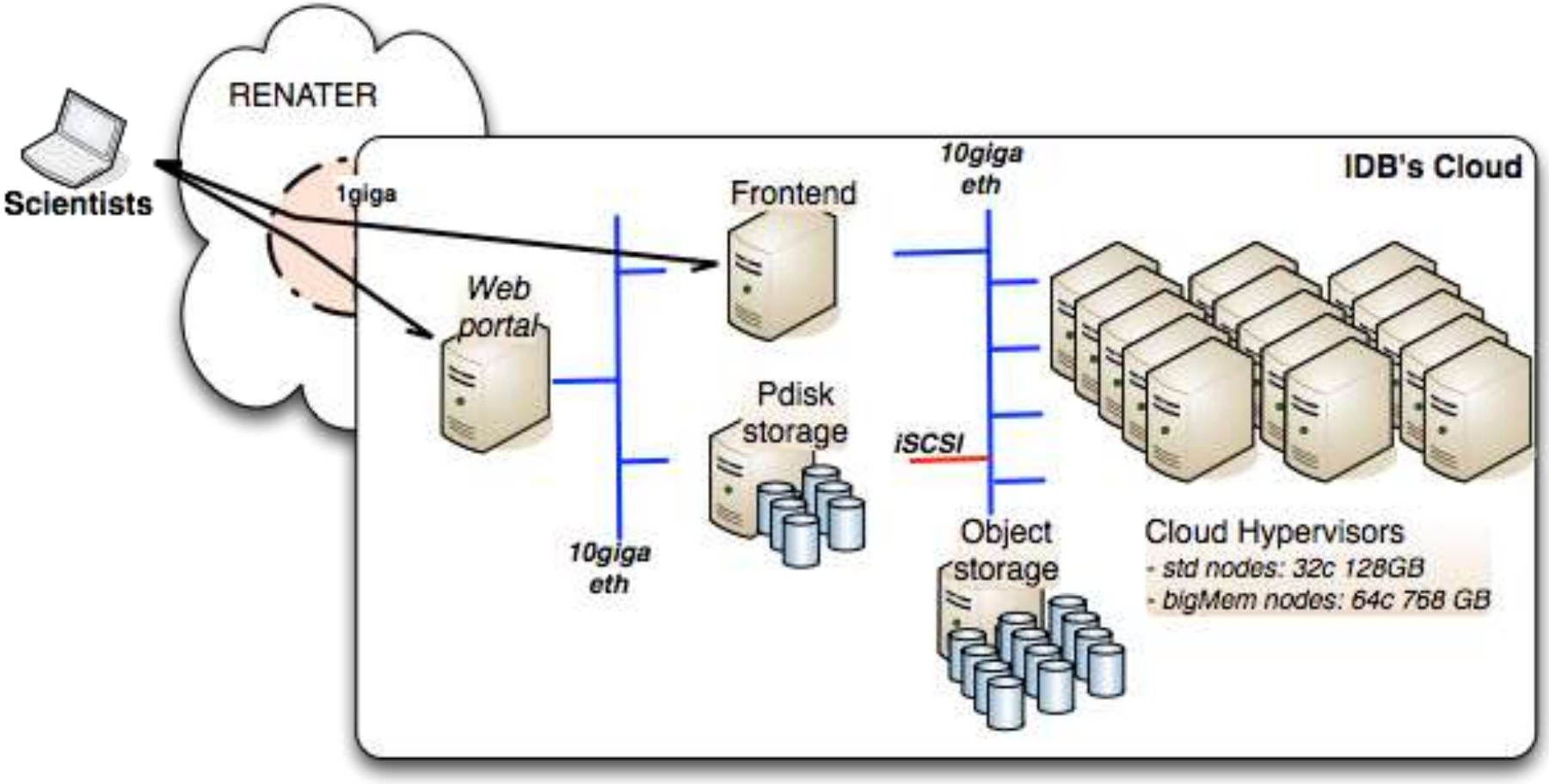
Miracle solution (?)

- Solution: deployment of Galaxy VMs on the cloud and treanings on the cloud.
- Already exist: Galaxy on AWS.
- Interest: deployment on academic cloud.
- Perpectives: use the academic cloud of IBCP for Migale trainings with Galaxy and later IFB

IDB Cloud and Bioinformatics Appliances

- **Cloud workbench for Biology**
  - http://idee-b.ibcp.fr/cloud.html
  - **Running since Sept. 2011**
  - **CNRS-IBCP** FR3302, Lyon, France
  - opened to **Biology community**
  - **14 bioinformatics appliances:** Galaxy portal,
  - standard compute nodes, proteomics, virtual desktop, structural biology, …
  - **+70 users** from all IFB regional centers
  - PRABI **16**, APLIBIO **28**, RENABI-NE **13**, -GO **7**, -SO **2**, -GS **5**
  - **VMs up to 32cores-768GB RAM**
- **Infrastructure**
  - **Compute +900cores +4TB ram**
    - **Standard** nodes (32c-128GB)
    - **Bigmen** nodes (**64c 768GB**)
    - Powered by **StratusLab**
  - **Storage +250TB**
    - Virtual disks, object storage (S3)

IDB's Cloud Infrastructure



https://idee-b.ibcp.fr/appliances.html

Bioinformatics Cloud Appliances

We provide different bioinformatics cloud appliances ready-to-run. A cloud appliance is a predefined virtual machine with pre-installed tools and workflows. Most of these appliances can be associated with one of your virtual disk.

You can get a description of each appliance by *clicking on their name* in the list below. *To run your own instances,* click on the corresponding power button. Then, you will be redirected to a pre-filled form to create your instances.

▸ Bioinformatics compute node

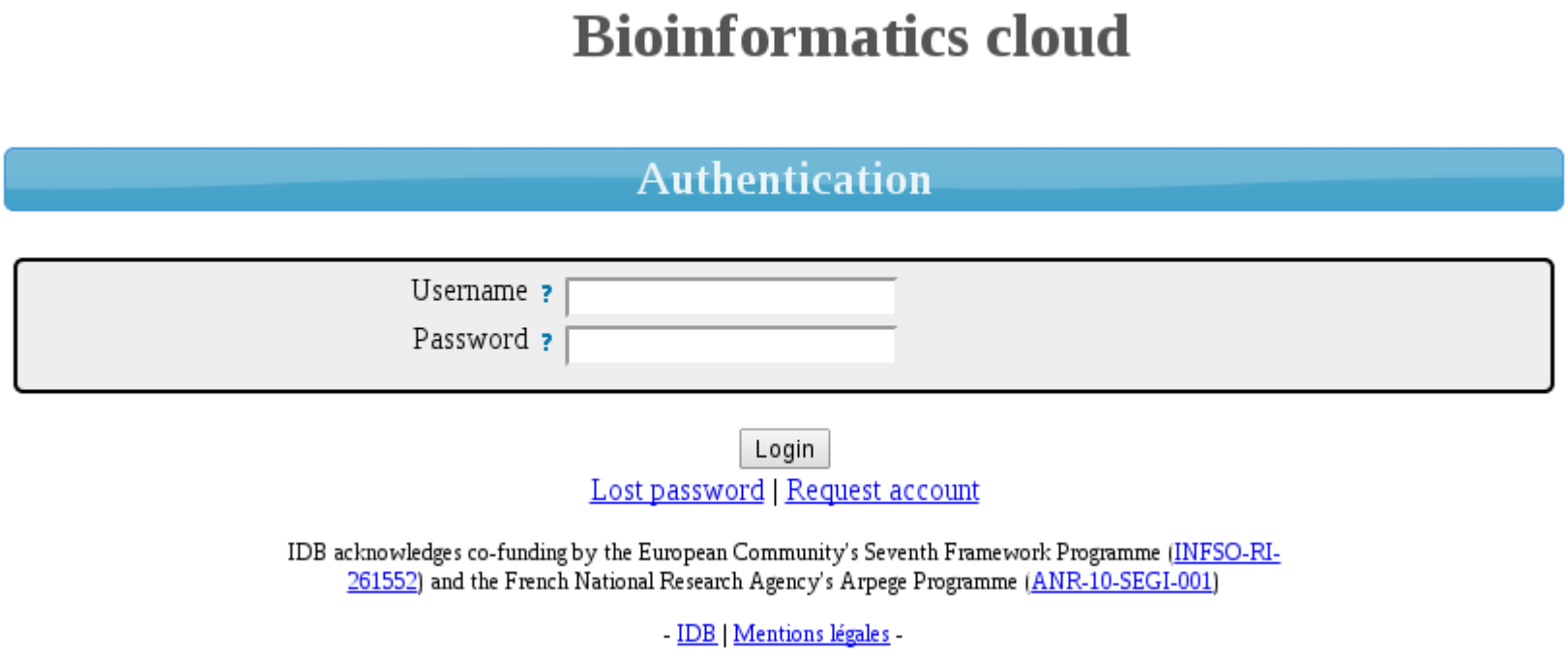
▾ Galaxy portal

Scientific gateway appliance configured with the well-known GALAXY portal. You have access to pre-installed standard bioinformatics tools and can connect to your own Galaxy portal with a standard web browser. Simply follow the link on the main IDB cloud interface and log in with the pre-defined user *user@cloud.idb.fr* (password *idbuser*).

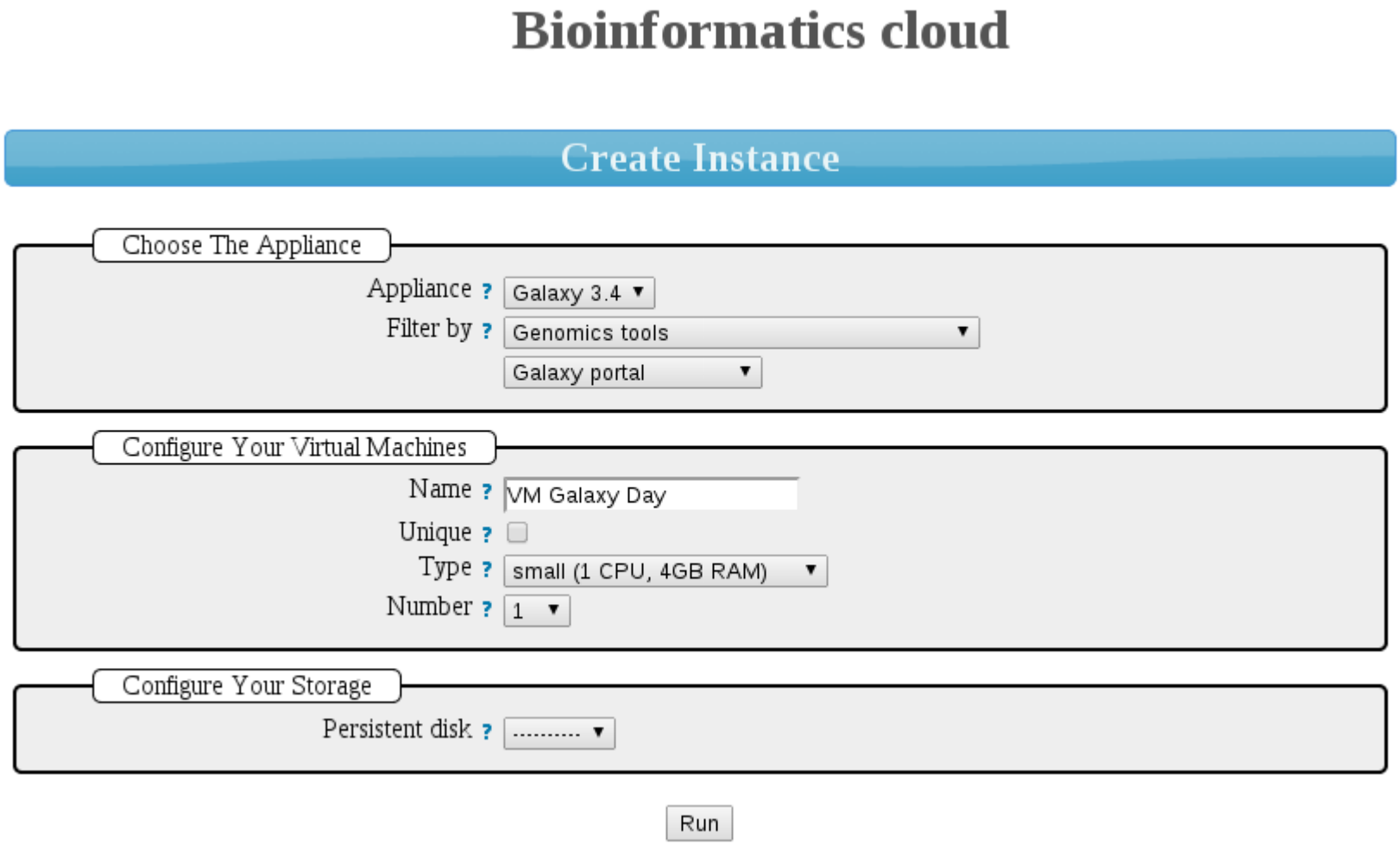
When this appliance is run in association with one of your virtual disks, the history and the data of your Galaxy portal is stored for a further execution. Don't forget to attach your favorite virtual disk in the 'Create instance' form.

Large files can be uploaded directly through the 'FTP upload method' in Galaxy. *Notice that you will use the SCP protocol instead of the FTP because of security reasons related to the cloud architecture*. To upload such large data files, use the command line scp or a graphical interface as described in the documentation of the IDB's cloud. For example to upload a file called `mydata.fastq` on a virtual machine accessible via the port 20198, in a terminal go to the directory containing your file and execute the following command line: `scp -P 20198 mydata.fastq root@idb-cloud.ibcp.fr:upload_dir/` . Once uploaded, the file will be

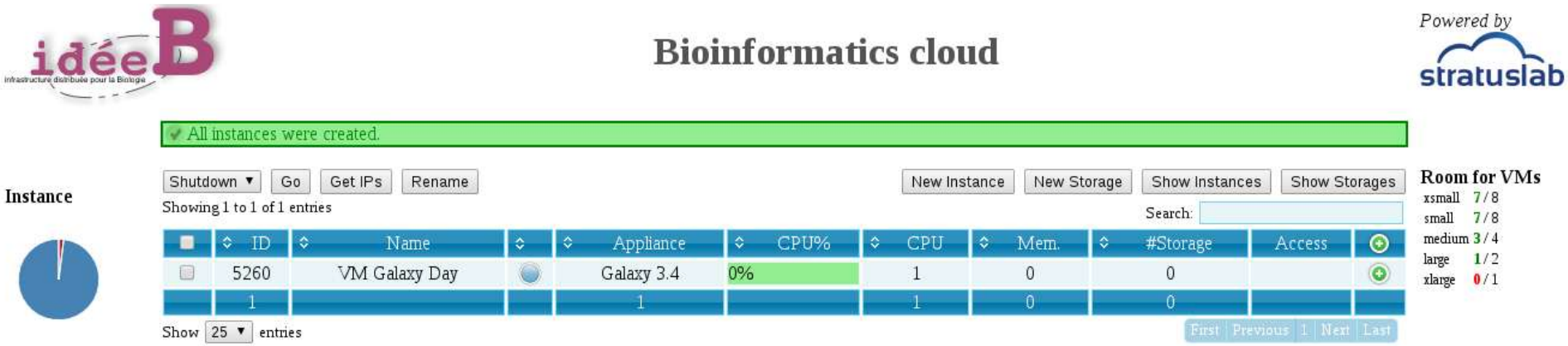
https://idee-b.ibcp.fr/appliances.html



https://idee-b.ibcp.fr/appliances.html



https://idee-b.ibcp.fr/appliances.html



Thanks

Christophe Blanchet,
Clément Gauthey

Infrastructure Distribuée pour la Biologie - IDB
CNRS - IBCP
LYON
FRANCE

Sandra Dérozier,
Véronique Martin,
Valentin Loux,
Franck Samson

MIGALE
Unité Mathématique Informatique et Genome (MIG)
Domaine de Vilvert
78352 Jouy en Josas

Thanks to :

- StratusLab members
- co-funding by the European Community's Seventh Framework Programme (INFSO-RI-261552) and by the French National Research Agency's Arpege Programme (ANR-10-SEGI-001).