Fourniture de services cloud pour la biologie

Exemple du cloud IDB-IBCP

Christophe Blanchet, Clément Gauthey

Institut de Biologie et Chimie des Protéines
Plateforme "Infrastructure Distribuée pour la Biologie - **IDB**"
CNRS-IBCP FR3302 - LYON - FRANCE - http://idee-b.ibcp.fr

IDB acknowledges co-funding by the **European Community's Seventh Framework Programme** (INFSO-RI-261552), the **French National Research Agency's Arpege Programme** (ANR-10-SEGI-001) and by the French Institute of Bioinformatics (**IFB**)

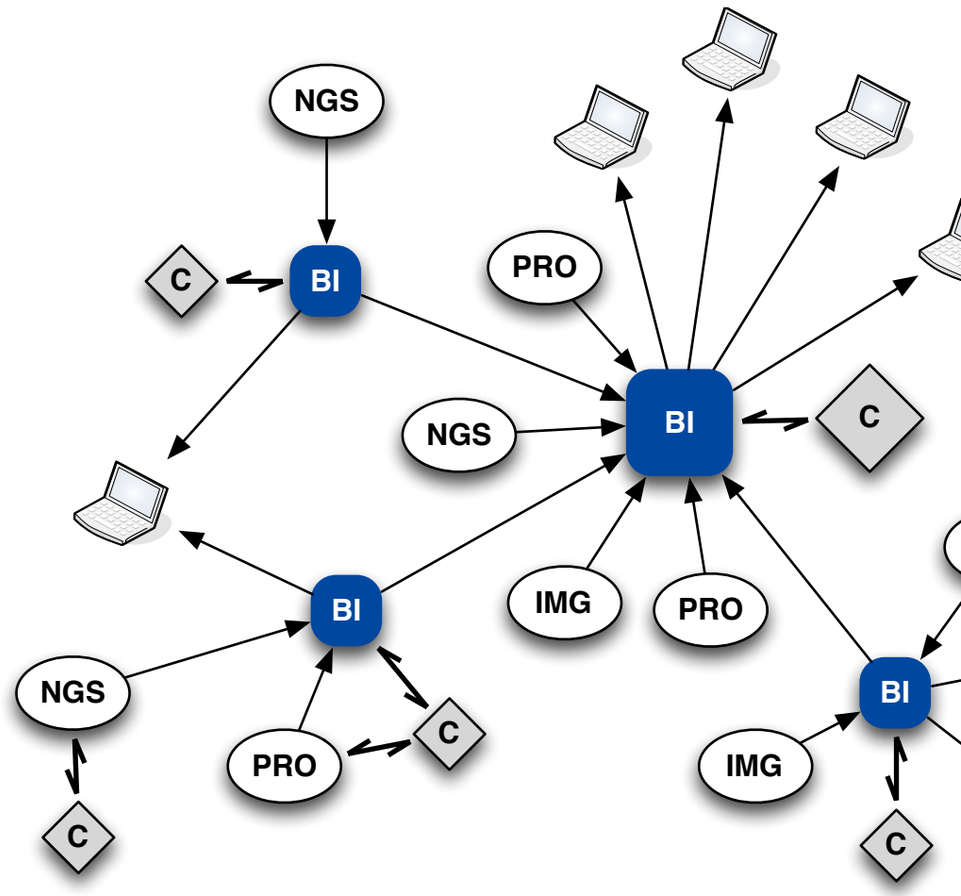
Bioinformatics Today

- Biological data are *big data*
  - 1552 online databases (NAR Database Issue 2014)
  - Institut Sanger, UK, 5 PB
  - Beijing Genome Institute, China, 7 sites, 20.6 PB

  ➡ **Big data in many places**
- Analysing such data became difficult
  - Scale-up of the analyses : gene/protein to complete genome/ proteome, ...
  - Lot of different daily-used tools
  - That need to be combined in workflows
  - Usual interfaces: portals, Web services,...

  ➡ **Datacenters with ease of access/use**
- Distributed resources
  - Experimental platforms: NGS, imaging, ...
  - Bioinformatics platforms

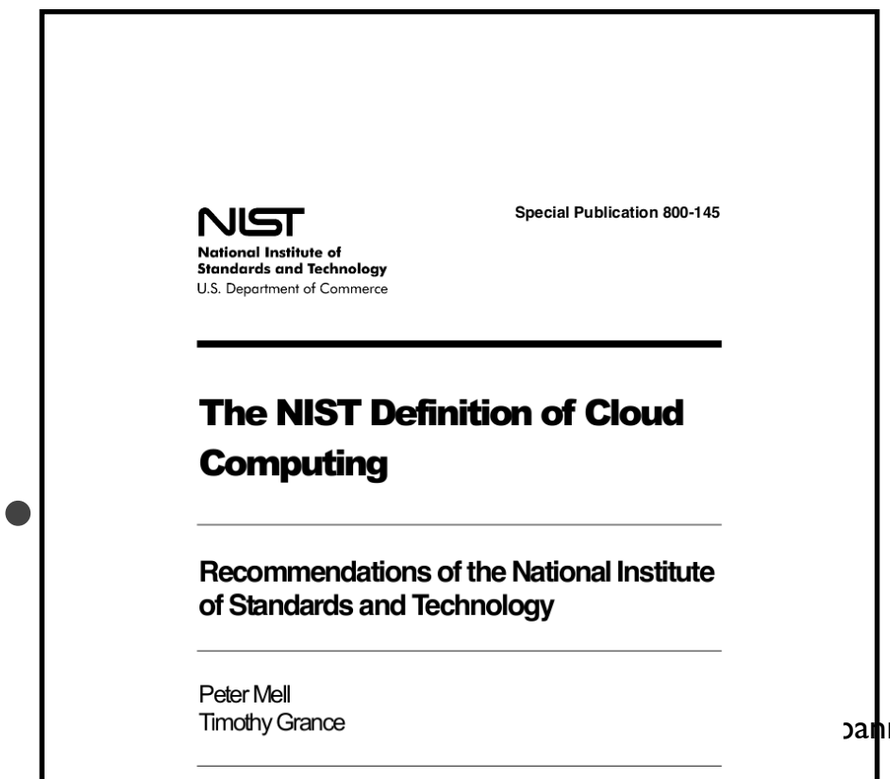
  ➡ **Federation of datacenters**



Cloud ?

- **Essential characteristics**
  - On-demand self-service
    - No human intervention
  - Broad network access
    - Fast, reliable remote access
  - Rapid elasticity
    - Scale based on app. needs
  - Resource pooling
    - Multi-tenant sharing
  - Measured service
    - Direct or indirect economic model with measured use
- **Deployment models**
  - Private
    - Single administrative domain, limited number of users
  - Community
    - Different administrative domains with common interests & proc.
  - Public
    - People outside of institute's administrative domain
  - Hybrid
    - Federation via combination of other deployment models
- **Service models**
  - Software as a Service (SaaS)
    - Direct (scalable) hosting of end user applications
  - Platform as a Service (PaaS)
    - Framework and infrastructure for creating web applications
  - Infrastructure as a Service (IaaS)
    - Access to remote virtual

http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf


NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Special Publication 800-145

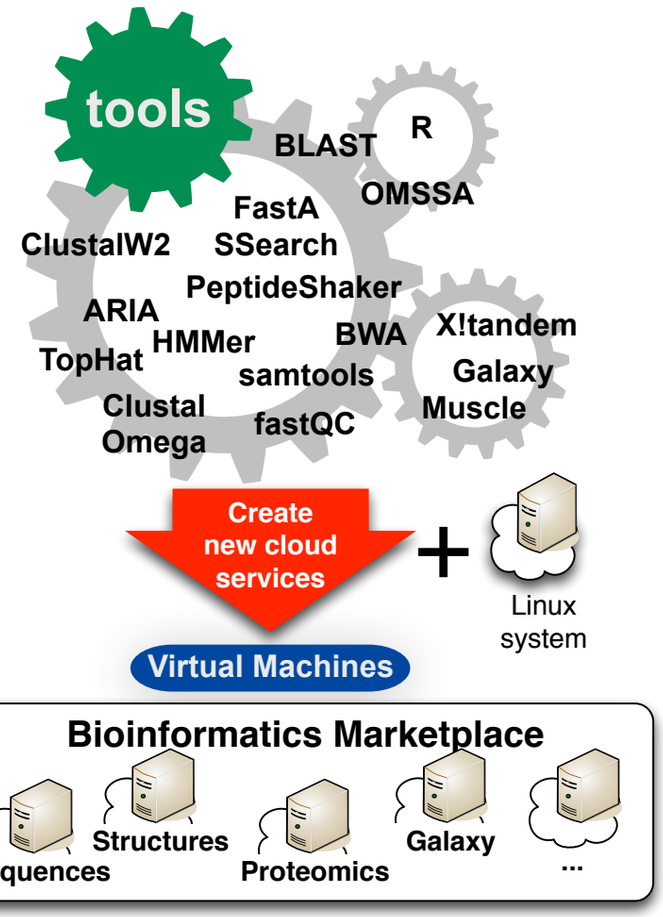
The NIST Definition of Cloud Computing

Recommendations of the National Institute of Standards and Technology

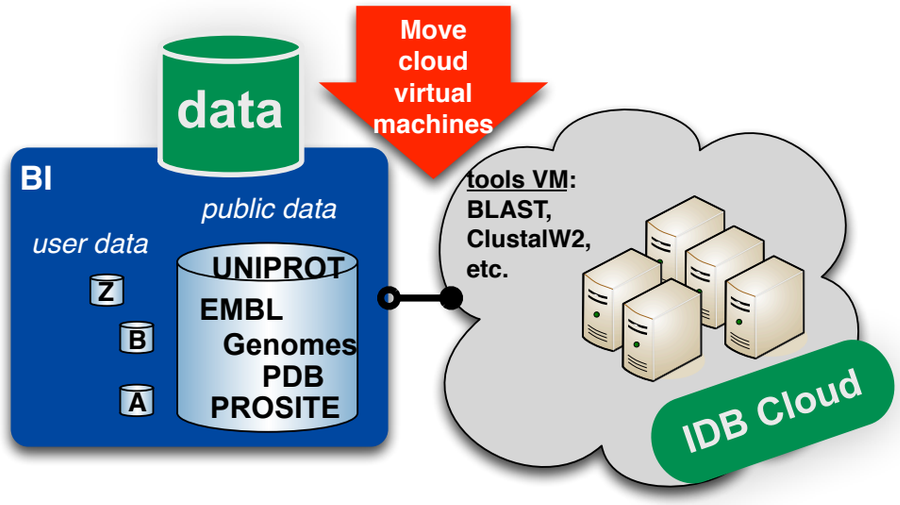
Peter Mell
Timothy Grance

IDB Cloud and Bioinformatics Appliances

- **Cloud workbench for Biology**
  - https://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**)
  - **Storage +250TB**
    - Virtual disks, large-scale object storage (S3)
  - Powered by **StratusLab** and **CEPH**



Cloud extended services

stratuslab

Native cloud services

- Authentication
- Virtual machine management
- Persistent disk service
- Client CLI
- *etc.*

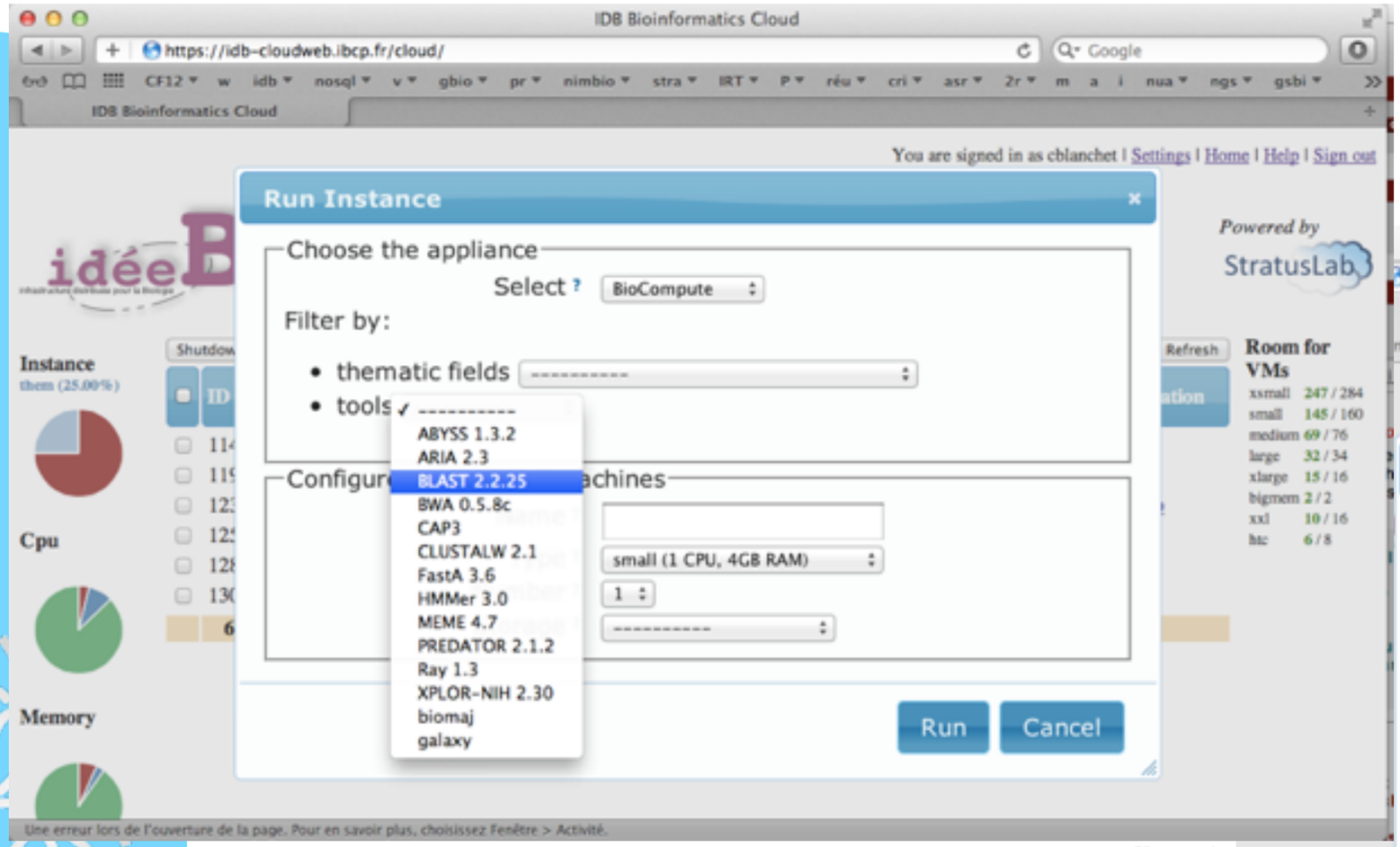
IDB

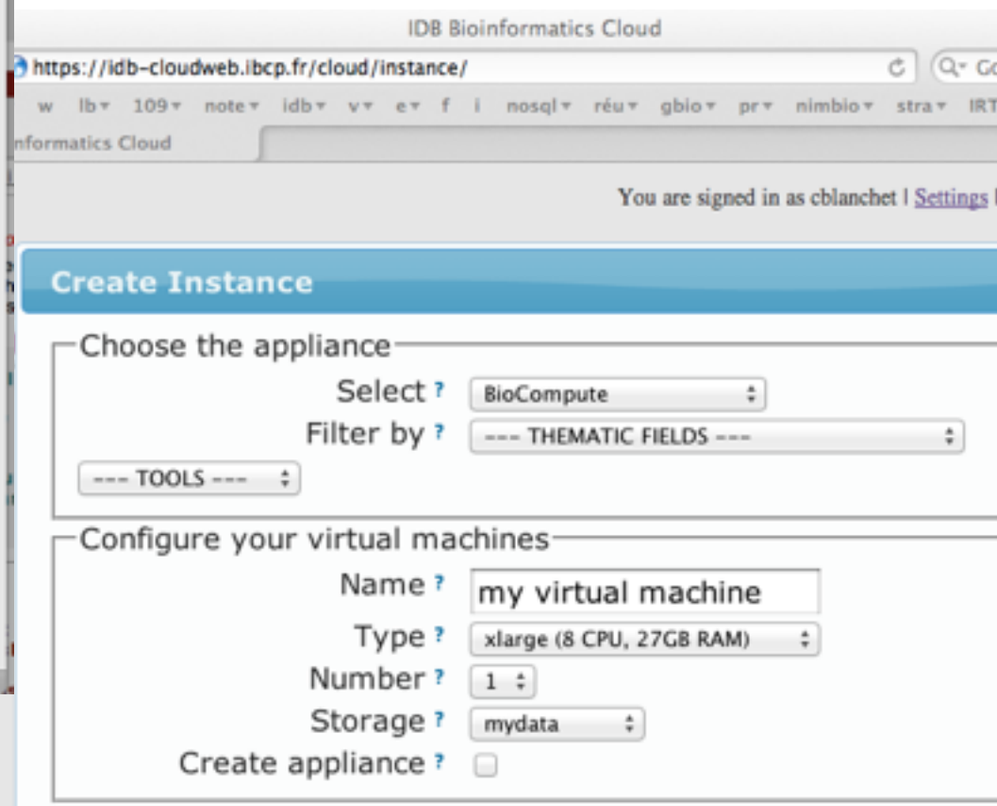
- **Bioinformatics Marketplace**
  - find appropriate appliances more easily.
  - reduce "noise" in the central Marketplace
  - respect visibility contraints for the bioinformatic appliances, such as confidentiality
- Bioinformatics **metadata "bio:tool"**
  - additional elements related to bioinformatics tools to annotate appliances
  - help users to search for the tools themselves or the type of analysis
  - select suitable bioinformatics appliances containing the required tools
- Integrated **Web interface**
  - VM & virtual disks management
  - browse bionformatics appliances with "bio:tool" MDz
- CEPH **storage** backend
  - large scale and distributed storage
  - reliable by replication
  - high-througput IO
  - single unified storage cluster for all interfaces: block, object and file system

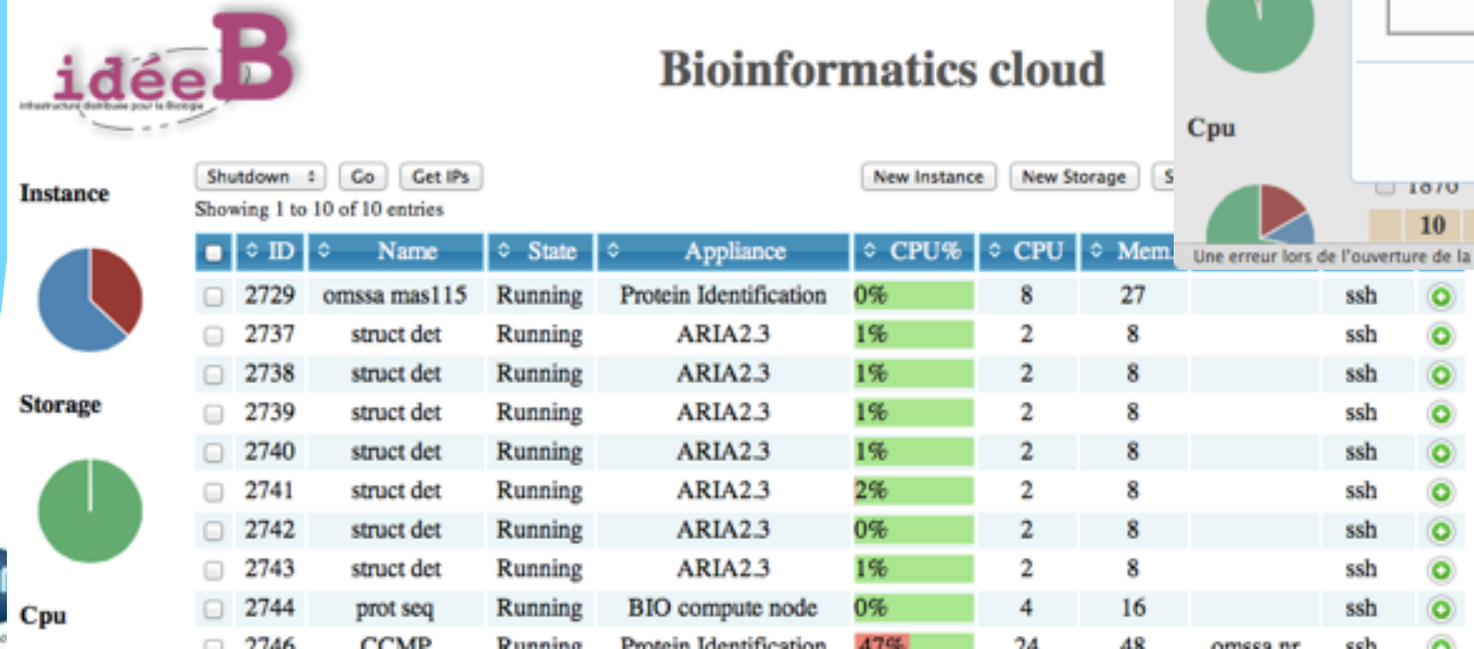
idéeB

IBCP cnrs Lyon 1

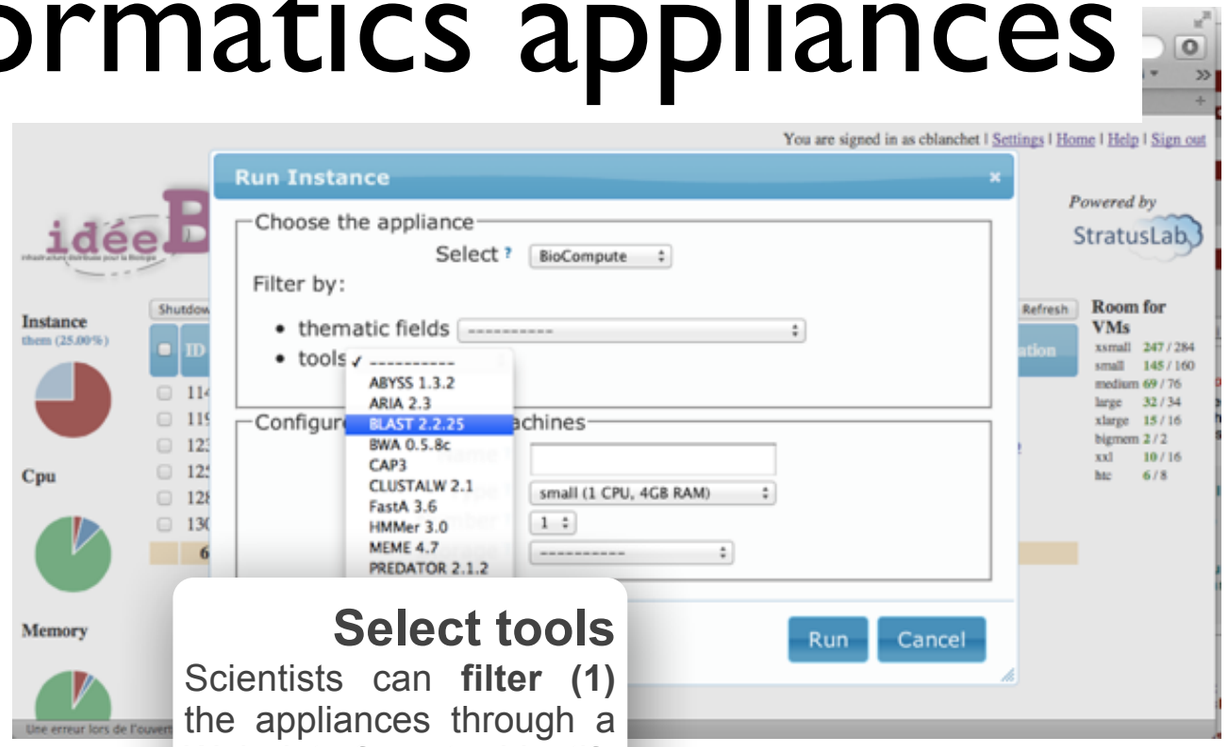
Driven throught a simple web interface





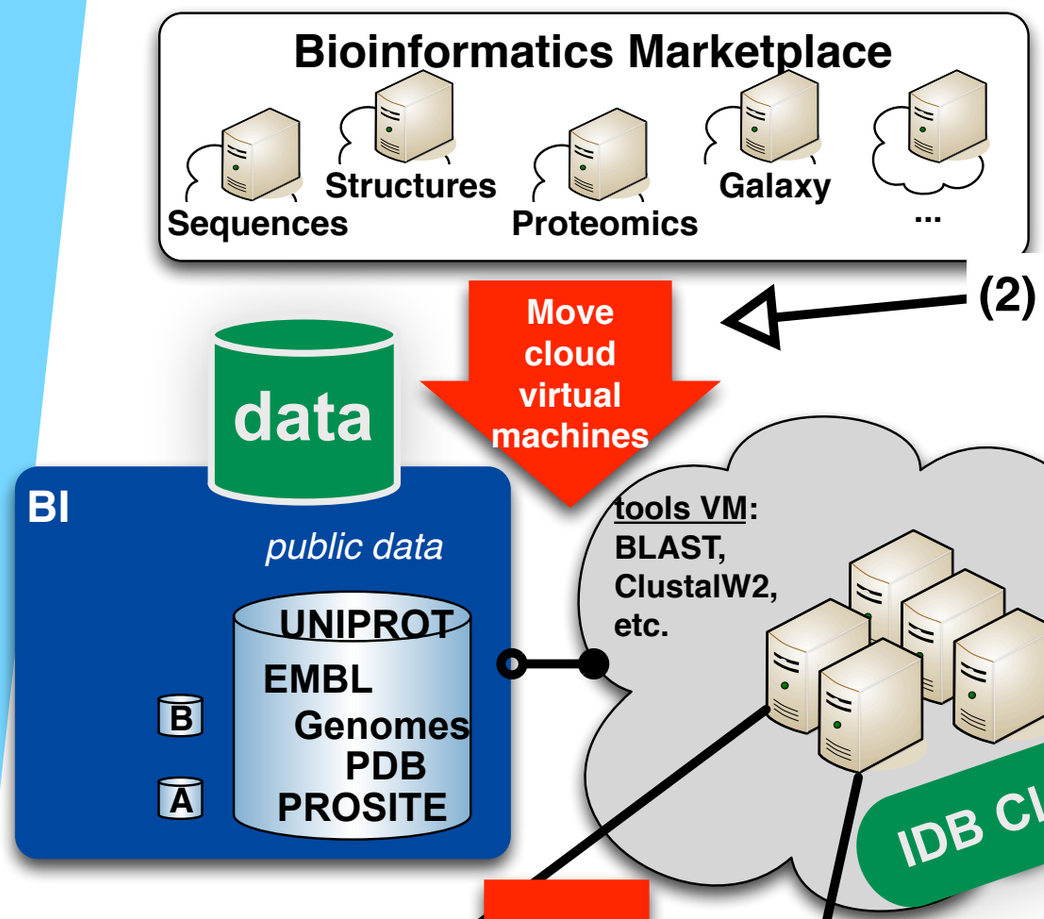


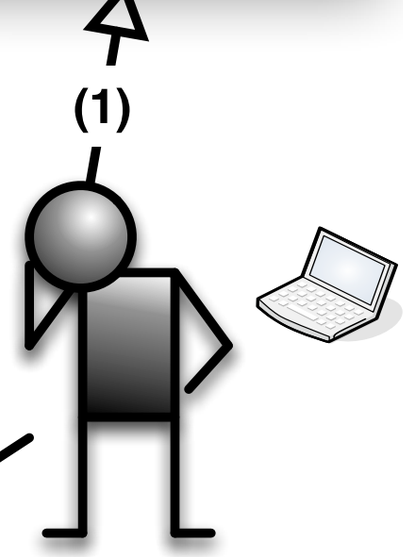
Thematic bioinformatics appliances

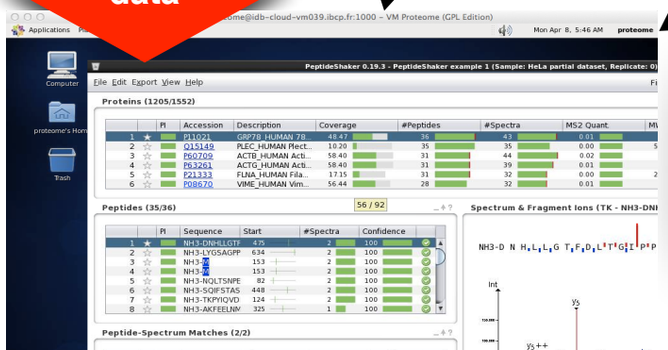


Select tools

Scientists can **filter (1)** the appliances through a Web interface to identify and **launch (2)** the appropriate ones.





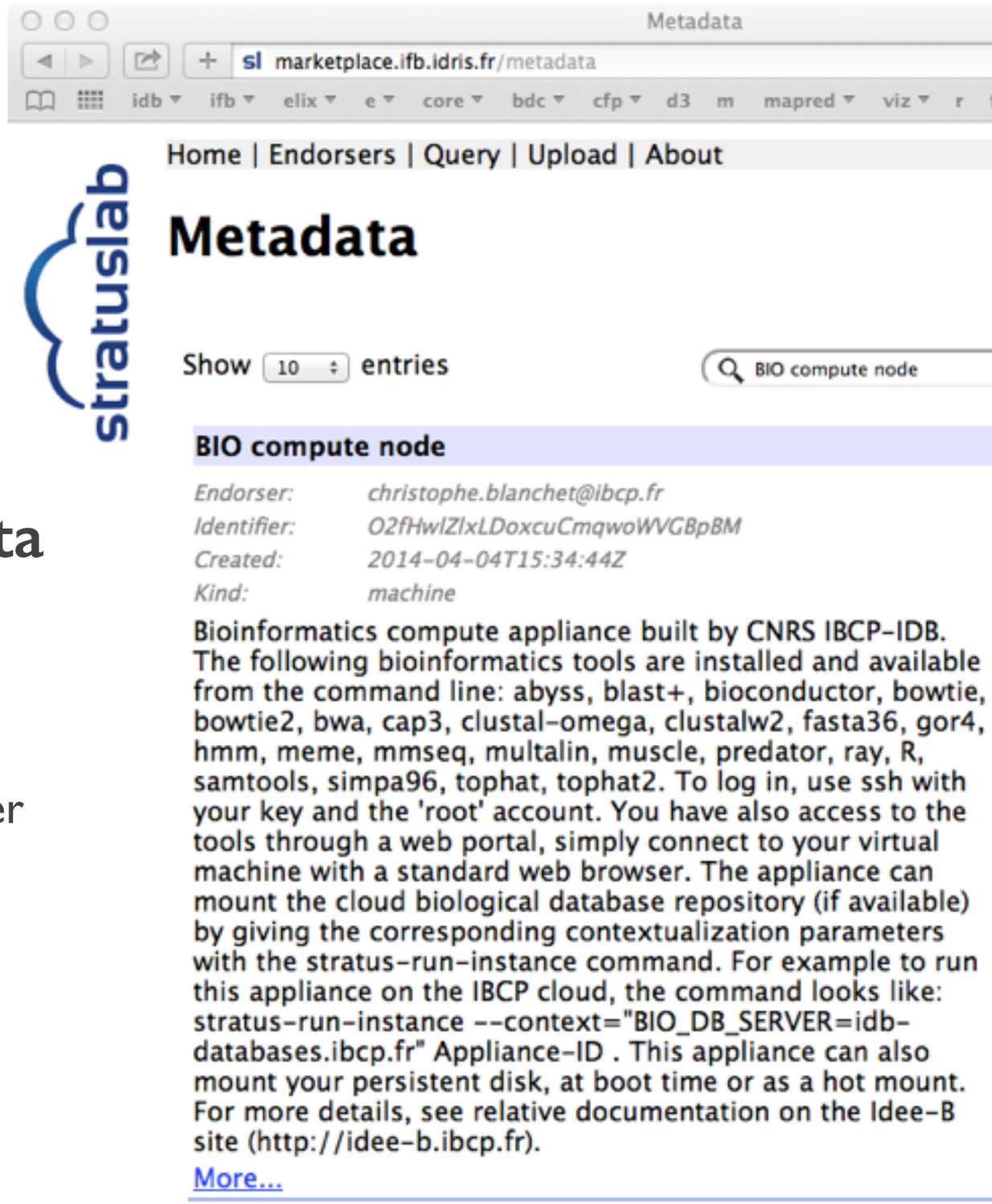


Use tools (3)

Scientists have access to their own **cloud resources** through web portal, remote virtual desktop or SSH.

Standard Bioinformatics node

- ‘Biocompute’ appliance
- Use **your own** instance(s)
- With **pre-installed standard bioinformatics tools**
  - BLAST, FastA, SSearch,HMM,...
  - ClustalW2, Clustal-Omega, Muscle,..
  - Bowtie(2), BWA, samtools, ...
  - MEME, R, etc.
- Connected to **public reference data**
  - Uniprot, EMBL, genomes, PDB, etc.
  - Automaticaly shared to the VMs
- **Cluster** mode
  - turn several instances in a single virtual cluster
  - shared file system
  - batch scheduling

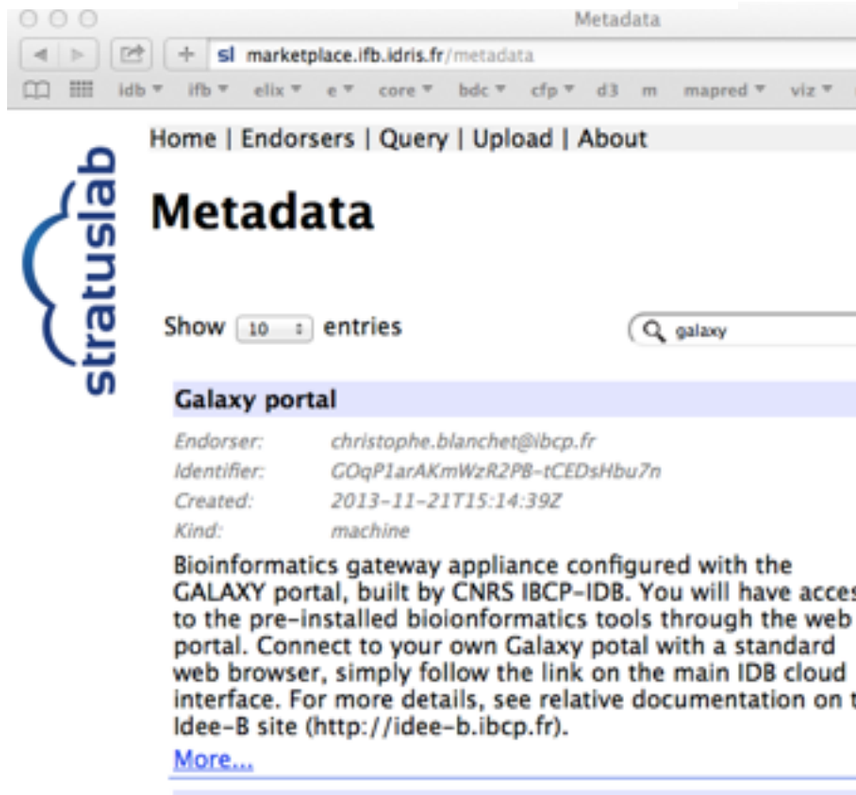


idéeB

IBCP cnrs Lyon 1

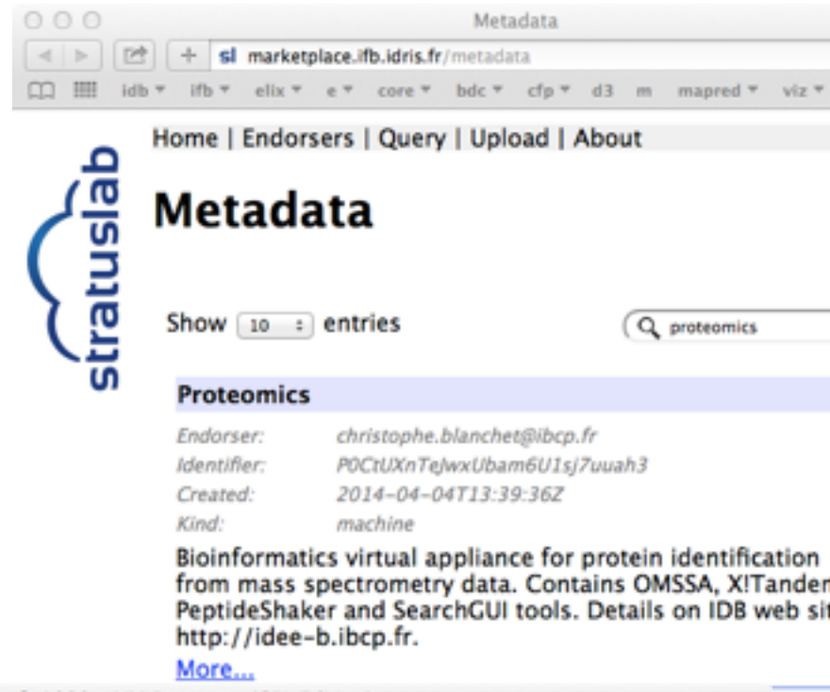
Cloud Galaxy portal for NGS analyses

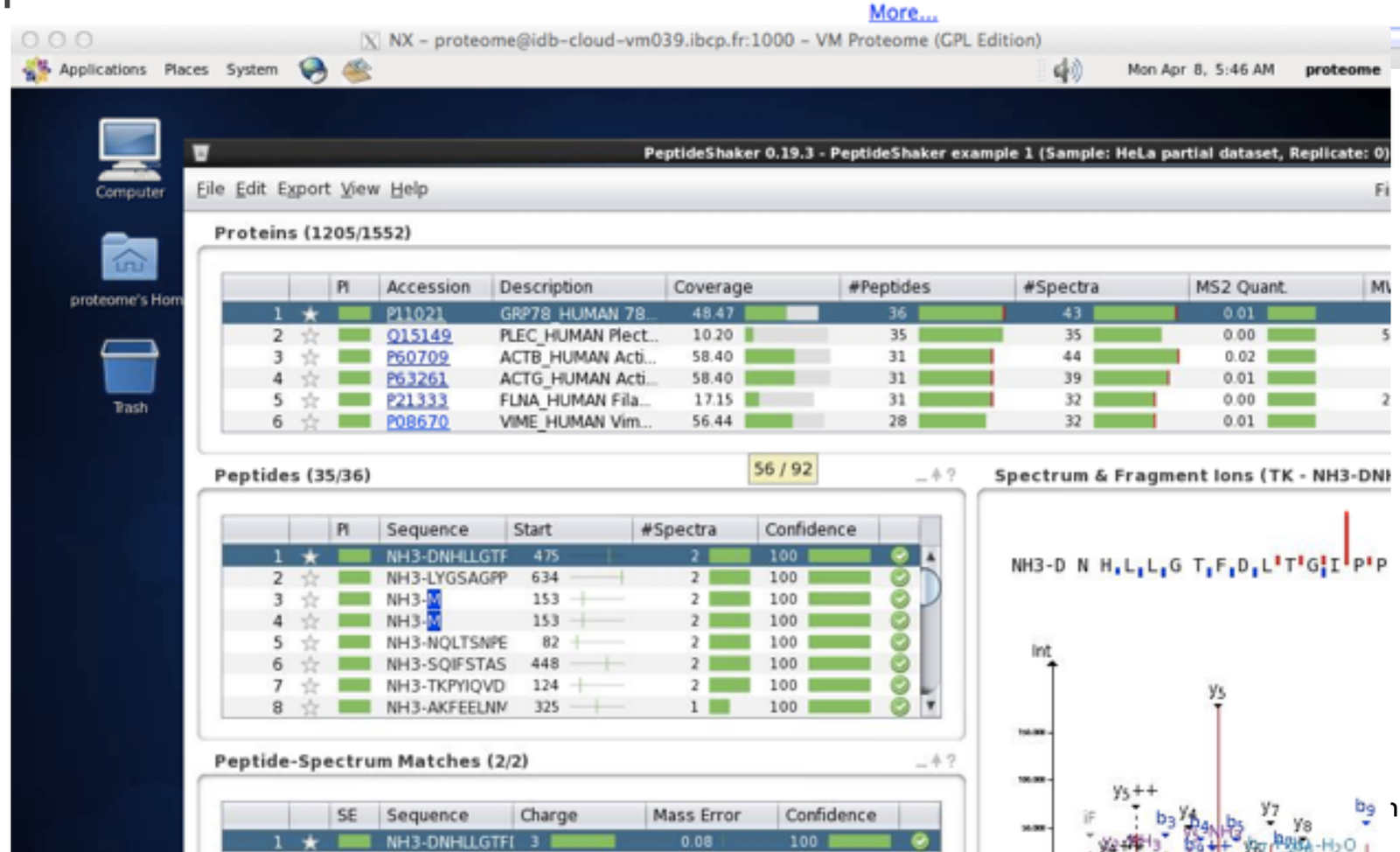
- Analyse NGS data
  - **portal Galaxy** is widely used in the community
  - connected to large public data: sequences and indexes
  - large user data (GBs)
- **Preserve workflows and results** (cloud virtual disk)
- Different **domain-specific instances** (RNAseq, CHIPseq, *etc.*)
- **For training**: create a special instance derived from the main instance but with dedicated datasets
- Help the integration of **monthly updates**



Proteomics virtual desktop

- Motivation
  - Collaboration with a mass spectroscopy platform
  - Running out of space on their local resources
- Protein identification tools
  - Mass experimental data
  - Reference databases : nr, Swiss-Prot
  - Reference screening tools:
    OMSSA, X!Tandem
- User interface
  - **Remote Virtual Desktop (NX)**
  - Reference GUIs
    - SearchGUI
    - PeptidShaker





OMSSA

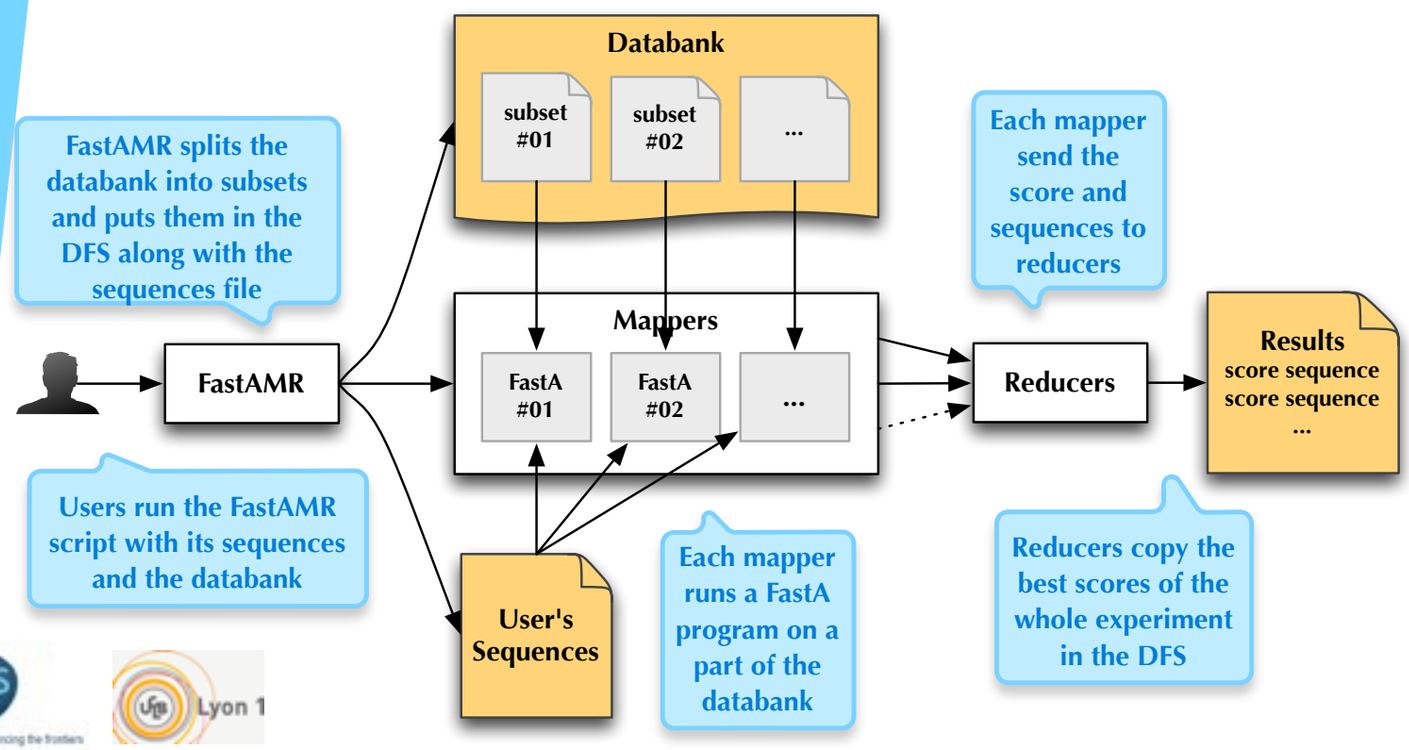
X!

IBCP

Lyon 1

Hadoop for Life Science

- Provide turnkey virtual machine with pre-configured mapreduce framework
  - **Accelerate bigadata analysis** with the two steps map & reduce paradigm
  - **Hadoop MapReduce** 1.0.4
- Appliances (2)
  - standard hadoop mapreduce
  - bioinformatics software integrated in hadoop
- Example of sequence similarity searching
  - FastA & SSearch
  - deploy database of sequences in HDFS
  - compare each structure to others



Developed in the context of the French project MapReduce, ANR ARPEGE

BIO MapReduce

Endorser: clement.gauthey@ibcp.fr
Identifier: J46wxrwGLdnoSskmb0JlfGv8UpY
Created: 2013-05-17T11:13:08Z
Kind: machine

This appliance provides an easy way to deploy a Hadoop MapReduce cluster (v1.0.4) with pre-installed bioinformat tools such as FastA. You just need to run the bash script hadoop-create-cluster with a nodes list and an username parameters and wait few minutes until the process is completed. Then you can login to the user account and submit your Hadoop jobs or interact with Hadoop filesyst You can extend a current cluster by submitting a list of n nodes to the script. A FastA MapReduce example is also provided under the directory /usr/local/share/fasta. (Cre for the French project MapReduce, ANR ARPEGE, 2010-20 mapreduce.inria.fr)

More...

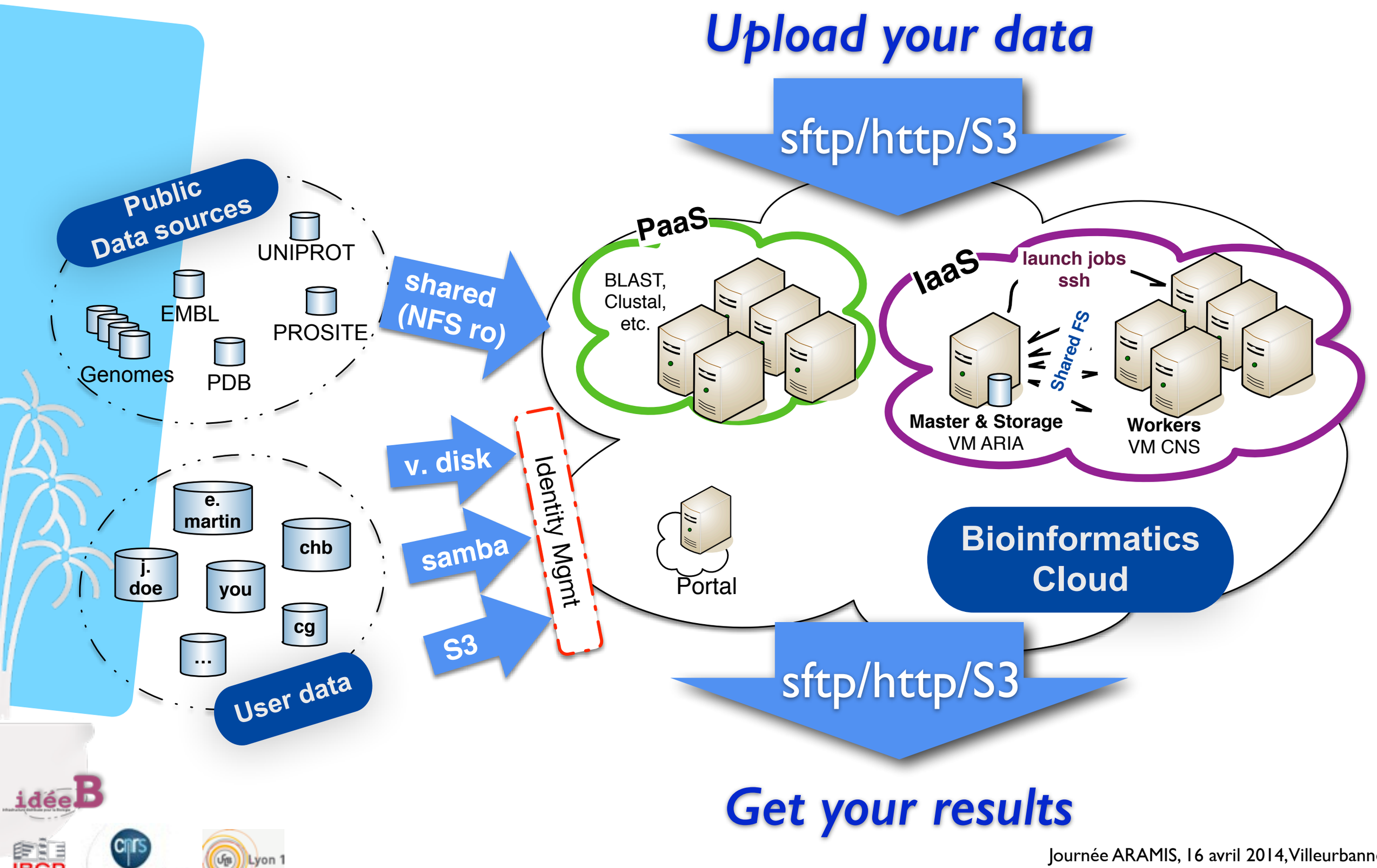
Hadoop MapReduce

Endorser: clement.gauthey@ibcp.fr
Identifier: BttU7uNM5UT1haUigVS7xySI2rr
Created: 2013-05-17T09:33:52Z
Kind: machine

This appliance provides an easy way to deploy an Hadoop MapReduce cluster (v1.0.4). You just need to run the bash script hadoop-create-cluster with a nodes list and an username in parameters and wait few minutes until the process is completed. Then you can login to the user acc and submit your Hadoop jobs or interact with Hadoop filesystem. Enjoy! In addition, you can extend a current cluster by submitting a list of new nodes to the command (Created for the French project MapReduce, ANR ARPEGE, 2010-2013, mapreduce.inria.fr)

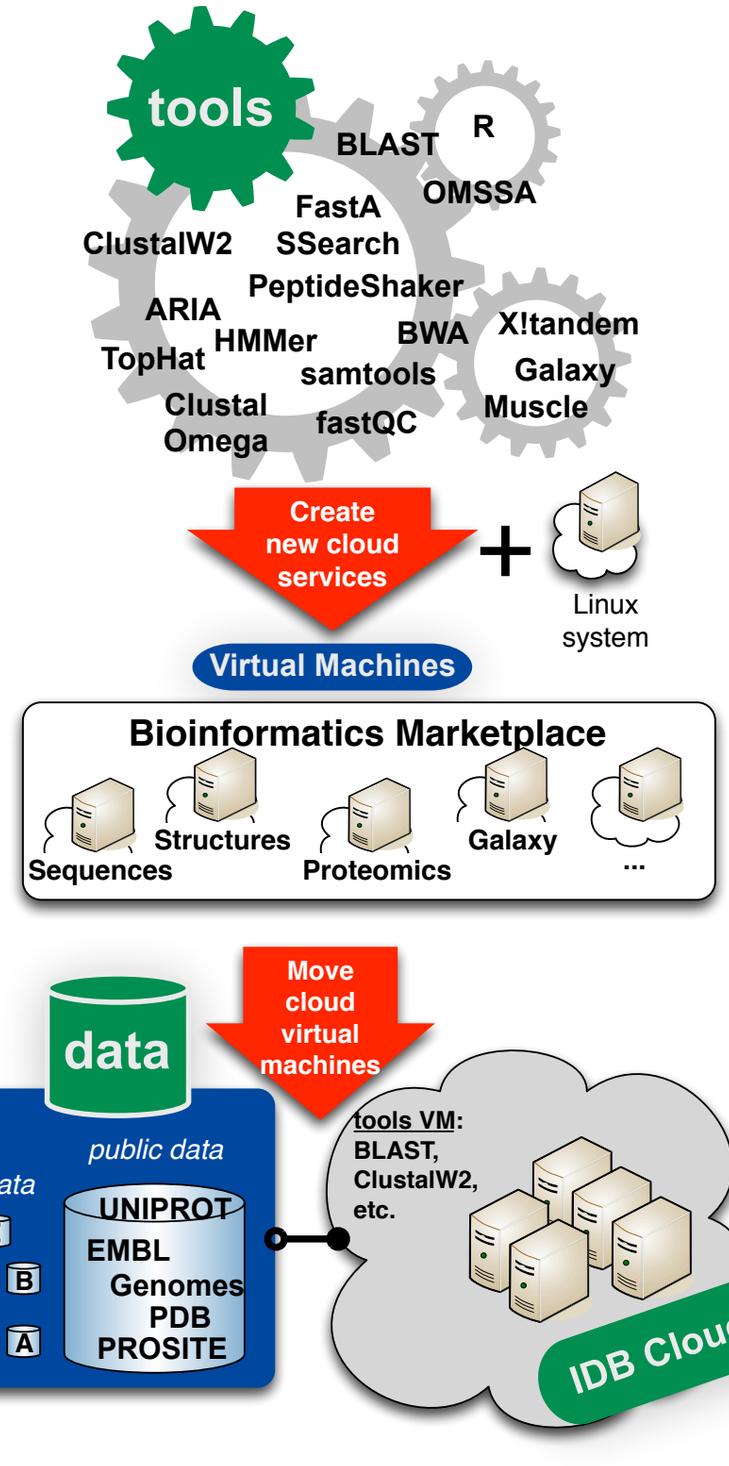
More...

Cloud Storage for Biological Data



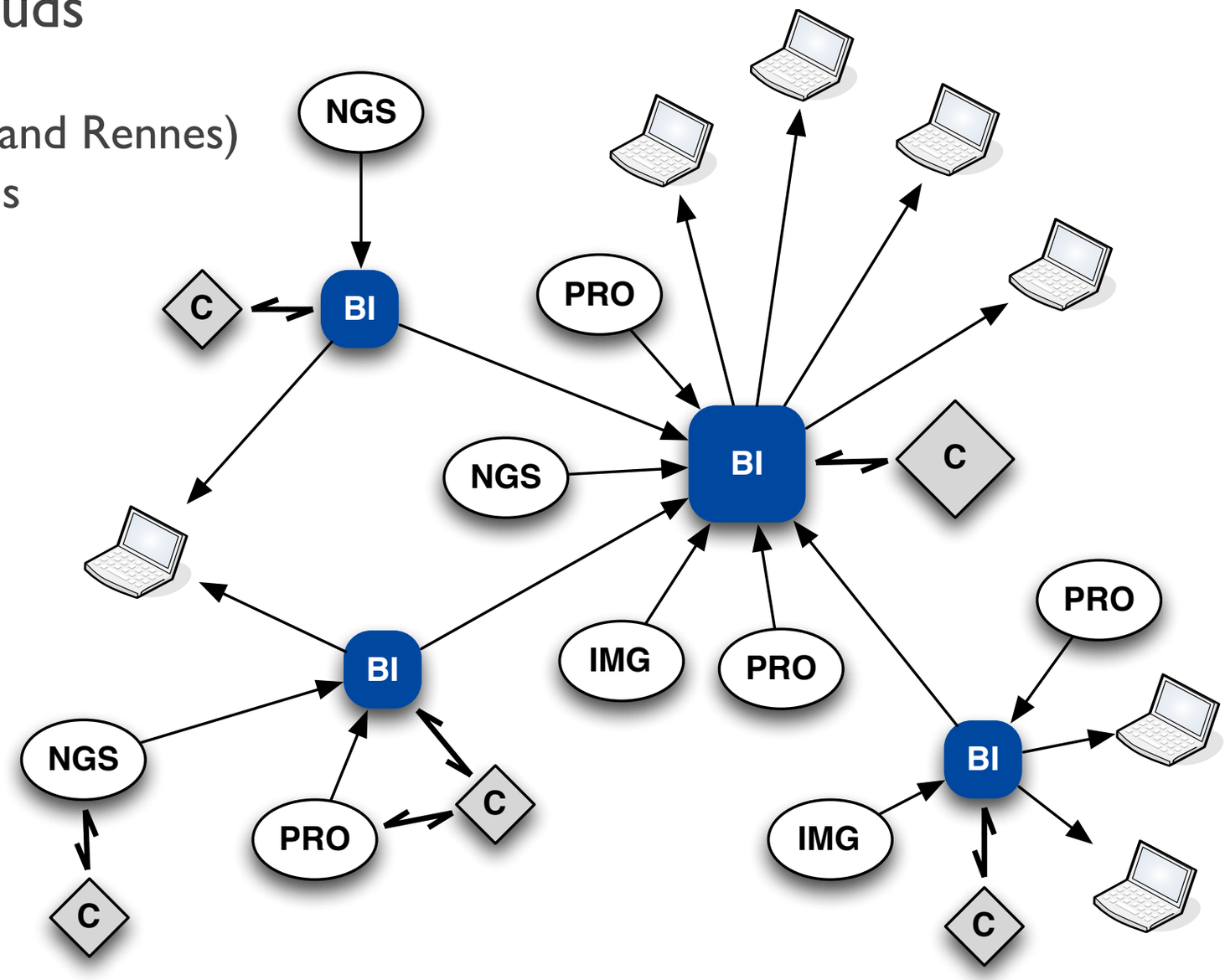
Conclusion

- Added value of cloud,
  - for scientific analyses: user-specific resources, isolated, different instances together
  - for training: create a special instance derived from the main but with dedicated datasets
  - for tools integration: semantic annotation, solve software dependencies
  - for development & operations (DevOps): different versions at the same time
- Provide turnkey bioinformatics appliances
  - Standard tools and pipelines
  - New developments
  - Ready to run on clouds
- Public bioinformatics cloud (e.g. IDB)
  - Tightly connected to existing bioinformatics resources
  - Linked to public biological databases
  - In collaboration with the French Institute of Bioinformatics



Perspectives

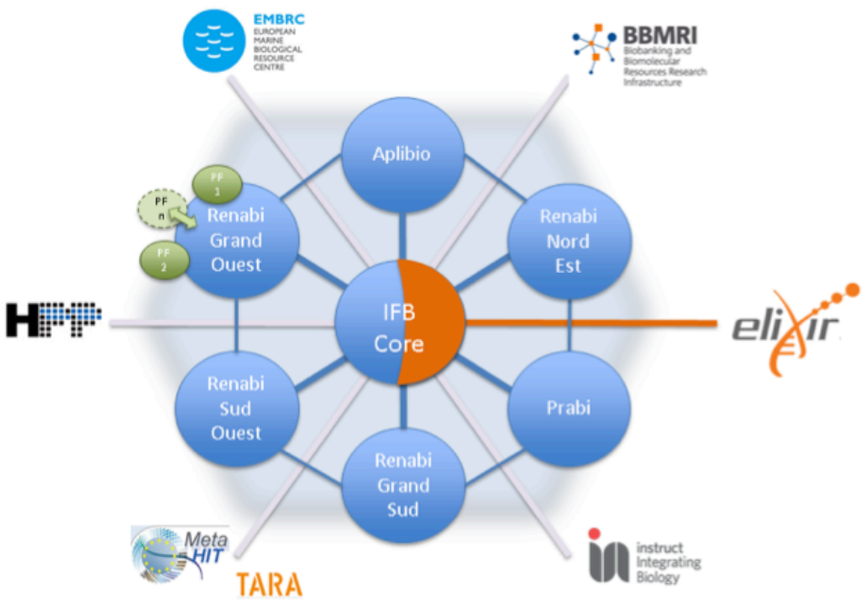
- **French Institute of Bioinformatics - IFB**
- Set up a national appliances catalogue for bioinformatics
- Deploy a national bioinformatics cloud
- Deploy mid-scale clouds
  - in regional centers
    (already done in Lyon and Rennes)
  - in life science platforms



French Institute of Bioinformatics - IFB

Mission : to make available core bioinformatics resources to the national/international life science research community.

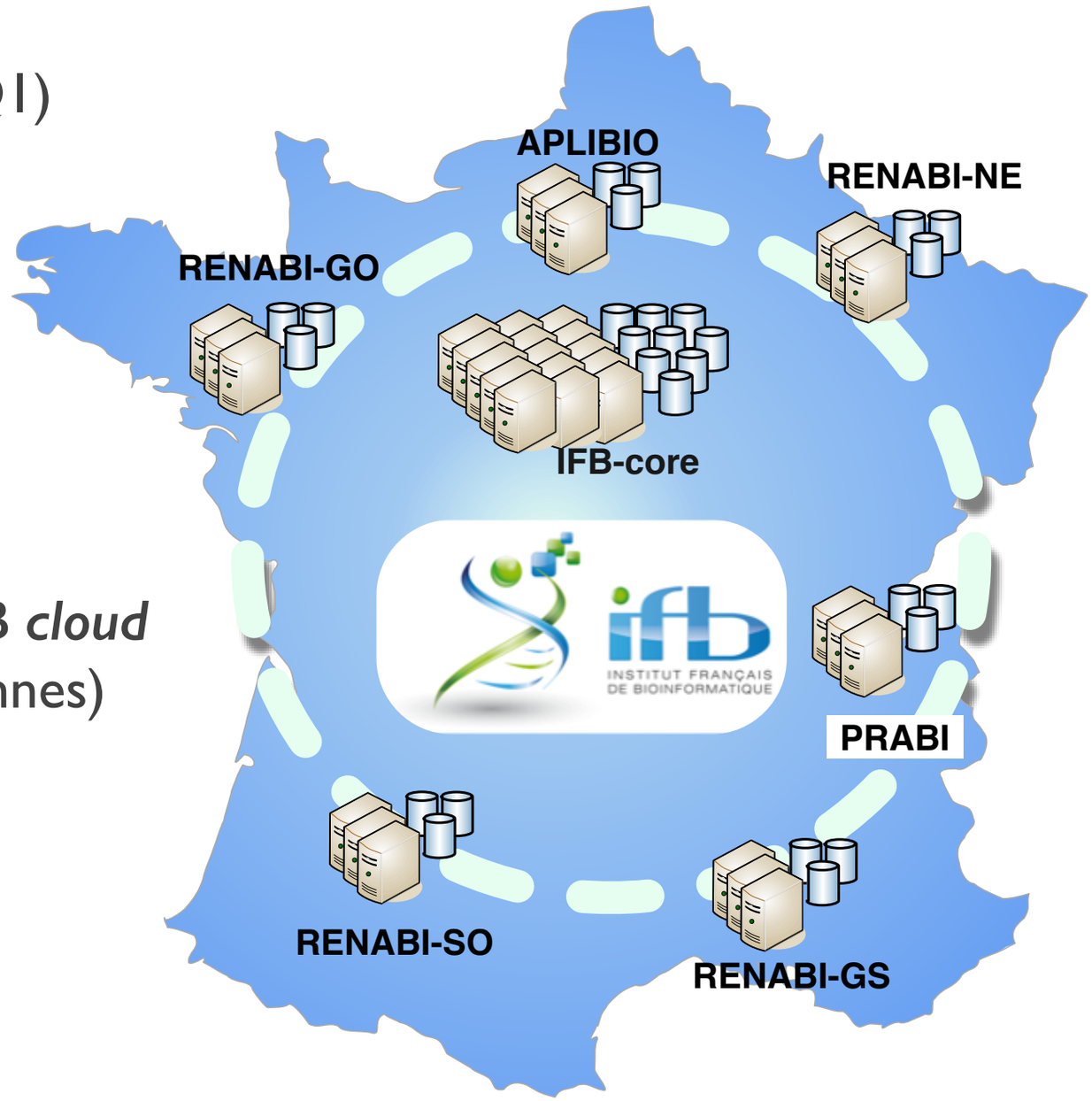
- To provide **support for national biology pro**grams
  - supporting projects
  - training users
- To provide an **IT infrastructure** devoted to management and analysis of biological data
  - material resources : CPUs, disks, etc.
  - availability of biology data collections
  - deployment of bioinformatics tools
- To act as a middleman between the life science community and the bioinformatics/ computer science research community



- **ELIXIR French Node**
  - optimizing the **interactions and coordination** between the national level and ELIXIR and other ESFRI infrastructures in biomedical and environmental field,
  - promoting **consistency and complementarities** between the components offered by the ELIXIR French node and those of other European nodes

Infrastructure

- IFB-Core resources
  - Academic **cloud for life science**
  - Will be hosted at **CNRS IDRIS** supercomputing center (PARIS)
    - A **pilot** infrastructure (2014-Q1)
    - **Production infrastructure**
      +5,000cores 1PB (2014-S2)
- + Regional resources
  - 6 regional bioinformatics centers
  - +6,000 cores ~1PB
  - but 20 bioinformatics platforms
  - 2 existing clouds: PRABI-IBCP *IDB cloud* (Lyon) & Genouest *genocloud* (Rennes)
- Deploy a federation of clouds



Questions ?

Acknowledgments

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

http://idee-b.ibcp.fr