

Institut de Biologie et Chimie des Protéines

Bioinformatique sur grille et cloud

Christophe Blanchet

Institut de Biologie et Chimie des Protéines

IBCP - LYON - FRANCE

christophe.blanchet@ibcp.fr

Institute of Biology & Chemistry of Proteins



- Dir. Prof. G. Deléage, LYON, FRANCE
- ~180 people. Associated to CNRS and University of Lyon
- «Protein Science » center
- Study of proteins in their biological context
- Approaches used include : integrative cellular (cell culture, various types of microscopies) and molecular techniques, both experimental (including biocrystallography, and nuclear magnetic resonance) and theoretical (structural bioinformatics)
- 3 departments, 14 groups
- Topics such as cancer, extracellular matrix, tissue engineering, membranes, cell transport and signalling, bioinformatics and structural biology

<http://www.ibcp.fr>



Lyon
France



christophe.blanchet@ibcp.fr



TIDRA, 14 déc. 2010, Lyon

Bioinformatics Requirements

- Etat de l'art
 - Deluge of data often complex and heterogenous, from different locations
 - Most common analyses have evolved to a larger scale,
 - study of a single gene/protein to a whole genome/proteome, a single metabolic pathway to Systems Biology...
 - Numerous tools with different behaviours (I/O, RAM)

=> Needs of distributed computing infrastructures for the storage, transfer and analysis of the biological data (fast access and short computing time!)

National Infrastructure RENABI GRISBI

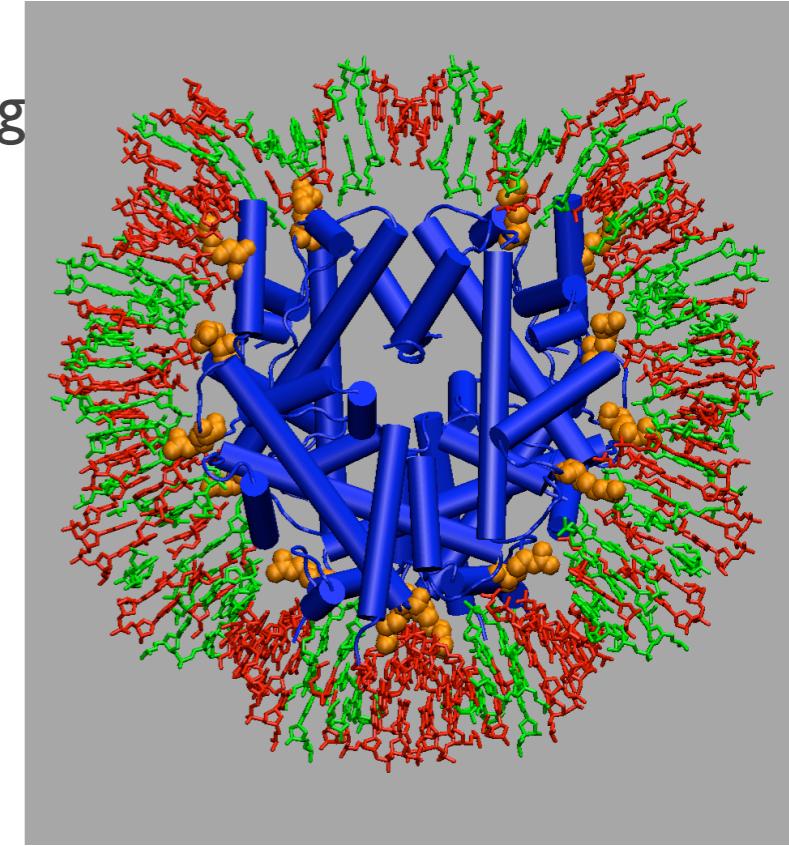
=> To give structure to the French community and propose solutions to satisfy the essential needs of biologists



- Bioinformatics distributed infrastructure
 - Financial support: **RENABI , IBISA 2008-2010, Institut des Grilles 2009-2010**
 - Resources: **1500 cores, 220 TB**
 - **www.grisbio.fr**
- **6 national platforms RENABI**
 - production PF for Bioinformatics
 - Approved by RIO / IBISA
 - **8 sites, with 7 CNRS / 40 participants**
- Collaboration with national computing infrastructures: Institut des Grilles, GENCI.

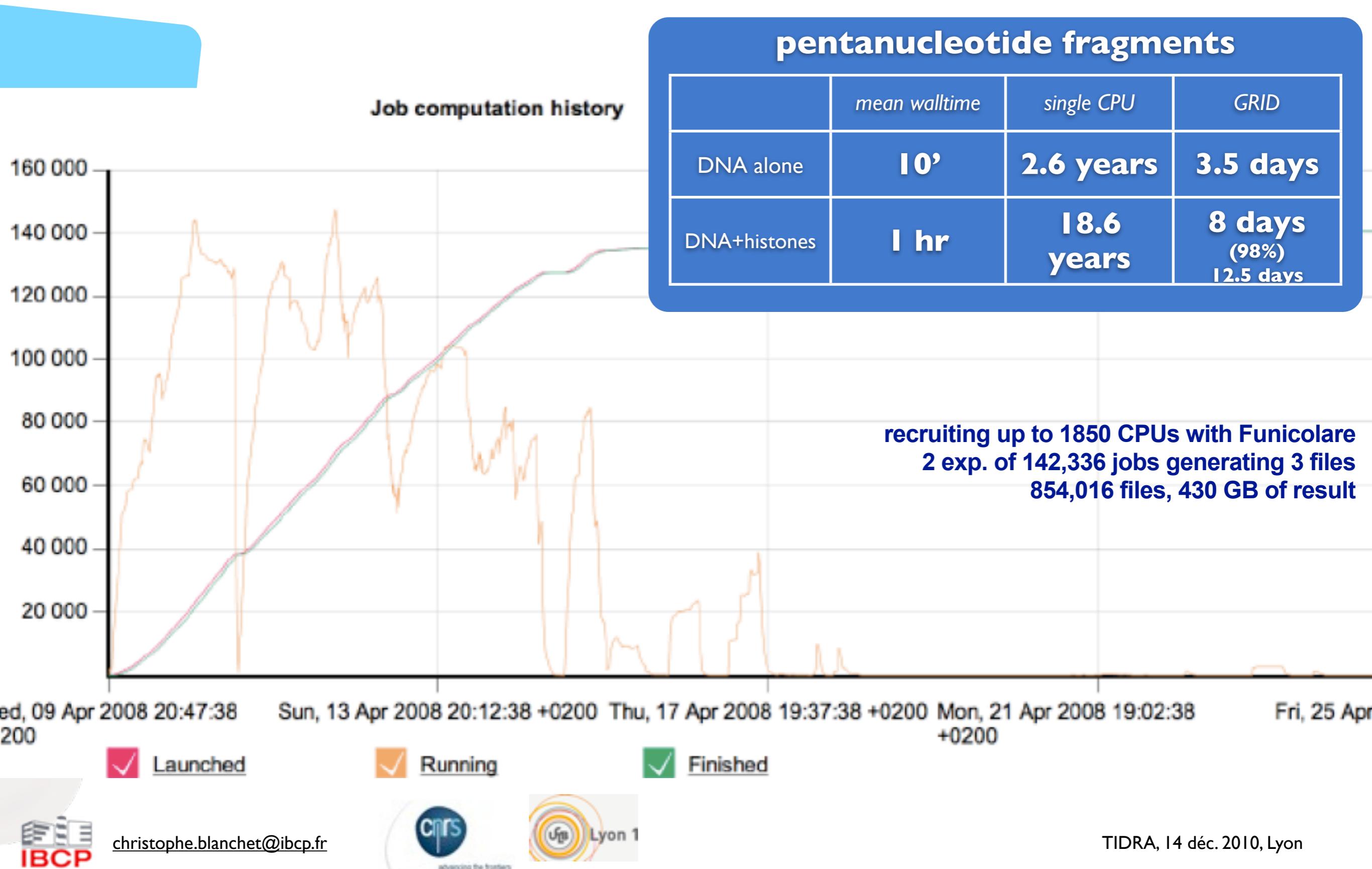
Study the Nucleosome

- Scientific Goal
 - Experimental data shows that proteins often find their target sites on DNA faster than simple diffusion would allow.
 - nucleosome involves an eight protein complex binding to roughly 140 bp DNA fragments: 1086 potential sequences
 - Determine the preferred nucleic base patterns for fixation of DNA on Histones
- Methods: ADAPT
 - JUMNA program developed in our team
 - Fortran g77 libraries (g77 compiler)
 - Reduction in combinatorial: overlapping fragments, "Nplets", involving N nucleotide pairs
 - moving one step along DNA: 143 positions
 - N=4, 4⁴=256 sequences, 35,840 simulations
 - N=5, 5⁵=1024 sequences, 146,432 simulations



Contact: richard.lavery@ibcp.fr

Nucleosome: Grid added value



Systems Biology on the Grid

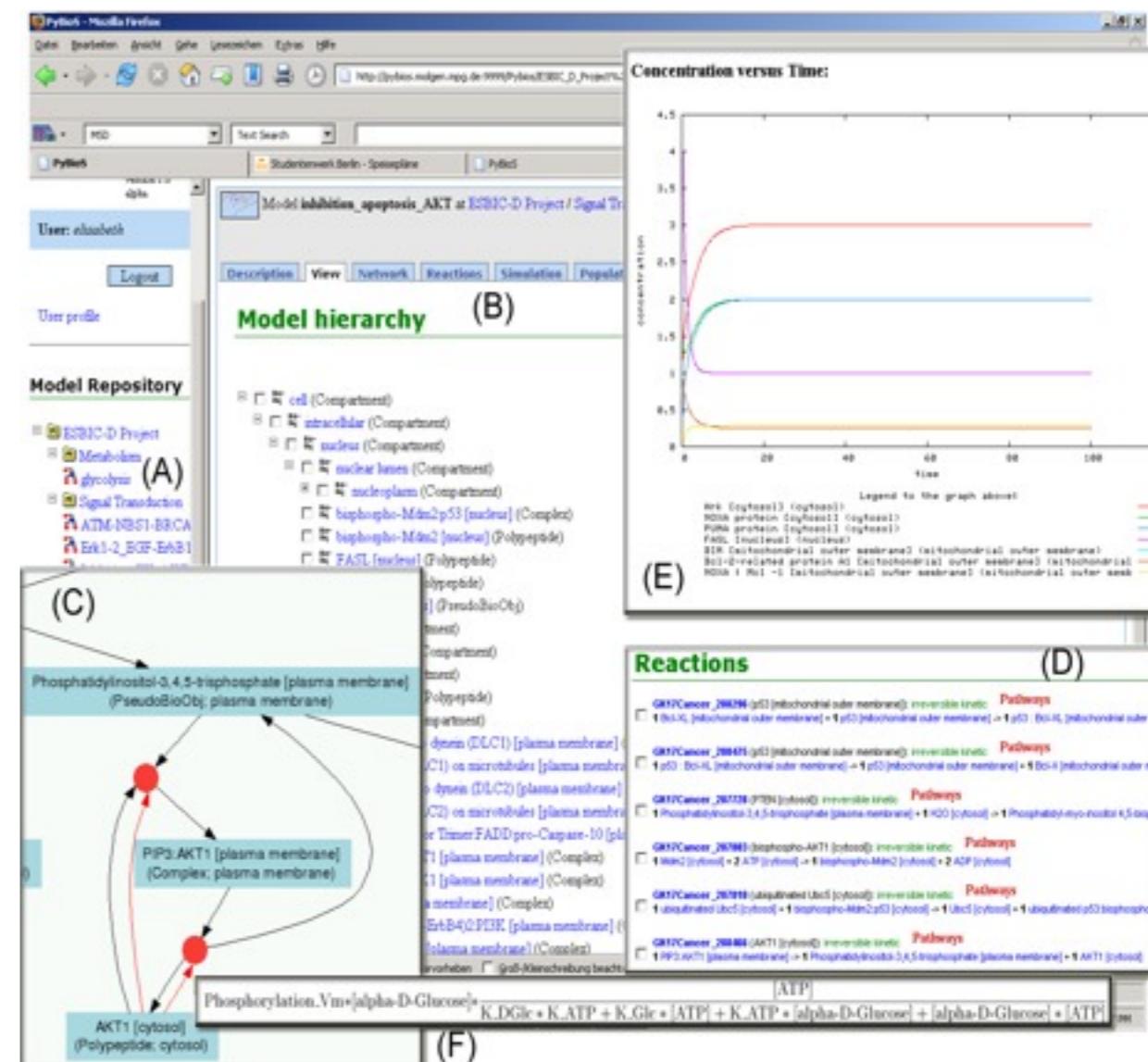
- Scientific objectives
 - Systems Biology: modelling and simulation of biological systems
 - Analysis of system behaviour in the context of experimental high-throughput data
 - Monte-Carlo simulation on the Grid
 - Applications on
 - Human melanoma cell-lines
 - Type-2 diabetes
 - Cancer Expression Data

Contact:

R. Herwig (Max-Planck Institute MG, Berlin)

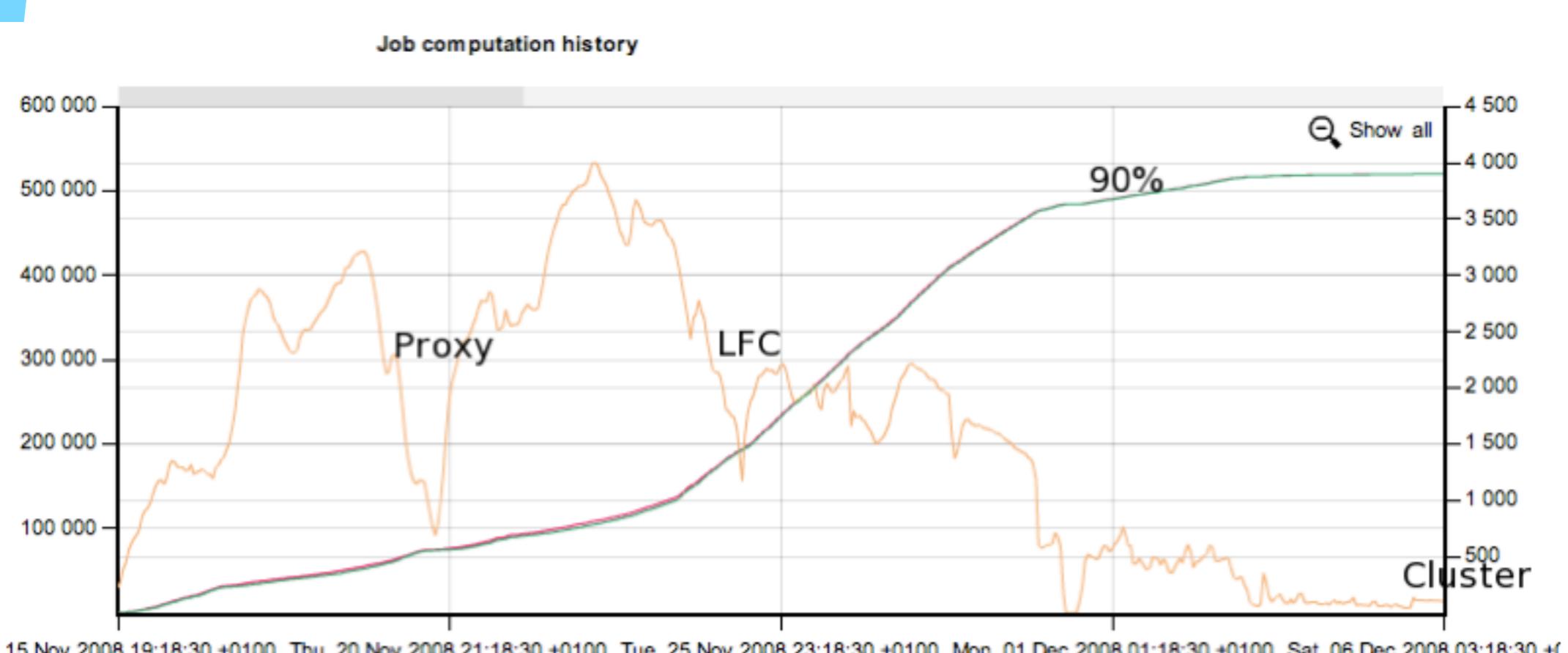
PyBioS

(<http://pybios.molgen.mpg.de/>)

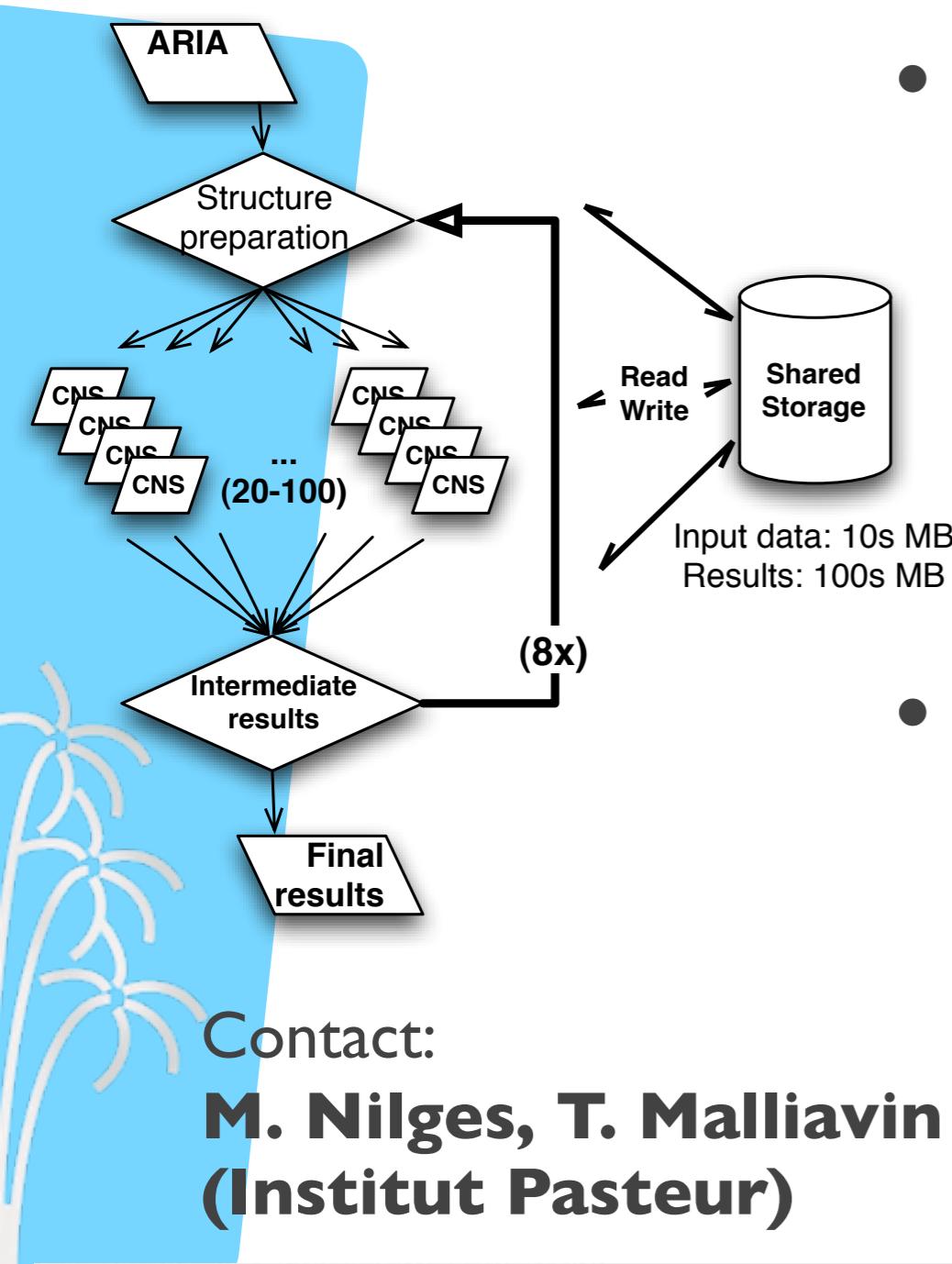


PyBioS: Grid added value

- Execution time **17 days** (speed-up 2200x)
 - Single CPU : **102 years** (910,000 hours)
- Recruiting up to 4,000 CPUs
- ~ 1,590,000 result files => 1,35 TB of data



Structural Biology: ARIA on GRID

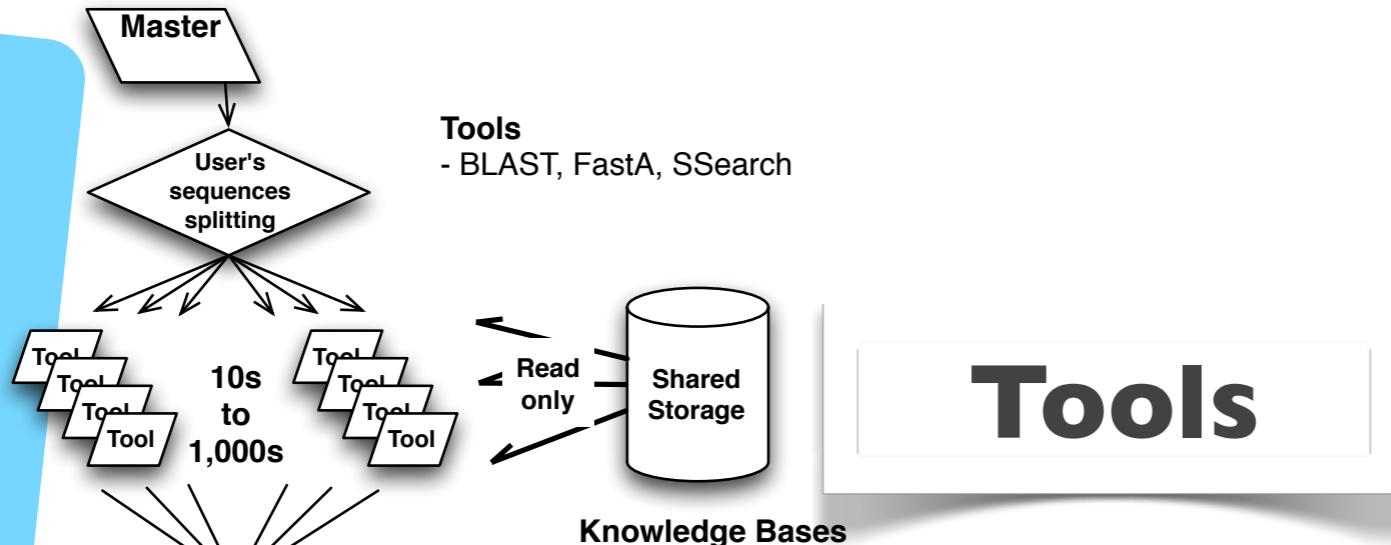


- ARIA GRID Mode
 - from liquid RMN data
 - Requirement about executable: CNS has problem with x86_64
 - CNS compiled on a centos 5 is not supported by a ScientificSL 4.6
 - InputSandbox: cns_solve, csh script, tarball with CNS working dirs
 - run2/structure/it , run2/cns aria_temp.../run_cns_ and eventually pdb
 - OutputSandbox: tarball with run2 and aria_temp...
- ARIA Job management modifications
 - Submits job with glite, check if the job submission is successful.
 - If proxy is not define: stop aria ; If not success : resubmit, If success : write the JobID into a variable
 - monitors job with the JobID and gLite commands:
 - If job is aborted : resubmit ; If job is Done but not successfully : resubmit ; If job is Done and successfully : download archive of job

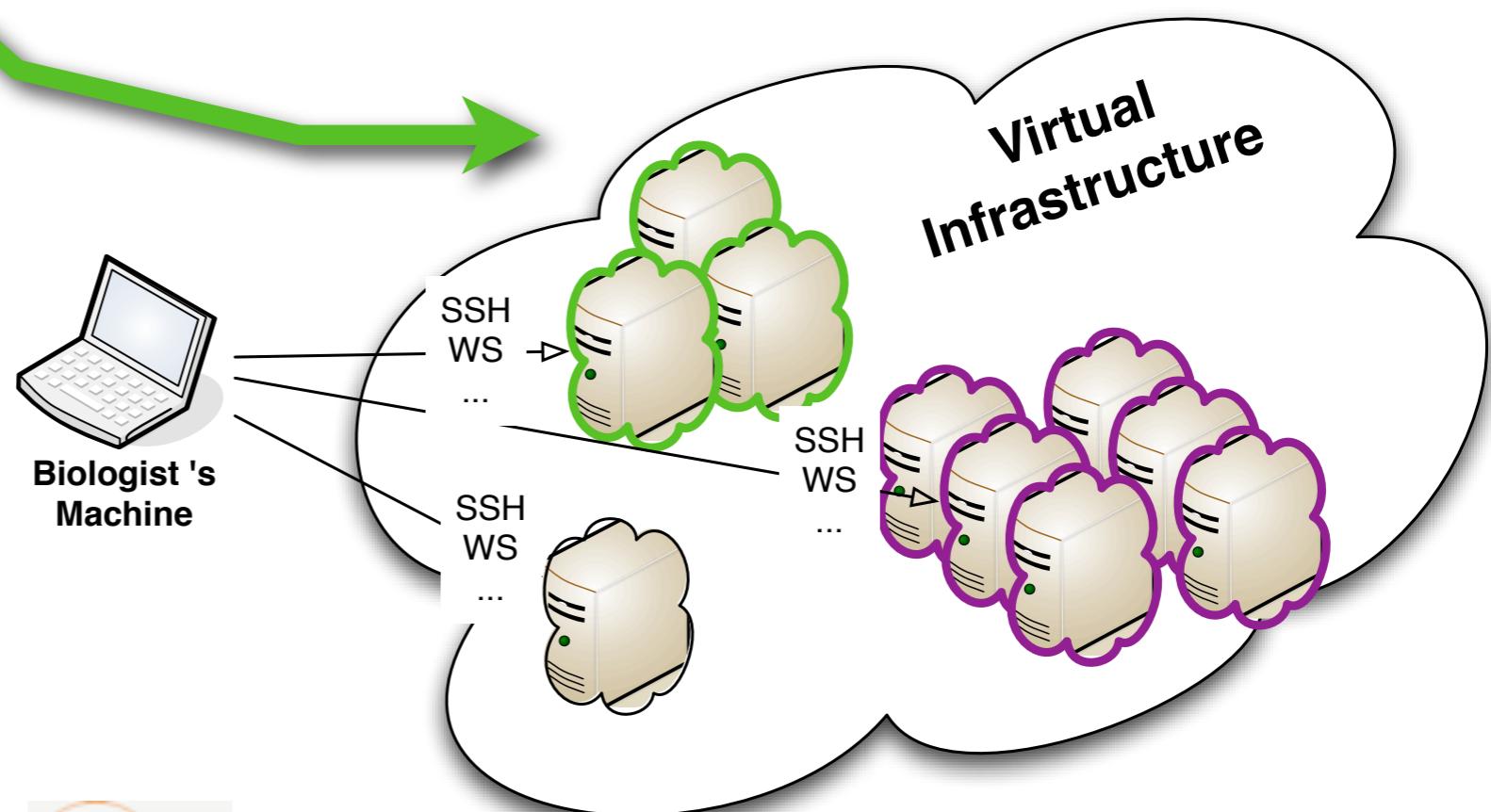
Contact:
M. Nilges, T. Malliauin
(Institut Pasteur)

```
Executable = "refine.csh";
Requirements = (other.GlueHostArchitecturePlatformType == "x86_64");
Rank = other.GlueCEStateEstimatedResponseTime;
InputSandbox = {"/home/grisbi/fmareuil/aria/examples/dimer/aria_temp/tmpNmUHwL1269614909/run_cns_28/refine.csh"};
OutputSandbox = {"aria_run_cns_28.tar.gz"}
```

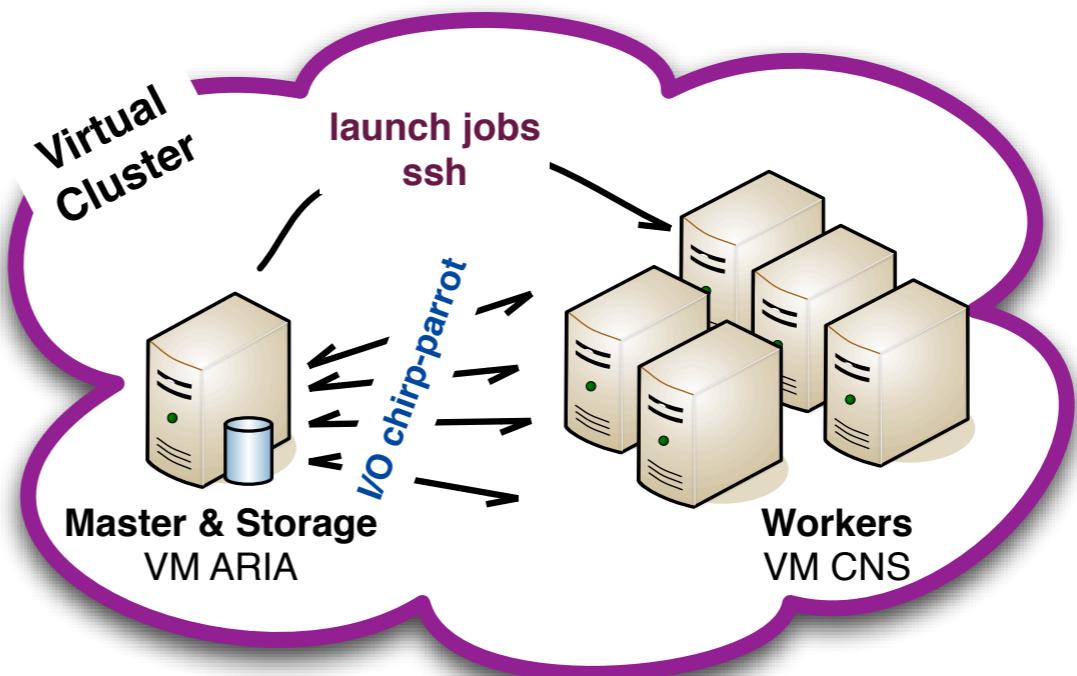
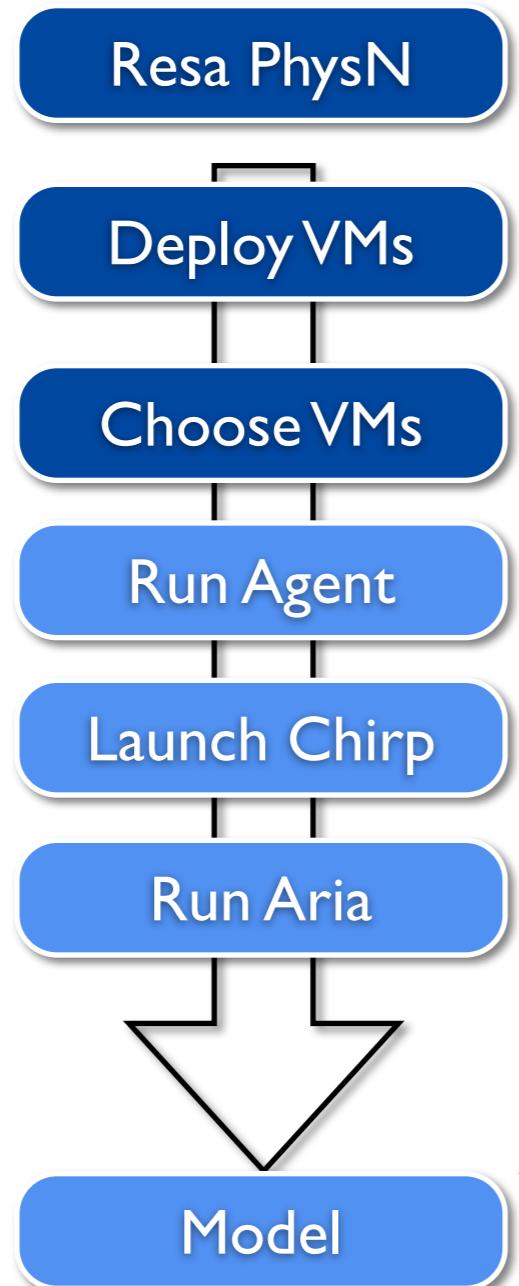
Bioinformatics on Cloud



IBCP's cloud
5 servers
40 cores, 160GB



Structural Biology on Cloud



2 VM types
ARIA; CNS

ARIA application: M. Nigels,
T. Malliavin, Institut Pasteur Paris

20 struct.

2 steps

Select Structures

Write Structures

Calculate Struct.
(CNS)

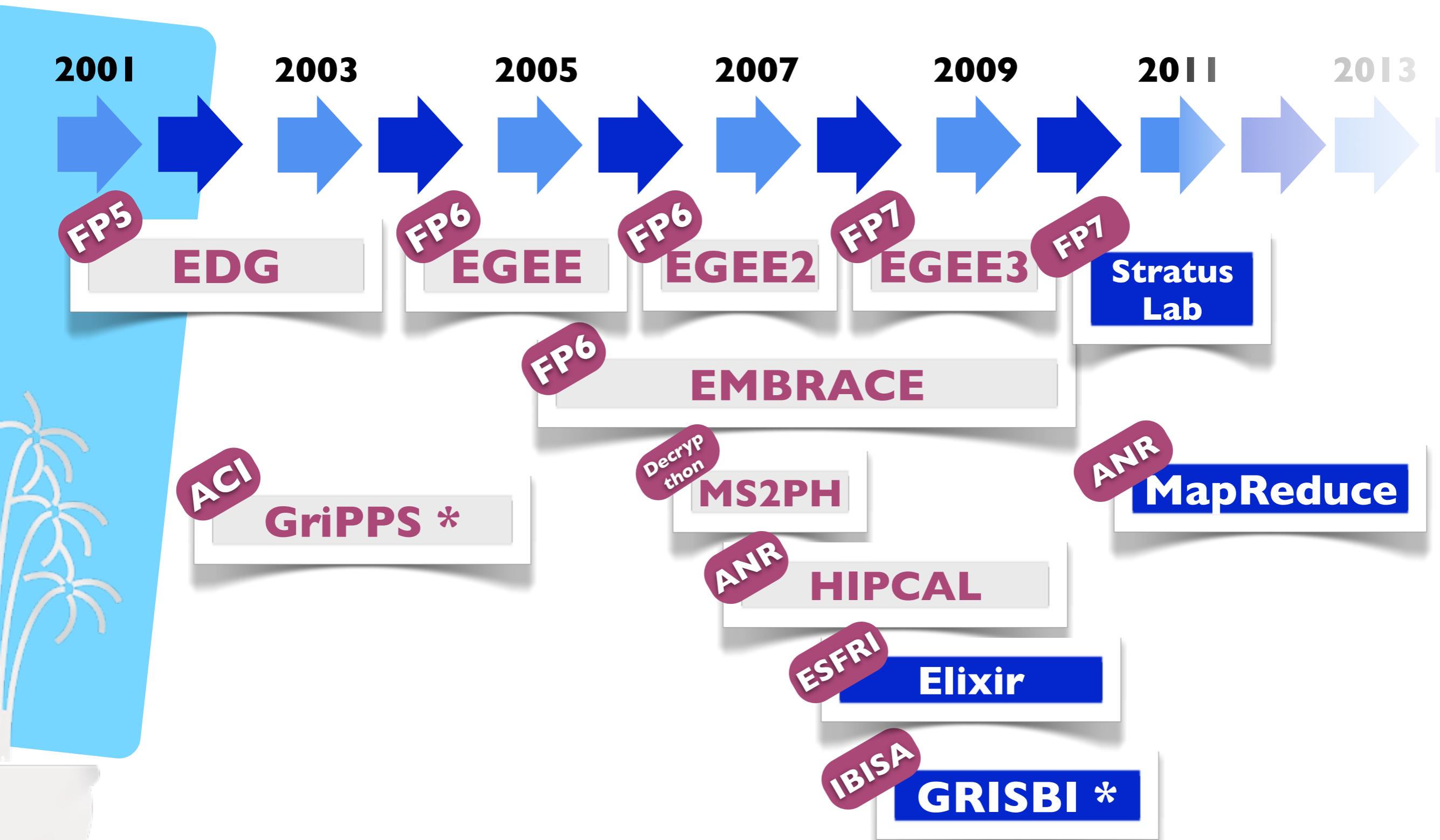


Bioinformatics in StratusLab

- Deploy sites ‘without’ grid/cloud expertise
 - have base images: UI, SE, CE, WNs
 - only have to put local param in context
 - enoughly validated to go in production mode
- Bioinformatics WNs on-demand
 - with pre-defined bioinformatics apps
 - specific: 24cores 16GB vs 2cores 96GB
 - coupled with grid jobs WMS
- Bioinformatics UI
 - with pre-defined bioinformatics apps
 - and configuration (RAM intensive, GUI, ...)
 - with their own certificate

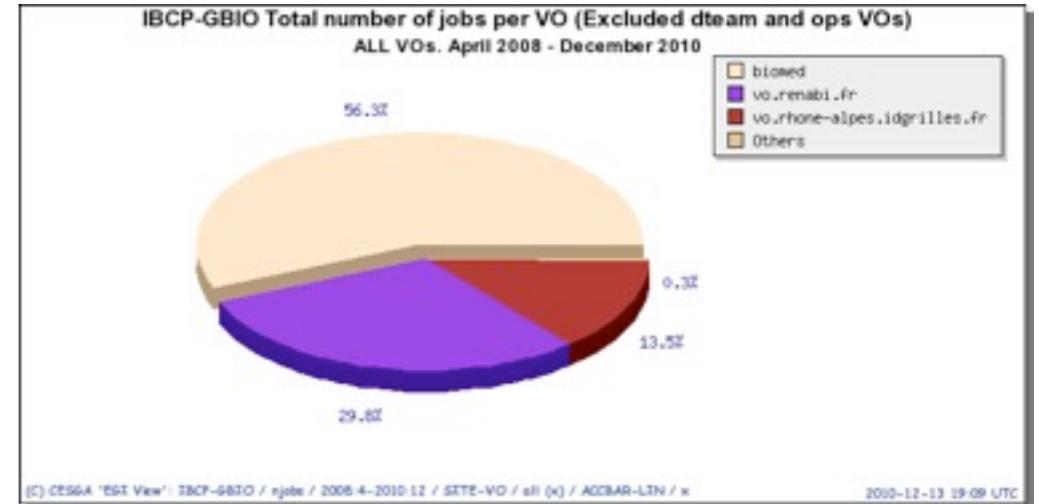


IBCP's Grid and Cloud projects

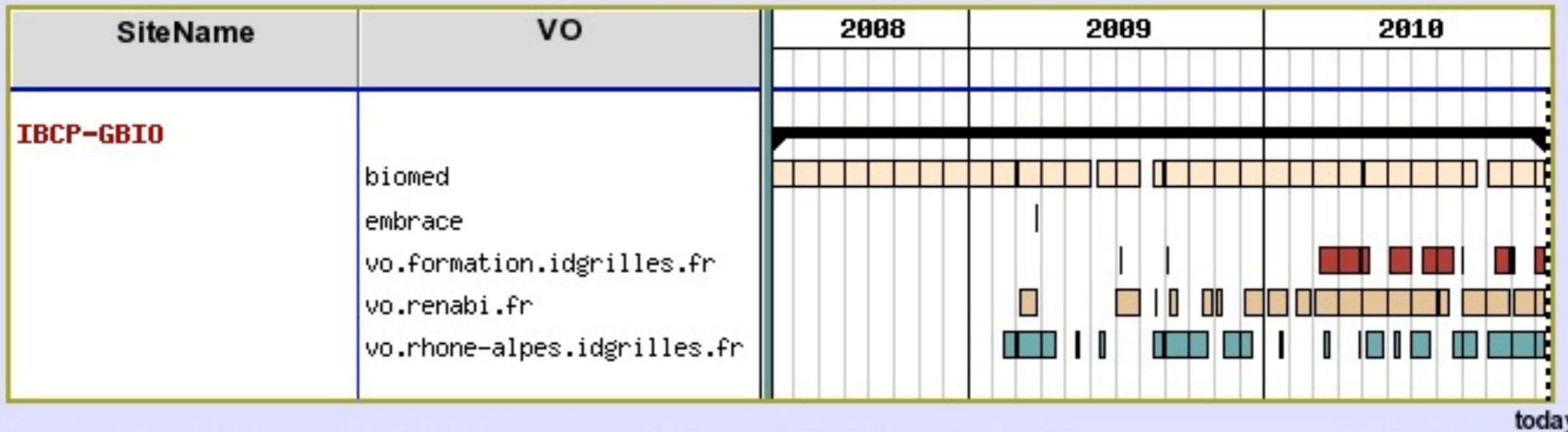


Grid Site in a Biology Lab

- 150 cores, 10 To
- Specific nodes (in p.)
 - 48 cores/128 GB RAM
 - GPU

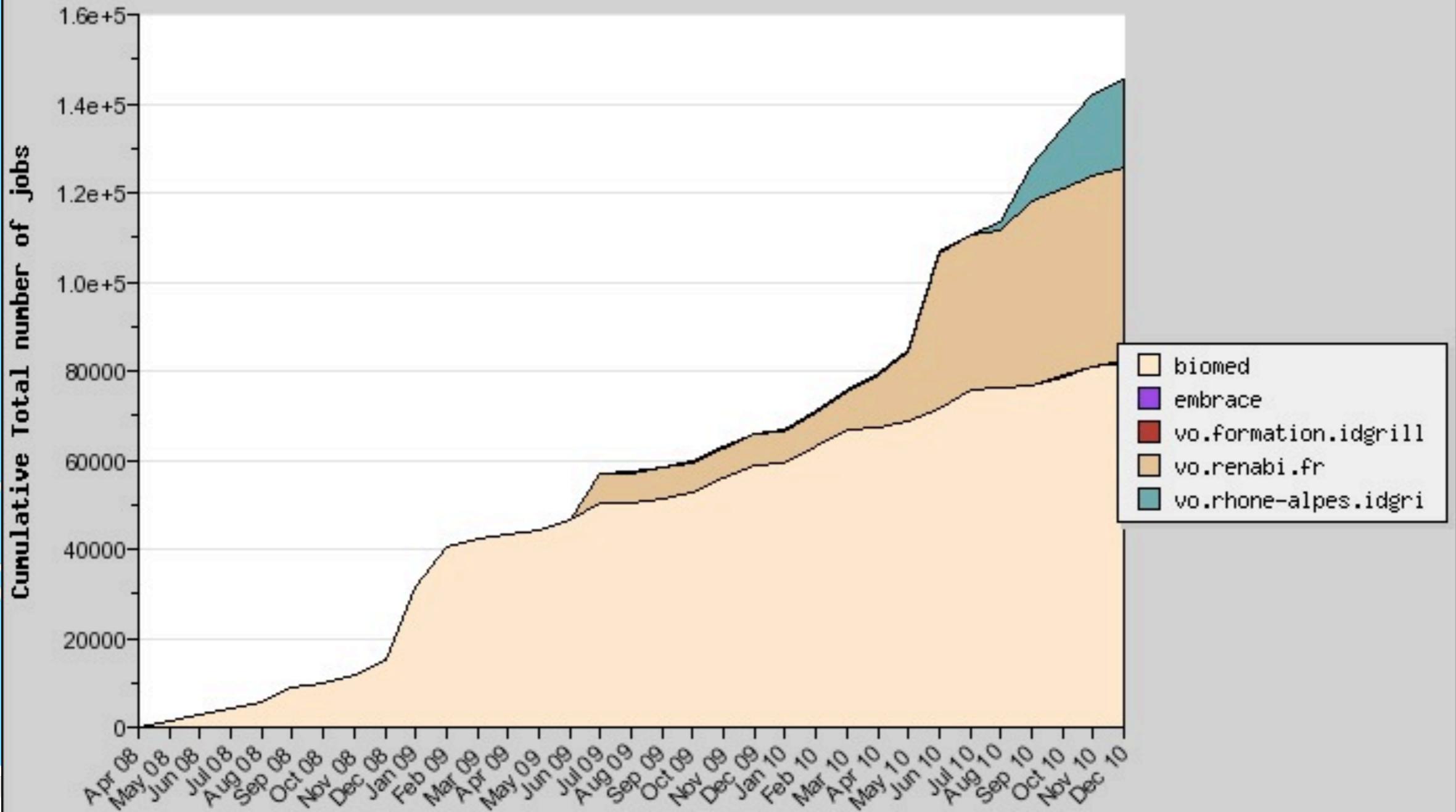


Accounting Data for IBCP-GBIO (Excluded dteam and ops VOs)
ALL VOs. April 2008 - December 2010



Providing Resources

IBCP-GBIO Cumulative Total number of jobs by VO and DATE (Excluded dteam and ops VOs)
ALL VOs. April 2008 - December 2010



Remerciements

CNRS - Centre National de la Recherche Scientifique

University of Lyon

French GIS IBISA through the project GRISBI PF 2008

French ANR through the HIPCAL project

The European Commission through the projects

- FP7 StratusLab -- 2010-12 -- RI-261552
- FP5 EMBRACE -- 2005-10 -- LHSG-CT-2004-512092
- FP7 EGEE-III -- 2008-10 -- INFSO-RI-222667



advancing the frontiers



CNRS IBCP: C. Eloto, C. Gauthey, A. Joseph, A. Michon, E. Bettler, A. Bockmann, C. Combet, G. Deléage, R. Lavery, F. Penin, K. Zakrzewska

RENABI GRISBI: C. Caron, O. Collin, and partners

CNRS IBCP

Institute of Biology and Chemistry of Proteins
7 passage du Vercors, 69007 LYON, FRANCE

Christophe.Blanchet@ibcp.fr

04 72 72 26 71

http://gbio-pbil.ibcp.fr/cblanchet

- **Domaine d'application**
 - Science de la Vie - Bioinformatique
- **Site de grille dans un laboratoire de Biologie (04/2008)**
 - Expertise dans la mise en oeuvre des solutions informatiques
- **Expertise pour le portage de vos applications**
 - Identification des besoins et contraintes liées à la grille
 - Conseils sur l'adaptation de vos applications
 - Aide à la mise en Oeuvre
- **En lien avec nos collègues locaux : LBBE, IN2P3**