cgfb
CENTRE BORDEAUX
GÉNOMIQUE FONCTIONNELLE

LaBRI

GRISBI

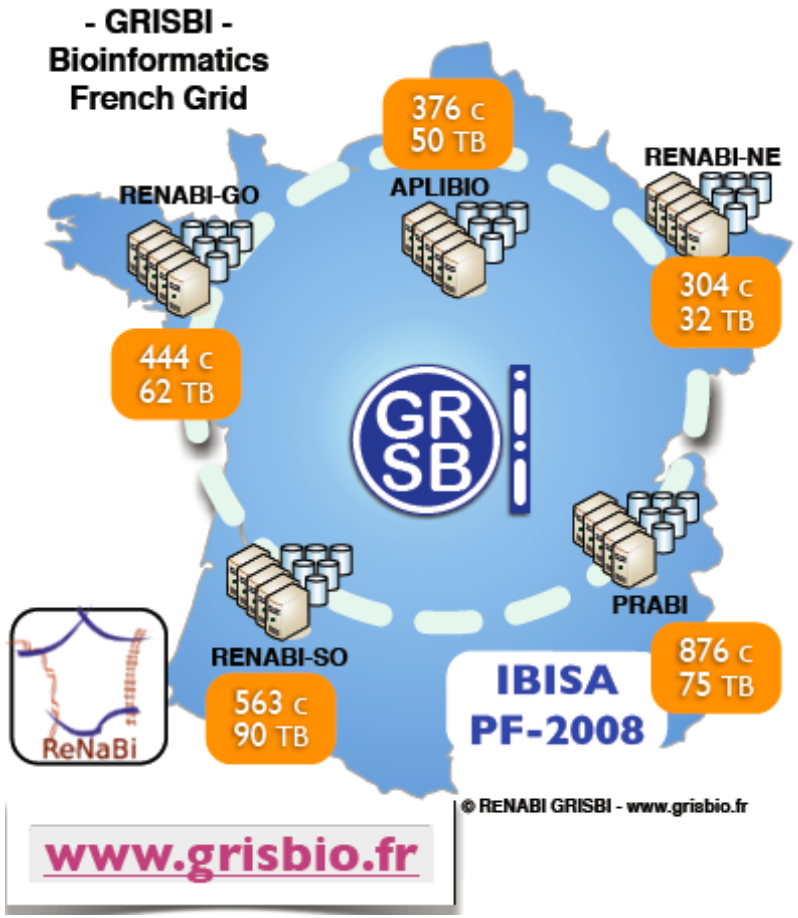
Use of a computing grid (GRISBI) for NGS data processing

French Grid, 19 Septembre 2011 - Lyon

Plan

- RENABI-GRSIBI
- Members
- Objectives
- Tools
- Data
- Installation/Use
- Results
- Comments
- Perpectives

- ➔ To structure the community and offer answers to the needs of biologists



.Think tank on organisation and technology

.distributed Infrastructure in bioinformatics

.5 regional Centers RENABI

.Label RIO/IBISA

.9 sites : 7 CNRS and 2 INRA

.Financial support : RENABI, IBISA 2008-2011, Institut des Grilles 2009-2010

.In PFs : 2700 nodes, 440 To storage

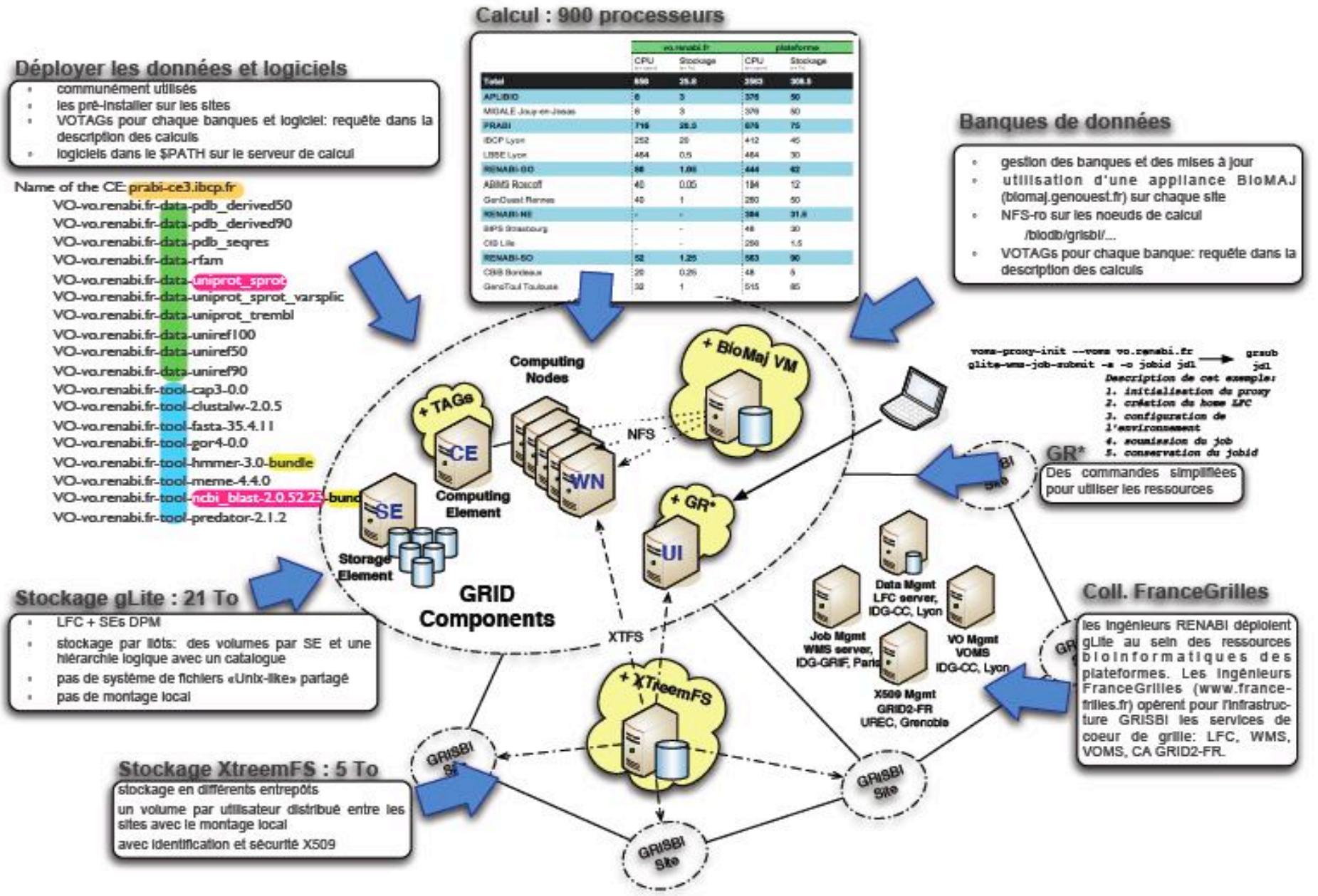
.In GRISBI : 860 nodes, 26 To storage

. ~70 registred members

.Collaboration with national infrastructures :

. Institut des Grilles, Genci, Mésocentres, GRID'5000, RENATER

Infrastructure RENABI-GRISBI



- Answer to biology needs
  - At the moment on data
- Pool existing infrastructures
  - Develop procedures for implementation in other centers
    - Renabi (CE gateways), FranceGrille (vo.renabi.fr)
- Adapt uses for biology community
  - Command GR*, biomaj, xtreemfs
- Training
  - Ecole CNRS 2010, tutorial 2011, …

- Study of cloud computing
  - Bring calculations to data : bioinformatic « appliances »
  - Collaboration with StratusLab project

Members of NGS project

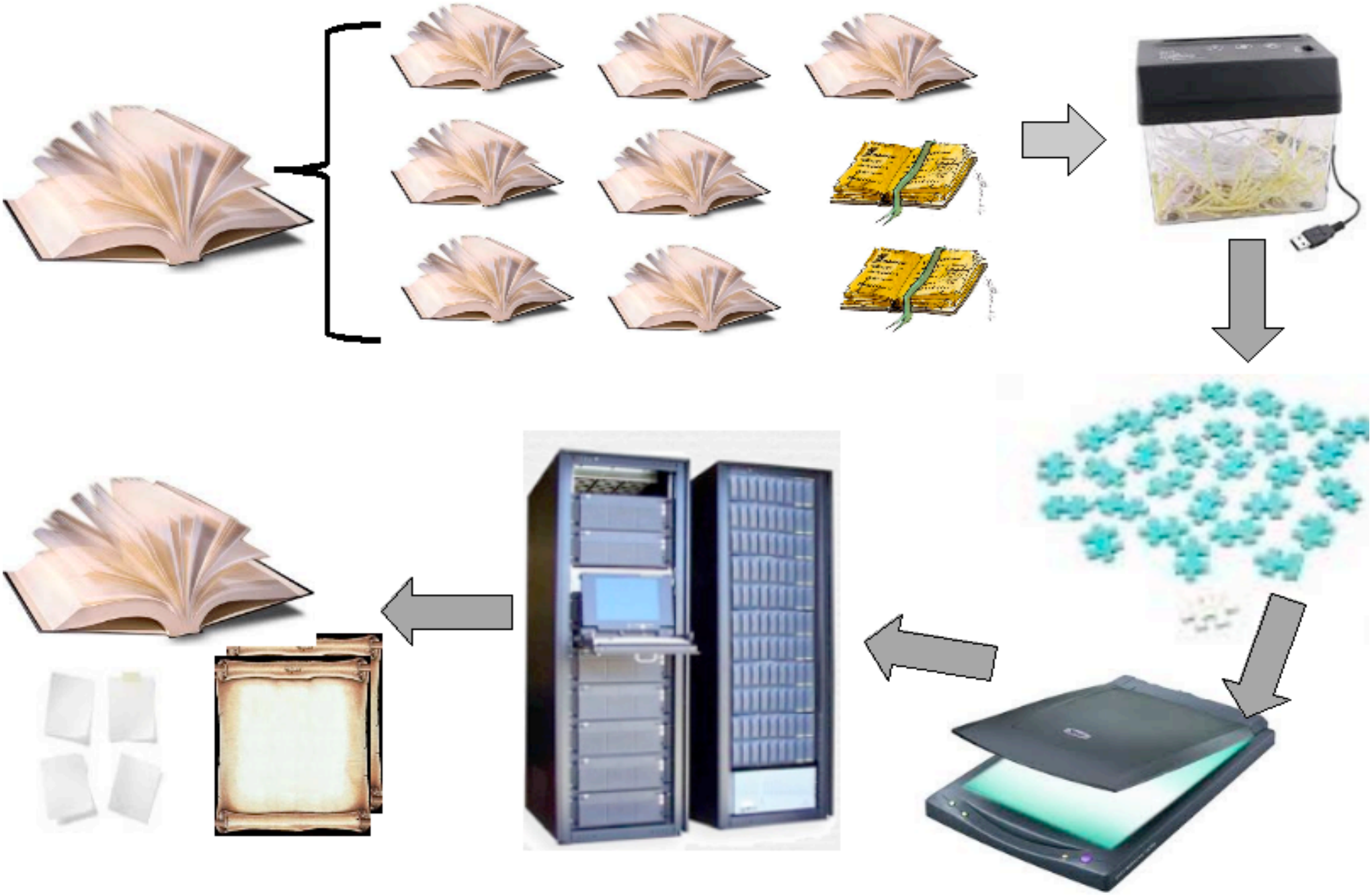
- Bordeaux, CBIB, LaBRI : Project leader, Ergatis*, PLAFRIM
  - Tiphaine Martin, Florence Maurier, Alexis Groppi, Aurélien Barré
- Rennes, IRISA Symbiose : NGS Expertise
  - Delphine Naquin, Aurelien Roult
- Lyon, Centre "Infrastructure Distribuée pour la Biologie" IBCP FR 3302 : Grid Expertise
  - Clément Gauthey, Christophe Blanchet

- *Toulouse, GenoToul : Expertise on Ergatis
  - Jérome Mariette, Céline Noirot, Christophe Klopp

Objectives

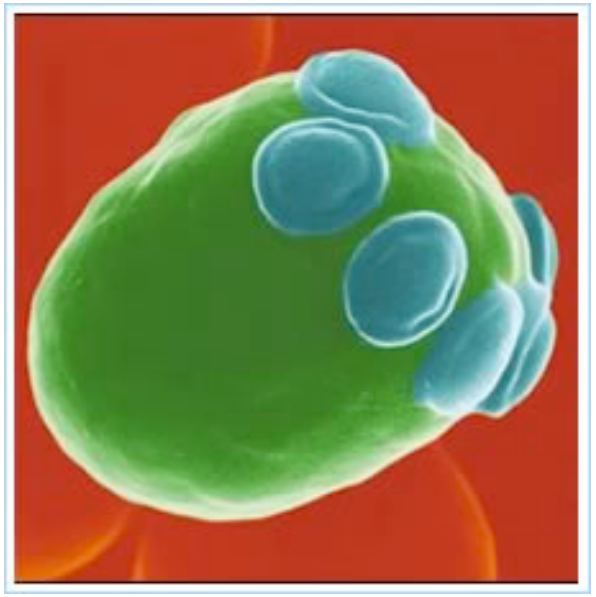
- **Feasibility study of the NGS data processing on grids**
  - **Alignment on reference genome**
  - ***de novo* Assembly**
- **Comparison with local server and experimental cluster**
- **Start jobs automatically from a workflow management system like Ergatis**

Book analogy



- **Bwa : (bwa-0.5.8c)**
  - Read alignment on reference genome
  - http://bio-bwa.sourceforge.net/
- **Abyss : (ABySS 1.2.6)**
  - *de novo* assembler of reads in single-end, mate-pair, paired-end from 454 and Illumina technologies
  - http://www.bcgsc.ca/platform/bioinfo/software/abyss
- **Ray : (Ray-1.2.4)**
  - *de novo* assembler (Solid, 454, Illumina) using MPI 2.2
  - http://sourceforge.net/projects/denovoassembler/

- **Genome : *Saccharomyces cerevisiae***
  - 16 chromosomes
  - Genome size : 12 272 524 pb
- **Simulation reads with Metasim (version 0.9.5) (cover 30X) in**
  - 454 and Illumina technologies and,
  - single-end and paired-end



- On Grid : GRISBI (http://www.grisbio.fr/bioapps/votags)
  - Bwa installation OK
  - Installation more complexe ABySS and Ray but ok
- On experimental cluster :PLAFRIM (https://plafrim.bordeaux.inria.fr/doku.php?id=start)
  - BWA and ABySS no problem (in module)
  - Problem for Ray with mpich2
- Launch *via* Ergatis (http://www.cbib.u-bordeaux2.fr/ergatis/)
  - BWA installed and inserted in a Ergatis module
  - ABySS et Ray no installed because need more memory
  - Creation a module to prepare data for grid job (zip archive containing data and scripts)

- **Simulated data of *Saccharomyces cerevisiae* : Illumina single-end**

Tool	Local*	PlaFRIM**	GRISBI***
BWA	index : 20 s aln : 553 s samse : 80 s	Index : 7s aln : 306 s samse : 45 s	Index : 15 s aln : 485 s samse : 70 s
ABySS	8 min	7 min	10 min
Ray	« std ::bad alloc »	Problem mpich2	85 min

•RAM 3Go, Processor 2,67 GHz ** RAM 24Go, Processor 2,93GHz

•*** IBCP : 4 sites from 32 to 252 Go RAM and from 8 to 200 CPUs

•Out transfer time and queue time for GRISBI and PLAFRIM

- **Simulated data of *Saccharomyces cerevisiae* with ABySS assembler**

Data	Local*	PlaFRIM**	GRISBI***
Illumina single-end	8 min	7 min	10 min
Illumina paired-end	67 min	33 min	81 min
454 single-end	« std ::bad alloc »	60 min	162 min

•RAM 3Go, Processor 2,67 GHz ** RAM 24Go, Processor 2,93GHz

•*** IBCP : 4 sites from 32 to 252 Go RAM and from 8 to 200 CPUs

•Out transfer time and queue time for GRISBI and PLAFRIM

- **Simulated data of *Saccharomyces cerevisiae* : Illumina single-end for 10 executions with different k-mer**

Tools	Local* (successives executions)	PlaFRIM** (array job for pbs)	GRISBI *** (parametric job)
ABySS	8 min x 10 ~ 80min	< 15 min (8 nodes)	<60 min (2 nodes)
Ray	« std ::bad alloc » (85 min X 10 ~ 850min)	Problem mpich2	~ 180 min

•RAM 3Go, Processor 2,67 GHz ** RAM 24Go, Processor 2,93GHz

•*** IBCP : 4 sites from 32 to 252 Go RAM and from 8 to 200 CPUs

•Out transfer time and queue time for GRISBI and PLAFRIM

Advantages

- Accession evrywhere and different types of work nodes
  - But need to GRID2FR certificate
- Centralized installation and maintenance of tools and databases
- Relatively simple use of tools already controlled on other infrastuctures (local server)

- Gain Grid *vs* local server:
  - possibility to launch parallel jobs
- No advantage grid over cluster

- Accession of job models (bash, JDL) on GRISBI site (cf. Tutorial June 2011)

- Need to be familiar with tools because difficult to obtain explicite error return
  - ➔ test locally before tool with data (format files input / output command to execute)
- Check in bash scripts presence of data befor processing (verbose)
- If job lasts more 12 hours, open long proxy
  - myproxy-init

- Impossible to define size of desired memory
- Bandwidth limitation for some GRISBI nodes
  - ➔ Compress data
  - Need bandwidth (1Gbps, voire 10Gbps)

- Problem for parametric jobs :
  - If one of them is stuck in waiting/submitted state, not back data in Outputsandbox
  - ➔ Replace grout with glite-wms-job-output <job_id>
- Verify state of CE befor launching job (fault CE (CPU free = 0) or blocked CE (Running/Waiting jobs higth)),
  - Reinforce Requirement in JDL
    - Requirement list : lcg-info --list-attrs
  - Launch again job *via* grsub
  - Last resort choose CE during submission
    - grsub -r <CE_name> <JDL_file>
- Only by command lines
  - ➔ need workflow management system, web service and web interface

Perspectives

- Installation and test Ray on PLAFRIM cluster
- Further comparison GRISBI GRID *vs* PLAFRIM cluster
- Update ABySS 1.3,0
- Install ABySS on other nodes if possible (need RAM)
- Launch automatically jobs via Ergatis on Grid

- Need workflow management system
- Need to launch jobs by web service or/and web interface
- Need to define size of desired memory
- Improve network

- **Participate in Assemblathon 2 project (june – sept 2011)**
  - http://assemblathon.org/
  - Use GRISBI and PLAFRIM with ABySS assembler
  - Obstacle on volumetric data (raw data 140 Go for *red tailed constrictor boa* genome)
  - ➔ Allow to validate full-scale projects of this scale and solutions that must accompany

- **Projects awaiting funding by CBIB partners :**
  - 20 genomes of spiroplasmes
  - 5 genomes of apricot trees

- www.grisbio.fr

- Acknowledgements :
  - Christophe Blanchet, Christophe Caron, Olivier Collin, Stéphane Delmotte, Christelle Eloto, Gael Even, Pierre Gay, Clément Gauthey, Daniel Jacob, Didier Laborie, Nouridine Melab, Alexis Michon, Frédéric Plewniak, Aurélien Roult, Franck Samson, Bruno Spataro