Grid assisted structure calculation of large protein systems in solid-state NMR context

C Blanchet(1), F Mareuil (2), A Joseph (1), A Loquet (1),
TE Malliavin (2), M Nilges (2), A Bockmann (1)

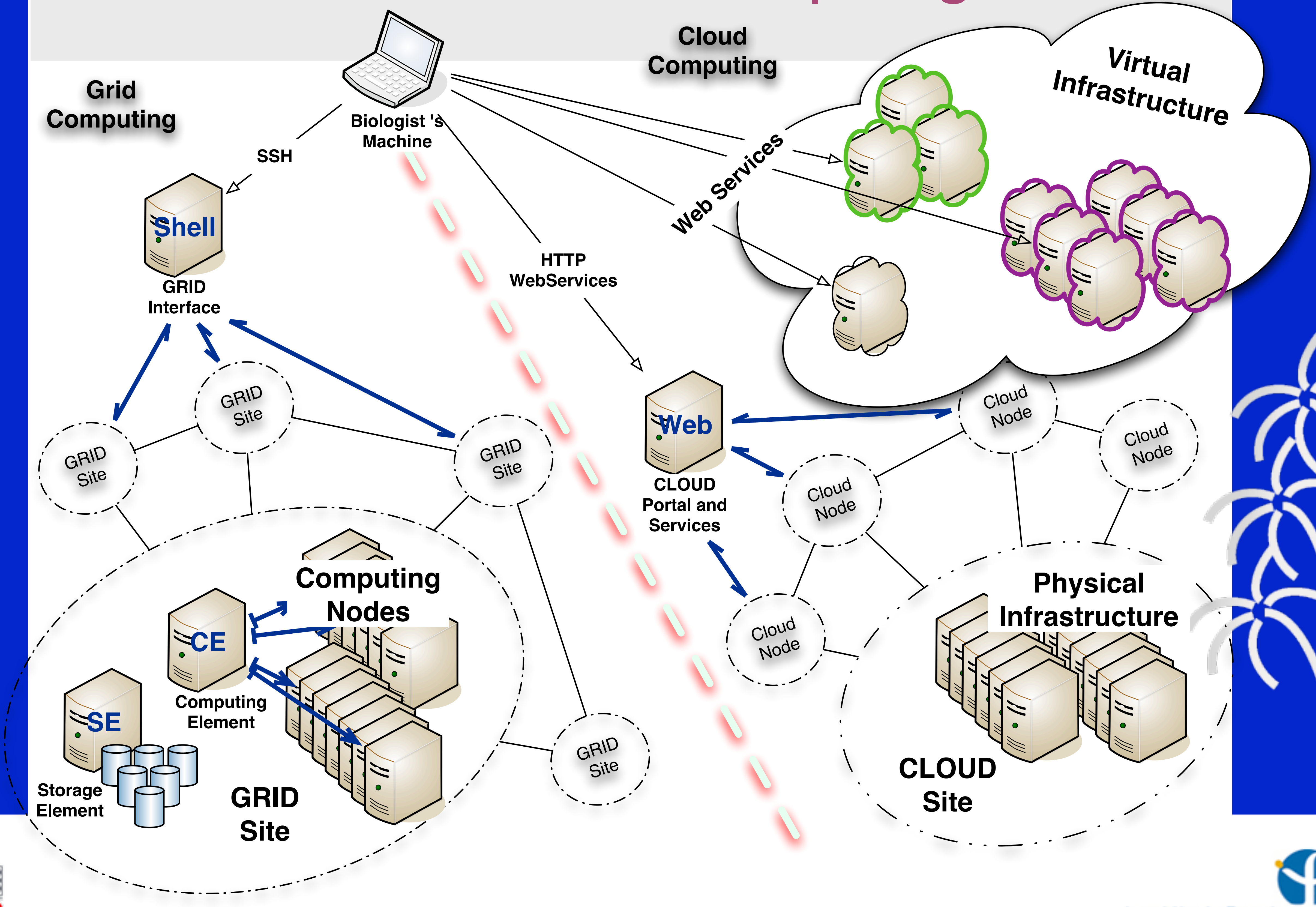
(1) Institut de Biologie et Chimie des Protéines, CNRS IBCP, LYON, FRANCE
(2) Unite de Bioinformatique Structurale, URA CNRS 2185 et Institut Pasteur, PARIS, FRANCE
christophe.blanchet@ibcp.fr

13 April 2010, EGEE User Forum 5, Uppsala

Solid-state NMR

- Solid-state NMR (ssNMR) spectroscopy
  - can address questions on structure, dynamics and interactions of insoluble proteins
  - valuable alternative to X-ray crystallography and solution NMR
- 3D structure calculation of proteins
  - Using NMR spectra as a source for structural constraints
  - ARIA (Ambiguous Restraints for Iterative Assignment)
    - Iterative assignment methods, based on successive simulation procedures , reliably assign NMR cross-peaks, calculate protein structures by using ambiguous distance restraints derived from the NMR cross-peaks intensities
- ssNMR increases demand of computing power
  - because of the low spectral resolution
  - increase the number of integration steps in the SA procedure, the number of protein conformations generated and the number of possible assignments explored
- Contact: A. Bockmann

Grid et Cloud Computing

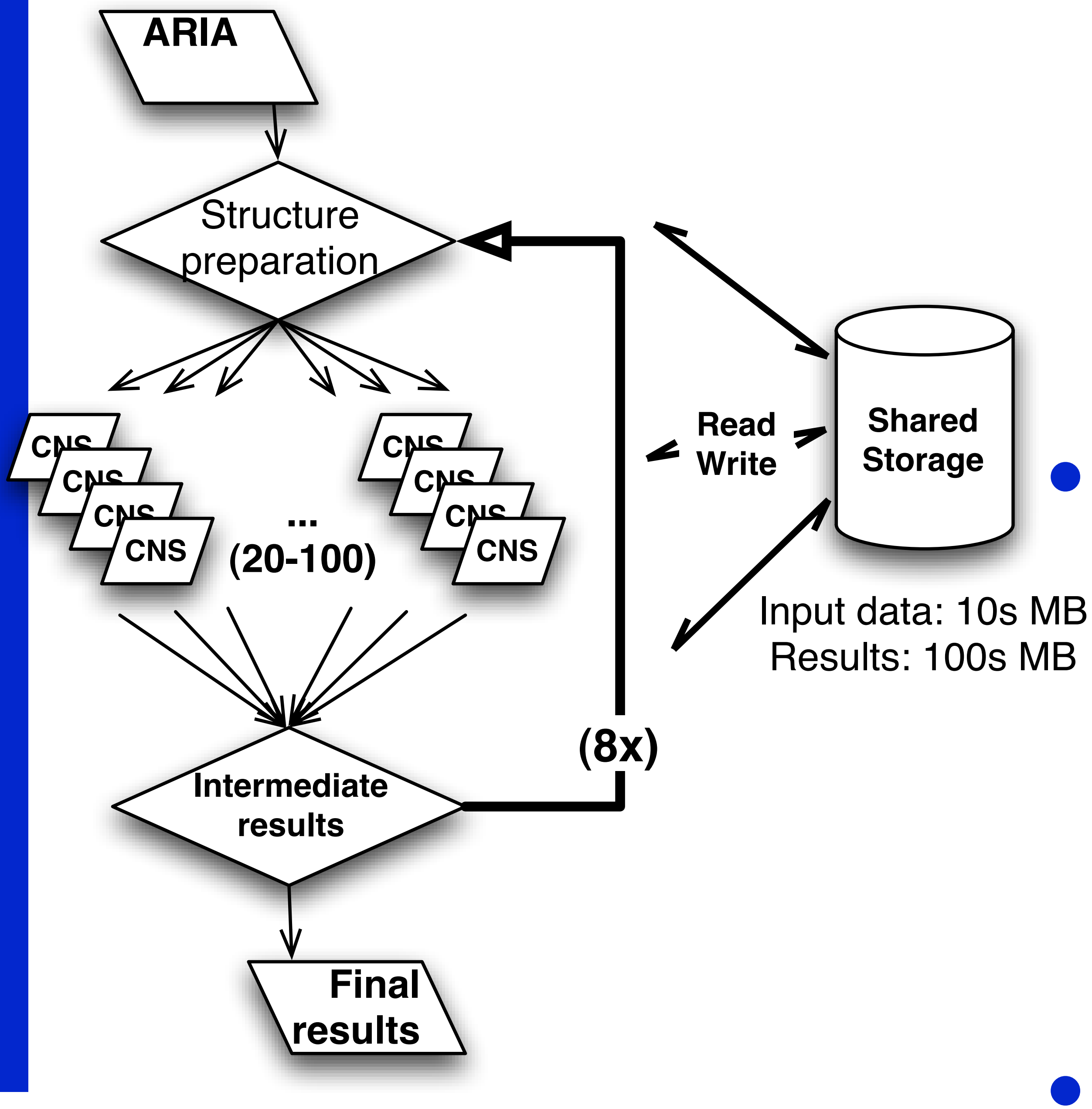


Christophe Blanchet, CNRS IBCP

Outline

- ARIA Software
- Deploying on GRID
- Deploying on CLOUDs

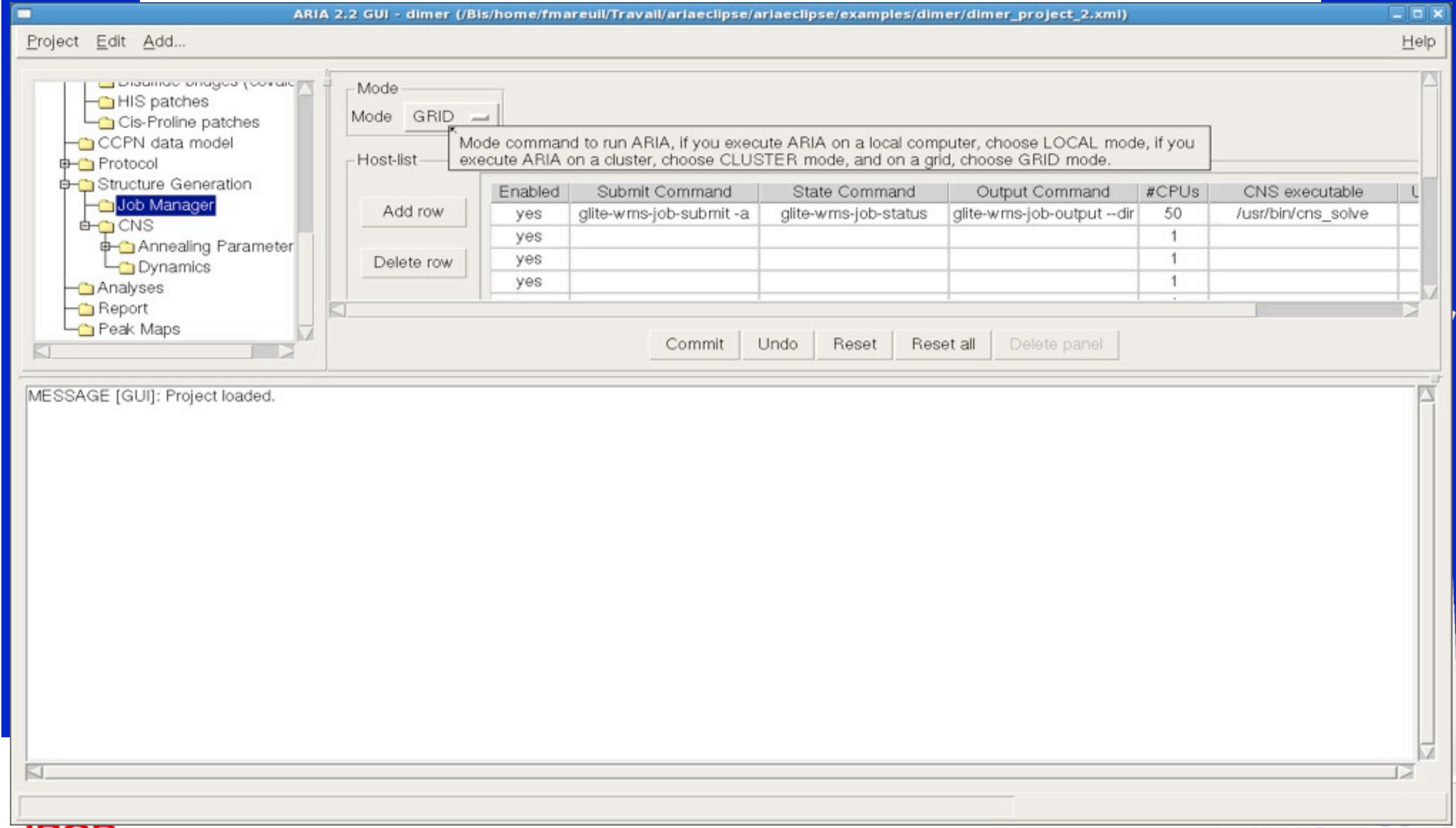
Software ARIA



- Typical ARIA procedure
  - 8 steps
  - conformations ranging from 20 to100 instances
  - between two steps, analyse of the calculated structures and definition of the new restraints
- Distributed computing benefits
  - run in parallel several structure determination
  - increase the capabilities of structure and assignment procedures on large systems
  - several experiments/users
- Contact: M. Nilges

Rieping W., Habeck M., Bardiaux B., Bernard A., Malliavin T.E., Nilges M. (2007) ARIA2: automated NOE assignment and data integration in NMR structure calculation. Bioinformatics 23, 381-382.

ARIA Graphic interface



Mode: GRID

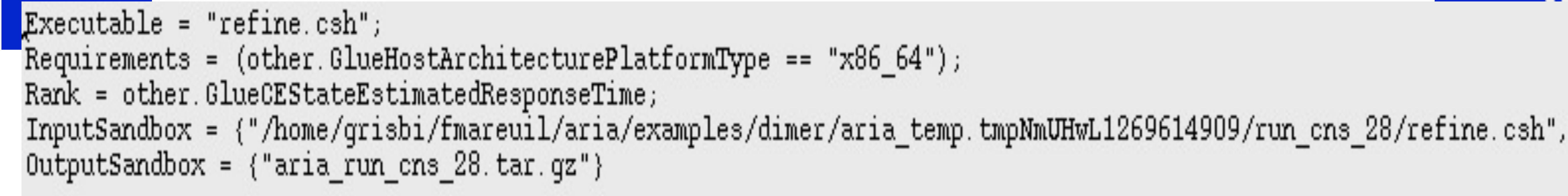
Mode command to run ARIA, if you execute ARIA on a local computer, choose LOCAL mode, if you execute ARIA on a cluster, choose CLUSTER mode, and on a grid, choose GRID mode.

Enabled	Submit Command	State Command	Output Command	#CPUs	CNS executable
yes	glite-wms-job-submit -a	glite-wms-job-status	glite-wms-job-output --dir	50	/usr/bin/cns_solve
yes				1	
yes				1	
yes				1	

```
MESSAGE [GUI]: Project loaded.
```

Deploying 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

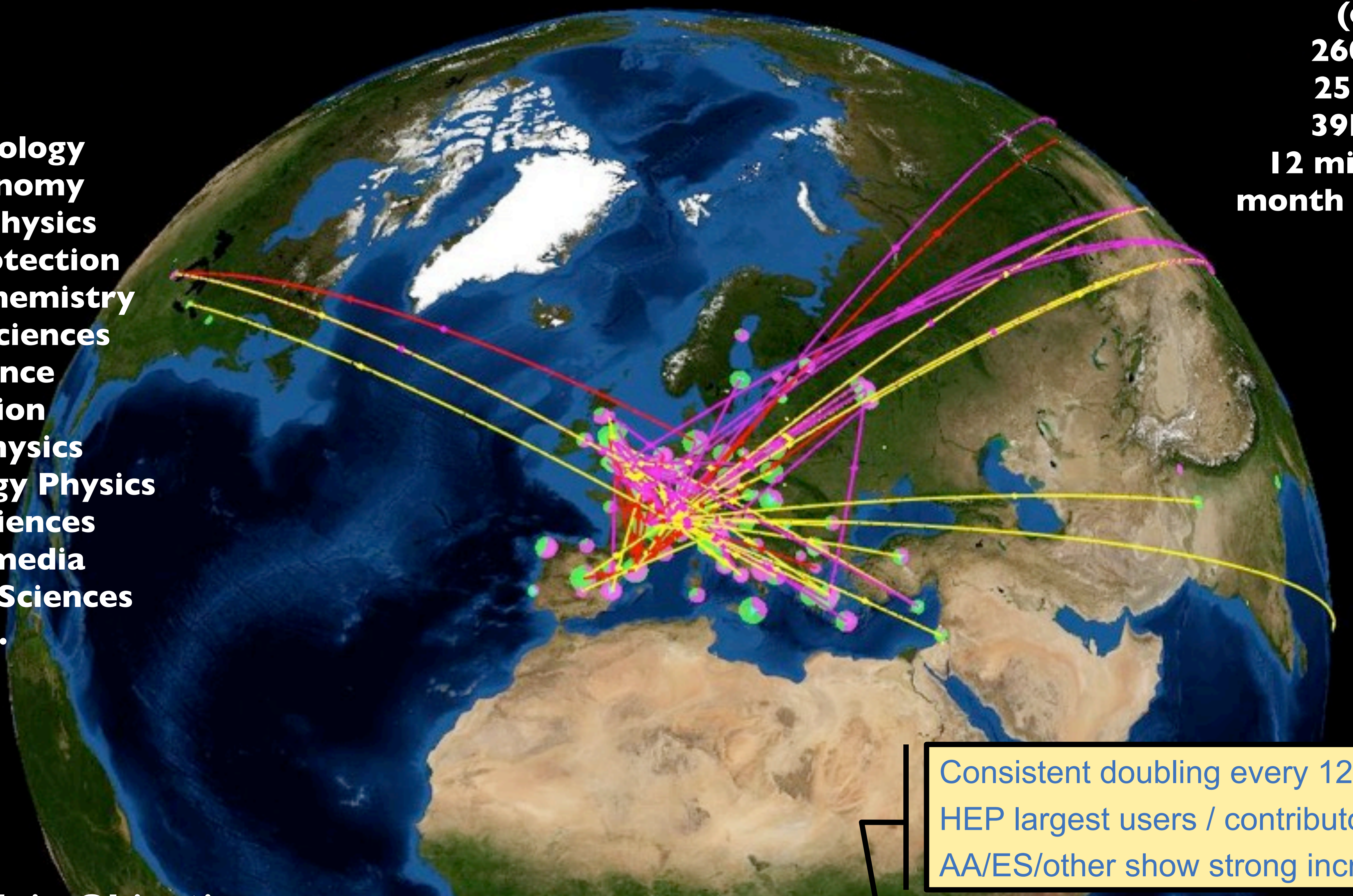


```
Executable = "refine.csh";
Requirements = (other.GlueHostArchitecturePlatformType == "x86_64");
Rank = other.GlueCEStateEstimatedResponseTime;
InputSandbox = {"/home/grisbi/fmareuil/aria/examples/dimer/aria_temp.tmpNmUHwL1269614909/run_cns_28/refine.csh",
OutputSandbox = {"aria_run_cns_28.tar.gz"}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

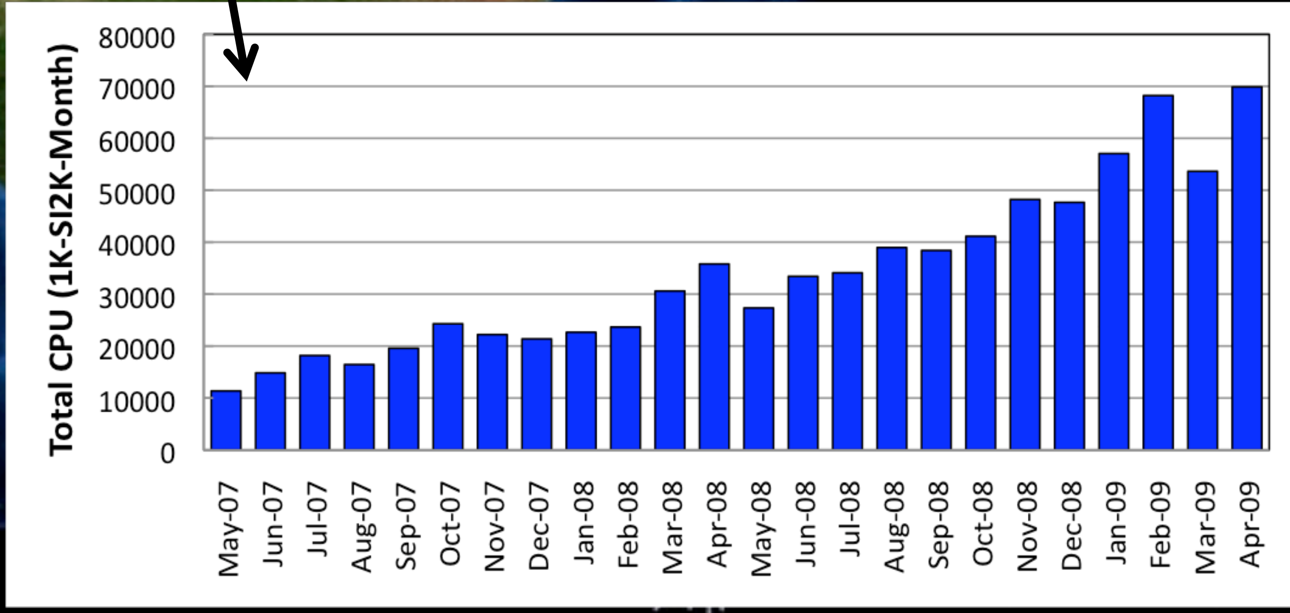
egee
Enabling Grids
for E-sciencE

15,000 users
140,000 CPUs
(cores)
260+ sites
25Pb disk
39Pb tape
12 million jobs/
month +45% in one
year

Archeology
Astronomy
Astrophysics
Civil Protection
Comp. Chemistry
Earth Sciences
Finance
Fusion
Geophysics
High Energy Physics
Life Sciences
Multimedia
Material Sciences
...



Consistent doubling every 12-18 months.
HEP largest users / contributors
AA/ES/other show strong increase



Main Objectives

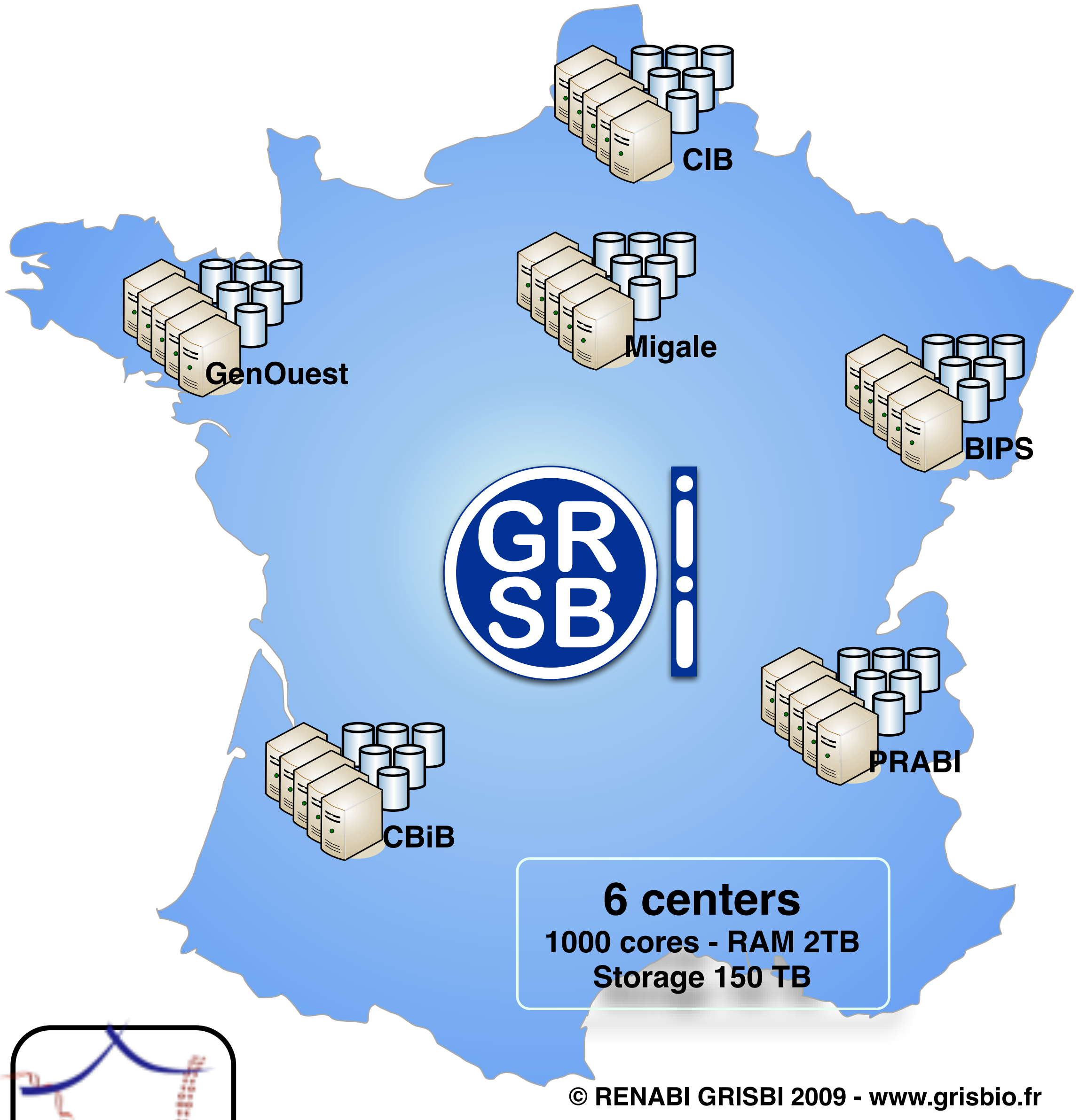
Expand/optimise existing EGEE infrastructure, include more resources and user communities

Prepare migration from a project-based model to a sustainable federated infrastructure based on National Grid Initiatives

GRISBI

- Grid Support to Bioinformatics -

Make possible challenging bioinformatics applications dealing with large scale biological systems



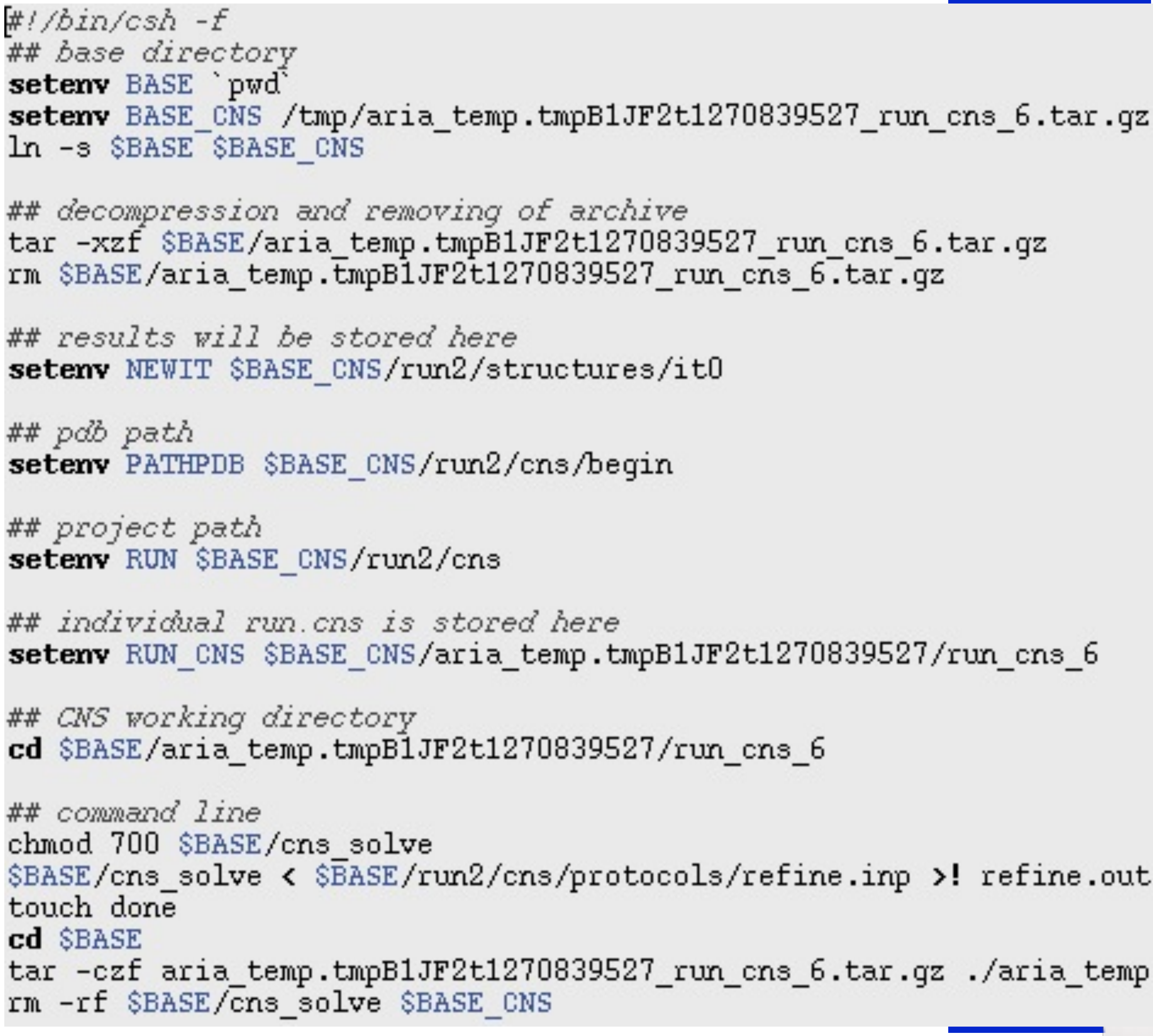
© RENABI GRISBI 2009 - www.grisbio.fr

- National Production infrastructure
  - RENABI, IBISA 2008-2010, Institut des Grilles 2009-2010
- 6 centers from RENABI
  - PRABI, MIGALE, GenOuest, CBIB Bordeaux, BIPS, CIB
  - 8 sites, with 7 CNRS institutes **IBCP Lyon, SBR Roscoff, CBiB Bordeaux, CIB Lille, IRISA Rennes, LBBE Lyon, MIGALE Jouy-en-Josas, BIPS Strasbourg**
  - 40 participants
- Computig resources
  - 1200 cores, 220 TB storage

ARIA agent

ARIA on GRID

- run.cns template is modified to use only environment variables PATHPDB (initial_pdb) and NEWIT (out_dir)
- csh template is written with relative path and environment variable

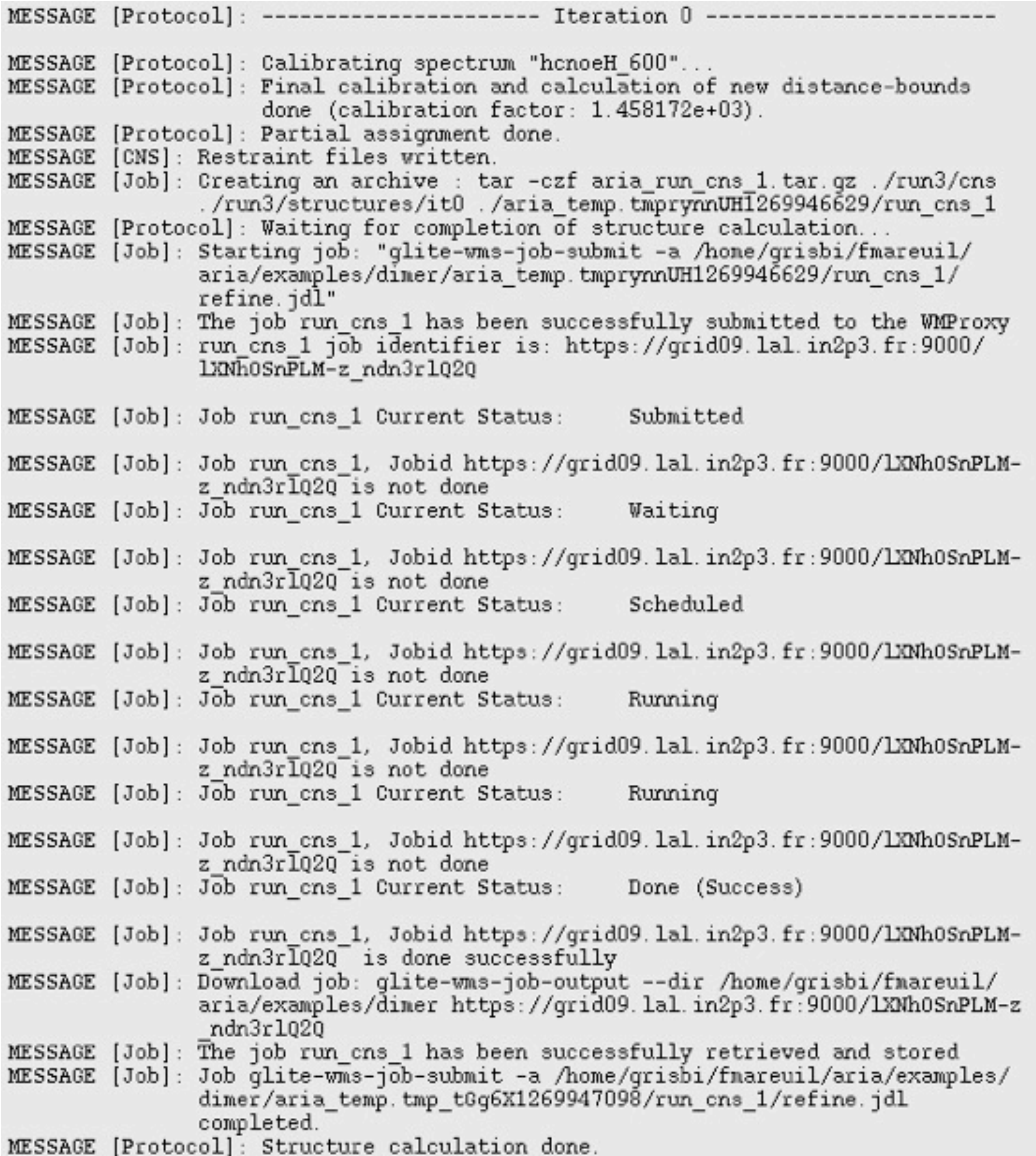


```
#!/bin/csh -f
## base directory
setenv BASE `pwd`
setenv BASE_CNS /tmp/aria_temp.tmpB1JF2t1270839527_run_cns_6.tar.gz
ln -s $BASE $BASE_CNS

## decompression and removing of archive
tar -xzf $BASE/aria_temp.tmpB1JF2t1270839527_run_cns_6.tar.gz
rm $BASE/aria_temp.tmpB1JF2t1270839527_run_cns_6.tar.gz

## results will be stored here
setenv NEWIT $BASE_CNS/run2/structures/it0

## pdb path
setenv PATHPDB $BASE_CNS/run2/cns/begin

## project path
setenv RUN $BASE_CNS/run2/cns

## individual run.cns is stored here
setenv RUN_CNS $BASE_CNS/aria_temp.tmpB1JF2t1270839527/run_cns_6

## CNS working directory
cd $BASE/aria_temp.tmpB1JF2t1270839527/run_cns_6

## command line
chmod 700 $BASE/cns_solve
$BASE/cns_solve < $BASE/run2/cns/protocols/refine.inp >! refine.out
touch done
cd $BASE
tar -czf aria_temp.tmpB1JF2t1270839527_run_cns_6.tar.gz ./aria_temp
rm -rf $BASE/cns_solve $BASE_CNS
```

IBCP

Institut Pasteur

Example of a run

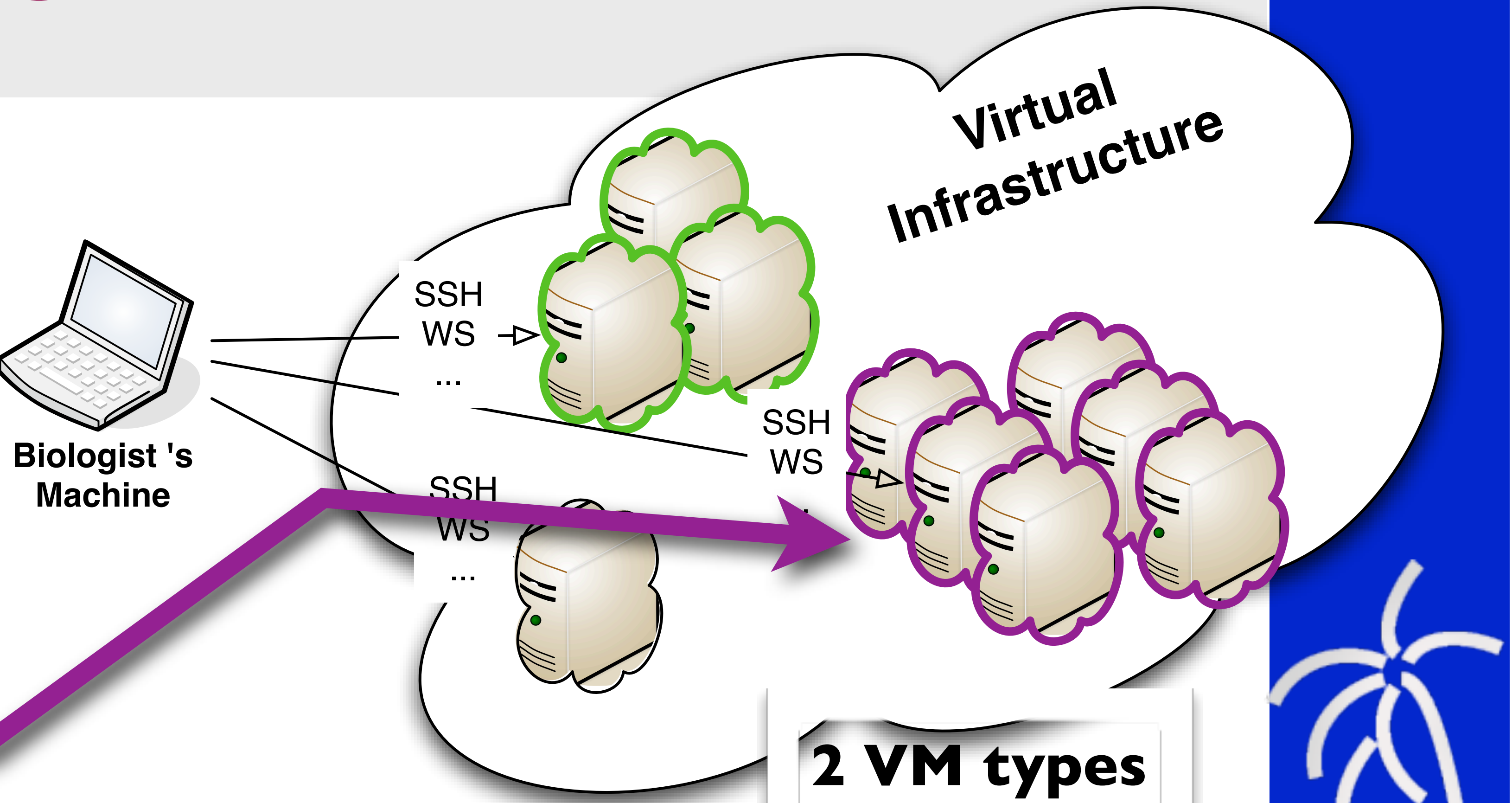


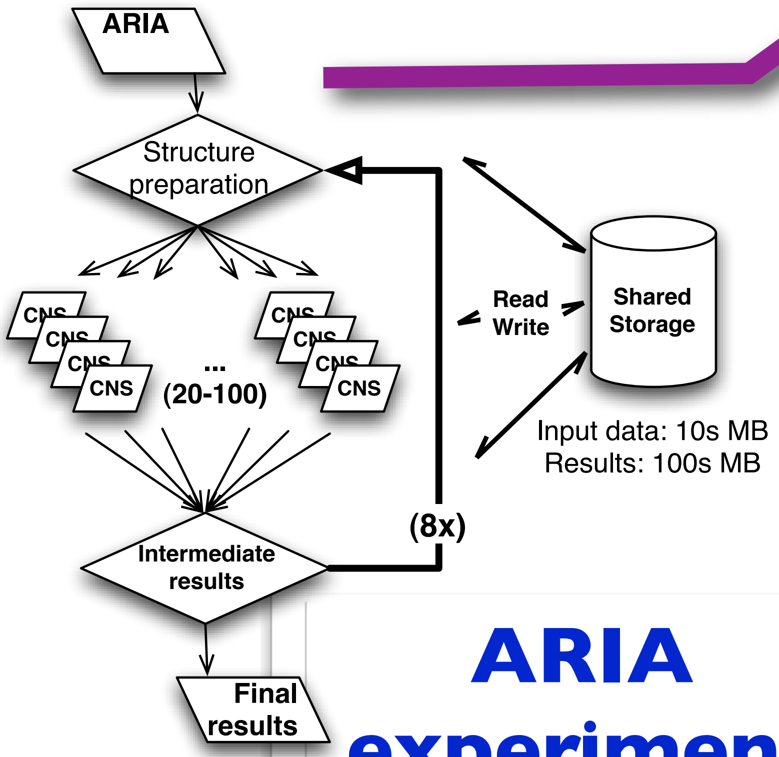
```
MESSAGE [Protocol]: ---------------------- Iteration 0 -----------------------

MESSAGE [Protocol]: Calibrating spectrum "hcnoeH_600"...
MESSAGE [Protocol]: Final calibration and calculation of new distance-bounds
                    done (calibration factor: 1.458172e+03).
MESSAGE [Protocol]: Partial assignment done.
MESSAGE [CNS]: Restraint files written.
MESSAGE [Job]: Creating an archive : tar -czf aria_run_cns_1.tar.gz ./run3/cns
               ./run3/structures/it0 ./aria_temp.tmprynnUH1269946629/run_cns_1
MESSAGE [Protocol]: Waiting for completion of structure calculation...
MESSAGE [Job]: Starting job: "glite-wms-job-submit -a /home/grisbi/fmareuil/
               aria/examples/dimer/aria_temp.tmprynnUH1269946629/run_cns_1/
               refine.jdl"
MESSAGE [Job]: The job run_cns_1 has been successfully submitted to the WMProxy
MESSAGE [Job]: run_cns_1 job identifier is: https://grid09.lal.in2p3.fr:9000/
               lXNhOSnPLM-z_ndn3rlQ2Q

MESSAGE [Job]: Job run_cns_1 Current Status:     Submitted

MESSAGE [Job]: Job run_cns_1, Jobid https://grid09.lal.in2p3.fr:9000/lXNhOSnPLM-
               z_ndn3rlQ2Q is not done
MESSAGE [Job]: Job run_cns_1 Current Status:     Waiting

MESSAGE [Job]: Job run_cns_1, Jobid https://grid09.lal.in2p3.fr:9000/lXNhOSnPLM-
               z_ndn3rlQ2Q is not done
MESSAGE [Job]: Job run_cns_1 Current Status:     Scheduled

MESSAGE [Job]: Job run_cns_1, Jobid https://grid09.lal.in2p3.fr:9000/lXNhOSnPLM-
               z_ndn3rlQ2Q is not done
MESSAGE [Job]: Job run_cns_1 Current Status:     Running

MESSAGE [Job]: Job run_cns_1, Jobid https://grid09.lal.in2p3.fr:9000/lXNhOSnPLM-
               z_ndn3rlQ2Q is not done
MESSAGE [Job]: Job run_cns_1 Current Status:     Running

MESSAGE [Job]: Job run_cns_1, Jobid https://grid09.lal.in2p3.fr:9000/lXNhOSnPLM-
               z_ndn3rlQ2Q is not done
MESSAGE [Job]: Job run_cns_1 Current Status:     Done (Success)

MESSAGE [Job]: Job run_cns_1, Jobid https://grid09.lal.in2p3.fr:9000/lXNhOSnPLM-
               z_ndn3rlQ2Q  is done successfully
MESSAGE [Job]: Download job: glite-wms-job-output --dir /home/grisbi/fmareuil/
               aria/examples/dimer https://grid09.lal.in2p3.fr:9000/lXNhOSnPLM-z
               _ndn3rlQ2Q
MESSAGE [Job]: The job run_cns_1 has been successfully retrieved and stored
MESSAGE [Job]: Job glite-wms-job-submit -a /home/grisbi/fmareuil/aria/examples/
               dimer/aria_temp.tmp_tGg6X1269947098/run_cns_1/refine.jdl
               completed.
MESSAGE [Protocol]: Structure calculation done.
```

- If downloaded archive is ok:
  - Unpack archive and aria continues to run normally

Deploying ARIA on CLOUD

ARIA on CLOUD





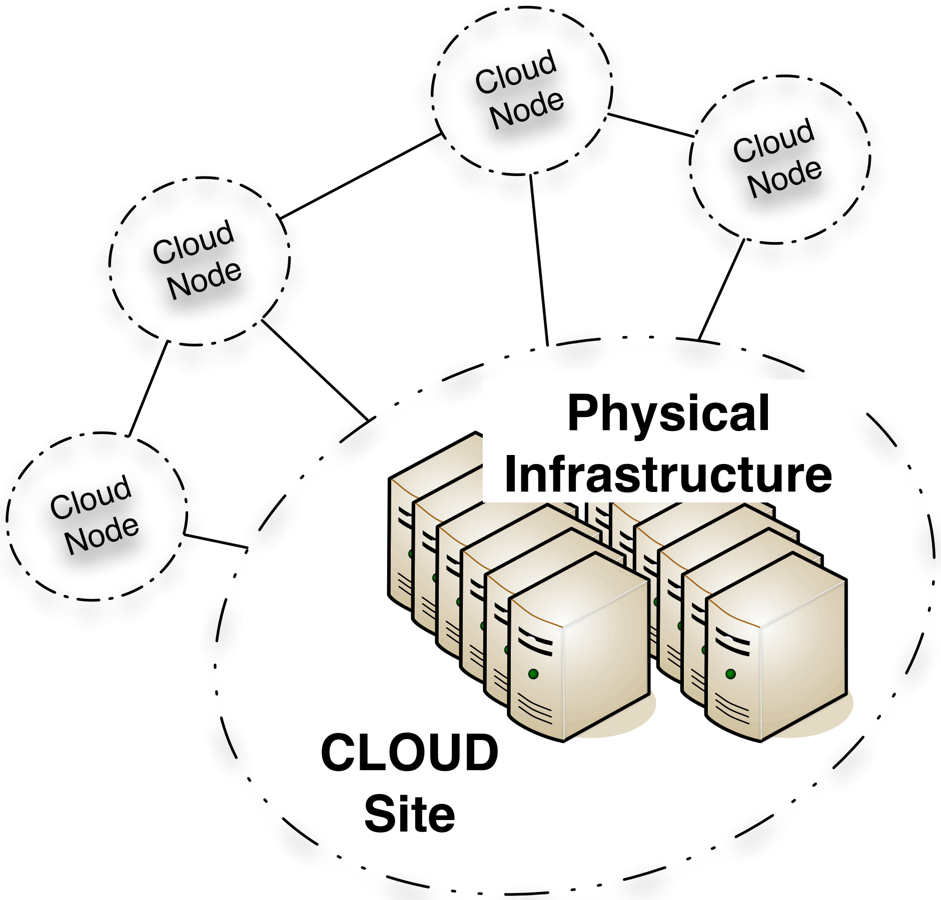
2 VM types
ARIA; CNS

ARIA experiment

Eucalyptus

Christophe Blanchet, CNRS IBCP

Cloud Infrastructures

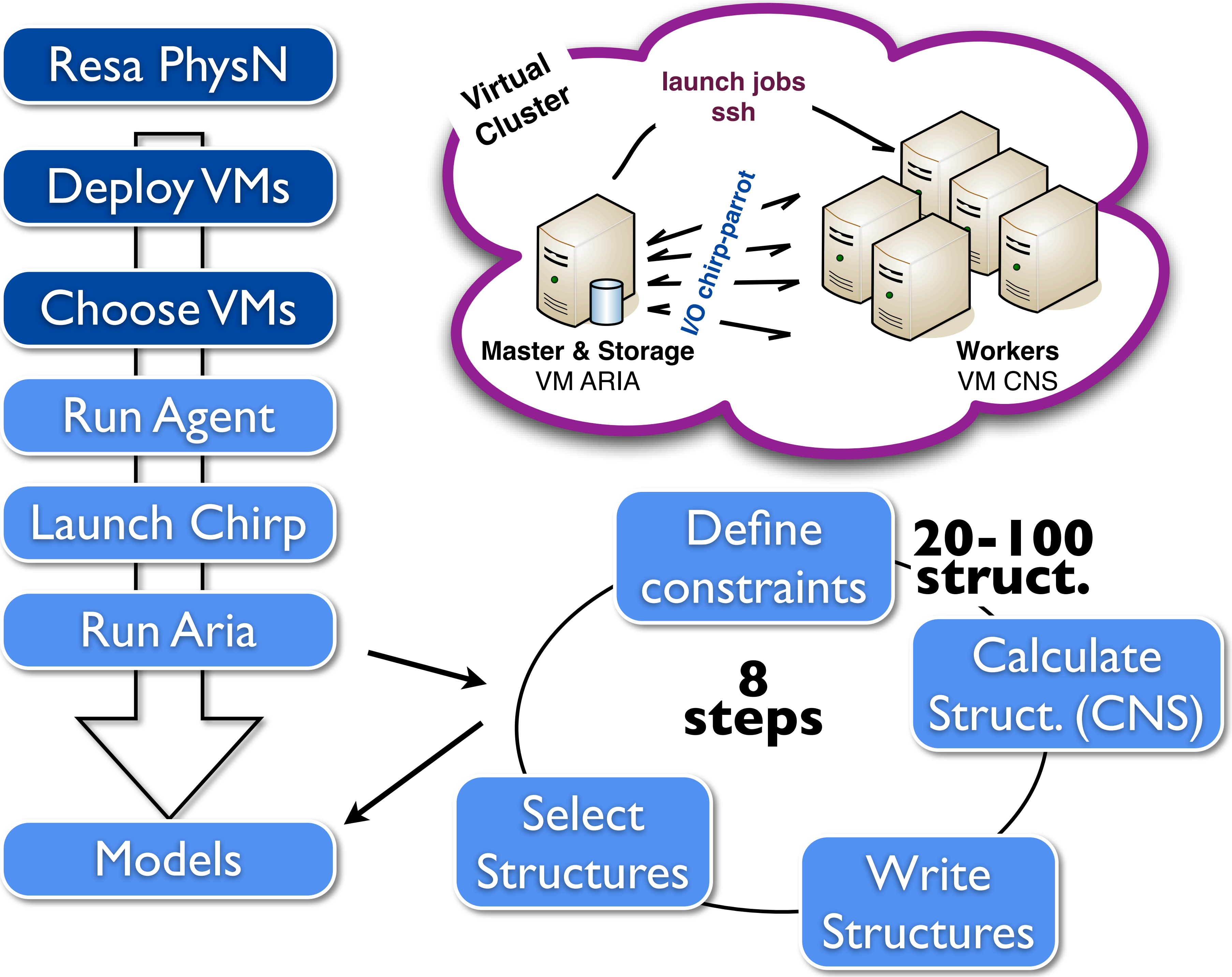


- HIPerNET and Grid'5000
  - 9 sites, 5000 cores
  - HIPerNET 0.6

- Eucalyptus and IBCP
  - 1 site , 40 cores
  - Eucalyptus 1.6.2

Eucalyptus

Cloud Workflow



Resa PhysN

Deploy VMs

Choose VMs

Run Agent

Launch Chirp

Run Aria

Models

Virtual Cluster

launch jobs ssh

I/O chirp-parrot

Master & Storage
VM ARIA

Workers
VM CNS

Define constraints

20-100 struct.

Calculate Struct. (CNS)

8 steps

Write Structures

Select Structures

Conclusion

Demo on April 14
16:00-18:00
Stand French NGI

- ARIA (Ambiguous Restraints in Iterative Assignment)
  - GRID/CLOUD added-value
    - run in parallel several structure determinations and several experiments
    - increase the capabilities of structure and assignment procedures on large systems, as membrane proteins and protein fibrils with more efficiency and reliability
    - EGEE, RENABI GRISBI, Eucalyptus, HIPERNET
  - Ongoing issues
    - Proxy management is difficult to integrate in ARIA
    - Job error rate and submission delay
- Perspectives
  - Continue integration on GRID/CLOUD with StratusLab project
    - Perspective of Hybrid GRID/CLOUD Interface for Bioinformatics
  - Evaluate with large molecular system
  - Make it available to bioinformatics community

Acknowledgment

CNRS - Centre National de la Recherche Scientifique

University of Lyon 1

Institut Pasteur

ANR - Agence Nationale de la Recherche project HIPCAL (ANR-06-CIS6-005)

The **European Commission** project EU FP7 EGEE III (INFSO-RI-222667)

IBISA - Infrastructures Biologie Santé et Agronomie, project GRISBI PF 2008

ReNaBi - Réseau National des plateformes Bioinformatiques