“PORTAL access to Grid” A proposal to TCG

Christophe Blanchet
CNRS IBCP
Chair of EGEE Bioinformatics activity

PORTALwg Kick-off
May, 2nd-3rd, 2007, CERN

PORTALwg: Proposal to TCG

- **PORTAL working group**
  - chair: C. Blanchet
- **Goals:**
  - The main objective of this "PORTAL" working group is to propose "best-practice" rules for the access of portals to the grid. By access to the grid, we mean that the portal should be able to store data and run job on the grid by delegation of real users, or by itself with its own credentials. To do so, a portal responsible should be able to get a server certificate for his portal from the EUGridPMA CAs, and then to be able to register this portal certificate to a VO allowed on the grid. Once the portal have been accepted into the concerned VO, it should be able to store and access data inside the VO area, and also to run job on site accepting this VO. The usage of these grid resources by the portal can be done by delegation of a grid- and VO-registered user, or as a resource provider for an user not registered in the grid.

PORTALwg members

- **Having agreed**
  - Christophe Blanchet , Bioinformatics usecase and chair
  - Hugues Benoit-Catin , Medical imaging usecase
  - Massimo Lamana , repr. of HEP experiments (repr. tbd)
  - Gidon Moont , GridPP portal usecase
  - Cal Loomis , small site repr. and representative of TCG
  - Rolf Rumler , large site repr. (repr. tbd)
  - Peter Kunszt , large site repr.
  - Dave Kelsey , repr. of JPSG
  - David Groep , repr. of EUGridPMA
  - Ake Edlund , repr. of MWSG (repr. tbd)
  - Jens Jensen , expert about authN/authZ
  - Christoph Witzig , expert about Shibboleth in SWITCH
- **Others ?**
  - Generic portal: No answer neither from Pgrade nor Genius
  - Jim Basney (US)

Plan

- **Interest ?**
  - PORTALwg has been announced
  - EGEE Bioinformatics meeting 3 (Valencia, March 6th, 2007)
    - great interest of the Bioinformatics community
  - EGEE JPSG meeting, March, 13rd-14th, Cern
    - good interest on the site operation and the policy aspects
  - other communities: Geophysic, Finance, ...

- **Collaboration with other projects**
  - OSG : certainly, according to last MWSG (March 1st-2nd), Jim Basney to be contacted
  - SwissGrid: participation of CSCS and Switch

- **Timeline**
  - from May 2007 to May 2008 ?

PORTALwg
Kick-Off Meeting
May, 2nd-3rd, 2007
Cern

Goals of the kick-off meeting

- Detail participant contribution
- Finalize the mandate of the PORTAL group
- Define a timeline to propose to TCG
- Suggest other participants if required.
- Work :-)

Guidelines

- **Questions**
  - How does one obtain service certificates from the EUGridPMA CAs?
  - How to register those certificates in a VO?
  - How to handle proxies of this certificate in a portal (e.g. wanting to add VOMS groups)?
  - Which software should be allow to be used ?
    - § enabling users to put his own software
    - § restrict to only portal ones
  - Do sites need to be advised of the use of these certificates (if so, how to inform them)?
  - What are the accounting/logging practices to follow on the portal?
  - Does the use of these need to appear in the VO AUP or similar documents?
- **Other questions ?**

Agenda

- **May, 2nd, 14h00-17h00**

- **Introduction (C. Blanchet)**
- **Application usecases**
  - Christophe Blanchet "Bioinformatics GPSA portal"
  - Gidon Moont "GridPP portal"
- **Grid and Site operation**
  - local site and global platform constraints and security flaws
  - Cal Loomis "Grid site: CNRS LAL"
- **Discussion**

Agenda

- **May, 3nd, 9h30-12h00**

- **Authentication and authorization: portal certificate, certificate authority and credential delegation**
  - David Groep "EUGridPMA: certificate authority at the European scale"
  - Jens Jensen "Single Sign-On and Identity Management on Grid"
- **Best practice and policy agreement:**
  - David Kelsey "JSPG Best practice and policy agreement on EGEE "
- **Mandate and timeline agreement**
- **Discussion and AOB**

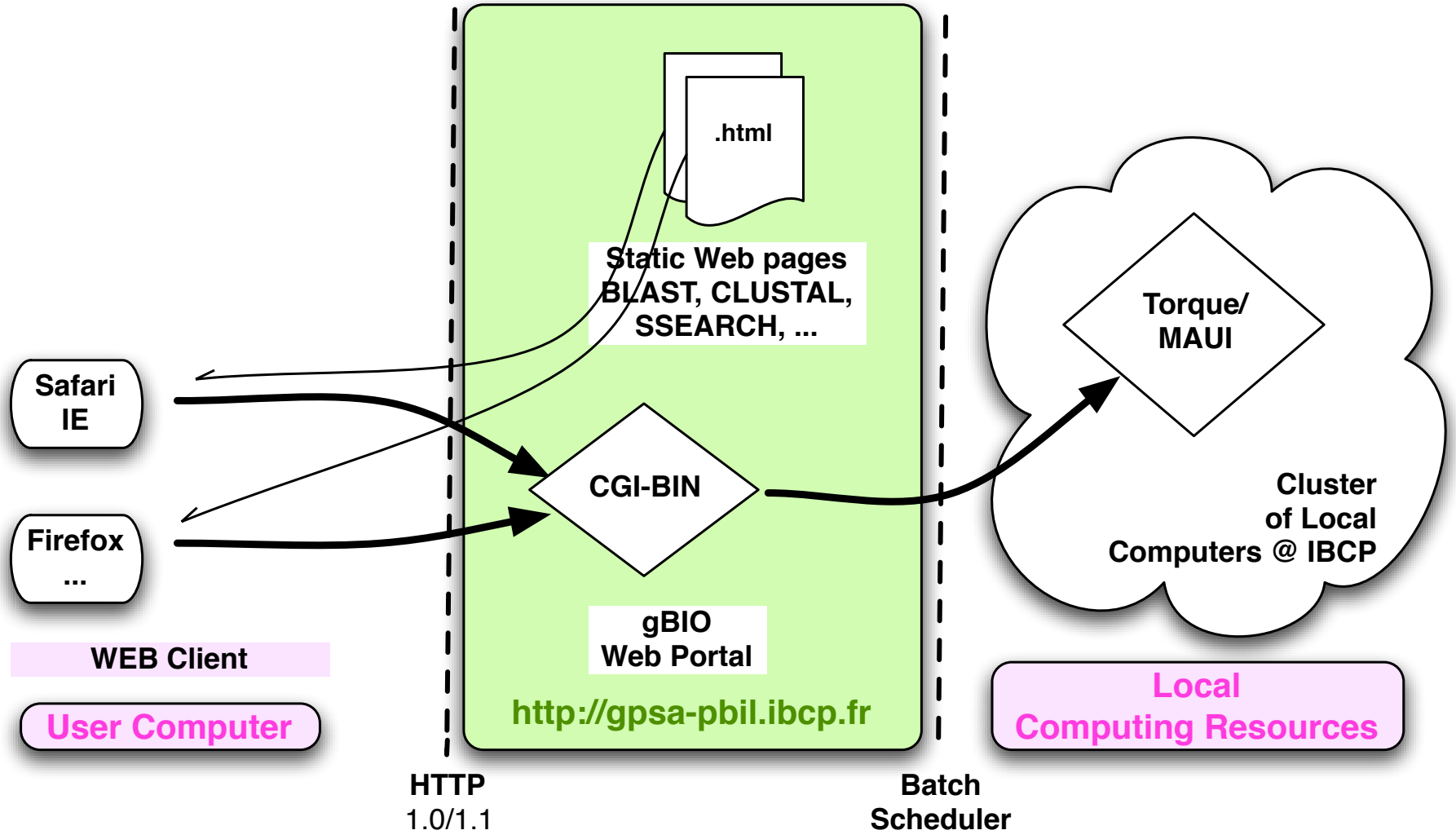
UseCase-1 : Bioinformatics Web Portal

Usecase1: Bioinformatics Portal

- Network Protein Sequence Analysis (NPS@ release 3)
  http://npsa-pbil.ibcp.fr
- Online since 1998
- 46 integrated methods for protein sequence analysis
- 12 Online up-to-date biological databanks

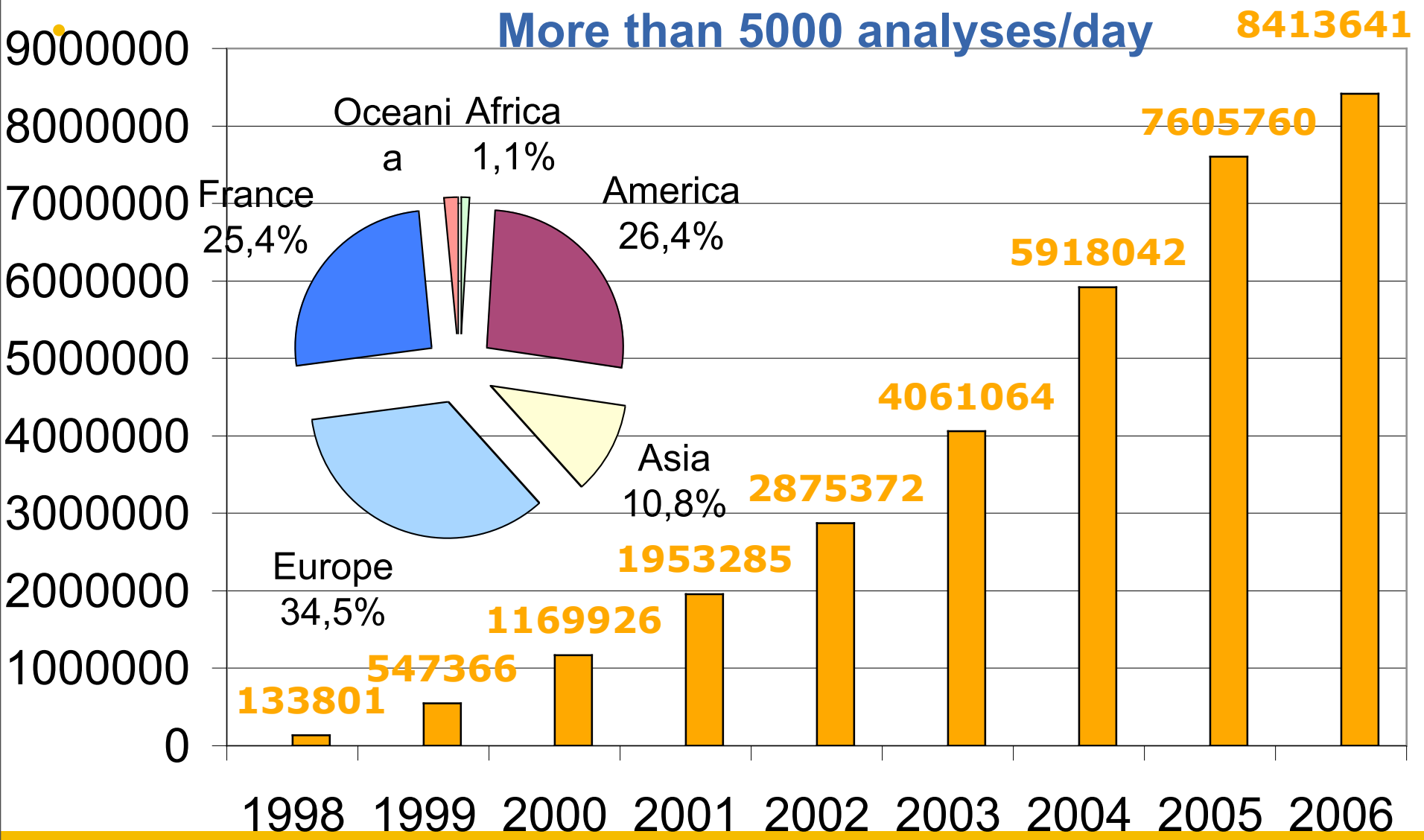
- International pointers: Expasy (Ch) , University of California, ...
- Ref.: " NPS@: Network Protein Sequence Analysis", Combet C., Blanchet C., Geourjon C. et Deléage G. Tibs, 2000, 25, 147-150.

Portal Computing



NPS@ hits

- More than 8 millions analyses since 1998
- More than 5000 analyses/day



GPSA: Bioinformatics Grid Portal

- **Scientific objectives**
  - –Molecular Bioinformatics: protein sequence analysis
  - –Analyse data from high-throughput Biology: complete genome projects, EST, complete proteomes, structural biology, ….
  - –Integration of biological data and tools
- **Method**
  - –Provide Biologists with an usual Web interface: NPS@
    - § NPS@ Web portal online since 1998
    - § 46 tools & 12 updated databases
    - § + 9,000,000 jobs & 5,000 jobs/day
  - –Ease the access to updated databases and algorithms.
  - –Protein databases are stored on grid storage as flat files.
  - –Legacy bioinformatics applications
    - § Wrapping usual binary in grid environment
    - § transparent remote access with local filesystem
  - –Display results in graphical Web interface.
- **Status:** Prototype

NPS@ : CLUSTALW multiple alignment

http://gpsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/np:

multiple alignment | CLUSTALW ALIGNMENT | Status page

Pôle Bioinformatique Lyonnais

Network Protein Sequence Analysis

GPS@ is the **grid port** of NPS@ from PBIL IBCP in Lyon, France

[HOME] [NPS@] [SRS] [HELP] [REFERENCES] [NEWS] [MPSA] [ANTHEPROT] [Geno3D] [SuMo] [Positic

February 27, 2006: First public release of GPS@ online at http://gpsa-pbil.ibcp.fr
Take advantage of the EGEE Grid platform for your bioinformatic analysis on the NPS@ portal.

Work supported in part by projects: French ACI Grid GriPPS , EU-FP6 EGEE and EU-FP6 EMBRACE.

CLUSTALW

[Abstract] [NPS@ help] [Original server]

Paste a nucleic sequence databank in Pearson/Fasta format below : help

```
>gnl|euhcvdb|AB016785_C1b Hepatitis C virus partial genome, complete
5'UTR, complete CDS (Contains: C, E1, E2, P7, NS2, NS3, NS4A, NS4B, NS5A,
NS5B), partial 3'UTR.
gccagccccctgatgggggcgacactccaccatagatcactccccctgtgaggaactactgtcttcacgcaga
aagcgtctagccatggcgttagtatgagtgtcgtgcagcctccaggcccccccctcccgggagagccatagt
ggtctgcggaaccggtgagtacaccggaattgccaggacgaccgggtcctttcttggatcaatcccgctcaa
tgcctggagatttgggcgtgcccccgcgagactgctagccgagtagtgttgggtcgcgaaaggccttgtgg
tactgcctgatagggtgcttgcgagtgcccccgg
```

All sequence names must be different !

Use the GRID resources from egee Enabling Grids for E-sciencE

SUBMIT CLEAR

Output width : 60

CLUSTALW Parameters

Output format :

Pôle Bioinformatique Lyonnais

Network Protein Sequence Analysis

GPS@ is the **grid port** of NPS@ from PBIL IBCP in Lyon, F

[HOME] [NPS@] [SRS] [HELP] [REFERENCES] [NEWS] [MPSA] [ANTHEPROT] [Geno3D

February 27, 2006: First public release of GPS@ online at http://gpsa-pbil.i
Take advantage of the EGEE Grid platform for your bioinformatic analysis on the N

Job **CLUSTALW** (ID: 06c442e86157) has been transfered on the
GPS@ Portal, an EGEE Grid interface for Bioinformatics
(started on 20070321-233721).
Results will be shown below. **Please wait and don't go back.**

In your publication cite :
NPS@: Network Protein Sequence Analysis
TIBS 2000 March Vol. 25, No 3 [291]:147-150
Combet C., Blanchet C., Geourjon C. and Deléage G.

CLUSTALW multiple alignment

Abstract Thompson, J.D., Higgins, D.G. and Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multi
sequence weighting, positions-specific gap penalties and weight matrix choice. (1994) Nucleic Acids Research, 22, 46

Residues repertoire

SHOW alignment with width : 60

all residues.

Conservation level of 0 % (works with "using conservation level of" option in list above, negative

View CLUSTALW in: [MPSA (Mac, UNIX) , About...] [AnTheProt (PC) , Download...] [HELP]

```
                                        10        20        30        40        50        60
                                        |         |         |         |         |         |
AB016785_C1b    GCCAGCCCCCTGATGGGGGCGACACTCCACCATAGATCACTCCCCTGTGAGGAACTACTG
AB031322_P1b    ------------------------------------------------------------
```

Status page

http://gpsa.ibcp.fr/php/html_status_frames.php?list_filename=/biodata/gpsa-pbil

multiple alignment | CLUSTALW ALIGNMENT | Status page

Warning: process() [function.process]: warning: in **/var/www/gpsa.ibcp.fr/pbil/servers/gpsa/w3-gpsa/html/php/html_task_list_generator.php** on line **19**

Warning: process() [function.process]: failed to load external entity "/opt/bio/tmp/gpsa.ibcp.fr/gpsa-pbil.ibcp.fr/070321.223721/bio_methods.xml" in **/var/www/gpsa.ibcp.fr/pbil/servers/gpsa/w3-gpsa/html/php/html_task_list_generator.php** on line **19**

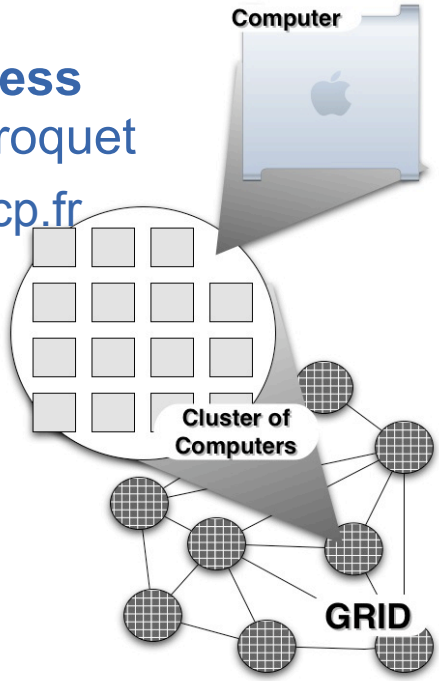
Jobs		
Task name	*Current status*	*Results*
clustalw	terminated	

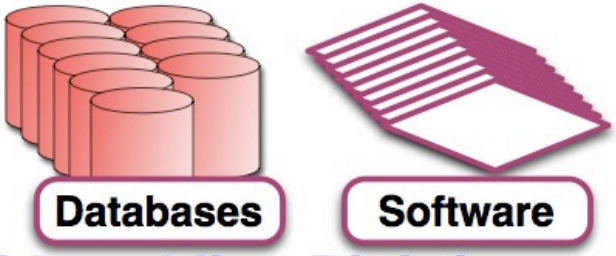
JobID : https://grid09.lal.in2p3.fr:9000/15ka5ab2a9rTbaC3lNIZUA

status	duration	total
submitted	00:01:08	00:01:08
waiting	00:00:00	00:01:08
scheduled	00:02:00	00:03:08
running	00:02:20	00:05:28
done	00:00:00	00:05:28
output_ready	00:00:07	00:05:35
terminated		

GPSA Results

- **Grid-enabled bioinformatics resources**
  - –9 algorithms
  - –3 protein databases
- **Bioinformatics** descriptors
  - –XML framework, WSRF
- **Encryption system:** EncFile
  - –Security: AES, Key sharing, M-of-N
- **Transparent access to grid files:** Perroquet
- **http://**gpsa-pbil.ibcp.fr





Grid-enabling Bioinformatics

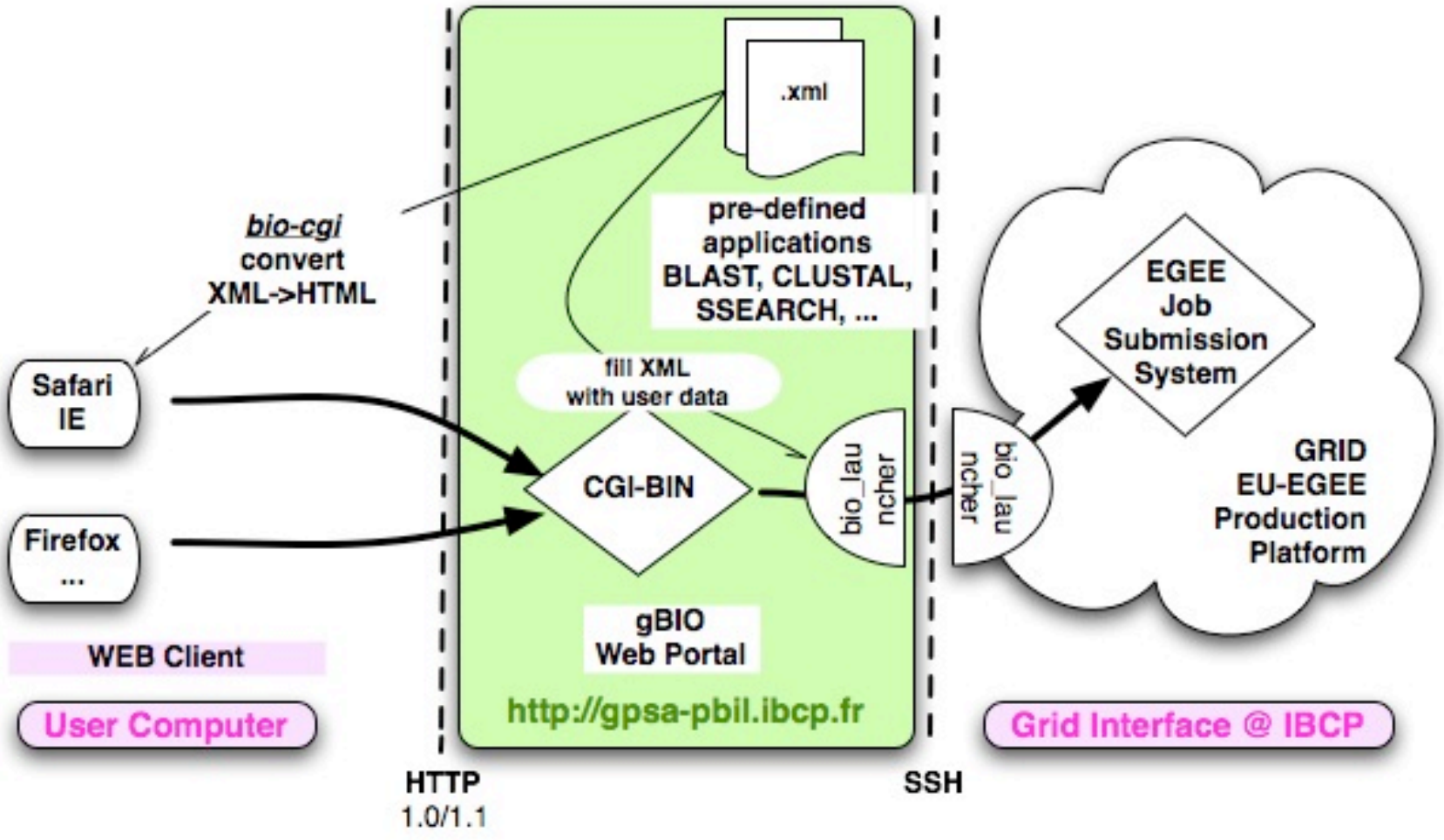
Legacy Bioinformatics Applications :
* Wrapping tools with XML descriptors
* BLAST, SSEARCH, FASTA, ClustalW, MultAlin, PattInProt, ...

Distributed databases:
* **Encrypting Data** with EncFile system
* **Swiss-Prot:**
lfn:/grid/biomed/db/swissprot/last/sprot.fas
Symlink to *(...)/swissprot/50/4/sprot.fas*
* **TrEMBL:**
lfn:/grid/biomed/db/trembl/33/4/sptr.fas

Details : http://gbio.ibcp.fr

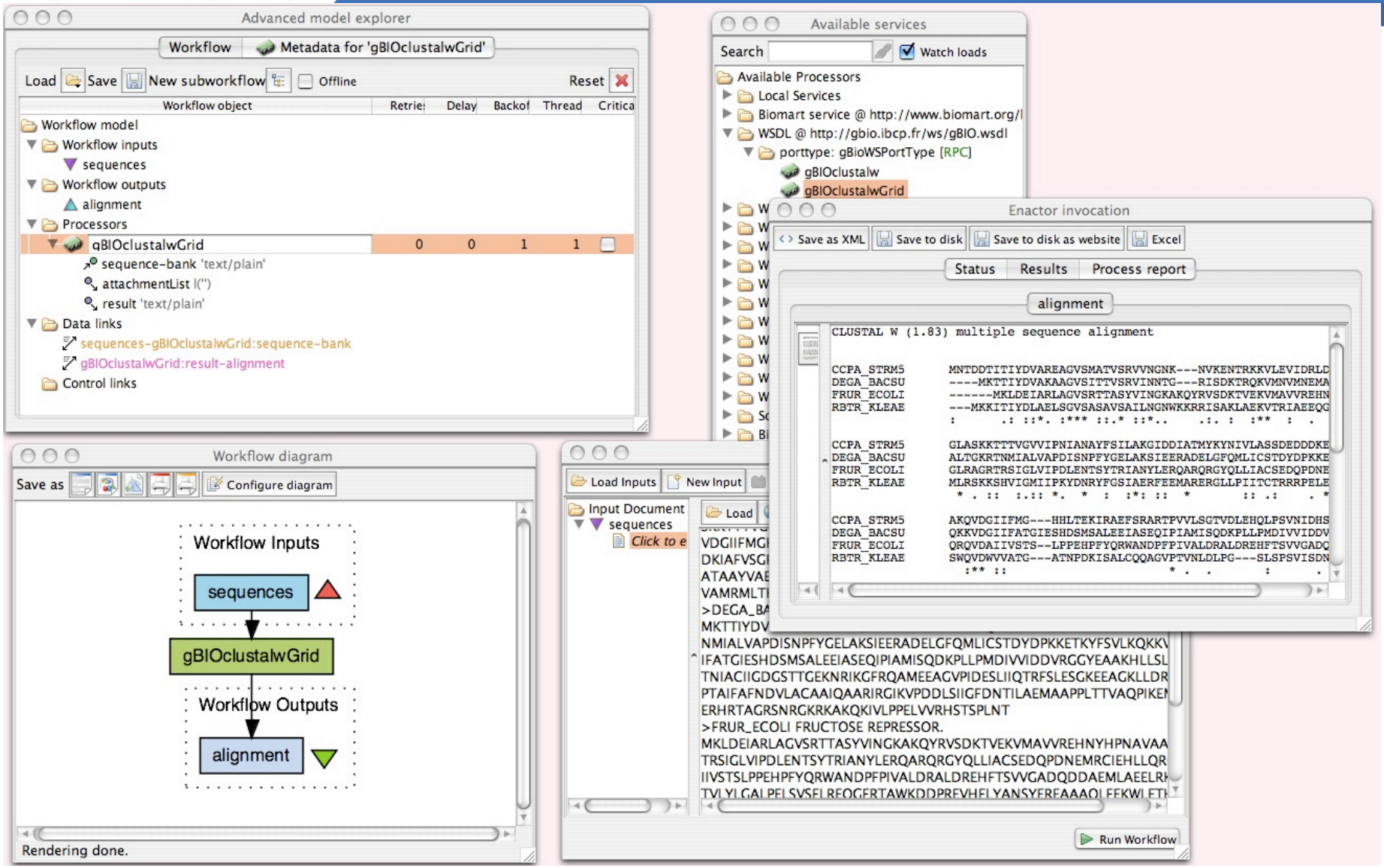
Institut de Biologie et Chimie des Protéines

Grid Portal Backend



UseCase-2 : Bioinformatics Web Services

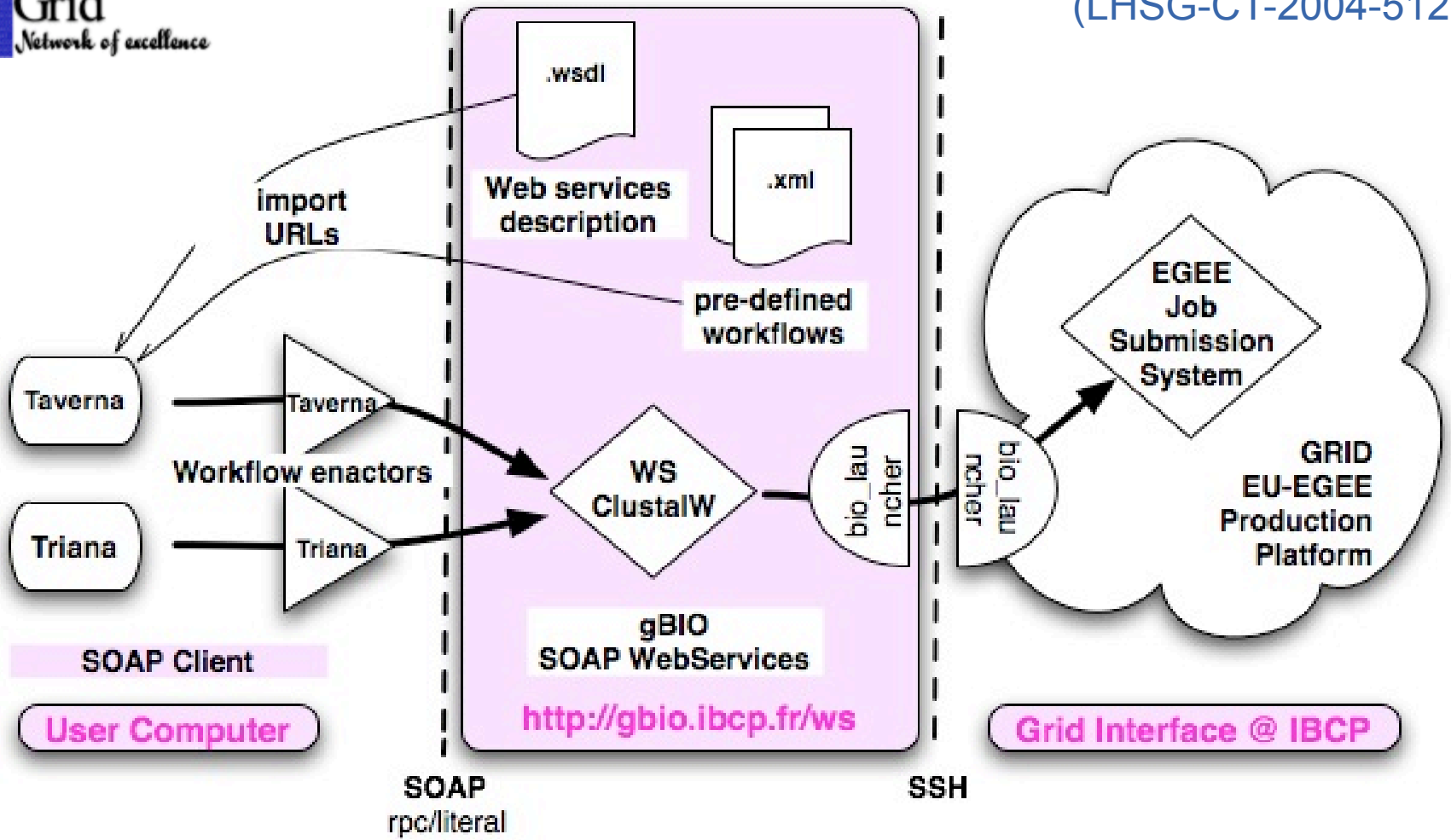
Usecase2: Bioinformatics WebServices



Call gBIO Web Services from Taverna, Triana, and other workflow tools

GPSA WSRF Web Services

Joined work with EU-FP6 EMBRACE (LHSG-CT-2004-512092)



Christophe Blanchet, Christophe Combet, Vladimir Daric and Gilbert Deléage
Web Services Interface to run Protein Sequence Tools on Grid, Testcase of Protein Sequence Alignment.
Lecture Notes in Computer Science : Biological and Medical Data Analysis (2006) 4345, 240-9