A Grid Infrastructure for Managing Workflows in Bioinformatics Applications

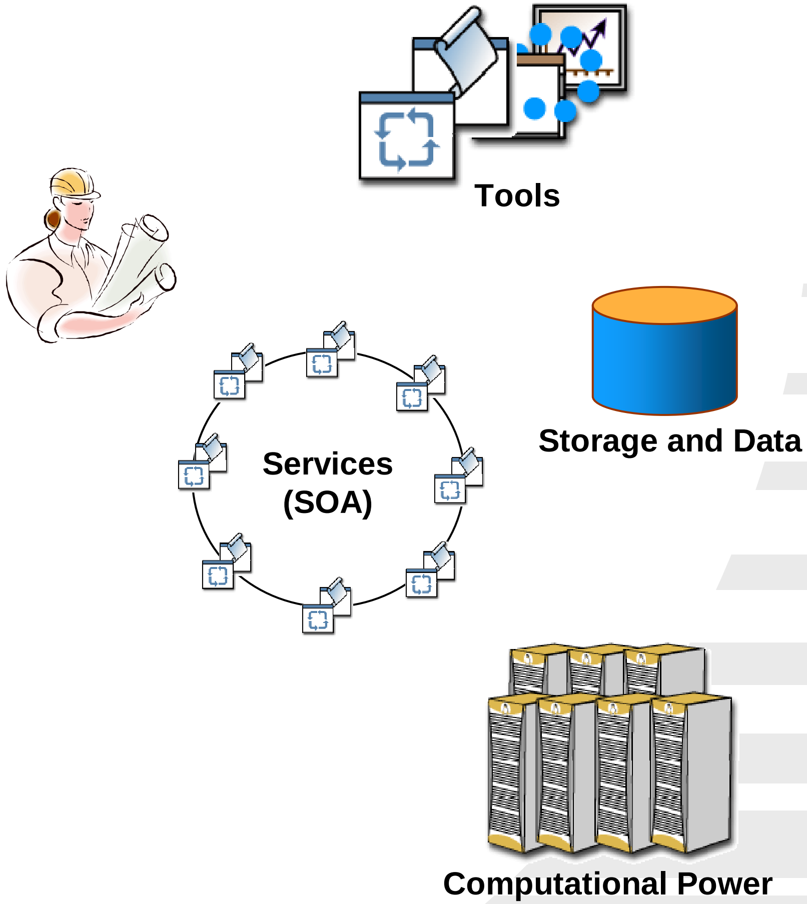
Maurizio Melato

<maurizio.melato@nice-italy.com>

NETTAB Workshop, July 10-13, 2006, Santa Margherita di Pula

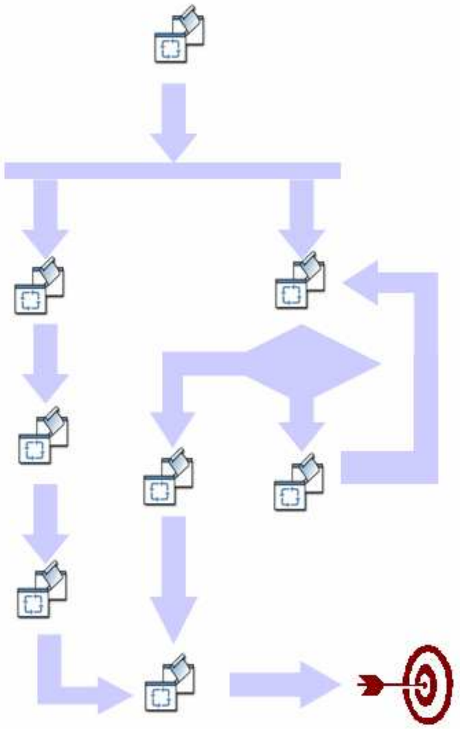
Scenario

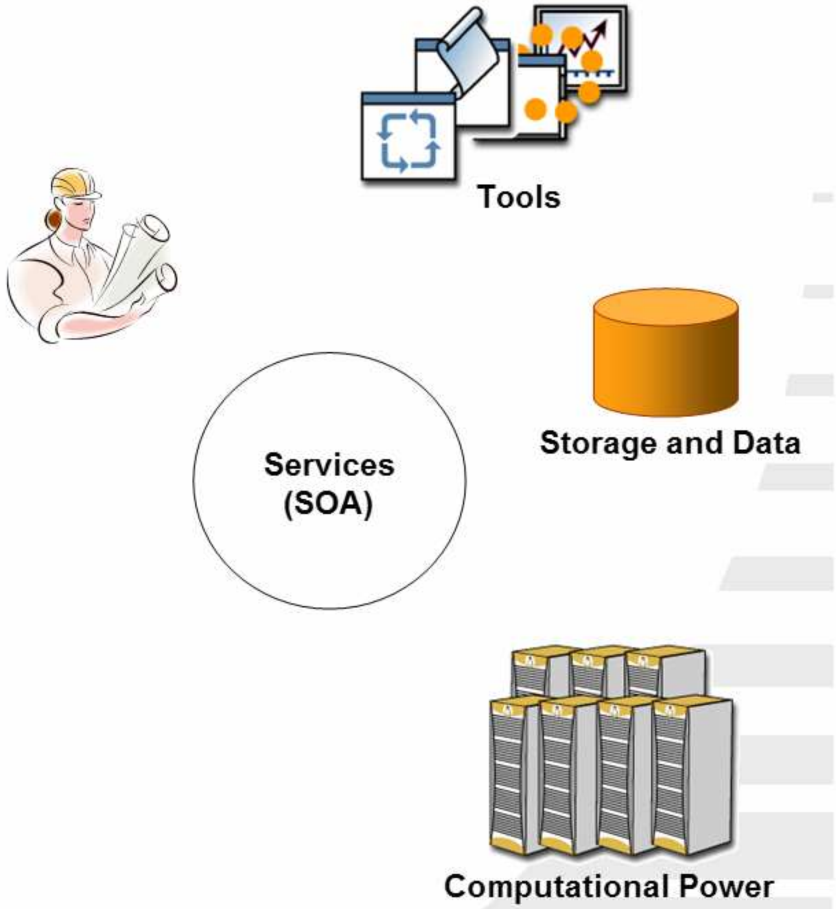
- Bioinformatics scientists have to execute complex tasks
- There is the need to orchestrate these services in workflows



Scenario

- Bioinformatics scientists have to execute complex tasks
- There is the need to orchestrate these services in workflows



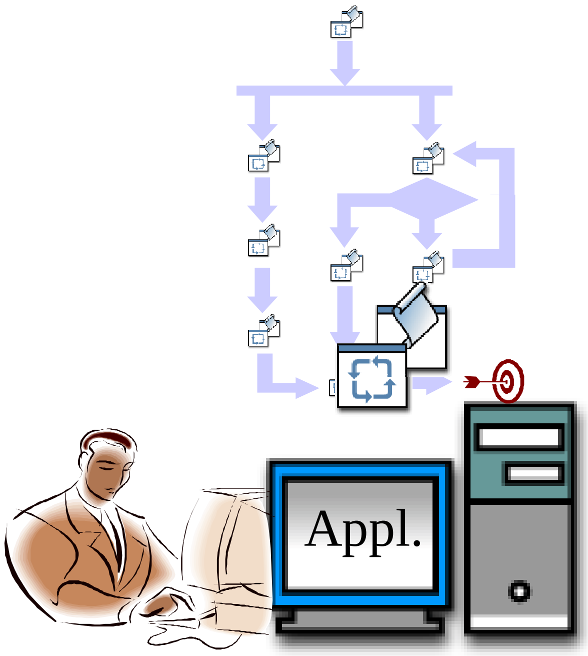


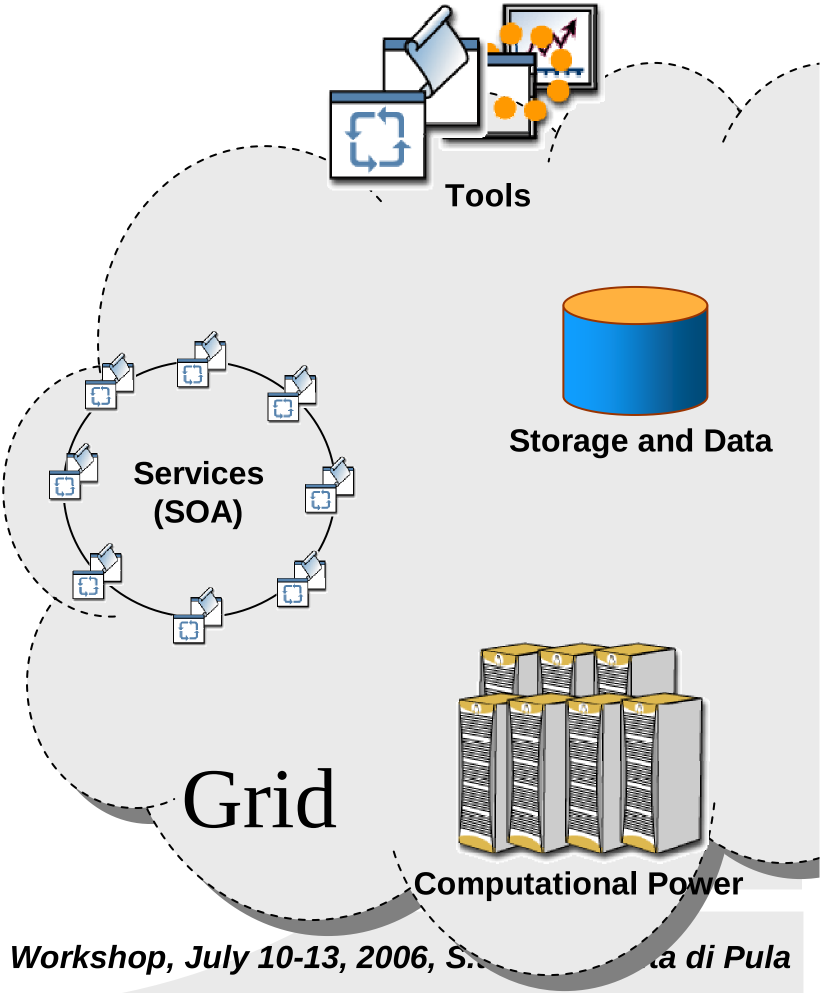
Scenario

- Today there is a high demand for workflow management
- Many current workflow management systems have strong limitations due to their ***client nature*** and *too many* standards
- With *client nature* we mean fat client application running on the user's workstation
  - low reliability
  - no fault tolerance
  - typically one workflow at a time
  - …
- Grid infrastructures play a very important role → enacting workflows based on Grid services

Grid*ified* Scenario

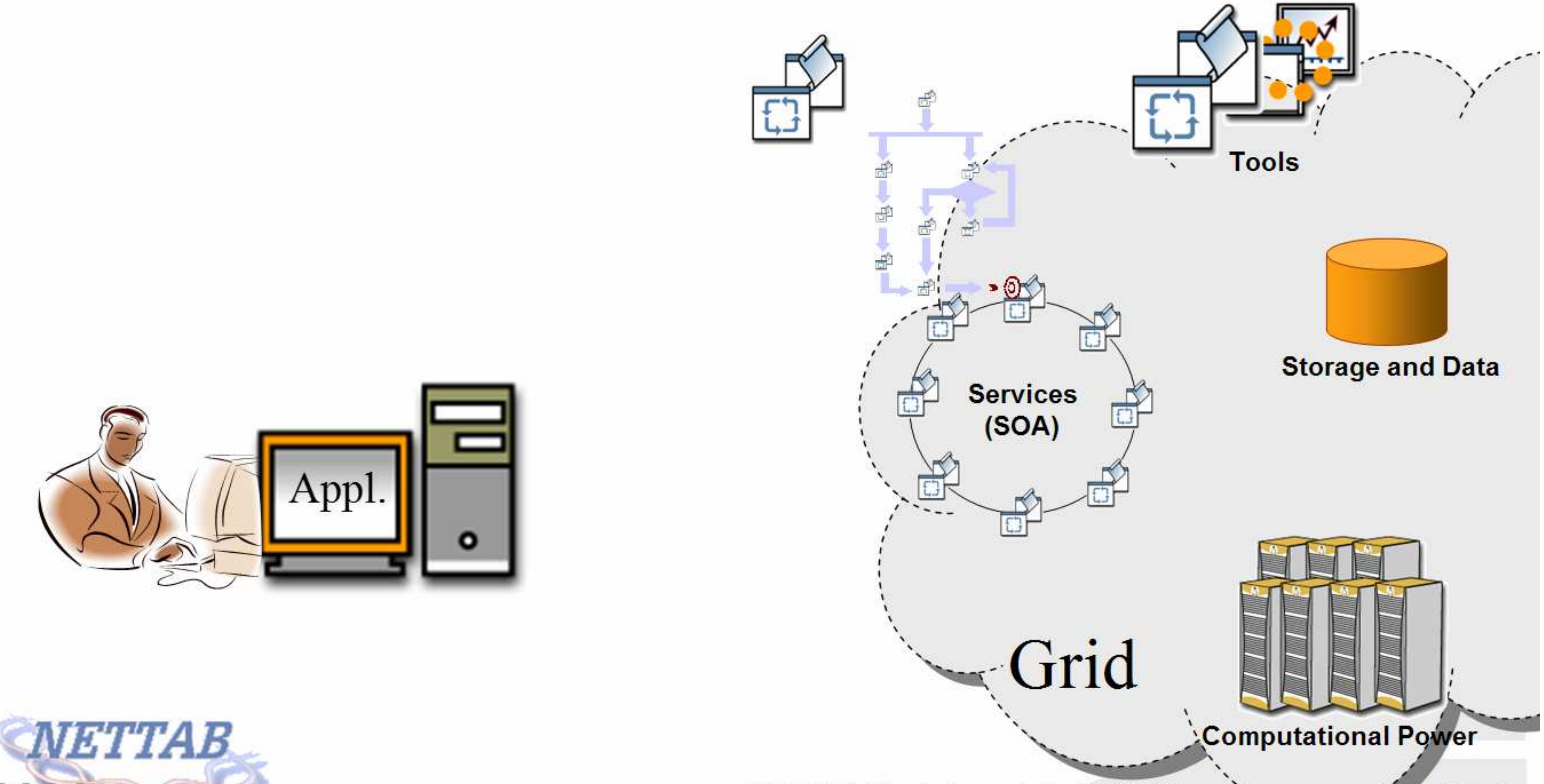
- Grid technology leverages both the *computational* and *data management* resources
- Providing optimisation, scalability, reliability, faul tolerance, QoS,…





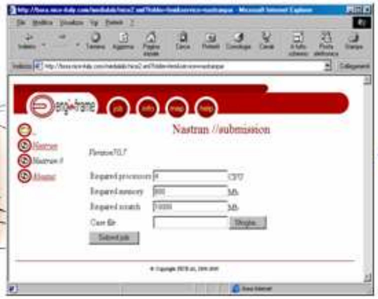
Grid*ified* Scenario

- Grid technology leverages both the *computational* and *data management* resources
- Providing optimisation, scalability, reliability, faul tolerance, QoS,…



Grid*ified* Scenario

- Grid technology leverages both the *computational* and *data management* resources
- Providing optimisation, scalability, reliability, faul tolerance, QoS,…



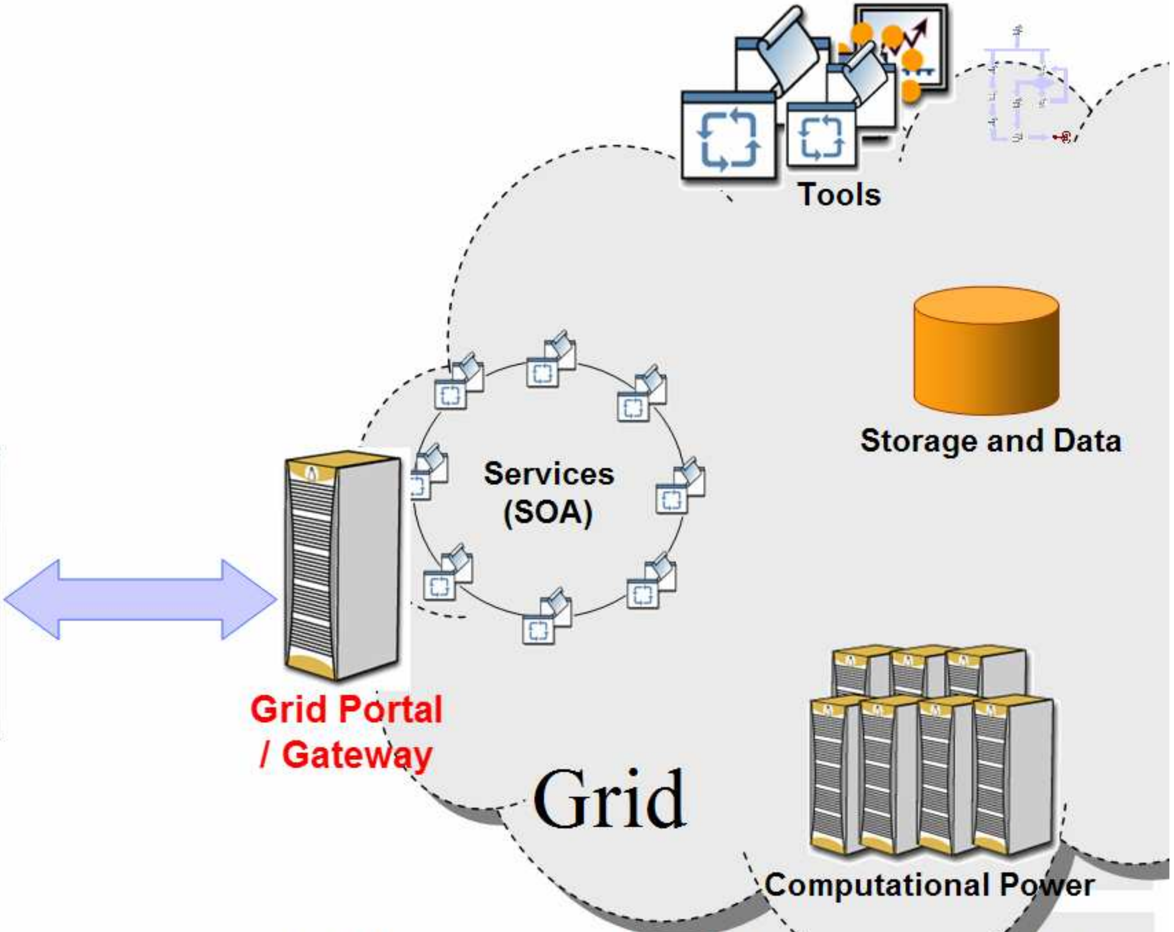
Nastran //submission

Required processors 8 CPU

Required memory 800 Mb

Required scratch 10000 Mb

Case file



Infrastructure proposal

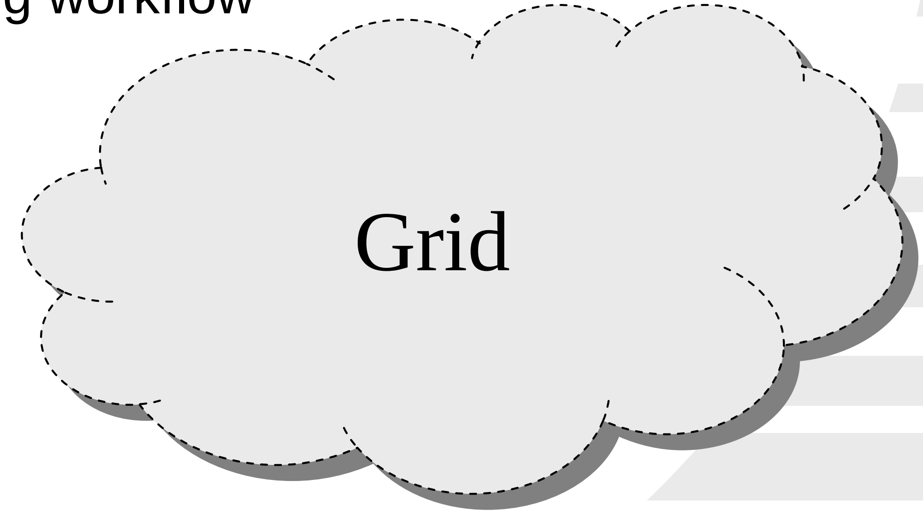
- **Goal**: proposal of a *Grid infrastructure* able to provide the basic building blocks for **composition** and **enactment** of bioinformatics workflows for the life sciences community

- Users can
  - build workflows on top of exposed Grid services
  - run and monitor workflows via a standard Web browser
  - Exploit in a transparent way the computational power and data access capabilities provided by the backend Grid infrastructure…

Tools

We used the following tools to build the proposed architecture:

- **EnginFrame or Genius:**
  - As Grid services provider
  - As Web interface for managing workflow
- **Taverna:**
  - As Workflow designer
- **Moteur:**
  - As Workflows enactor



Taverna & Moteur

- **Taverna** is a graphical workbench tool for both creating and running workflows -in Scufl language- that allows the integration of resources/services that are published as Web Services.

- **Moteur** is a service based Scufl workflow engine developed in the AGIR project and optimized for dealing with data intensive applications

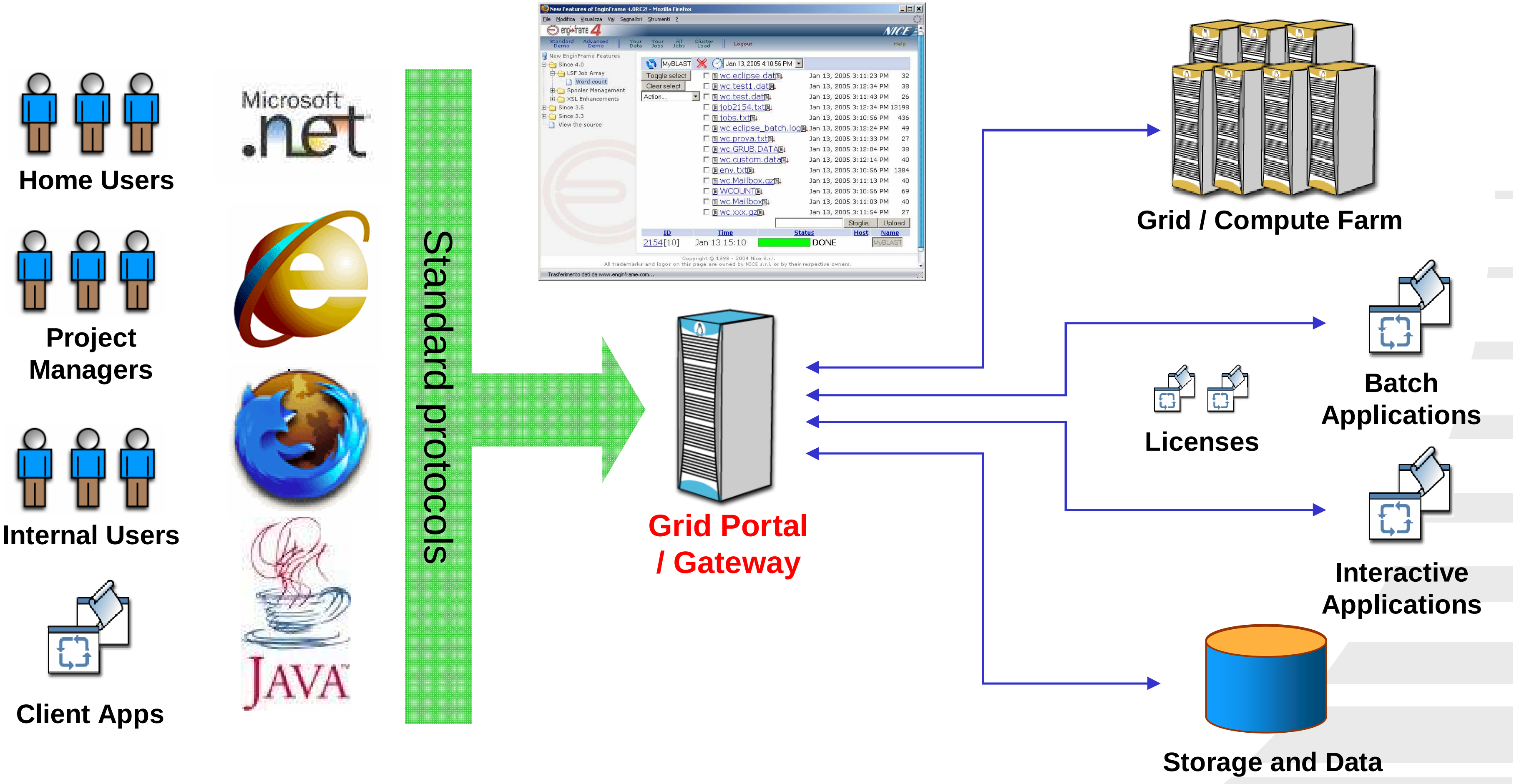
↓

The current prototype is able to use Moteur as batch enactor of Scufl workflows of standard Web Services

What is EnginFrame?

- It is a Web-based technology able to expose Grid services running on Grid infrastructures.
- It allows organizations to provide application-oriented computing and data services to both users (via Web browsers) and applications (via SOAP/WSDL and/or RSS), hiding all the complexity of the underlying Grid infrastructure
- It greatly simplifies the development of Web portals exposing computing services that can run on a broad range of different computational Grid systems

The Grid Portal / Gateway



New Features of EnginFrame 4.0RC2! - Mozilla Firefox

□ wc.eclipse.dat	Jan 13, 2005 3:11:23 PM	32
□ wc.test1.dat	Jan 13, 2005 3:12:34 PM	38
□ wc.test.dat	Jan 13, 2005 3:11:43 PM	26
□ job2154.txt	Jan 13, 2005 3:12:34 PM	13198
□ jobs.txt	Jan 13, 2005 3:10:56 PM	436
□ wc.eclipse_batch.log	Jan 13, 2005 3:12:24 PM	49
□ wc.prova.txt	Jan 13, 2005 3:11:33 PM	27
□ wc.GRUB.DATA	Jan 13, 2005 3:12:04 PM	38
□ wc.custom.data	Jan 13, 2005 3:12:14 PM	40
□ env.txt	Jan 13, 2005 3:10:56 PM	1384
□ wc.Mailbox.gz	Jan 13, 2005 3:11:13 PM	40
□ WCOUNT	Jan 13, 2005 3:10:56 PM	69
□ wc.Mailbox	Jan 13, 2005 3:11:03 PM	40
□ wc.xxx.gz	Jan 13, 2005 3:11:54 PM	27

ID	Time	Status	Host	Name
2154[10]	Jan 13 15:10	🟩 DONE		MyBLAST

Copyright © 1998 - 2004 Nice S.r.l.
All trademarks and logos on this page are owned by NICE s.r.l. or by their respective owners.

Trasferimento dati da www.enginframe.com...

EnginFrame adoption

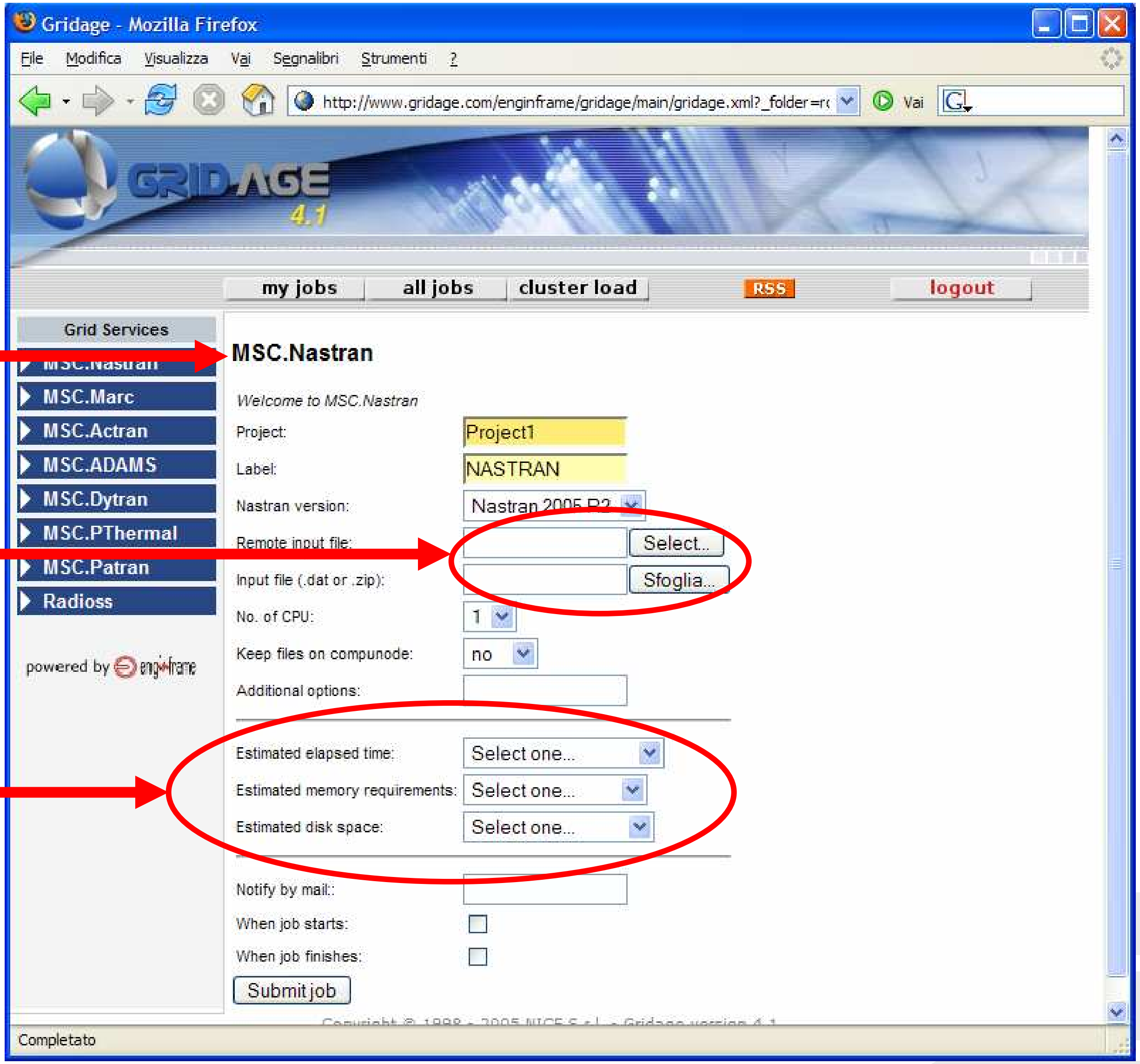
- Industries
  - **Mechanical**: Ferrari, Audi, BMW, FIAT Auto, Delphi, Elasis, Magneti Marelli, P+Z, Swagelok, Toyota, TRW
  - **Manufacturing:** Bridgestone, Procter & Gamble, Galileo Avionica
  - **Oil&Gas:** Slavneft, Schlumberger, TOTAL, VNIIGaz
  - **Electronics** :STMicroelectronics, Accent, SensorDynamics, Motorola
  - **Biotech:** ENEA, EGEE LS community
  - **Telecom:** Telecom Italia
- Research
  - CERN, INFN, ASSC, CCLRC, CILEA, CINECA, CNR, CNRS/IN2P3, ENEA, FzU, ICI, IFAE, ITEP, JSC G.G.M., KU Leuven, SSC-Russia,
- EnginFrame is the technology on which is based GENIUS, that's nowadays a standard of graphical user interface access to the EGEE grid infrastructure.

Usability & Input management

User friendly,
Application-oriented
Job submission

Flexible and efficient
Input file management

Hide complexity of
Underlying scheduler



MSC.Nastran

Welcome to MSC.Nastran

Project: Project1

Label: NASTRAN

Nastran version: Nastran 2005 R2

Remote input file: [Select...]

Input file (.dat or .zip): [Sfoglia...]

No. of CPU: 1

Keep files on compunode: no

Additional options:

Estimated elapsed time: Select one...

Estimated memory requirements: Select one...

Estimated disk space: Select one...

Notify by mail::

When job starts: □

When job finishes: □

[Submit job]

Copyright © 1998 - 2005 NICE S.r.l. - Gridage version 4.1

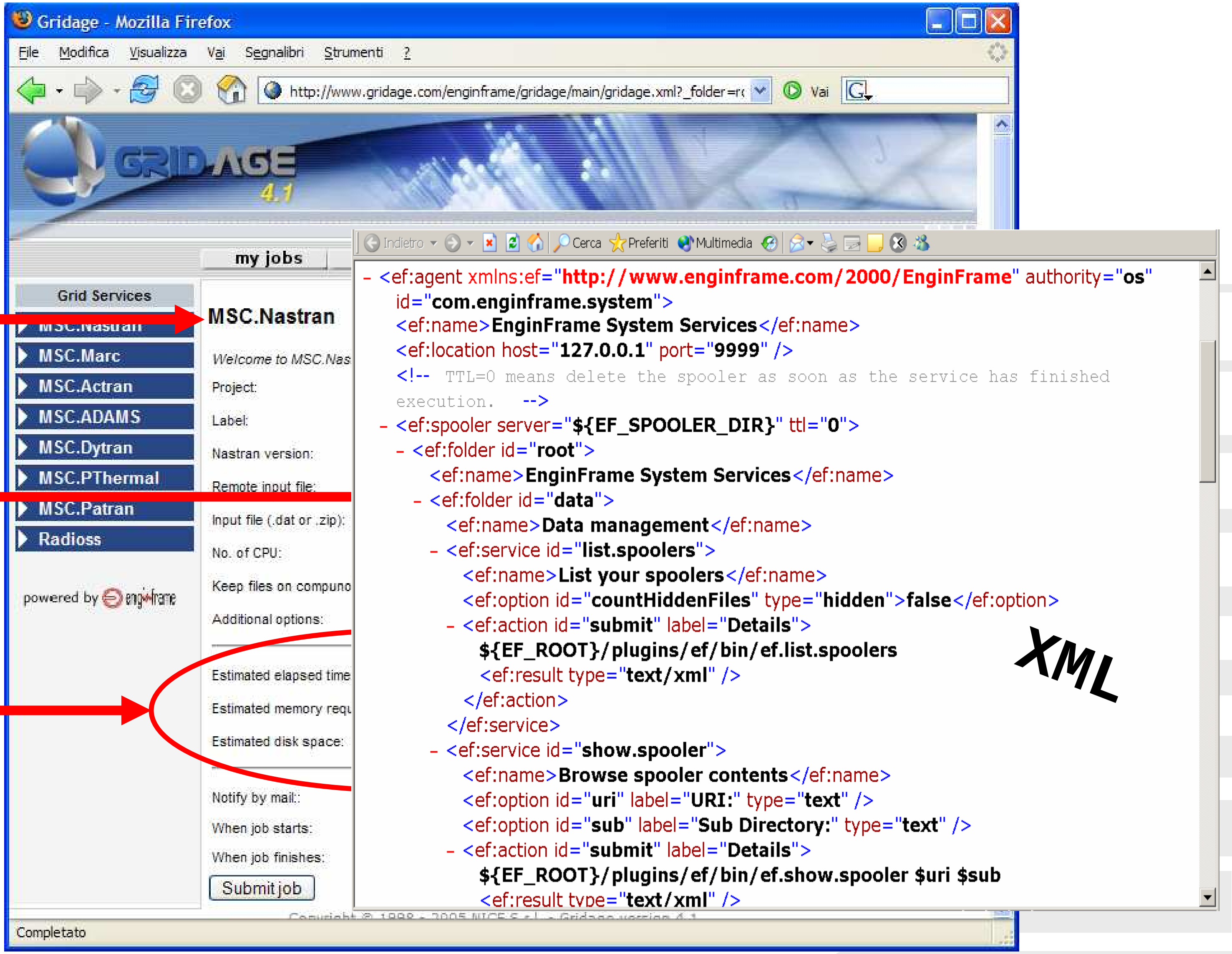
NETTAB 2006

Usability & Input management

User friendly,
Application-oriented
Job submission

Flexible and efficient
Input file management

Hide complexity of
Underlying scheduler


Gridage - Mozilla Firefox
File Modifica Visualizza Vai Segnalibri Strumenti ?
http://www.gridage.com/enginframe/gridage/main/gridage.xml?_folder=r
Vai
GRIDAGE 4.1
my jobs
Grid Services
MSC.Nastran
MSC.Marc
MSC.Actran
MSC.ADAMS
MSC.Dytran
MSC.PThermal
MSC.Patran
Radioss
powered by enginframe
Submit job
Completato
Indietro Cerca Preferiti Multimedia
XML

MSC.Nastran

Welcome to MSC.Nas

Project:

Label:

Nastran version:

Remote input file:

Input file (.dat or .zip):

No. of CPU:

Keep files on compuno

Additional options:

Estimated elapsed time

Estimated memory requ

Estimated disk space:

Notify by mail::

When job starts:

When job finishes:

Copyright © 1998 - 2005 NICE S.r.l. - Gridage version 4.1

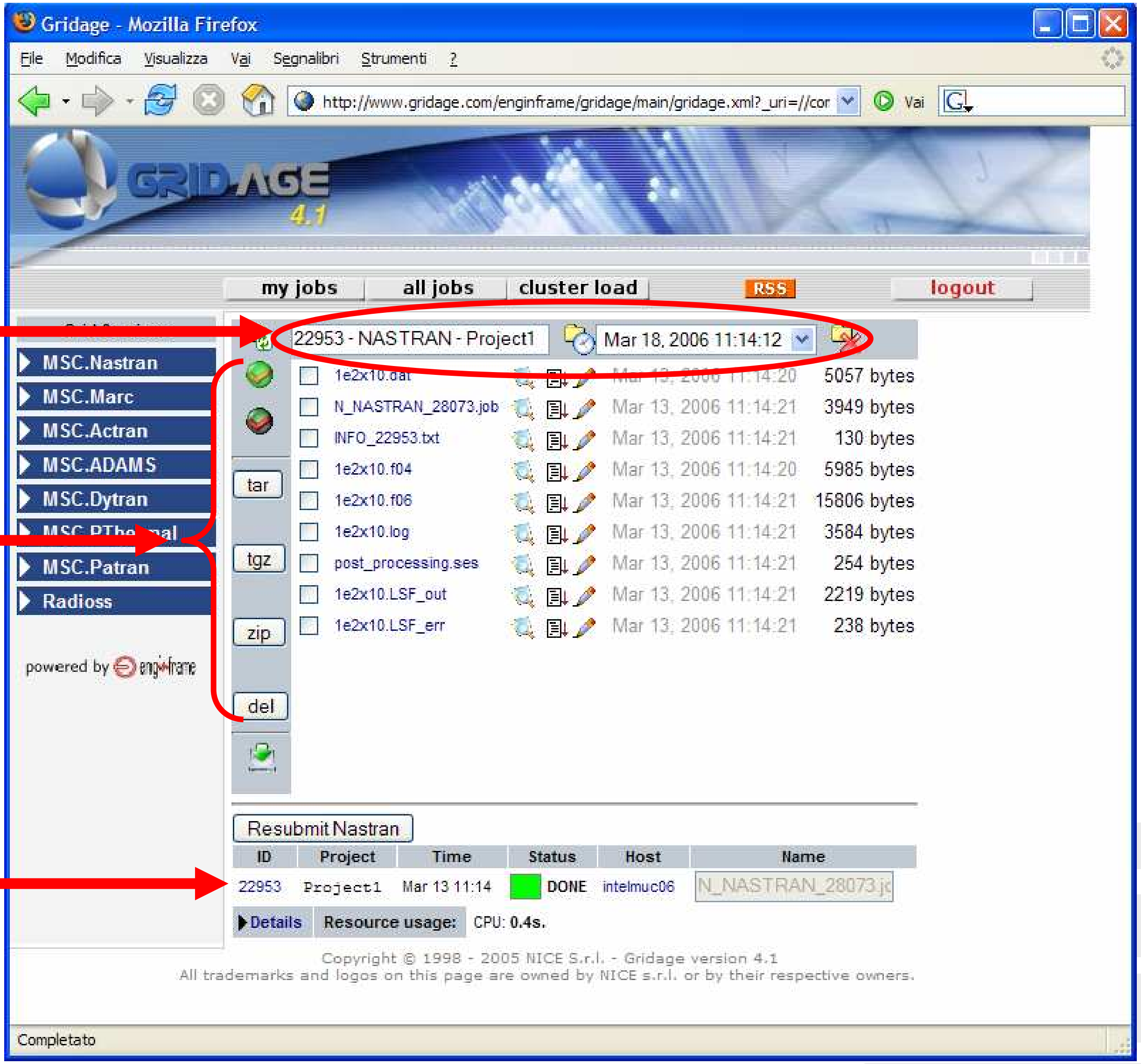
```
- <ef:agent xmlns:ef="http://www.enginframe.com/2000/EnginFrame" authority="os"
    id="com.enginframe.system">
    <ef:name>EnginFrame System Services</ef:name>
    <ef:location host="127.0.0.1" port="9999" />
    <!--  TTL=0 means delete the spooler as soon as the service has finished
    execution.   -->
  - <ef:spooler server="${EF_SPOOLER_DIR}" ttl="0">
    - <ef:folder id="root">
        <ef:name>EnginFrame System Services</ef:name>
      - <ef:folder id="data">
          <ef:name>Data management</ef:name>
        - <ef:service id="list.spoolers">
            <ef:name>List your spoolers</ef:name>
            <ef:option id="countHiddenFiles" type="hidden">false</ef:option>
          - <ef:action id="submit" label="Details">
              ${EF_ROOT}/plugins/ef/bin/ef.list.spoolers
              <ef:result type="text/xml" />
            </ef:action>
          </ef:service>
        - <ef:service id="show.spooler">
            <ef:name>Browse spooler contents</ef:name>
            <ef:option id="uri" label="URI:" type="text" />
            <ef:option id="sub" label="Sub Directory:" type="text" />
          - <ef:action id="submit" label="Details">
              ${EF_ROOT}/plugins/ef/bin/ef.show.spooler $uri $sub
              <ef:result type="text/xml" />
```

Monitoring and Output management

Data lifecycle managemnet

Comprehensive output File manipulation (view, edit, delete, zip, …)

Job details & control



Gridage - Mozilla Firefox

22953 - NASTRAN - Project1 Mar 18, 2006 11:14:12

□	1e2x10.dat	Mar 13, 2006 11:14:20	5057 bytes
□	N_NASTRAN_28073.job	Mar 13, 2006 11:14:21	3949 bytes
□	INFO_22953.txt	Mar 13, 2006 11:14:21	130 bytes
□	1e2x10.f04	Mar 13, 2006 11:14:20	5985 bytes
□	1e2x10.f06	Mar 13, 2006 11:14:21	15806 bytes
□	1e2x10.log	Mar 13, 2006 11:14:21	3584 bytes
□	post_processing.ses	Mar 13, 2006 11:14:21	254 bytes
□	1e2x10.LSF_out	Mar 13, 2006 11:14:21	2219 bytes
□	1e2x10.LSF_err	Mar 13, 2006 11:14:21	238 bytes

Resubmit Nastran

ID	Project	Time	Status	Host	Name
22953	Project1	Mar 13 11:14	🟩 DONE	intelmuc06	N_NASTRAN_28073.jo

▶Details Resource usage: CPU: 0.4s.

Copyright © 1998 - 2005 NICE S.r.l. - Gridage version 4.1
All trademarks and logos on this page are owned by NICE s.r.l. or by their respective owners.

Completato

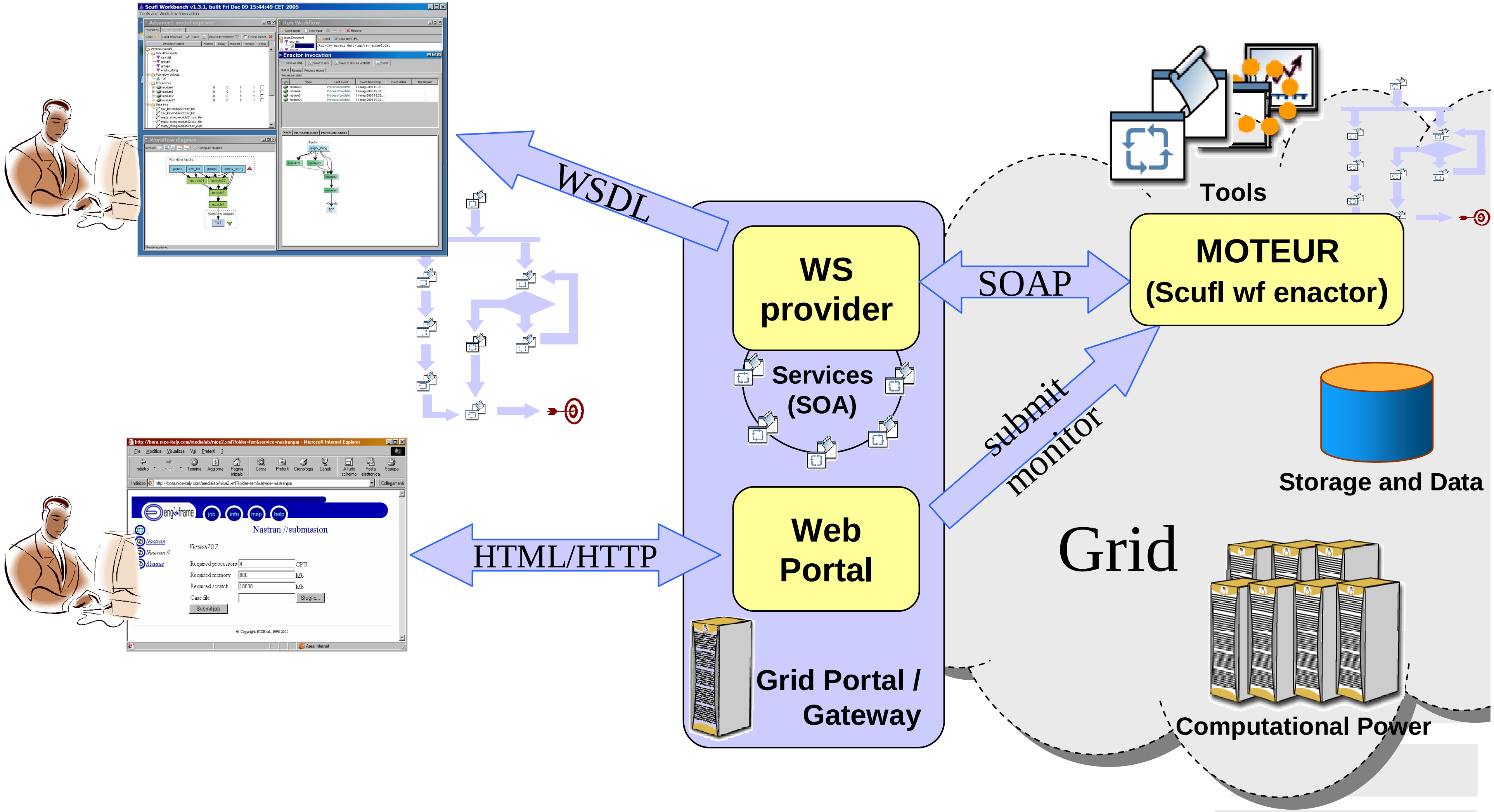
Double role of EnginFrame

EnginFrame plays a double role at two different levels of the proposed architecture:

1. Grid services provider: Grid services exposed as Web services by EnginFrame can be used in the workflow as standard nodes
2. Grid portal Web interface: for managing workflow submission and management on the Grid

In both of its roles EnginFrame exploits the computational power and data access capabilities provided by the backend Grid infrastructure.

The proposed Grid Infrastructure



Indirizzo: http://bora.nice-italy.com/medialab/nice2.xml?folder=fem&service=nastranpar

Nastran //submission

Version70.7

Required processors 4 CPU

Required memory 800 Mb

Required scratch 10000 Mb

Case file [ ] Sfoglia...

Submit job

© Copyright NICE srl, 1998-2000

Grid portal Web interface

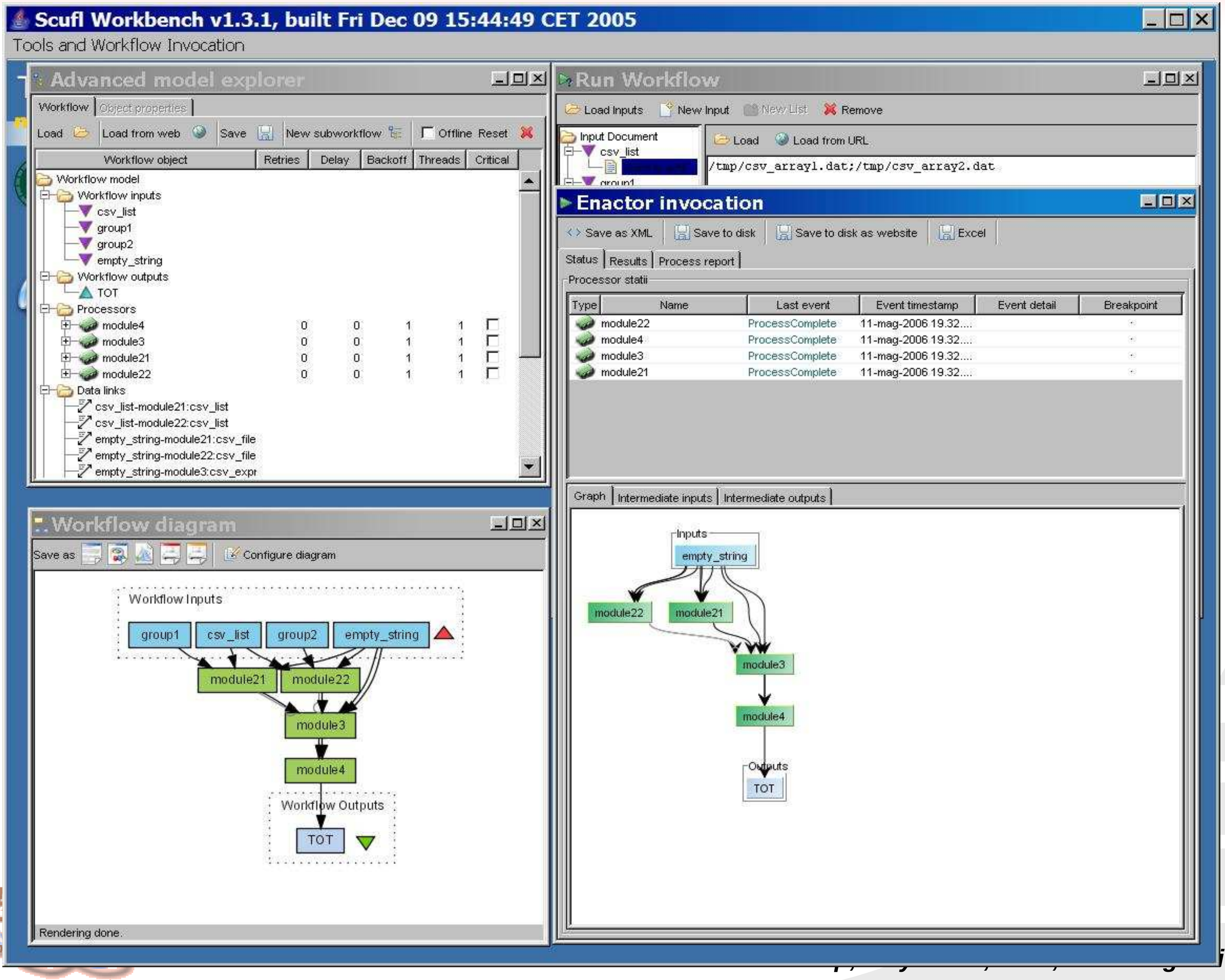
The service interface implemented provides the following functionalities:

- *Upload of Scufl workflow and related inputs:* users can upload their own workflows in Scufl language and insert input data for execution
- *Submission of Moteur as Grid job:* the workflow is executed by Moteur on the Grid infrastructure as a standard Grid job
- *Monitoring of workflow:* it is possible to check both Moteur submission and the workflow processing status together with data produced by intermediate results.
- *Results visualization:* when jobs are terminated, workflow results are staged in the EnginFrame spooler area and made available through the portal for visualization, post-processing or download.

Bioinformatics application

- Implementation of a bioinformatics application workflow exploiting the proposed architecture
- Workflow version of an innovative bioinformatics application developed by the Bio-Lab team of the University of Genoa
- Based on DChip, one of the most complete and diffuse free software for the microarray data analysis
- Composed of different modules:
  1. data set opening and normalization
  2. model based gene expression
  3. extraction of differentially expressed genes
  4. clustering

Application Workflow design



Scufl Workbench v1.3.1, built Fri Dec 09 15:44:49 CET 2005

Tools and Workflow Invocation

Advanced model explorer

Workflow object	Retries	Delay	Backoff	Threads	Critical
Workflow model					
Workflow inputs					
csv_list					
group1					
group2					
empty_string					
Workflow outputs					
TOT					
Processors					
module4	0	0	1	1	☐
module3	0	0	1	1	☐
module21	0	0	1	1	☐
module22	0	0	1	1	☐
Data links					
csv_list-module21:csv_list					
csv_list-module22:csv_list					
empty_string-module21:csv_file					
empty_string-module22:csv_file					
empty_string-module3:csv_expr					

Run Workflow

/tmp/csv_array1.dat;/tmp/csv_array2.dat

Enactor invocation

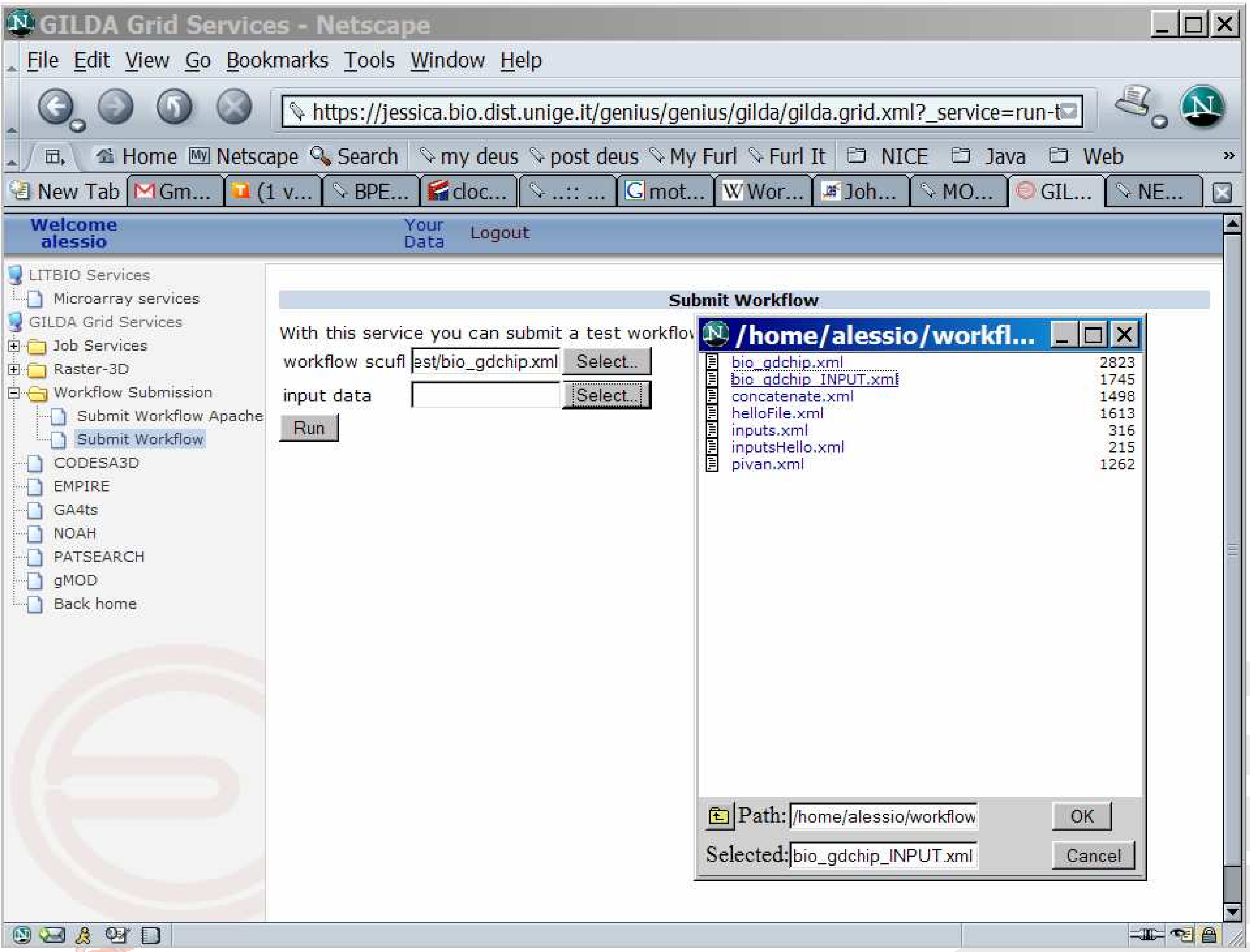
Processor statii

Type	Name	Last event	Event timestamp	Event detail	Breakpoint
	module22	ProcessComplete	11-mag-2006 19.32....		·
	module4	ProcessComplete	11-mag-2006 19.32....		·
	module3	ProcessComplete	11-mag-2006 19.32....		·
	module21	ProcessComplete	11-mag-2006 19.32....		·

Workflow diagram

Rendering done.

Scufl submission and enactment



GILDA Grid Services - Netscape
File Edit View Go Bookmarks Tools Window Help
https://jessica.bio.dist.unige.it/genius/genius/gilda/gilda.grid.xml?_service=run-t
Home Netscape Search my deus post deus My Furl Furl It NICE Java Web
New Tab Gm... (1 v... BPE... doc... ..:: ... mot... Wor... Joh... MO... GIL... NE...
Welcome alessio
Your Data
Logout
LITBIO Services
Microarray services
GILDA Grid Services
Job Services
Raster-3D
Workflow Submission
Submit Workflow Apache
Submit Workflow
CODESA3D
EMPIRE
GA4ts
NOAH
PATSEARCH
gMOD
Back home

Submit Workflow

With this service you can submit a test workflo

workflow scufl: est/bio_gdchip.xml Select..

input data: Select..

Run

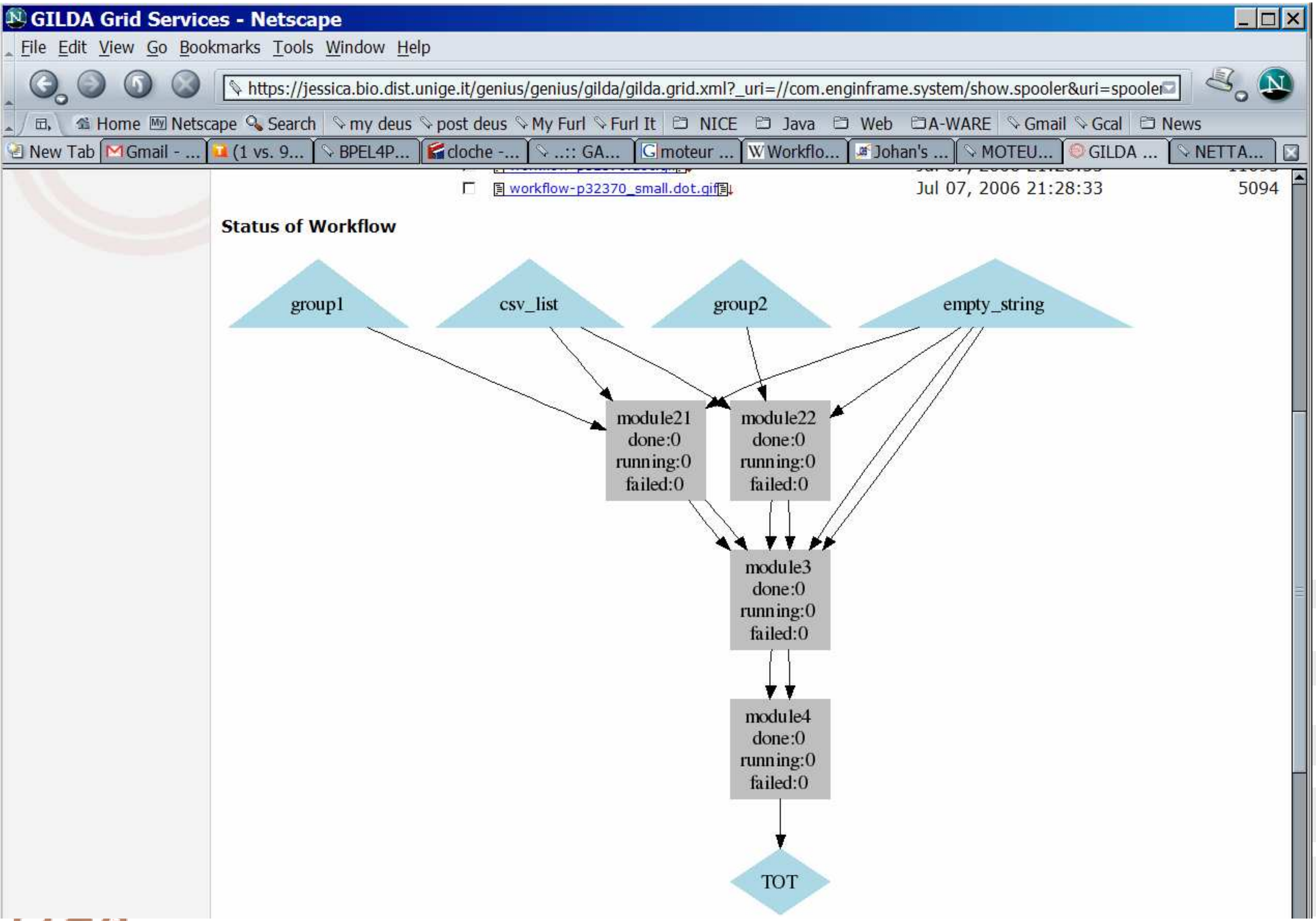
/home/alessio/workfl...

bio_gdchip.xml	2823
bio_gdchip_INPUT.xml	1745
concatenate.xml	1498
helloFile.xml	1613
inputs.xml	316
inputsHello.xml	215
pivan.xml	1262

Path: /home/alessio/workflow OK

Selected: bio_gdchip_INPUT.xml Cancel

Workflow status



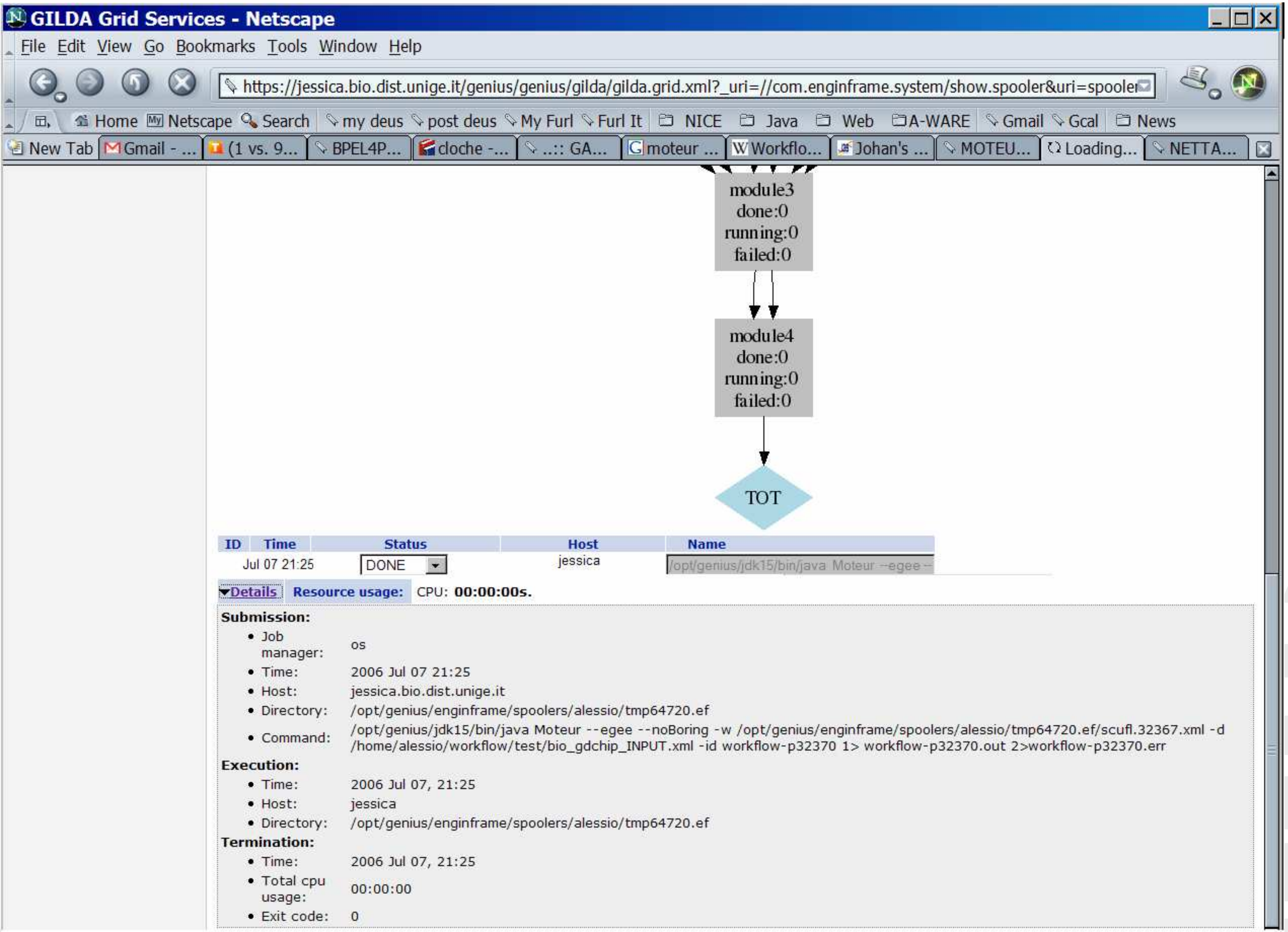
GILDA Grid Services - Netscape
File Edit View Go Bookmarks Tools Window Help
https://jessica.bio.dist.unige.it/genius/genius/gilda/gilda.grid.xml?_uri=//com.enginframe.system/show.spooler&uri=spooler
Home Netscape Search my deus post deus My Furl Furl It NICE Java Web A-WARE Gmail Gcal News
New Tab Gmail - ... (1 vs. 9... BPEL4P... cloche -... ..:: GA... moteur ... Workflo... Johan's ... MOTEU... GILDA ... NETTA...

□ workflow-p32370_small.dot.gif Jul 07, 2006 21:28:33 5094

Status of Workflow

group1
csv_list
group2
empty_string
module21
done:0
running:0
failed:0
module22
done:0
running:0
failed:0
module3
done:0
running:0
failed:0
module4
done:0
running:0
failed:0
TOT

Job details



ID	Time	Status	Host	Name
	Jul 07 21:25	DONE	jessica	/opt/genius/jdk15/bin/java Moteur --egee --

▾Details **Resource usage:** CPU: **00:00:00s.**

Submission:

- Job manager: os
- Time: 2006 Jul 07 21:25
- Host: jessica.bio.dist.unige.it
- Directory: /opt/genius/enginframe/spoolers/alessio/tmp64720.ef
- Command: `/opt/genius/jdk15/bin/java Moteur --egee --noBoring -w /opt/genius/enginframe/spoolers/alessio/tmp64720.ef/scufl.32367.xml -d /home/alessio/workflow/test/bio_gdchip_INPUT.xml -id workflow-p32370 1> workflow-p32370.out 2>workflow-p32370.err`

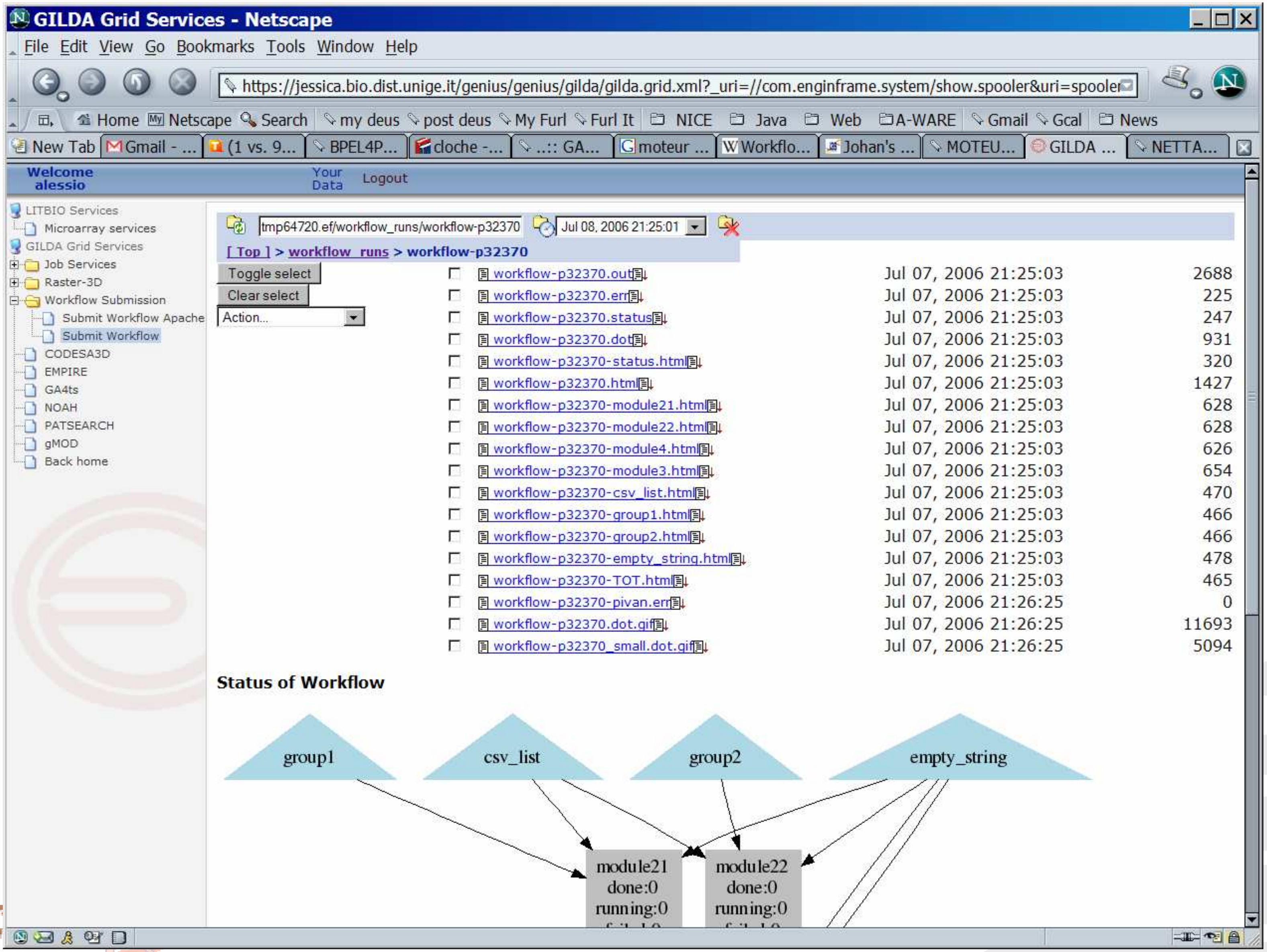
Execution:

- Time: 2006 Jul 07, 21:25
- Host: jessica
- Directory: /opt/genius/enginframe/spoolers/alessio/tmp64720.ef

Termination:

- Time: 2006 Jul 07, 21:25
- Total cpu usage: 00:00:00
- Exit code: 0

Output Management



GILDA Grid Services - Netscape

https://jessica.bio.dist.unige.it/genius/genius/gilda/gilda.grid.xml?_uri=//com.enginframe.system/show.spooler&uri=spooler

File Edit View Go Bookmarks Tools Window Help
Home Netscape Search my deus post deus My Furl Furl It NICE Java Web A-WARE Gmail Gcal News
New Tab Gmail - ... (1 vs. 9... BPEL4P... cloche -... ..:: GA... moteur ... Workflo... Johan's ... MOTEU... GILDA ... NETTA...
Welcome alessio
Your Data
Logout
LITBIO Services
Microarray services
GILDA Grid Services
Job Services
Raster-3D
Workflow Submission
Submit Workflow Apache
Submit Workflow
CODESA3D
EMPIRE
GA4ts
NOAH
PATSEARCH
gMOD
Back home

tmp64720.ef/workflow_runs/workflow-p32370 Jul 08, 2006 21:25:01

[Top] > workflow_runs > workflow-p32370

Toggle select
Clear select
Action...

□	workflow-p32370.out	Jul 07, 2006 21:25:03	2688
□	workflow-p32370.err	Jul 07, 2006 21:25:03	225
□	workflow-p32370.status	Jul 07, 2006 21:25:03	247
□	workflow-p32370.dot	Jul 07, 2006 21:25:03	931
□	workflow-p32370-status.html	Jul 07, 2006 21:25:03	320
□	workflow-p32370.html	Jul 07, 2006 21:25:03	1427
□	workflow-p32370-module21.html	Jul 07, 2006 21:25:03	628
□	workflow-p32370-module22.html	Jul 07, 2006 21:25:03	628
□	workflow-p32370-module4.html	Jul 07, 2006 21:25:03	626
□	workflow-p32370-module3.html	Jul 07, 2006 21:25:03	654
□	workflow-p32370-csv_list.html	Jul 07, 2006 21:25:03	470
□	workflow-p32370-group1.html	Jul 07, 2006 21:25:03	466
□	workflow-p32370-group2.html	Jul 07, 2006 21:25:03	466
□	workflow-p32370-empty_string.html	Jul 07, 2006 21:25:03	478
□	workflow-p32370-TOT.html	Jul 07, 2006 21:25:03	465
□	workflow-p32370-pivan.err	Jul 07, 2006 21:26:25	0
□	workflow-p32370.dot.gif	Jul 07, 2006 21:26:25	11693
□	workflow-p32370_small.dot.gif	Jul 07, 2006 21:26:25	5094

Status of Workflow

Conclusions

- We have proposed a Grid infrastructure supporting workflow design with Grid services as building blocks, workflow enactment and life cycle management.

- A workflow of a bioinformatics application has been described to show how it is possible to exploit this type of infrastructure in bionformatics area

- Limits: Moteur appears as a performing enactor tool, but it imposes some limitations about processors and data inputs → many standard incompatibilities with most common used bioinformatics services

- Future steps: the proposed infrastructure is still a prototype.
  - Stay tuned on the evolution of MOTEUR…
  - Improvements to DChip modules interfaces to make easier to wrap them as EF services
  - Improve Web interface to provide users with dynamically generated forms for workflow inputs definition

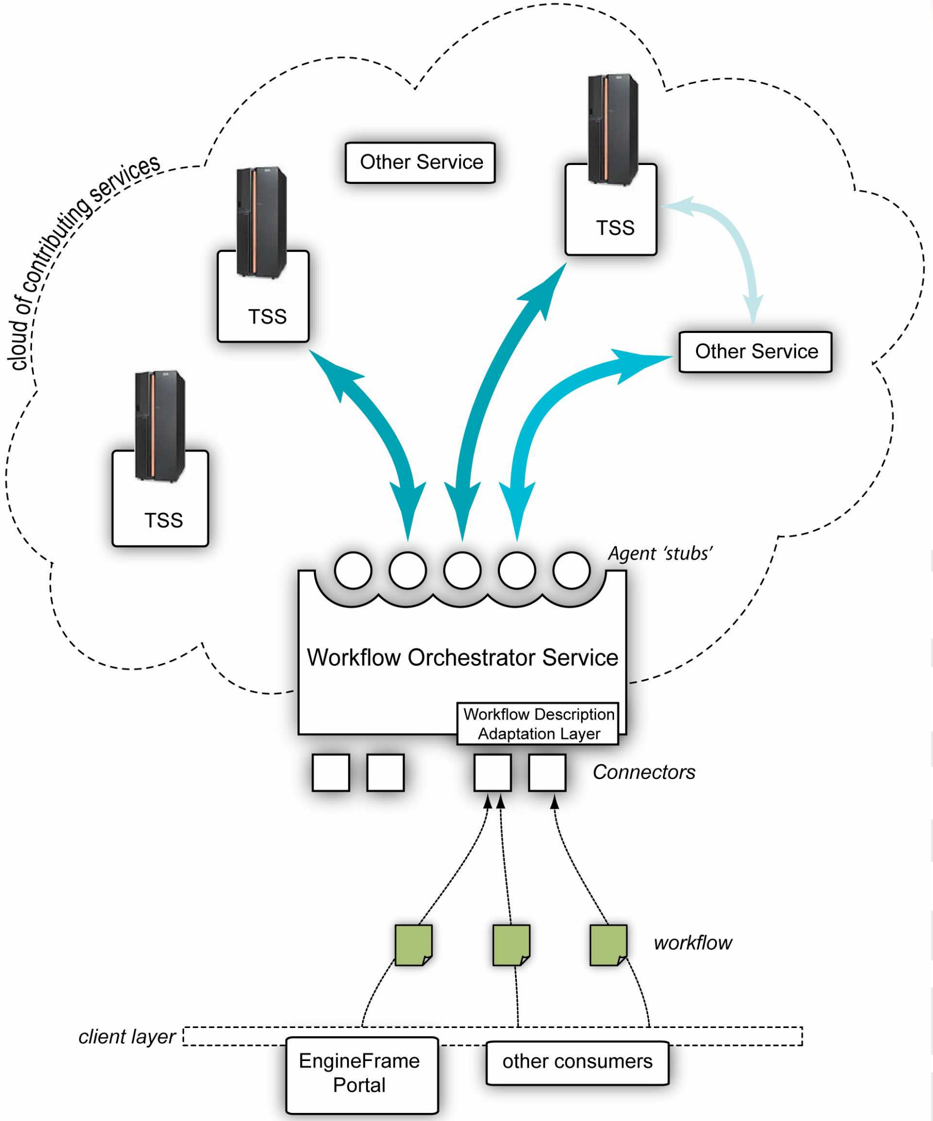
A-WARE FP6 Funded Project

- Project goals
  - Simplify users'
    - life (focus on problems)
    - way of perceiving the GRID
  - Fill an existing gap
    - between middleware and portals
  - EnginFrame + A-WARE + UNICORE/GS aim to be a completely integrated solution
  - Contribute to the standards

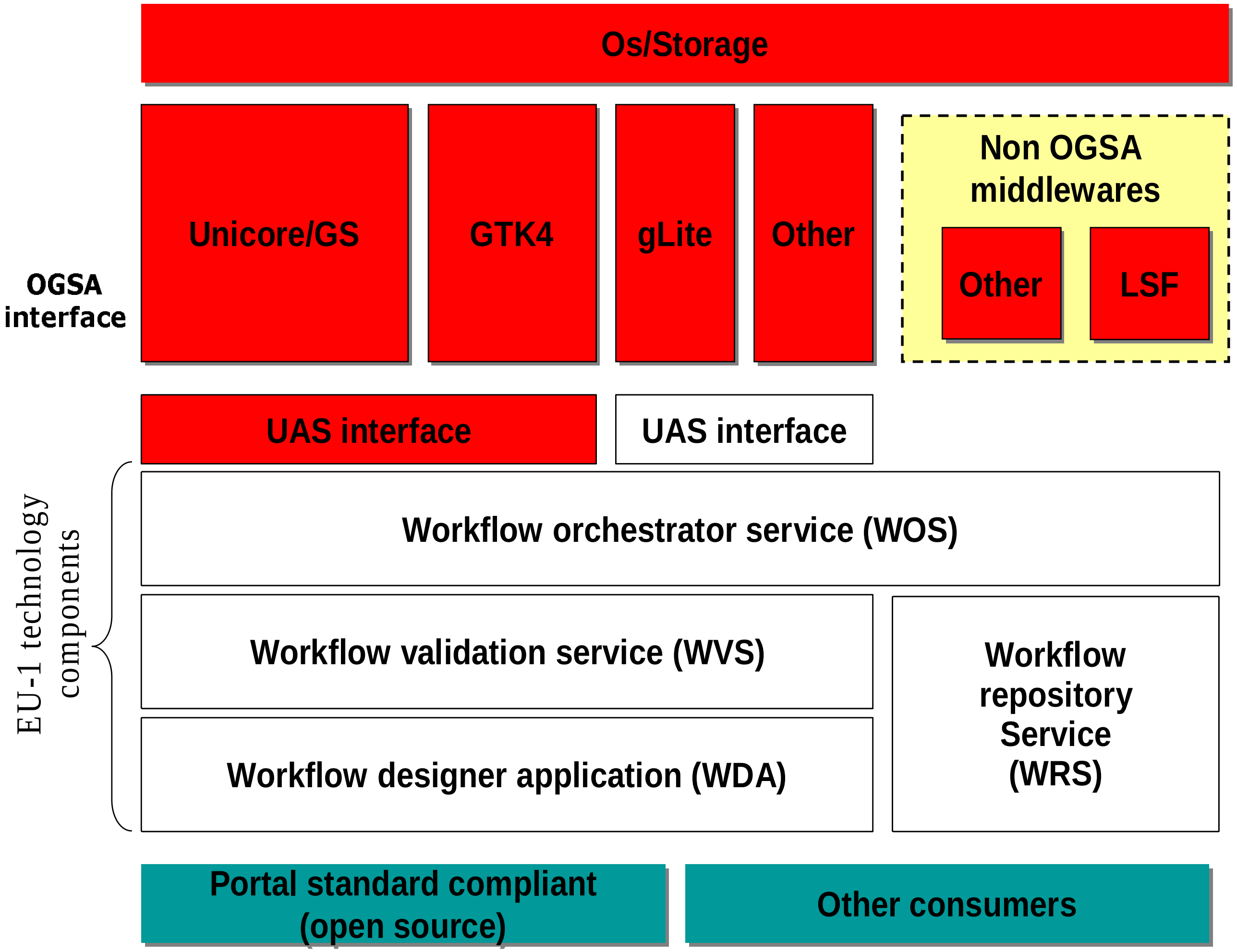
AIRBUS CINECA Consorzio Interuniversitario FUJITSU NICE

A-WARE: Key technology advancements

- Key Advancements
  There is a requirement for a higher level access point to the Grid, as an entry at the Target System is too low level. What is missing is a high-level service, which can manage the multiple invocations of TSS, and other services. The project will supply this as a Workflow Orchestrator Service (WOS).



A-WARE: Key technology advancements


Os/Storage
OGSA interface
Unicore/GS
GTK4
gLite
Other
Non OGSA middlewares
Other
LSF
UAS interface
UAS interface
EU-1 technology components
Workflow orchestrator service (WOS)
Workflow validation service (WVS)
Workflow designer application (WDA)
Workflow repository Service (WRS)
Portal standard compliant (open source)
Other consumers

Thanks for your attention!

Q&A