Executing complex computational workflows on EGEE Grid

Osvaldo Gervasi [1], Alessandro Costantini [1,2], Antonio Laganà [2]

[1]Dept. of Mathematics and Computer Science, University of Perugia (Italy)
[2]Dept. of Chemistry, University of Perugia (Italy)

EGEE 08 Conference

Istanbul (TR), September 22nd- 26th, 2008

Information Society and Media

Outline

- **The Molecular Science community**
- **The CompChem VO**
- **GEMS complex workflow**
  - scripts
  - Web portal
  - Workload Management System
- **Using PGrade: the ABC test case**
- **Conclusions**

The Molecular Science community

- **The understanding of the behavior of molecular systems is of great importance for the progress of**
  - the life sciences
  - several industrial applications
- **The Molecular Science community study the molecular systems performing simulations that are heavy demanding in terms of computational resources.**
- **It is mandatory to put together the competencies of various laboratories to achieve ambitious results:**
  - active collaboration between people with complementary expertise
  - interaction between various computational approaches

↳ **COST CMST Action D37, *GridChem***

http://www.cost.esf.org/index.php?id=189&action_number=D37

The Molecular Science community

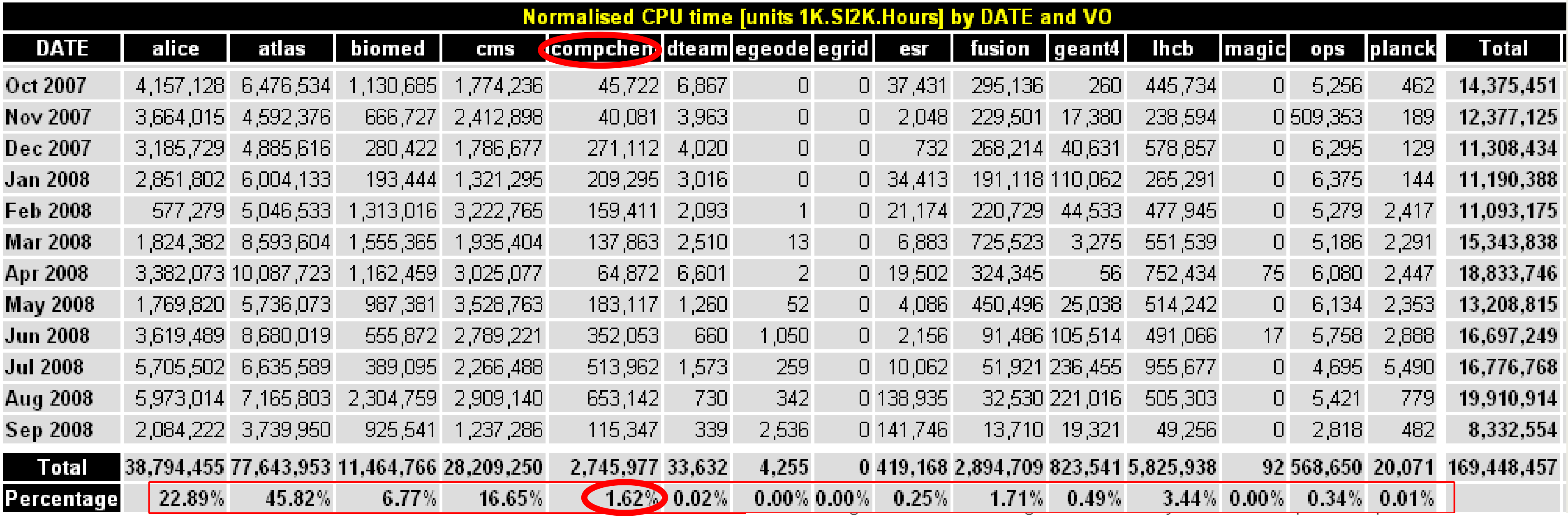
- **The EGEE Grid environment represents for this community an important infrastructure able to supply**
  - the necessary computational power
  - the proper middleware enabling people to collaborate and access the shared resources in a secure way.
- **Several EGEE sites are supporting the VO, in particular the Italian EGEE sites, CESGA (Spain), CYFRONET and POZNAN Supercomputing Center (Poland), Hellas Grid and GRNET (Greece), University of Cyprus (Cyprus).**
- **The users of CompChem VO have already performed some relevant intentive computational studies (N+N2, CcO, new materials design) on EGEE Production Grid.**

Compchem VO

- **CompChem VO is running on the EGEE production Grid from the end of 2004 to support Computational Chemistry applications (http://compchem.unipg.it)**
- **The VO contributes to the EGEE production grid with 2 small Clusters. A medium-sized cluster will be added at the end of the year.**
- **We made available a powerful User Interface (UI) to CompChem users: ui.grid.unipg.it**
  - New users of the VO are encouraged to use such host to start exploiting the CompChem VO facilities.
- **The porting in the EGEE Grid environment of the programs the user need to use, is one of the most crucial task for the user and the VO management.**
- **We are members of the EGEE project (the membership started in EGEE III).**

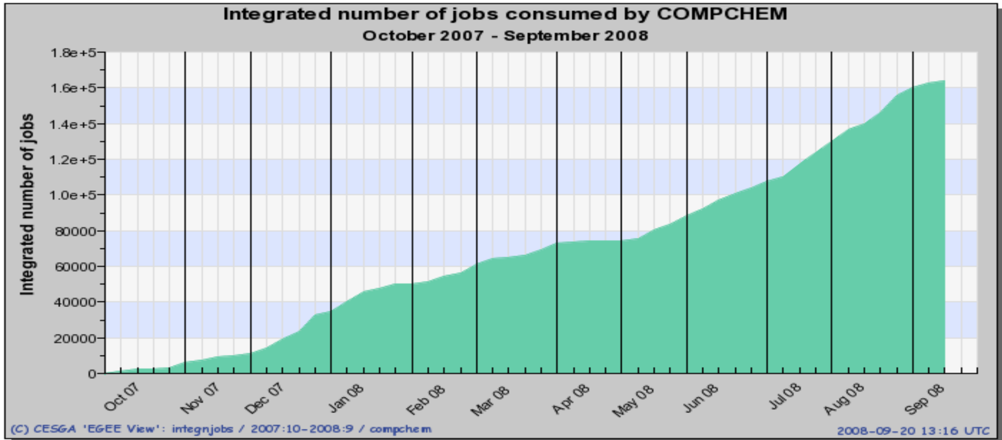
CompChem sustained activity

- The CompChem VO is one of the most active VOs of the “Generic Applications” of EGEE (rank 4th after HEP and Biomed and Fusion communities)



Normalised CPU time [units 1K.SI2K.Hours] by DATE and VO

DATE	alice	atlas	biomed	cms	compchem	dteam	egeode	egrid	esr	fusion	geant4	lhcb	magic	ops	planck	Total
Oct 2007	4,157,128	6,476,534	1,130,685	1,774,236	45,722	6,867	0	0	37,431	295,136	260	445,734	0	5,256	462	14,375,451
Nov 2007	3,664,015	4,592,376	666,727	2,412,898	40,081	3,963	0	0	2,048	229,501	17,380	238,594	0	509,353	189	12,377,125
Dec 2007	3,185,729	4,885,616	280,422	1,786,677	271,112	4,020	0	0	732	268,214	40,631	578,857	0	6,295	129	11,308,434
Jan 2008	2,851,802	6,004,133	193,444	1,321,295	209,295	3,016	0	0	34,413	191,118	110,062	265,291	0	6,375	144	11,190,388
Feb 2008	577,279	5,046,533	1,313,016	3,222,765	159,411	2,093	1	0	21,174	220,729	44,533	477,945	0	5,279	2,417	11,093,175
Mar 2008	1,824,382	8,593,604	1,555,365	1,935,404	137,863	2,510	13	0	6,883	725,523	3,275	551,539	0	5,186	2,291	15,343,838
Apr 2008	3,382,073	10,087,723	1,162,459	3,025,077	64,872	6,601	2	0	19,502	324,345	56	752,434	75	6,080	2,447	18,833,746
May 2008	1,769,820	5,736,073	987,381	3,528,763	183,117	1,260	52	0	4,086	450,496	25,038	514,242	0	6,134	2,353	13,208,815
Jun 2008	3,619,489	8,680,019	555,872	2,789,221	352,053	660	1,050	0	2,156	91,486	105,514	491,066	17	5,758	2,888	16,697,249
Jul 2008	5,705,502	6,635,589	389,095	2,266,488	513,962	1,573	259	0	10,062	51,921	236,455	955,677	0	4,695	5,490	16,776,768
Aug 2008	5,973,014	7,165,803	2,304,759	2,909,140	653,142	730	342	0	138,935	32,530	221,016	505,303	0	5,421	779	19,910,914
Sep 2008	2,084,222	3,739,950	925,541	1,237,286	115,347	339	2,536	0	141,746	13,710	19,321	49,256	0	2,818	482	8,332,554
Total	38,794,455	77,643,953	11,464,766	28,209,250	2,745,977	33,632	4,255	0	419,168	2,894,709	823,541	5,825,938	92	568,650	20,071	169,448,457
Percentage	22.89%	45.82%	6.77%	16.65%	1.62%	0.02%	0.00%	0.00%	0.25%	1.71%	0.49%	3.44%	0.00%	0.34%	0.01%	



Integrated number of jobs consumed by COMPCHEM
October 2007 - September 2008

(C) CESGA 'EGEE View': integnjobs / 2007:10-2008:9 / compchem

2008-09-20 13:16 UTC

Compchem users

Italy	١٦	Univ. Perugia, CNR-IMIP, ENEA, Univ. Palermo
Grece	١٤	FORTH Crete, Universities: Crete, Athens, Tessalonilki
Spain	٥	Univ. Basque Country, Univ. Barcelona, CESGA
Austria	٥	Univ. Vienna, Univ. Innsbruk
France	٣	CNRS
Poland	١	Cyfronet
Sweden	١	PDC, Royal Institute of Technology,Stockholm
UK	١	Imperial College London
Portugal	١	IRICUP
Cyprus	١	Univ. Cyprus
Lithuania	١	Univ Vilnius
Croatia	١	Univ Zagabria
Hungary	١	Hungarian Academy of Science
Total:	٥١	

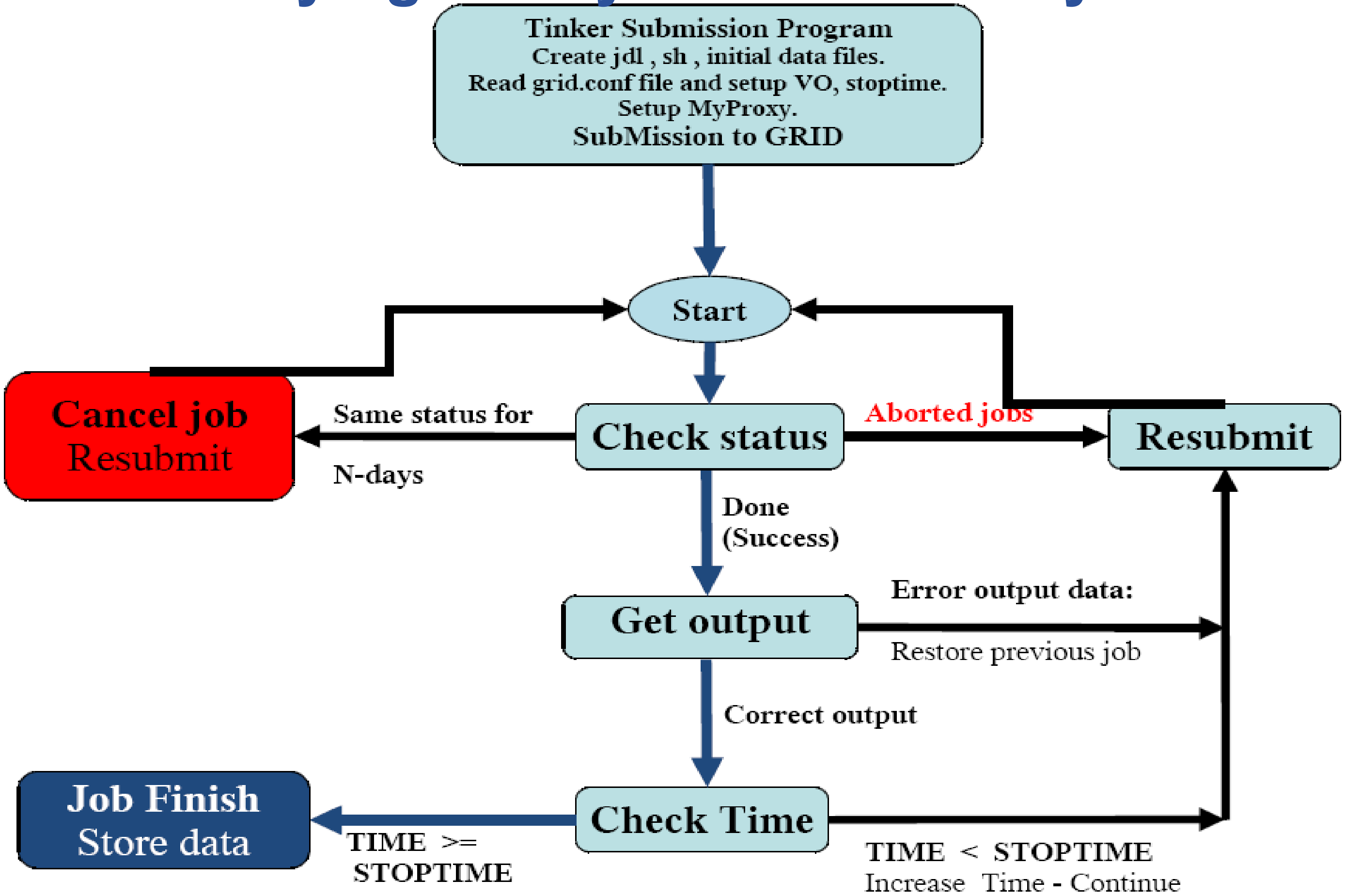
Most users are collaborating in the COST CMST D37 Action, GridChem

User's needs

- **When performing a computational campaign, the user needs an interface for managing jobs during the various phases:**
  - Submission of the job
  - Monitoring during the job execution
  - Retrieval of the output when the job completes
  - Management of failed jobs (aborted jobs for the instability of Grid services, time exeeded, hardware errors. etc) resubmitting them with the same input data
- **The user may also need:**
  - an high level interface to specify the details of a given computation
  - to control the execution of a series of programs under the user supervision (fine tuning of parameters, etc)
  - an high level access and representation of the results of the computation

An easy answer: scripts

- An easy answer to the needs expressed, could be implemented using a scripting language (sh, python, PHP, etc). Here is the diagram of the command-line scripts implemented for studying the Cytochrome c Oxydase:



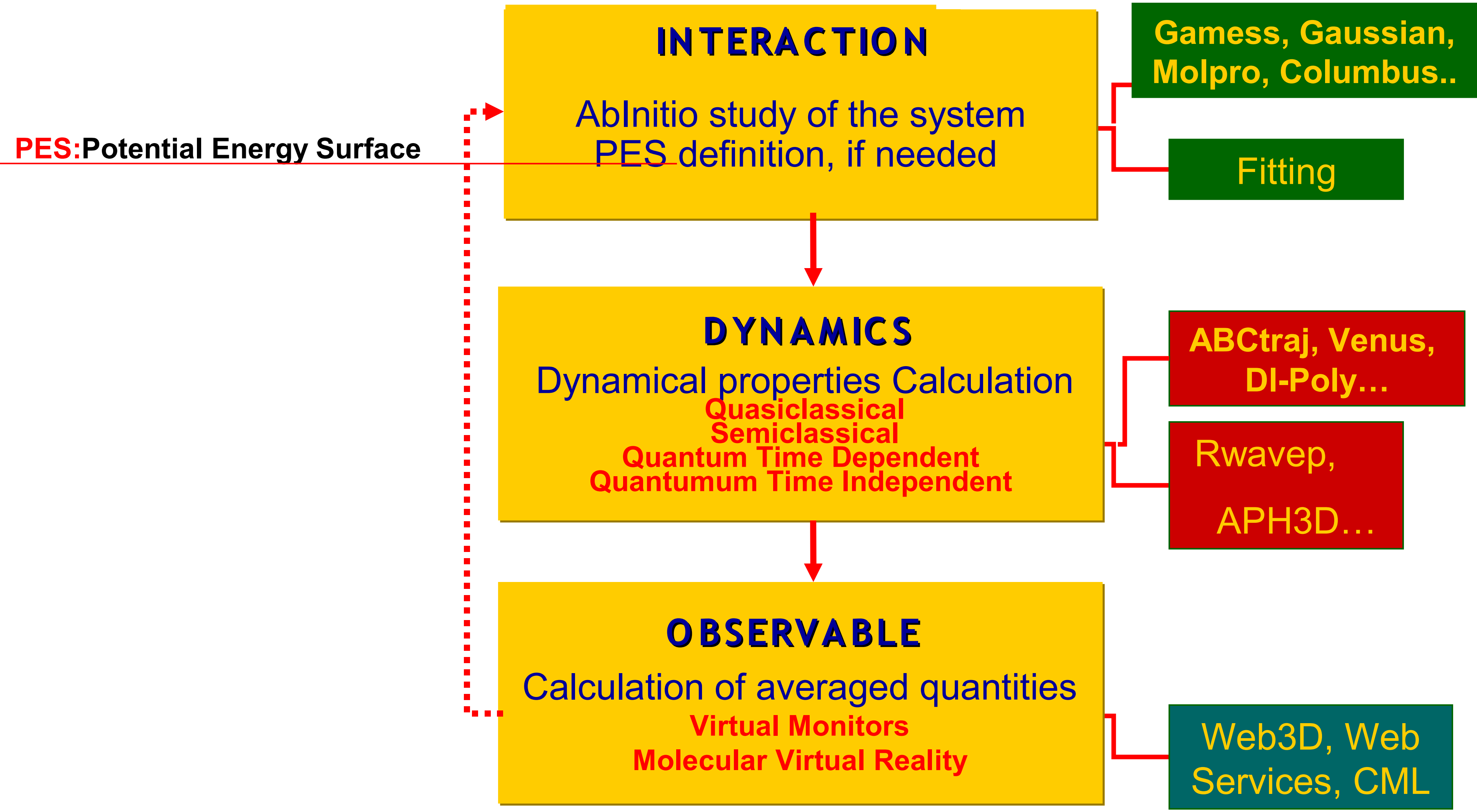
CompChem VO

Structure	١ar١A_B١٨	١arA_BGH١٨	١ar١A_DGH١٨
Execution time (days)	٢٩	٣١	٣٠
# of submitted jobs	١٨٥٤	١٩٠٢	١٨٦٤
# of aborted jobs	١٨٠	٢٧٣	٢٢٨
# of canceled jobs	٣٥	١٧	١٨
# of succeeded jobs	١٦٣٩	١٦١٢	١٦١٨
Efficiency	٨٨.٤%	٨٤.٨%	٨٦.٨%

# of succeded jobs	٤٨٦٩
# of aborted jobs	٦٨١
# of canceled jobs	٧٠
Efficiency	٨٦.٤%
Size (MB) /job	١٤
IO (GB)	٨٠
CPU time (days)	٣٠٤٣

Very good results!

Complex workflow: GEMS



PES: Potential Energy Surface

INTERACTION

AbInitio study of the system
PES definition, if needed

Gamess, Gaussian, Molpro, Columbus..

Fitting

DYNAMICS

Dynamical properties Calculation
Quasiclassical
Semiclassical
Quantum Time Dependent
Quantumum Time Independent

ABCtraj, Venus, DI-Poly…

Rwavep,

APH3D…

OBSERVABLE

Calculation of averaged quantities
Virtual Monitors
Molecular Virtual Reality

Web3D, Web Services, CML

GEMS: the INTERACTION module

- **The module related to the ab initio study of the molecular systems is still a prototype. We are working to make available to the users the following programs:**
  - GAMESS
  - COLUMBUS
- **Several activities planned in the COST Action GridChem are focused on the subjects of this module.**
- **Regarding the popular commercial programs Gaussian and Molpro we will collaborate with the Gaussian VO.**

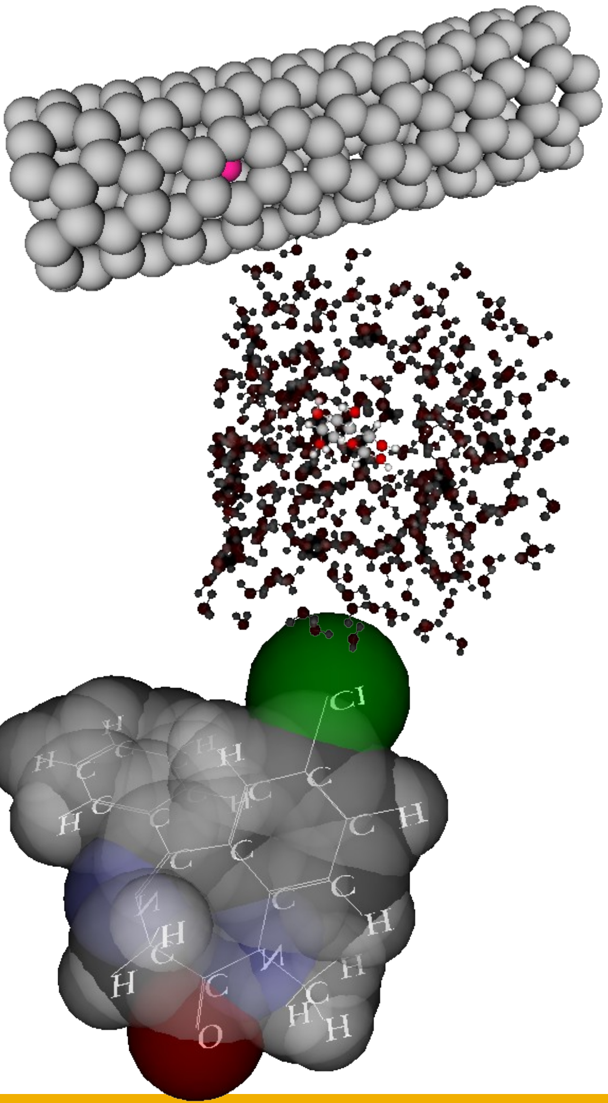
GEMS: the FITTING module

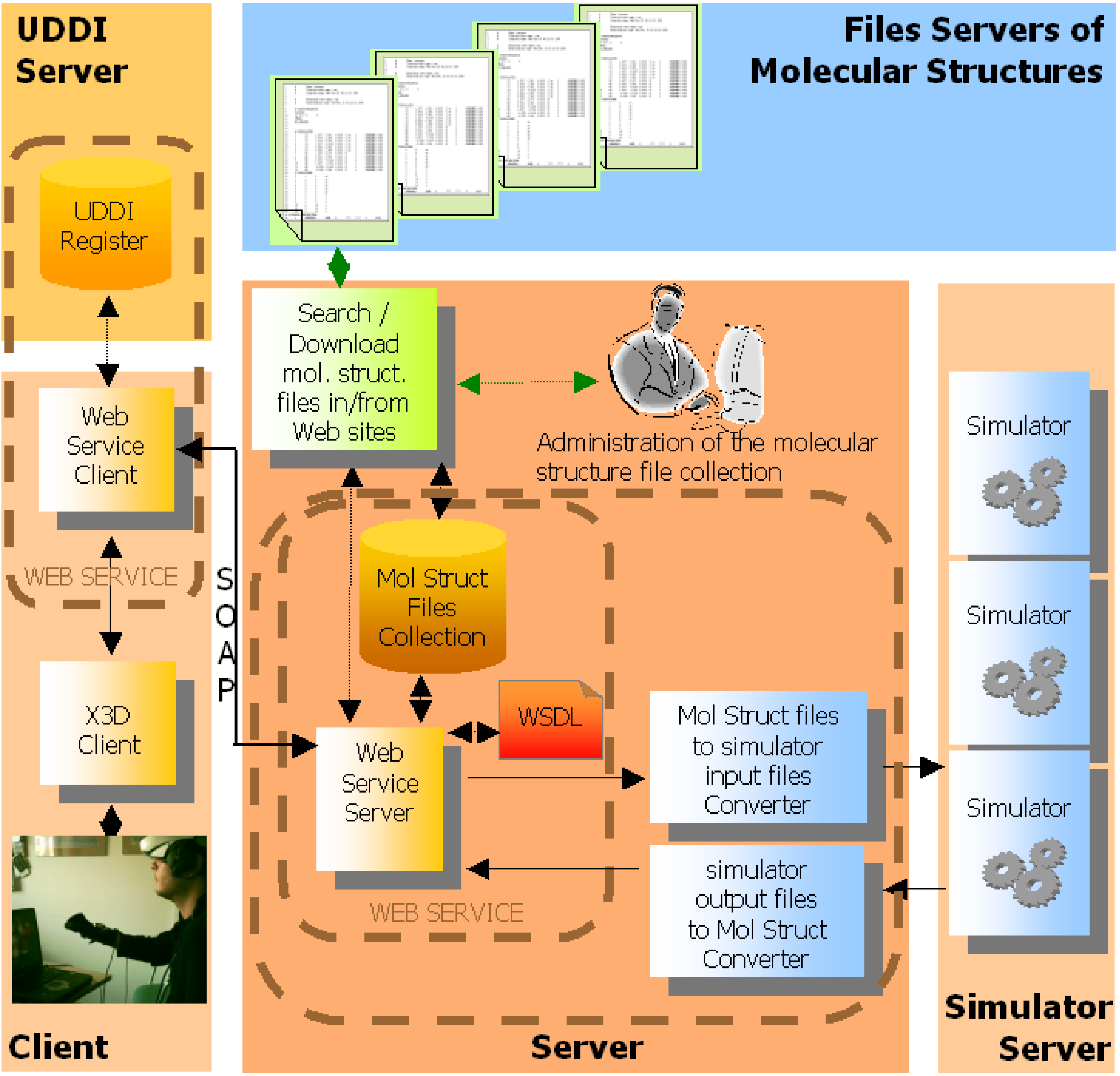
- The Ab initio study of the system produces a grid of values that need to be interpolated (best fitting process) in order to produce a functional form of the Potential Energy Surface (PES)
- Some constraints must be reproduced by the functional form (i.e.: spectroscopic and experimental information).
- The functional form have to satisfy some mathematical constraints of the computational model used to perform the dynamical study of the system (i.e.: the PES function must be derivable, etc)
- Several types of functional forms are used. They have different properties and degree of accuracy (LEPS, Bond Order, Rotating Bond Order, etc).
- This module is working as a prototype

GEMS: the DYNAMICS module

- **In the Dynamics modules we consider all programs that perform the dynamical simulations, using (if required by the method) a given functional representation of the PES and providing the estimate of the reaction observables.**
- **The following programs have been implemented:**
  - Quasiclassical approach
    - ABCtraj (atom + diatom)
    - Venus (many atoms)
    - DL-Poly (complex and biological systems)
  - Quantum approach
    - Rwavepr (time dependent approach)
    - ABC (time independent approach)

GEMS: the Observable module

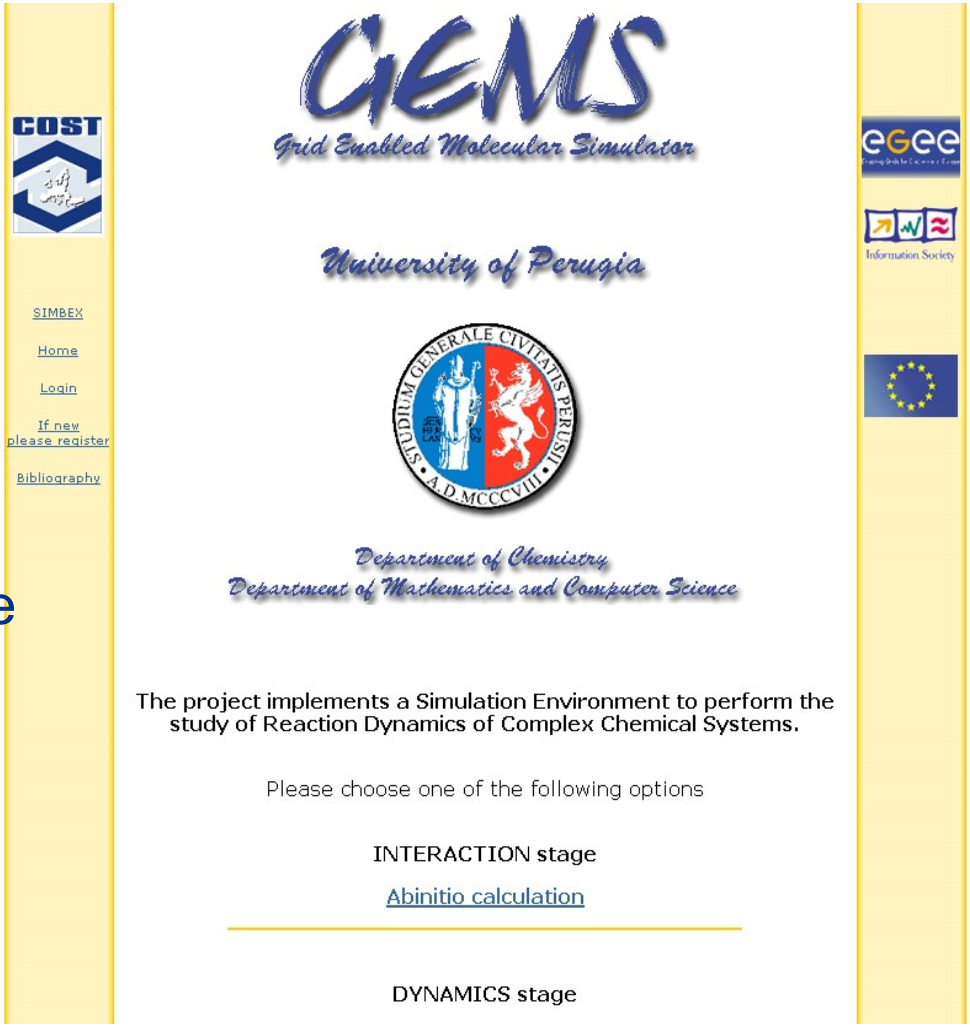



UDDI Server
UDDI Register
Files Servers of Molecular Structures
Search / Download mol. struct. files in/from Web sites
Administration of the molecular structure file collection
Web Service Client
WEB SERVICE
X3D Client
SOAP
Mol Struct Files Collection
WSDL
Web Service Server
WEB SERVICE
Mol Struct files to simulator input files Converter
simulator output files to Mol Struct Converter
Simulator
Simulator
Simulator
Client
Server
Simulator Server

WFM implemented on a web portal

- The current implementation of the Grid Enabled Molecular Simulator:
  - is based on a set of PHP scripts implemented on Apache server and MySQL RDBMS
  - Very efficient and easy to use
  - Requires inbound/outbound connectivity on ports 25000-25999* in the Grid site where the computation is performed and nodes accessible with public IP addresses
  - The implementation of the interface for new programs is relatively complicate
  - Limited to the specific context

*EGEE recommended ports



GEMS

Grid Enabled Molecular Simulator

University of Perugia

Department of Chemistry

Department of Mathematics and Computer Science

The project implements a Simulation Environment to perform the study of Reaction Dynamics of Complex Chemical Systems.

Please choose one of the following options

INTERACTION stage

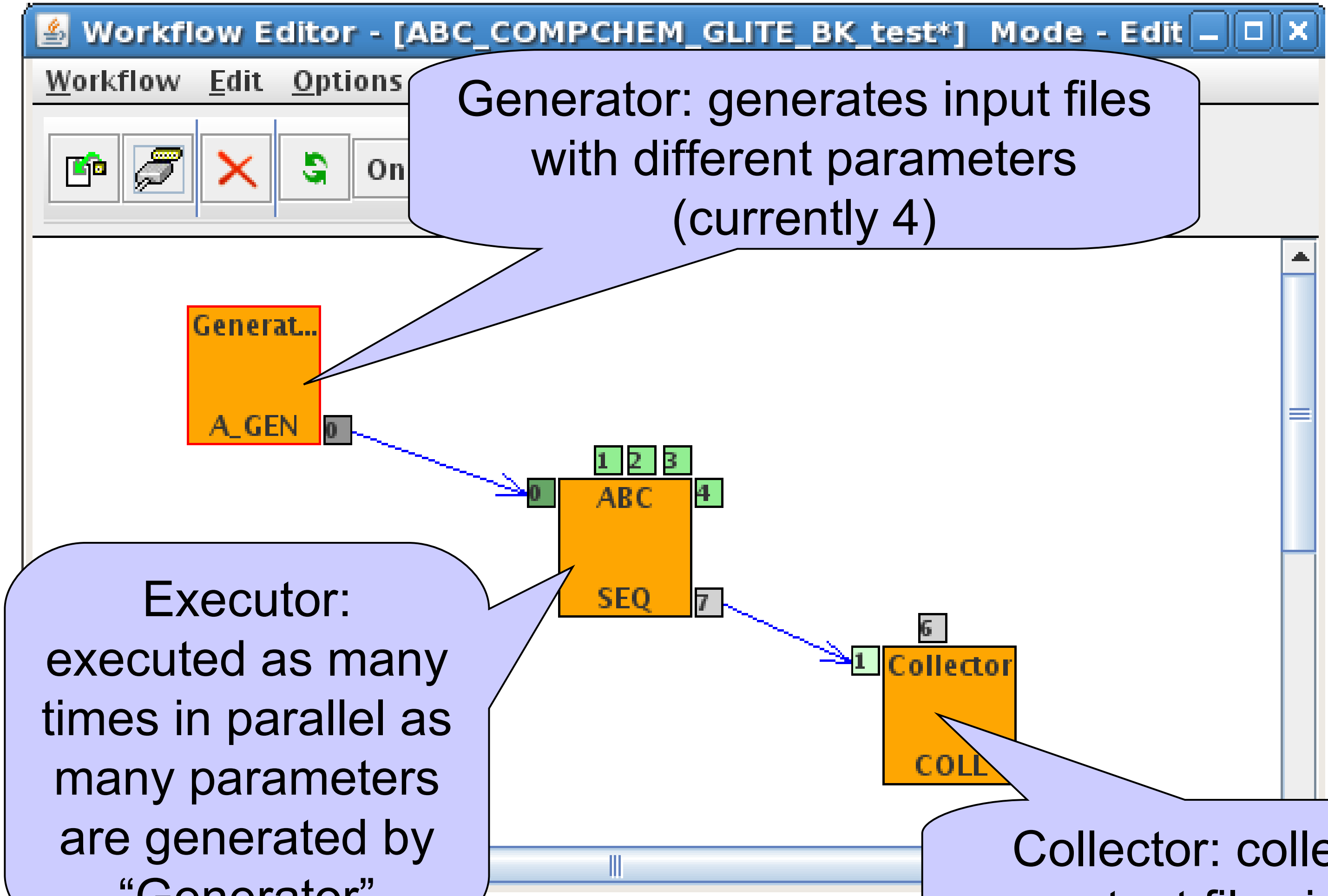
Abinitio calculation

DYNAMICS stage

WFM on P-Grade

- We ported one of the most recent codes developed for GEMS, ABC (time independent quantum reactive scattering code), on PGrade Portal 2.7.
- Using PGrade we are able to specify the WF related to ABC program, customizing the input data, the execution environment and the visualization of the results.
- The environment is really easy to implement and use!
- The porting has been carried out with the help of GASuC (MTA-SZTAKI) using Pgrade Grid Portal 2.7 installed on the CompChem UI.
- This effort is an outcome of Alessandro Costantini's Short Term Mission at CESGA, an activity of the COST D37 Action (collaboration of QDYN and ELAMS working groups).

ABC workflow



Generator: generates input files with different parameters (currently 4)

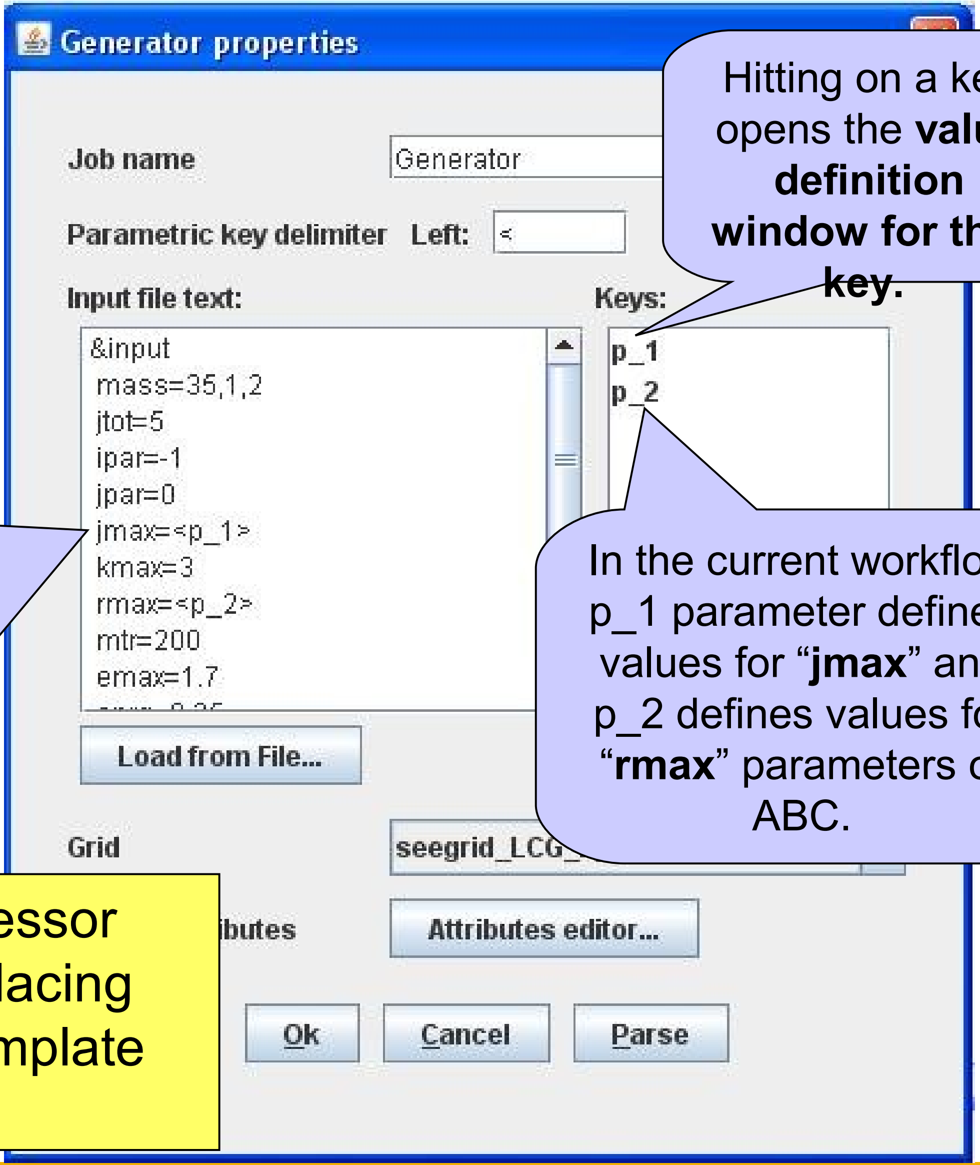
Executor: executed as many times in parallel as many parameters are generated by “Generator”

Collector: collects all output files into a single .tar file

ABC WF: modify input parameters

Template text with keys. Keys will be replaced with actual numbers by the Generator during the execution of the workflow.

All the possible combinations of the replaced template are written into separate files.



Generator properties

Job name: Generator

Parametric key delimiter Left: <

Input file text:

```
&input
 mass=35,1,2
 jtot=5
 ipar=-1
 jpar=0
 jmax=<p_1>
 kmax=3
 rmax=<p_2>
 mtr=200
 emax=1.7
```

Keys:
- p_1
- p_2

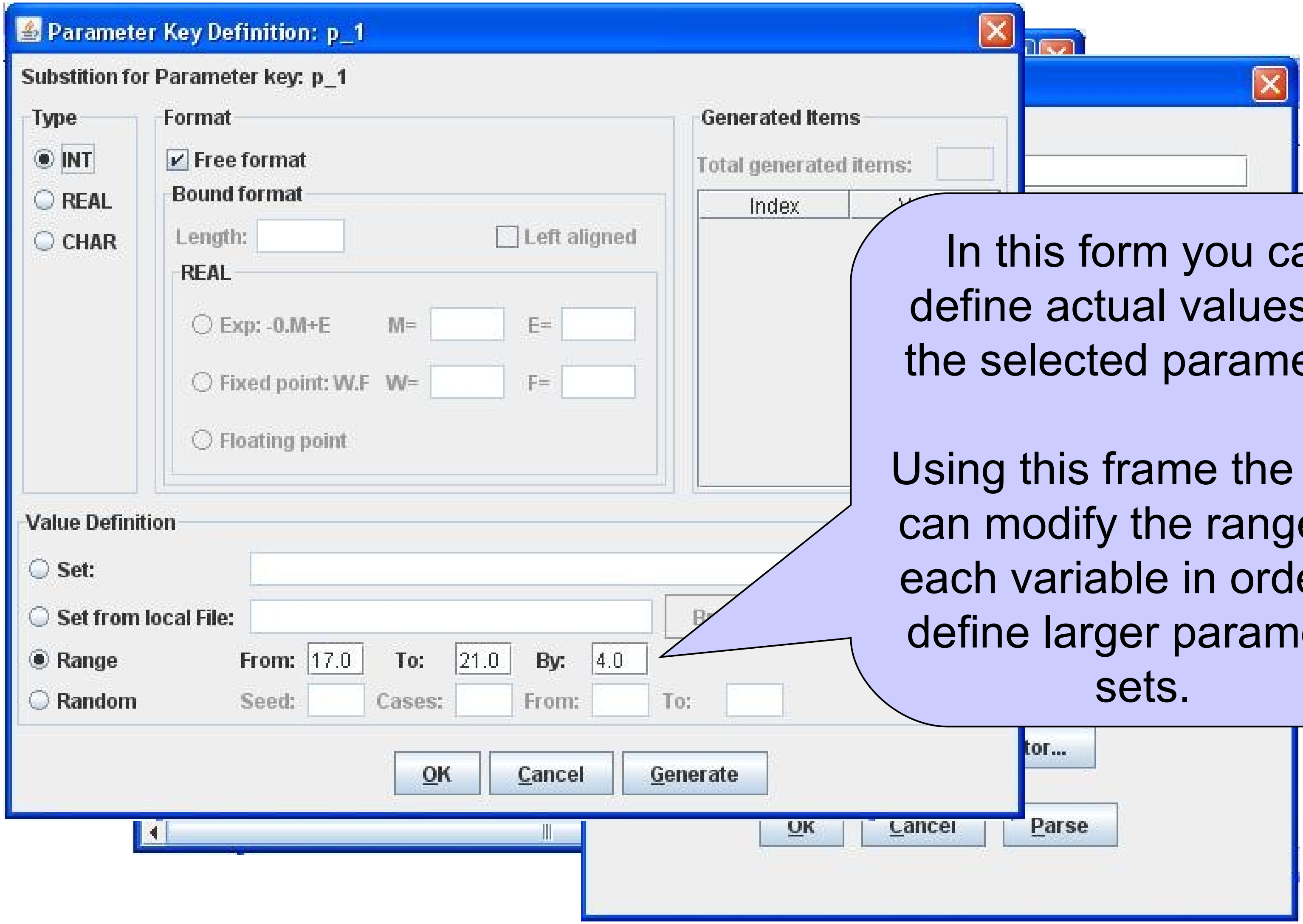
Grid: seegrid_LCG

Hitting on a key opens the **value definition window for that key.**

In the current workflow p_1 parameter defines values for “**jmax**” and p_2 defines values for “**rmax**” parameters of ABC.

Generator job is a macro processor that generates text files by replacing keys with actual values in a template which is defined by the user.

ABC WF: modify input parameters



Parameter Key Definition: p_1

Substition for Parameter key: p_1

Type
- ● INT
- ○ REAL
- ○ CHAR

Format
- ☑ Free format

Bound format
- Length:
- □ Left aligned

REAL
- ○ Exp: -0.M+E M= E=
- ○ Fixed point: W.F W= F=
- ○ Floating point

Generated Items
- Total generated items:

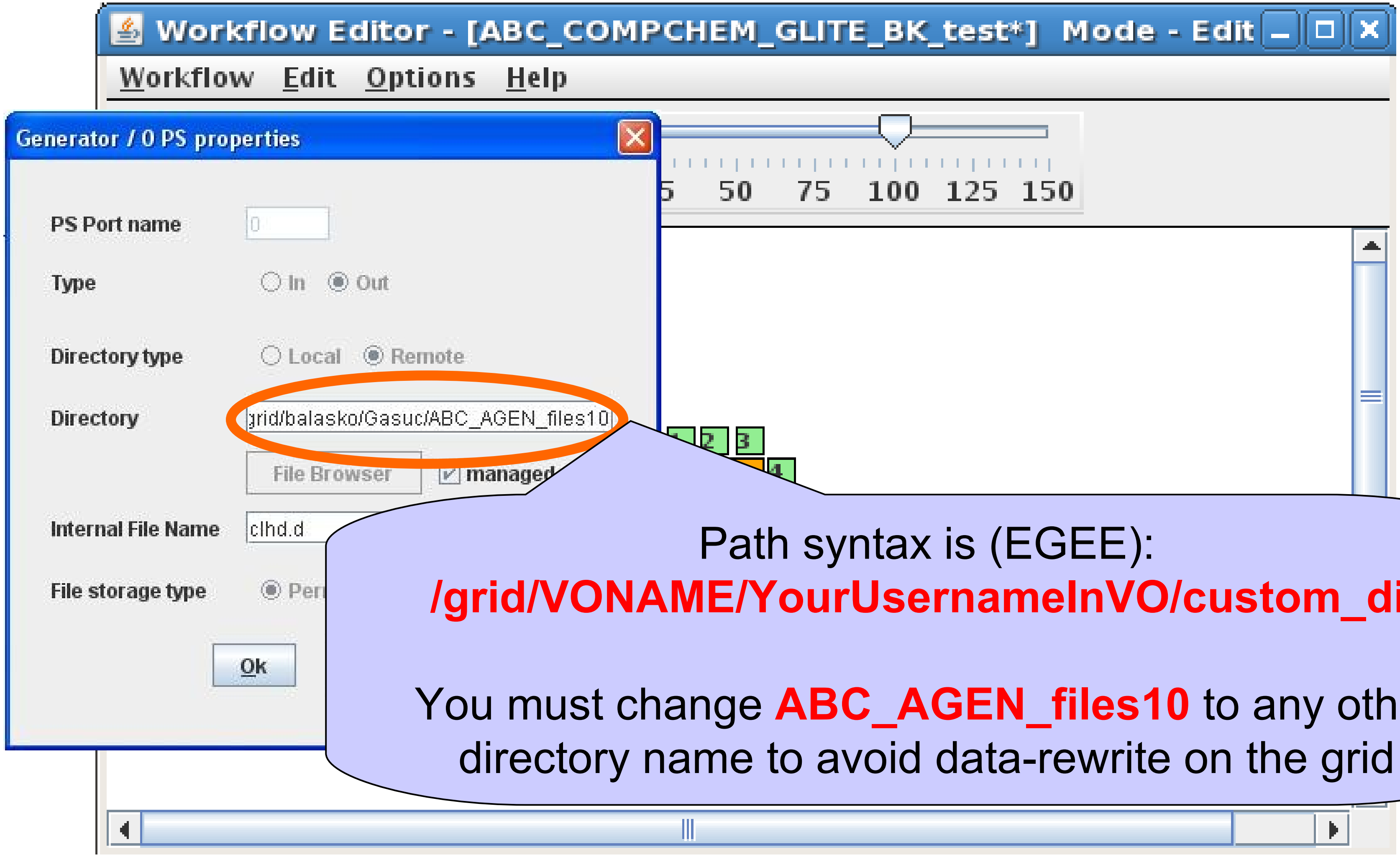
Index	V

Value Definition
- ○ Set:
- ○ Set from local File:
- ● Range From: 17.0 To: 21.0 By: 4.0
- ○ Random Seed: Cases: From: To:

In this form you can define actual values for the selected parameter.

Using this frame the user can modify the range for each variable in order to define larger parameter sets.

ABC WF: modify Grid parameters



Workflow Editor - [ABC_COMPCHEM_GLITE_BK_test*] Mode - Edit
Workflow Edit Options Help
50 75 100 125 150
1 2 3
4

Generator / 0 PS properties

PS Port name: 0

Type: ○ In ● Out

Directory type: ○ Local ● Remote

Directory: grid/balasko/Gasuc/ABC_AGEN_files10

File Browser ☑ managed

Internal File Name: clhd.d

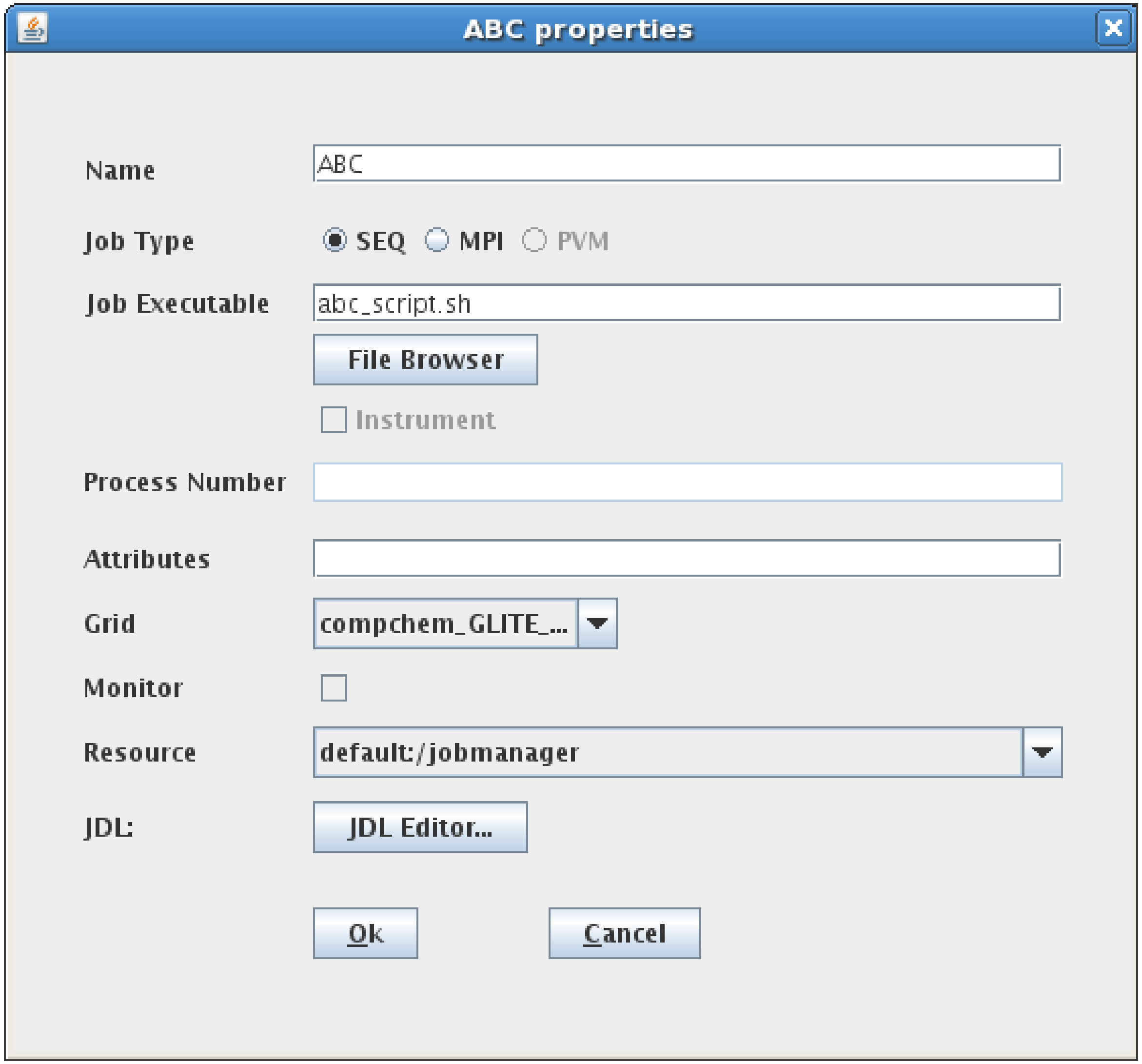
File storage type: ● Per

Ok

Path syntax is (EGEE):
/grid/VONAME/YourUsernameInVO/custom_dir

You must change **ABC_AGEN_files10** to any other directory name to avoid data-rewrite on the grid

ABC WF: modify Grid parameters



ABC properties

Name: ABC

Job Type: ● SEQ ○ MPI ○ PVM

Job Executable: abc_script.sh

□ Instrument

Process Number:

Attributes:

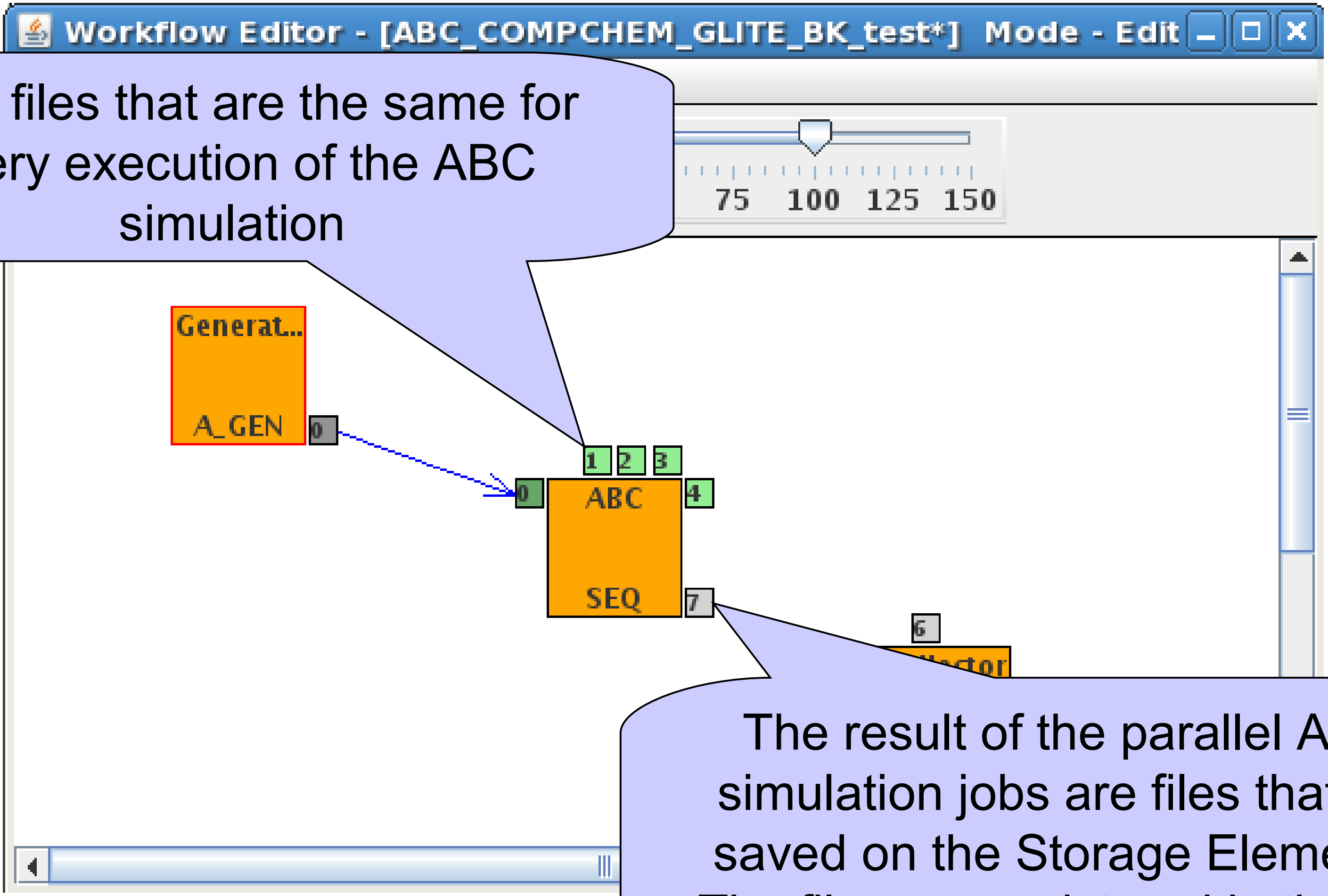
Grid: compchem_GLITE_...

Monitor: □

Resource: default:/jobmanager

JDL:

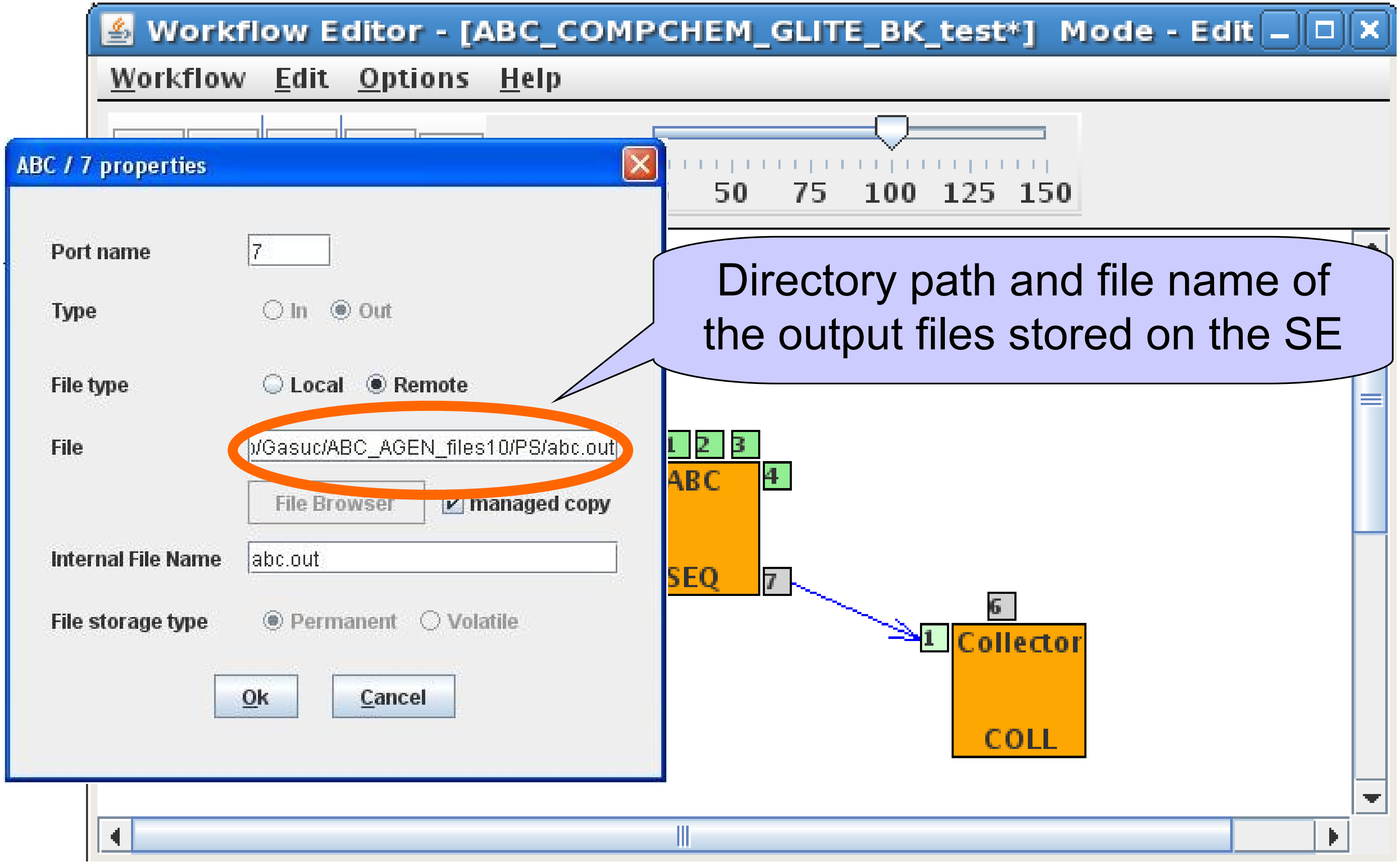
ABC WF: input and output files



Input files that are the same for every execution of the ABC simulation

The result of the parallel ABC simulation jobs are files that are saved on the Storage Elements. The files are registered in the File Catalog with Logical File Names.

ABC WF: modify Grid parameters



Port name: 7

Type: ○ In ● Out

File type: ○ Local ● Remote

File: /Gasuc/ABC_AGEN_files10/PS/abc.out

File Browser ☑ managed copy

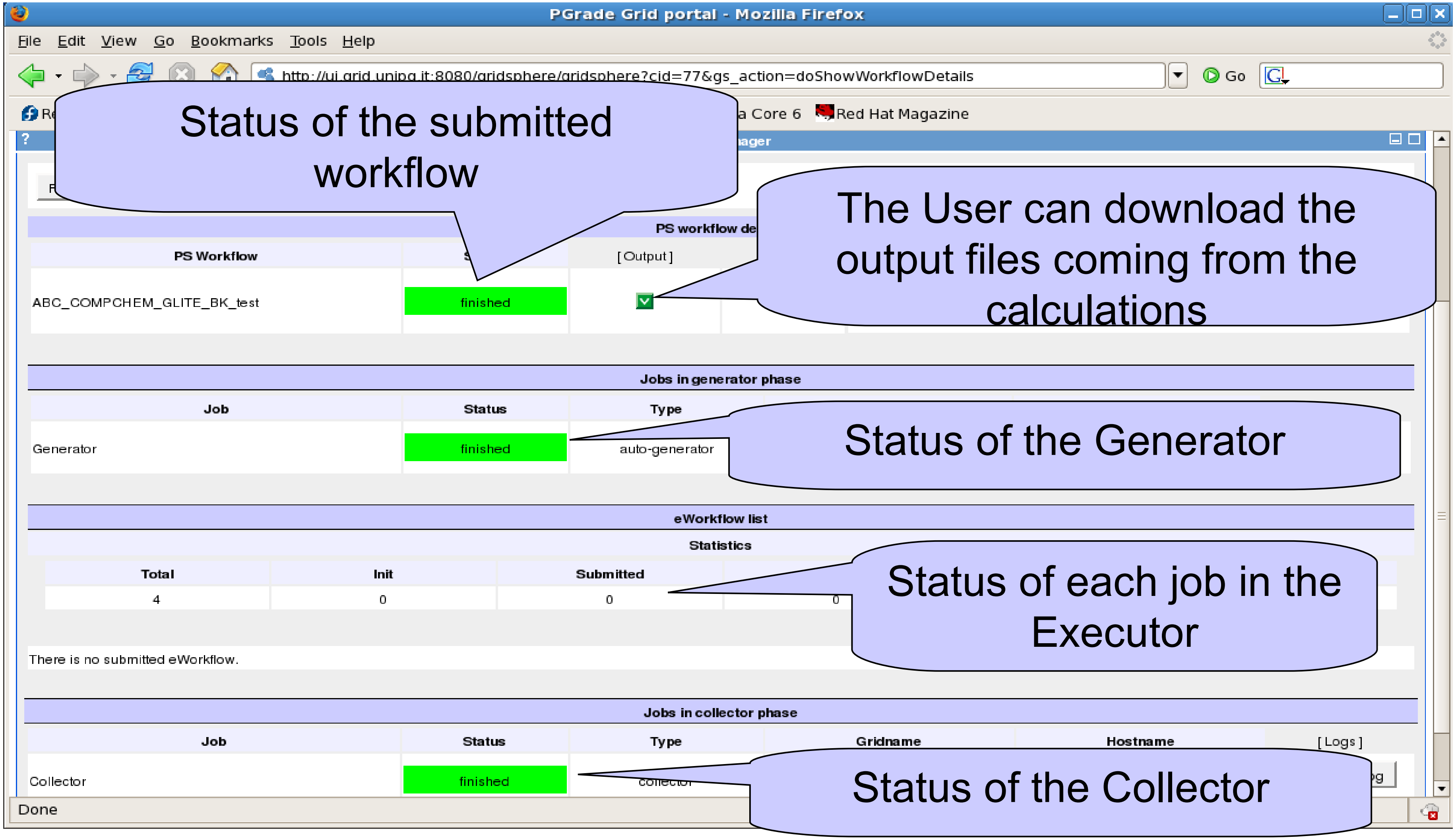
Internal File Name: abc.out

File storage type: ● Permanent ○ Volatile

Ok Cancel

Directory path and file name of the output files stored on the SE

ABC WF: job status


PGrade Grid portal - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Go
Re
a Core 6
Red Hat Magazine
?
ager
Done

http://ui.grid.unipg.it:8080/gridsphere/gridsphere?cid=77&gs_action=doShowWorkflowDetails

PS workflow de

PS Workflow	S	[Output]
ABC_COMPCHEM_GLITE_BK_test	finished	✓

Jobs in generator phase

Job	Status	Type
Generator	finished	auto-generator

eWorkflow list

Statistics

Total	Init	Submitted	
4	0	0	0

There is no submitted eWorkflow.

Jobs in collector phase

Job	Status	Type	Gridname	Hostname	[Logs]
Collector	finished	collector			g

Status of the submitted workflow

The User can download the output files coming from the calculations

Status of the Generator

Status of each job in the Executor

Status of the Collector

Future developments

- The porting of ABC program in PGrade has been really successful: we will extend this approach to other computational procedures, in order to simplify the user's production activity on EGEE.
- The computational procedures related to GEMS WF can be easily ported in PGrade adopting the same approach.
- The work done for restructuring ABC, may enable the inclusion of GEMS components in various open surce WFM Systems, like BPEL.
- We acknowledge the excellent support from GASuC group, MTA-SZTAKI, Budapest (HU).
- This work has been made possible thanks to the collaboration in the COST CMST D37 Action GridChem