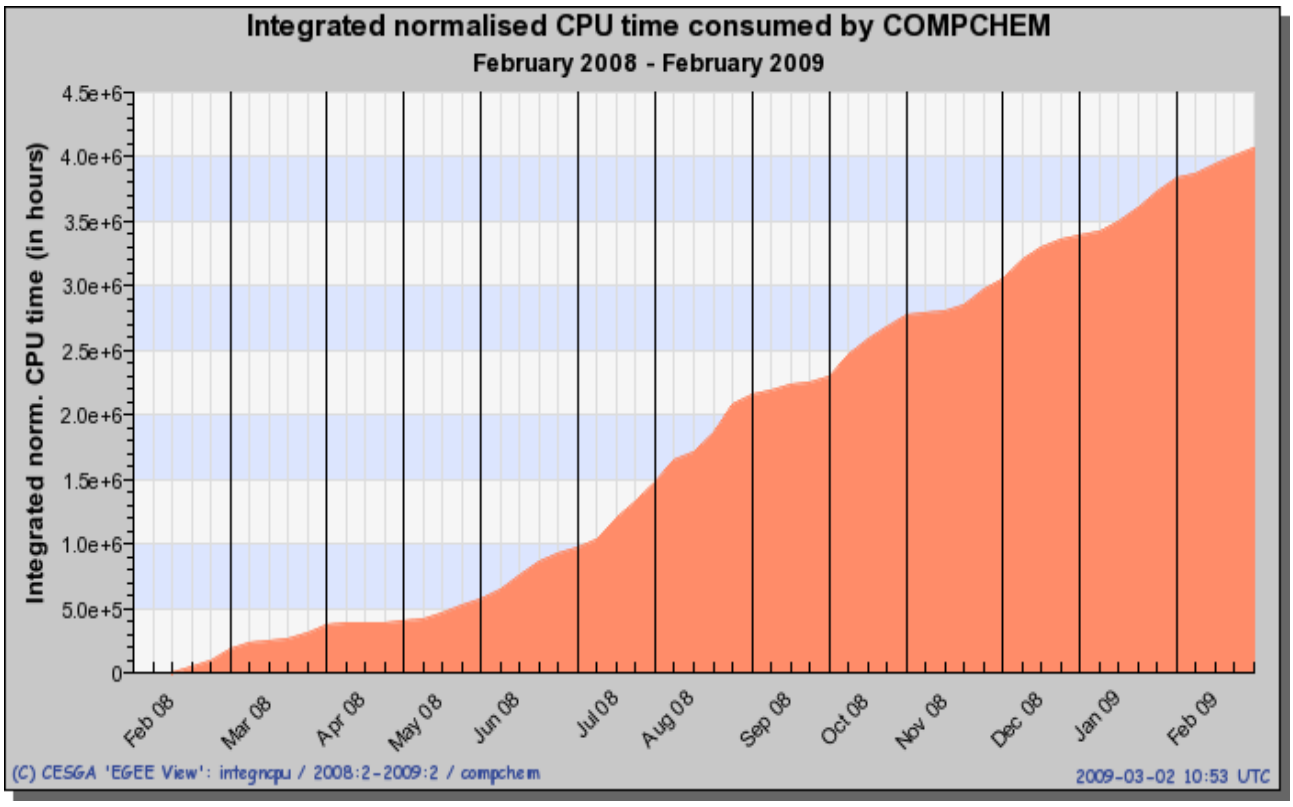
A NEW BOTTOM UP APPROACH TO THE CMMST-VRE

A. Laganà, C. Manuali, L. Pacifici and S. Rampino, Dipartimento di Chimica, Biologia e Biotecnologie, Università di Perugia, IT

A. Costantini, CNAF-INFN, Bologna, IT

1 - INTRODUCTION

From the convergent activities of COST D23 (D23 | Metachem | 17 October 2000 - 18 July 2005) and D37 (D37 | Grid Computing in Chemistry: GRIDCHEM | 06 July 2006 - 05 July 2010) Actions (made by a set of networked laboratories operating in: Computational photochemistry and photobiology (coordinated by Hans Lischka, University of Vienna, AT), Quantum dynamics engines for grid empowered molecular simulators (coordinated by Antonio Laganà, University of Perugia, IT), E-science for learning approaches in molecular science (coordinated by Osvaldo Gervasi, University of Perugia, IT), Code interoperability in computational chemistry (coordinated by Elda Rossi, CINECA, IT), Computational chemistry workflows and data management (coordinated by Thomas Steinke, Zuse Institute Berlin, DE)) [1] and EGEE III (that created the largest collaborative infrastructure in the world for e-science involving high energy physics, astronomy, astrophysics, computational chemistry, earth science, fusion and computer science [2]) the Computational Chemistry Virtual Organization (VO) COMPCHEM (https://www3.compchem.unipg.it/compchem/) was established. COMPCHEM, though formed by a limited number of laboratories of the Chemistry, Molecular & Materials Sciences and Technologies (CMMST) community became immediately the European molecular science VO most active in EGEE III (see Fig. 1)



ISSN: 2279-8773

Figure 1 – Integrated normalized cpu hours utilized by COMPCHEM in the first year of activity in EGEE III (EGEE Accounting Portal at the Centro de Supercomputación de Galicia, http://www3.egee.cesga.es/gridsite/accounting/CESGA/egee_view.html).
As a matter of fact, the Figure shows that the amount of integrated normalized cpu hours of the Grid infrastructure utilized by COMPCHEM in the first year raised to about 5 millions (during the same period COMPCHEM utilized 86% of the overall cpu time utilized by the three chemistry based VOs with the other two being GAUSSIAN (2%) and CHEM.VO. IBERGRID (12%)). As shown in Fig. 2, such result was achieved by utilizing almost exclusively the compute resources of the National Grid Infrastructures (NGI) of Italy, France, Greece, Iberia (Spain + Portugal) and Poland.

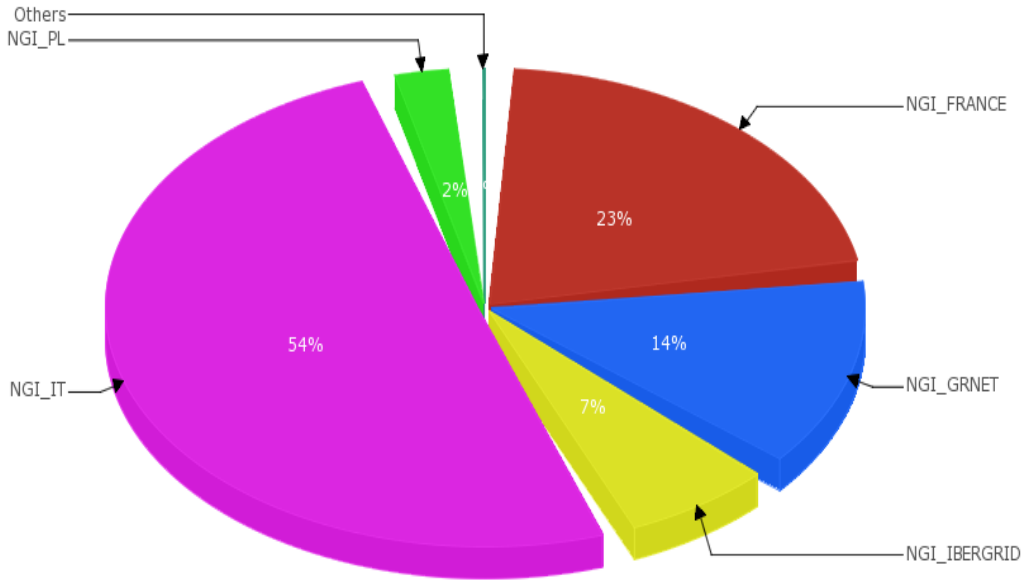


Figure 2 – As in Fig. 1 for the percentage of the overall Normalised CPU time (kSI2K) consumed by the three (COMPCHEM, GAUSSIAN and CHEM.VO.IBERGRID.eu) VOs on the grid platforms of the various NGIs in the period ranging from January to December 2013.

During the next EGI.eu project EGI-Inspire (https://www.egi.eu/about/egi-inspire/) the COMPCHEM yearly utilization of cpu time increased by a factor of 5 [3]. Moreover, in order to extend the initiative to a larger share of the community, a Virtual Team (VT) was formed (https://wiki.egi.eu/wiki/VT_Towards_a_CMMST_VRC) to the end of gathering together the CMMST VOs GAUSSIAN, CHEM.VO.IBERGRID and COMPCHEM with minor participation of TRGRID (a Turkish VO) in a single Virtual Research Community (VRC). The Virtual Team finally managed to establish in May 2014 the CMMST VRC with the purpose of generalizing the use of the so called Grid Empowered Molecular Simulator (GEMS) [3,4] to various innovative applications on Distributed Computing Infrastructures (DCI). Due to the associated high demand of High Performance Computing (HPC) specific effort was paid to connect experimentally the Grid node of Perugia to the supercomputer networks of PRACE and XSEDE.

In order to provide the CMMST VRC with a more solid ground a Virtual Research Environment (VRE) proposal was submitted to the H2020-EINFRA-2015-1 call of January

2015 [5]. The CMMST VRE proposal was aimed at establishing a specific environment in which HW and SW resources, tools and applications for research, education, innovation and market oriented activities specifically fitting the nature of the CMMST user communities were aggregated to assemble several independent academic research laboratories sharing common scientific goals without being necessarily grafted either on a common distributed ICT platform, or on a common operating or experimental frame while being potentially suited to benefit from the highly collaborative environment of a VRE.
The specific goals of the CMMST VRE were those of:
1- providing its user communities with efficient access, selection and utilization of high throughput and high performance computing resources,
2-supplying appropriate software tools and portals enabling a user friendly inter community crossed composition and development of higher level of complexity new applications and data sets,
3-training the users to take full advantage of distributed data management and computing,
4-utilizing quality metrics for both resource usage and user behaviour to ground synergistic (collaborative/competitive) networked computing in research, innovation, education and training,
5-providing a common frame for supporting shared challenging endeavours of different communities,
6-deploying a cloud based environment for user support by integrating an e-assesment platform within the production environment,
7-building a service oriented system of credits reward & redemption in terms of allocation of e-infrastructure resources (computing and financial) able to support a business model enhancing sustainability.

2 – FROM THE OLD TO A NEW PROPOSAL

The mentioned CMMST-VRE proposal gathering together about 40 laboratories was evaluated as follows (the text of the proposal is given in ref. [5]):

EXCELLENCE Score 3.50
The proposal aims at a VRE for molecular and material sciences, which fits in the scope of the work programme; the objectives are clearly stated and are strongly related to the call.
The scientific approach is credibly explained and justified.
It is not clear how Networking activities will foster a culture of cooperation. For instance, training and workshop events mentioned in WP2 are more intended for outreach and dissemination.
The involvement of entities such as EGI and PRACE will provide suitable support and expertise to users to help them conduct excellent research.
The use of a credit based economy to encourage sharing and cooperation is a useful and innovative concept as this stimulates users to contribute data to the VRE rather than just make use of it.
The concept is sound for the development of VRE in chemistry/material sciences. However the proposed work focuses mainly on material science, while the trans-disciplinary aspects are somewhat limited.
The proposed development could support novel atomic and molecular dynamics simulations, as well as multi-scale simulations relevant to combustion and materials applications. The proposed JRA are excellent. The proposal does not consider activities to experimental data, limiting the overall outcome.

The propose work extends the current state-of-the-art and the proposed science and technology
The research for handling metadata and ontologies is not sufficiently described.

IMPACT Score 3.50
The proposed project will support synergies and collaboration between research teams in the fields of computational chemistry and materials, and will substantially promote innovation and the creation of new knowledge, as well as the use of standards.
Foreseen innovation from the project results are formulated. The proposal does not have a clear plan for commercial exploitation e.g. possible patents and their market implementation are not envisaged further.
It does not describe how the proposed work will strengthen competitiveness or support the growth of companies, beyond those immediately involved
KPIs are clearly identified.
The communication of the project is good and an effective plan for data management is described in WP5. However the ownership of the data has not been sufficiently addressed.

QUALITY AND EFFICIENCY OF THE IMPLEMENTATION Score 3.50
The work plan is adequate and contains valuable elements of scientific nature.
The structure of the work plan is simple and tasks are feasible.
The allocation of tasks and resources is appropriate: however WP7 is overloaded with activities making it difficult to manage.
The quality and relevant experience to the participants in order to comply with the overall project ambitions is good and is well supported by an experienced coordinator.
The participants collectively possess the skill and the experience to carry out the proposed work and complement each other well.
The management structure is appropriate for the project size, including the External Advisory Board
The risk management approach is generally sufficient; however, the technical risks are underestimated and the proposed mitigation measures are not adequate.
The proposal lacks a coherent plan for innovation management.

From the above quoted evaluation we have extracted the following suggestions for an improvement of the new proposal:

Criterion1-Execellence
1.1 It is not convincing how Networking Activities will foster a culture of cooperation. For instance training and workshop events mentioned inWP2 are more intended for outreach and dissemination.
WP2 (NA) and WP3 (SA) are designed to work together by sharing effort and competences. Such cooperation should be better emphasized and more strongly projected beyond dissemination and training.

1.2 The concept is sound for the development of VRE in chemistry/material sciences. […]
A key feature of the project is to span over the various steps of GEMS from ab-initio/fundamental research to materials/bioapplications. We should better emphasize the advantage of implementing GEMS as a service for innovative applications leveraging on the synergistic nature of Distributed Compute Infrastructure DCI. Therefore, for the WP devoted to data management, more emphasis should be given about the data reuse in

ISSN: 2279-8773

research and e-learning.

1.3 The use of a credit-based economy to encourage sharing and co-operation […]
This credit-based economy should be further emphasized as vital for enhancing cooperation.

Criterion2-Impact
2.1 The proposal does not have a clear plan for commercial exploitation, […]
This is a further step forward in the use of the credit-based economy that should be headed also towards implementing business-related activities.

2.2 […] However the ownership of the data has not been sufficiently addressed.
This was clearly a missing item in the previous proposal. We need to define ownership and licenses for data and intellectual properties!

Criterion3-Quality and efficiency of the implementation
3.1 The risk management approach is generally sufficient, however the technical risks are underestimated and the proposed mitigation measures are not adequate.
The previous proposal was underestimating the risks associated with the project activities and tasks. A true expert in such field should be involved in order to be competitive.

3.2 The proposal lacks a coherent plan for innovation management.
This sentence has to be understood (like the ones above) in terms of a need for more business oriented behaviors and targets.

In order to better take into account the above suggestions and properly implement them into a new proposal we decided to meet the potential partners in a meeting to be held in Fulda (DE) during the EUCO CC 10 (http://www.euco-cc-2015.org/) conference on Tuesday September 1. On that day a session on Computational Environments is going to be held in the morning. This will offer us an ideal ground for a bottom up elaboration of a new synergistic scheme for the CMMST-VRE and its implementation on the to be defined DCI of the project.

3 – SUGGESTED SCHEME OF THE PROPOSAL

Considering that in the proposed synergistic model the members of the CMMST community will provide as-a-service support for the use of their data and programs to the other members in return for credits redeemable as a better share of the community resources (compute time, hardware and software tools and interfaces, links to workflows for multi-scale applications, research funds, collaborative projects, etc.), we expect that also the proposal will be also assembled collaboratively. For this reason what we propose here is only a basic scheme and all suggestions by potential partners will be considered for the assemblage of the proposal.

MANAGEMENT (coordination, administration, external advisors, risk assessment, business plan, ..)

DISTRIBUTED COMPUTING TECHNOLOGIES (compute resources, middleware, search for existing utilities, tools, ..)

ISSN: 2279-8773

MULTISCALE SIMULATIONS OF GAS DYNAMICS (specialization of GEMS for complex gas phase systems dynamics)

MOLECULAR PROCESSES ON NANOSTRUCTURES (specialization of GEMS for chemical processes on nanostructures)

CHEMODYNAMICS (specialization of GEMS for pharmaceutical and biological processes)

TRAINING AND KNOWLEDGE MANAGEMENT (Distributed data formats, knowledge on distributed repositories, e-tools for training)

The above schemes prompts the following questions

- What is the driving force of the proposal
- Who and how will provide human effort and compute resources to the community
- Which Middleware services will be adopted for an optimal reuse of data, tools and services
- How evaluation/crediting mechanisms will be implemented for establishing a business model
- Which applications will be adopted for aggregating partners and achieving the objectives of the considered H2020 call
- What type of recruitment and training of new members will be chosen for consolidating the VRC.

More in detail:

a) GEMS (Grid Empowered Molecular Simulator): the competitive-collaborative driving force of the proposal

The central common service of the CMMST-VRE (to be used as a support to all applications of the project) is GEMS whose synthetic scheme is given in Fig. 3

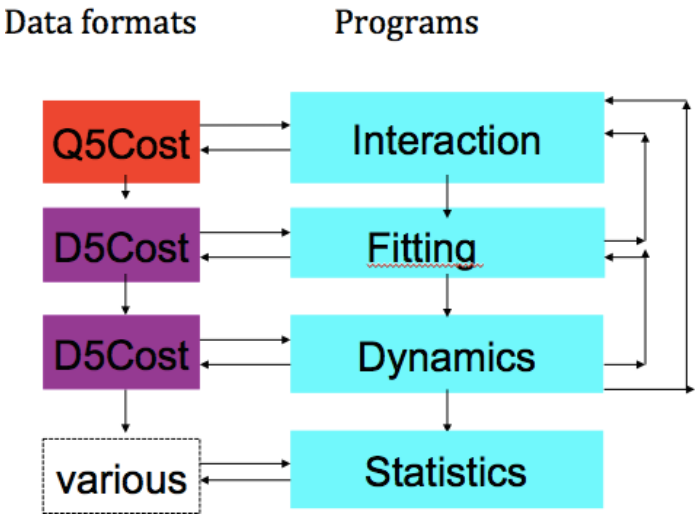


Figure 3 – The synthetic scheme of GEMS and related data formats.

The scheme is articulated in a block (lhs column) of data formats specifications and a block of programs (rhs column). In the data formats block the de facto standard formats of

Q5COST (single molecular geometry), D5COST (single layer of potential energy values for different geometries plus various other property specific formats. In the programs block, whose detailed flowchart is given in Fig. 4, the INTERACTION, FITTING, DYNAMICS and STATISTICS sections are exploded. In the section INTERACTION the flowchart is channelled either to the use of existing PES (if available) or to the use or production ab initio calculations (if either available or feasible). If ab initio calculations are to be run the proper procedure is adopted (SUPSIM in the figure) and then already available and the afresh computed values are fitted. Otherwise Force Field solutions are adopted. Once a PES is assembled (or imported) detailed exact or reduced dimensionality quantum, semi- or quasi-classical dynamical calculations can be performed, and detailed **S** or **P** matrix elements evaluated. Out of such detailed data cross sections, rate coefficients can be calculated and used for higher level multi-scale studies.

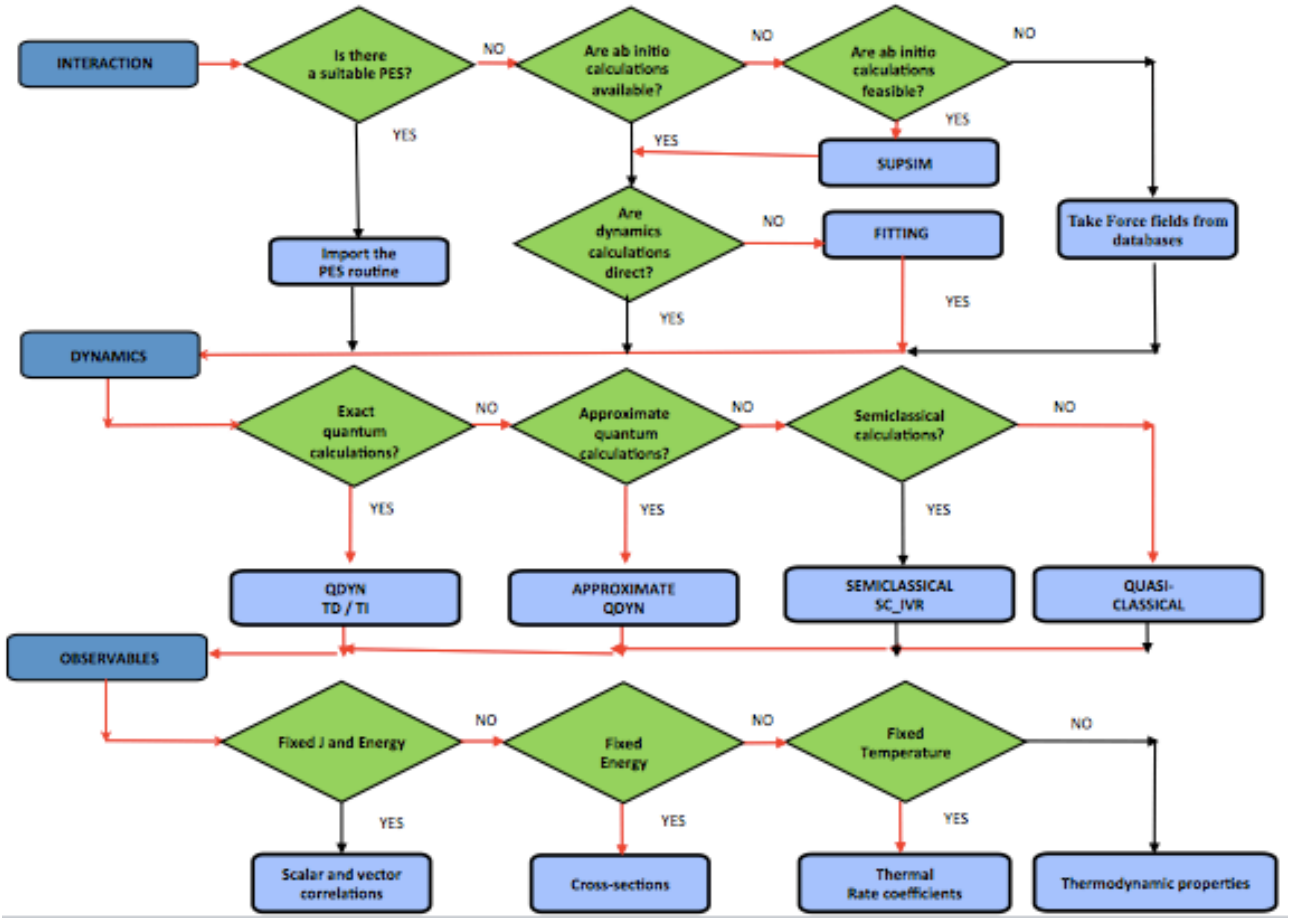


Figure 4 – The general flowchart of GEMS in its present version.

b) who and how will provide human resources

The synergistic model of the proposed bottom up project will encourage the community members (especially those of the partner VOs) to provide services to the others leveraging on the computational engines of GEMS and its components. For this is important that the community members can grade its involvement and scale up in level as detailed in the table below:

Membership Level	Short Description
1. User	*Passive*: Runs programs implemented by other VO members.
	Active: Implements at least one program for personal usage.
2. Software Provider	*Passive*: Implements at least one program for use by other members.
	Active: Manages at least one implemented program for collaborative usage.
3. Infrastructure Provider	*Passive*: Confers to the infrastructure at least a small cluster of processors.
	Active: Contributes to deploy and manage the infrastructure.
4. Manager (Stakeholder)	Takes part to the development and the management of the VO.

Table 1 – Levels of membership in the proposed CMMST community.

c) who and how will provide compute resources and middleware services

On the ground of the past experience of the member VOs, two main types of compute resources will be considered for the CMMST-VRE: High Performance Computing (HPC) of large scale facilities and High Throughput Computing (HTC) of distributed networked computers. For the former the proposed approach is to apply for cumulative community grants with supercomputer networks (talks are on-going with some PRACE (EU) and XSEDE (US) partners). For the latter the proposed approach is to integrate in the proposal already structured DCI resource providers from different platforms (EGI.eu NGIs, research centres facilities, individual users members of the project, etc.) rewarded through quality evaluation of the services provided on the project funds.
As to the middleware, an agreed common environment will be adopted by involving as members of the project the related providers, leveraging on existing centres of competence and relying on providers already equipped with the necessary expertise and infrastructures.

d) how services provided and user behaviours will be monitored

Relying on the past experience of the member VOs, the use of the resources can be managed, monitored, and turned to account using the **GriF** tool (see Fig. 3).

ISSN: 2279-8773

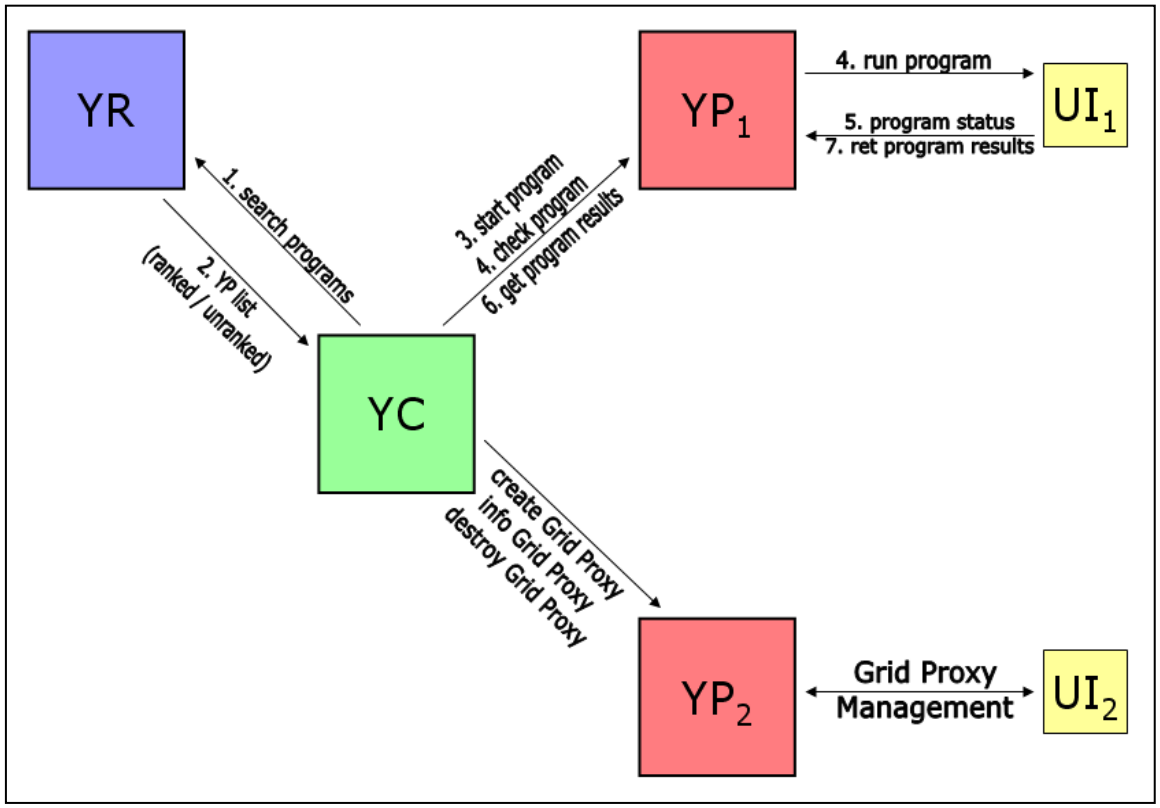


Figure 5 – Sketch of the GriF flowchart and of its servers and User interfaces.

GRIF is a Framework made of two Java servers (YC and YR the Consumer and the Registry servers) and one Java client (YP the Provider). The entry-points to the computational platforms are the User Interfaces (UI) which are able to capture, out of the data supplied by the monitoring sensors of the DCI, the information relevant to properly manage the computational applications of interest and articulate them in sequential, concurrent or alternative quality paths by adopting a Service Oriented Architecture (SOA) and Web Services. This allows at the same time the guided search of the compute resources on the DCI and the Evaluation of the Quality of the Users (QoU). The computational services provided, are analysed and used to compose the submission, the monitoring, and the results recollection of molecular science simulations.

e) how credit mechanisms will be implemented

Credits will be assigned using a community agreed **METRICS** for evaluating QoS (Quality of Service), QoU (Quality of User) and RR (Resources Ranking) of the activities offered as-a-service using the GriF (Grid Framework) and GCreS (Grid Credit System) tools developed by the CMMST Virtual Research Community (VRC) and its Virtual Organizations (VO)s. A scheme of this type has been already implemented by the COMPCHEM VO and can be adopted by CMMST and further adapted to the specificities of the VRC. Relying on the past experience of the member VOs, the use of the Grid Credit System **GCreS** tool to reward both the QOU (the Quality of the users exploiting the compute resources and applications) and the QoS (the Quality of the services provided by the users to the other members of the community) it will be possible to assign to the users a congruent amount of Credits (according to agreed mechanisms). Such credits, redeemable in terms of a preferential utilization of the Community resources (selection of

compute systems, DCI services, low and high-level capabilities, memory size, cpu/wall time, storage capacity plus financial resources) will not only foster collaboration among the members of the community but will also entitle related researchers to participate to the multi-competence teams which apply for the most challenging bids. This will result in an enhancement of the competition among different teams (the so called competitive collaboration).

f) gas dynamics applications

Gas dynamics applications impact several innovative technologies [6] based on the modelling of rarefied gas flows in which the mean free path of a body is of the same order (or greater) than a representative physical length scale often expressed in terms of the Knudsen number (Kn). Usually these processes (of interest for plasmas, space shuttle re-entry aerodynamics, the modeling of micro-electronic-mechanical systems, etc.) are simulated using phenomenologically modelled Direct Simulation Monte Carlo (DSMC) techniques [7] in which the Boltzmann equation for finite Knudsen number fluid flows is solved by moving colliding bodies through a realistic simulation of the physical space that is directly coupled to physical time. In order to obtain higher accuracy inter-body and body-surface treatments GEMS services will be used to the end of massively producing high level of theory data and storing them in a distributed repository. This kind of ab initio-last mile approach has been already proposed for the rationalization of Crossed Molecular Beams (CMB) experiments of colliding OH and CO [8] by properly defining the initial conditions, the characteristics of the experimental apparatus, the nature and the properties of the ensemble of bodies considered.

g) molecular processes on nanostructure applications

The use of GEMS services finds also appropriate application to the simulation of the formation of innovative condensed phase materials. The major part of this activity is achieved by enabling accessibility of community members to the applications of an advanced material laboratory for materials' design. This is accomplished by adapting the interfaces and workflows of some already operating procedures at the material laboratory considered for partnership and by activating the parallel approaches of related high performance computing (HPC) platform. This will allow the assemblage of applications using the running GEMS version and various other specialized popular packages including the parallelized plane wave/pseudopotential formulation of the interaction. Such a synergistic use of GEMS will enable accurate materials modelling at the nano-scale and the study of related atomistic and molecular simulations of solid and liquid state, of molecular and biological systems in order to create a large shared open data basis.

h) chemodynamics

Large use is made at present of statistics for determining the structure of large molecular systems and related pharmaceutical and biological activity. Using GEMS it is possible to ground such endeavour on workflows based on high level ab initio calculations, quantum and classical dynamics. This activity consists of porting on the e-infrastructure a set of codes exploiting the features of distributed environments for pharmacological and medicinal chemistry applications. This will significantly enhance the possibility for feeding the related databases and producing information on new safer drugs production,

crystallographic breakthroughs enabling receptor function and drug design studies with easy to use data visualisation tools and extension of the current residue diagrams and phylogenetic trees. This will include also accurate multiscale calculations of kinetic parameters and elucidating plausible catalytic mechanism(s) of enzyme processes of interest. The procedure will facilitate the evaluation of related effects, including inhibition, point mutations, protonation states, external stimulus and tunnelling and support an improved understanding of enzyme action.

i) training and knowledge management

This activity consists of developing modern tools for training members of the project and disseminating its activities. The methodologies adopted are those of the European Chemistry Thematic Network ECTN) that include a description of the General Chemistry syllabus based on a Europe-wide analysis of the curricula for Higher Education and ECTN Eurolabel standards and the adoption of the learning ability using its EChemTest self evaluation sessions. On this ground porting on distributed computing environments of a set of tools and procedures developed for the handling of Molecular and materials science knowledge both under the form of Learning objects and under the form of self evaluation sessions will be made. This implies the development of specific methodologies and taxonomies to handle the project's contents as well as the dissemination of the related best practices on a European scale.

References

1] COST Action D37 "Grid Computing in Chemistry: GRIDCHEM, Final scientific Report, Dimensione Grafica Editrice, Spello, Italy, COST Office, 2011, A. Laganà and H.P. Luethi Ed. ISBN 978-88-97228-03-5

2] http://eu-egee-org.web.cern.ch/eu-egee-org/index.html; http://eu-egee-org.web.cern.ch/eu-egee-org/fileadmin/documents/publications/EGEEIII_Publishable_summary.pdf

3] Laganà A., Manuali C. and Costantini A. (2013) Grid Computing in Computational Chemistry. In: Reedijk, J. (Ed.) Elsevier Reference Module in Chemistry, Molecular Sciences and Chemical Engineering. Waltham, MA: Elsevier. 30-Jun-14 doi:[Doi] 10.1016/B978-0-12-409547-2.10933-3.

4] S. Rampino, N. Faginas Lago, A. Laganà, F. Huarte-Larrañaga, An extension of the Grid empowered Molecular Simulator GEMS to quantum reactive scattering J. Comput. Chem. 33, 708–714 (2012).

5] A. Laganà, Research and Innovation actions. Chemistry, Molecular & Materials Sciences and Technologies Virtual Research Environment (CMMST-VRE), VIRT&L-COMM.6.2014.1

6] Molecular Physics and Hypersonic Flows
M.Capitelli Ed., Kluver Academic Publishers, Norwell, M.A., 1996

7] S. Dietrich, I. Boyd, Scalar and parallel optimized implementation of the direct Monte Carlo method, J. Comput. Physics 126, 328-342 (1996)

[8] A. Lagan\`a , E. Garcia; A. Paladini, P. Casavecchia, N. Balucani, The last mile of molecular reaction dynamics virtual experiments: the case of the OH(N=1-10)+CO(j=0-3) reaction, Faraday Discussion of Chem. Soc. 157, 415 - 436 (2012)

ISSN: 2279-8773