## FOREWORDS: A MANIFESTO FOR THE COMPUTATIONAL CHEMISTRY COMMUNITY

As Editor of the VIRT&L-COMM magazine, that has achieved in its first year of life a stable stream of production, it is my pleasure to pinpoint that the issue of virtual research communities (VRC) is rapidly developing and that VRCs are deepening their roots into science. This is largely dependent on the European effort to innovate the way researchers exploit the progress of distributed computing platforms that has led to the assemble of High Throughput Computing (HTC) and High Performance Computing (HPC) networks (typically represented by the European Grid Infrastructure (EGI, www.egi.eu) and Partnership for Advanced Computing in Europe (PRACE, www.prace-ri.eu) organizations). In particular, the present issue of VIRT&L-COMM hosts contributions illustrating the lines along which the Chemistry, Molecular and Materials Science and Technologies (CMMST) computational community is committed to advance in the field of research and research based education. Similar efforts have been already undertaken (or are being undertaken) by other scientific communities at both national and international level. For this reason the present forewords are an attempt to outline a kind of manifesto of the CMMST Virtual Community.

PRESENT CMMST COMPUTATIONAL APPLICATIONS

CMMST computing activities cover a wide range of theoretical and application areas. A list of the most popular CMMST programs (mainly aimed at determining molecular structures) is given in http://en.wikipedia.org/wiki/List_of_quantum_chemistry_and_solid_state_physics_software. Most of them are based on Hartree-Fock (HF) and some post Hartree-Fock methods. Density functional theory (DFT), Molecular Mechanics (MM) and semi-empirical quantum chemistry are the other methods involved. Such programs are either open source or commercial software. Most of them are large packages, often embodying different techniques and computational approaches, and have developed over many years. On top of this layer of theoretical chemistry software, several CMMST computational procedures have been built to the end of investigating light and matter interaction with matter and are mainly aimed at rationalizing experimental measurements (spectroscopy, crystallography, NMR, reactive and non reactive beam scattering, macromolecules properties, biological process, etc.). At higher level of complexity are various multi-scale applications designed to investigate macroscopic observables under realistic conditions. These applications superimpose to the description of molecular structures and processes coarser grain approximate treatments like fluid-dynamics and large scale statistical treatments. Most of these programs and procedures are built in-house and are not properly catalogued and made available as services to the CMST community members.

ACCESS TO COMPUTING RESOURCES

For these computer applications the request of computing power increases with the size and complexity of the system as well as with the need for combining and coordinating different competences and packages. As to the access to adequate computer resources, the most popular approach adopted by the members of the CMMST community is that of relying on a locally managed cluster of cpus and/or of applying for computer time grants at national and international level. According to a recent (still incomplete) survey carried out by the Italian theoretical and computational chemistry division, the local approach is adopted by 60% (group facilities), 40% (personal desktop), 25% (university or departmental facilities) of the CMMST members while the second is adopted by 40% (CINECA), 20% (former CASPUR), 17% (International computer centres) of the members implying a certain extent of mixed usage. As to the use of local resources, they bear the advantage of being highly customized and fully utilizable (yet they require specific hardware and system software management). However, the possibility of having the machines under full control, of optimizing codes and algorithms at will, of freely planning the resources usage are usually highly appreciated by the members of the community (especially those who prefer to develop new codes and are keen to get involved in the management of computers) despite the risk of letting hardware and software obsolesce. As to the use of remote large scale facilities, the management policies adopted in general by the computer centres to award computer time grants are based on the wish of keeping the machines as busy as possible and offering computer time (on the ground of applications evaluated by a panel of experts) only to "first time" or "record breaking" projects rather than

ISSN: 2279-8773

on the wish of supporting of long term research. Accordingly, the grants set also constraints on the size of per core memory, on the minimum number of cores per run, on the maximum time per run, on the period of the facility accessibility, on the inclusion in the code of check points/restart features, on the total storage/hard disk needed for the whole duration of the project, on the total storage/hard disk needed for production runs, on the applications and libraries availability, on the parallelization tools, on the I/O intensity. Applications to the grant systems requires also preliminary benchmarks and tests of the code and of the middleware used, ex ante evaluations of the quality of the used software before getting access to the computing resource (from the point of view of the computer centre rather than from the validity of the subtended research line). Al this results in an increased difficulty in long term planning research activities and in a distortion of research development.

## THE DISTRIBUTED COMMUNITY COMPUTING MODEL

In order to overcome the limitations of both (the full local and the only external) computing models, the adoption of a grid infrastructure is proposed. At present the European and national computer grids are used by less than 10% of the CMMST users. Yet, the networking in grid of the above mentioned 60% of local usage of CMMST computing resources could bring enormous benefits to its users. This is due not only to the gathering together of the overall computer power of the above mentioned local resources but also to the connectivity to the already connected hundred thousands of the extended HTC distributed grid platform of EGI and to the already experimented possibility of bridging to HPC resources through a transparent access. An advantage of such distributed model consists in the possibility for the user to choose the better platform for his/her applications and for the computer centres to qualify the usage of their resources. Another advantage of the model comes from the fact that, in addition to passive users (those using from a suitable portal services and products made available by some providers on the grid) CMMST users are often keen to be active users (those developing and implementing their own programs and packages either for personal usage or, what is even more productive for the community, for making them available to the other members by providing a stable running version of their under development programs building so far an evolving advanced cooperative library). This offers to the CMMST members the possibility of combining different pieces of software into single workflows (or workflow of workflows) to the end of assembling quite complex realistic simulators (like GEMS [A. Costantini, O. Gervasi, C. Manuali, N. Faginas Lago, S. Rampino, A. Laganà, COMPCHEM: progress towards GEMS a Grid Empowered Molecular Simulator and beyond, Journal of Grid Computing, 8(4), 571-586 (2010)]) and undertake more ambitious research projects. Such possibility of building workflows of shared programs has, for example, stimulated the setting of proper (de facto) standards of data in quantum chemistry and quantum dynamics [E. Rossi, S. Evangelisti, A. Laganà, A. Monari, S. Rampino, M. Verdicchio, K. Baldridge, G.L. Bendazzoli, S. Borini, R. Cimiraglia, C. Angeli, P. Kallay, H.P. Lüthi, K. Ruud, J. Sanchez-Marin, A. Scemama, P. Szalay, A. Tajti, Code Interoperability and Standard Data Formats in Quantum Chemistry and Quantum Dynamics: the Q5/D5cost Data Model submitted to the J. Comp. Chem.]. This has prompted the development of tools (like the framework GriF [C. Manuali, A. Lagana' GRIF: A New Collaborative Framework for a Web Service Approach to Grid Empowered Calculations Future Generation of Computer Systems, 27(3), 315-318 (2011)]) enabling the redirecting of computer applications to run on the best suited sites (including HPC machines) by properly redirecting the jobs to the most appropriate architecture. This enhances cooperative compute capabilities by opening the perspective of combining different complementary know how into single (higher level of complexity) realistic applications paves the way for applying for more ambitious research grants to be shared within the community. After all, some existing grid tools (like GriF and GCreS) already allow the evaluation of the quality of the services provided to the end of grounding the award of credits to the members carrying out activities useful to the community.

## FROM VIRTUAL ORGANIZATIONS TO VIRTUAL RESEARCH COMMUNITIES

Grid users are usually clustered in virtual organisations (VO)s. VOs are groups of researchers with similar scientific interests and requirements, who are able to work collaboratively with other members and/or share resources (e.g. data, software, expertise, CPU, storage space), regardless of geographical location. Researchers must join a VO in order to use grid computing

resources provided by EGI (how to join a VO). Each virtual organisation manages its own membership list, according to the VO's requirements and goals.
EGI provides support, services and tools to allow VOs to make the most of their resources. EGI currently hosts more than 200 VOs for communities with interests as diverse as Earth Sciences, Computer Sciences and Mathematics, Fusion, Molecular and Life Sciences and High-Energy Physics, Computational Chemistry (COMPCHEM and GAUSSIAN are the VOs belonging to the CMMST computational community). However, as stated in the EGI website (http://www.egi.eu/community/vrcs/), EGI is articulating itself in VRCs that are groups of like-minded individuals (organised by discipline or computational model) which have an established presence in their field and represent a well-defined scientific or research community (VRCs are usually assembled out of one or more VOs).
EGI establishes partnerships with individual VRCs through a Memorandum of Understanding (MoU). Following the accreditation process and final agreement, VRCs can access the computing resources and data storage provided by the EGI community through open source software solutions. VRC members can store, process and index large datasets and can interact with partners using the secured services of EGI's production infrastructure.
International scientific communities can draw many benefits from a strong partnership with EGI. EGI offers an open process to the improvement of its user community oriented services (a register of grid-ready applications and training resources), workshops and forums to collect and refine community input, help and support on resolving specific technical issues, as well as involvement in the evolution of EGI's production infrastructure. In turn, VRCs provide EGI with their technical and service requirements, which are then fed into the overall development of the infrastructure as a consumer-driven resource. VRC representatives sit on the User Community Board and are encouraged to advise EGI on its planning and operational priorities, based on requirements collected from its members.
VRCs having already signed a MOU or a letter of intent (LOI) are:
**WeNMR - A worldwide e-Infrastructure for NMR and structural biology**
**LSGC – The Life-Science Grid Community**
**HMRC – Hydro-Meteorology Research Community**
**wLCG - Worldwide Large Hadron Colliders Computing Grid**
**CLARIN – Common Language Resources and Technology Infrastructure**
**DARIAH – Digital Research Infrastructure for the Arts and Humanities**
which can be considered as prototypes for CMMST.

THE ROLE OF NATIONAL INFRASTRUCTURES
An important line of intervention of EGI is through National organizations (the so called National Grid Initiatives (NGIs)) aimed at preventing fragmentation and duplication of effort across the communities. Through NGIs EGI provides:

- Coordination of management services

Manage operations to coordinate and perform the activities required to deliver to the users services at an agreed level and supervise the ongoing management of technology deployment and technical support services.

Support and coordinate communities to source users' requirements

Provide technology to upgrade software through outsourcing technology developments, negotiating with potential technology providers and assessing the new software's quality

Design and implement strategy, policy and collaborations fostering, preparing and supervising at the European Union level the preparation of related documents, papers and multimedia communication

Build community and enhance outreach through the organisation of flagship events like the Community Forum in Spring, the Technical Forum in Autumn, other smaller events and

ISSN: 2279-8773

workshops throughout the year. Promote the latest news and scientific achievements of the research communities using EGI.

- Infrastructure services

Coordinate the integrated and seamless computing e-infrastructure contributed by the members of the federation spanning more than 30 countries and support to its users

Supply information to Operations Centres, VO managers and other interested parties, as well as related services such as the VO registration tool, the broadcast and downtime tool and the regional dashboard.

Take care of the monitoring infrastructure and of the performance tests of grid services

Present through the central accounting portal a homogeneous, user friendly view of the data and supply a good understanding of resource utilisation. Validate the gathered data and supervise the publication process.

Provide the support to users and operations staff through the GGUS distributed helpdesk with central coordination.

Maintain the Grid Configuration Database containing general information about the resource centres that make up the grid infrastructure.

Take care of main and auxiliary core (including middleware) services providing information discovery, authentication, workflow management, file cataloguing and other services that can vary, depending on the scale of the NGI commitments.

- Support services

Provide support services to allow new and current researchers to make the most of the infrastructure, enable researchers to organise themselves into sustainable virtual research communities

Assure support to resource centres to manage daily operations, train marketplace, feed the application database, track requirements, assemble science gateways, implement workflows and produce web gadgets

THE ROLE OF VRCs AT NATIONAL LEVEL
A key contribution of the VRCs to the implementation of the above mentioned cooperative grid model and functions is the strengthening through their activities of the national roots of EGI. In particular the CMMST members should foster through their national association the creation of a certain number of suitably equipped local computing sites will be gathered together. This means the hinging on the grid of some research laboratories able to support a sufficiently robust distributed HTC platform, the bridging of this platform to some large scale HPC (using low level instruments like SSH or more advanced middleware products (like those of EMI)) machines, the maintaining on the nodes of the related middleware (Glite, Arc, Unicore or their evolutionary products), the adopting or developing of the tools needed to facilitate the usage of the grid.
Already established EGI tools for grid work are:
DASHBOARDS

- EDMS (Experiment Dashboard Monitoring System) a system to monitor, transfer data, commission sites and provide as well assistance and Virtual Organizations management. EDMS can operate on several Grid.

USER INTERFACES AND FRAMEWORKS

ISSN: 2279-8773

- GANGA an easy to use front end for job definition and management offering a uniform environment across multiple distributed computer systems;
- DIANE a lightweight task processing framework utilizing an application aware scheduler allowing an efficient and robust execution of large number of computational tasks on heterogeneous computing infrastructure.

WORKFLOWS

- Tools developed to govern complex ensembles of data, models and programs of an increasing number of applications and to offer a unified user friendly way of composing related tools. Among them are ASKALON, KEPLER, K-WF GRID, MOTEUR, PEGASUS, P-GRADE, TAVERNA ,TRIANA, UNICORE WORKFLOW**.**

GATEWAYS

- Tools offering the service of routing packets outside the local network providing not only the basic functions but also a series of services which are often specific of a community. Among them SOMA2 is specific of the molecular science community).

DATA MANAGEMENT

- GREIC (Grid Relational Catalog) a tool providing a set of advanced data grid services aimed at transparently, efficiently and securely managing databases on the Grid;
- HYDRA a file encryption/decryption tool developed as part of the gLite middleware,
- MPI (Message Passing Interface) is a library of routines providing concurrent execution of parallel programs**;**
- FTS (File Transfer Service) is a lightweight but fully functional set of services supporting data management;
- DPM (Disk Pool Manager);
- LFC (LCG File Catalog).

In this respect the CMMST should get acquainted with the mentioned tools and develop their utilization by its members with the aim of making their usage more user friendly and suitable for CMMST applications.

Other EGI supports to choose from for use in the CMMST community are:

- Catalogues of existing solutions: http://go.egi.eu/sciencegateways
  - Gateway components repository like the SCI-BUS portlet repository http://*www.**sci-bus**.eu*
  - "How-to" documentation to develop an EGI science gateway primer (through a new EGI Virtual Team project and portal-community@mailman.egi.eu)
  - Portal & gateway workshops like the portal workshops of the various NGIs and community workshops on portals
- Collaboration with workflow system developers like:
  - Various VOs and VRCs
  - the SHIWA project;
  - the ER-Flow proposal "Building a European Research Community through Interoperable Workflows and Data' (http://*www.youtube.com/user/**ERFLOW)***
  - Requirements, workshops (e.g. http://go.egi.eu/workflowworkshops)
  - Middleware API table: https://wiki.egi.eu/wiki/Service_APIs
  - Applications Database (EGI software catalogue) aimed at storing information, organising software into clusters, performing subscriptions and notifications, integrating the catalogue with the EGI platform (e.g. Info system) and with other catalogues
- Training Marketplace allowing to
  - Advertise and search for events, web gadgets and various materials like the information available on the Indico page (http://indico.egi.eu/indico)
  - Assemble or utilize online courses and Webinar presentation about EGI tools and services. Such services can be roughly categorised as
    - Tools to integrate scientific applications and models with grids and clouds (Portals, workflow frameworks, APIs,...)
    - Tools to facilitate collaboration inside the CMMST research community (Applications Database, Training Marketplace, Document server, Wiki pages, Email lists, ...)
    - Tools to operate software services for CMMST researchers (Monitoring tools, accounting tools, helpdesk,...)

ISSN: 2279-8773

ROADMAP FOR ASSEMBLING THE VRC

The assemblage of a CMMST VRC is the goal of the homonymous Virtual Team (VT). The VT roadmap is first the assemblage of a document analysing the advantages that membership as a VRC within EGI will bring. The VRC status could help the CMMST community satisfy the requirements of its members concerning the access and use of national computing resources that are federated in EGI. The document will document:

i. the structure that such a VRC should have to represent the CMMST community in EGI;
ii. the technologies, resources and services that already exist within EGI and could be used to satisfy the requirements of the CMMST VRC;
iii. the technologies that need to be developed or brought into EGI, then integrated with the production infrastructure so the VRC members can efficiently manage and use resources from EGI.

and should represent the ground for

i. developing a plan aimed at assembling a VRC out of the existing CMMST oriented EGI VOs and from the applications, tools and other resources and services that NGIs and projects of EGI provide.
ii. identifying tools, services and resources that the VRC needs to develop or bring into EGI in order to operate as a sustainable entity for the CMMST scientific community.
iii. developing, on the basis of the outcomes of the above mentioned two items, a proposal to establish a CMMST VRC in EGI. Besides the technical aspects, the proposal will have to define the organisational and funding models for the VRC.

To cope with the above mentioned objectives it is important to extend the basis of *passive* and *active users*. Yet, it is even more important to allow the number of active users which represent a significant fraction of the community members, to be equipped with user friendly tools (like the already mentioned GRIF that is a Workflow Management System (WfMS) designed to help the user on the ground of QoS and QoU evaluators designed on top of the experience of the COMPCHEM VO) and similar tools aimed at facilitating job management and optimizing resource selection.
Obviously users can exploit the advantage of utilizing either more generic or more specialized tools to get better organized including the adoption of some portals or simply fish in the application Data Bases (AppDB) of EGI (http://*appdb.egi.eu/*). To this end the assemblage of a specific portal is vital. The feeding of AppDB is of paramount important to enhance the possibility of composing complex applications out of programs maintained by different active users. Some of the already implemented programs in the perspective use in the GEMS simulator are listed in Table 1.

| **Application** | **Description** | **License** |
|---|---|---|
| ABC | Solve the Schrodinger equation for triatomic systems using the time independent quantum method | Academic |
| MCTDH | MultiConfigurational Time Dependent Hartree method | Academic |
| FLUSS | Lanczos iterative diagonalization | Academic |
| VENUS96 | Quasi-classical dynamics of reactive collisions | |

ISSN: 2279-8773

| | | |
|---|---|---|
| DL_POLY | Classical Molecular Dynamics | |
| NAMD | Classical Molecular Dynamics | Academic |
| GAMESS-US | General Atomic and Molecular Electronic Structure Package | Academic |
| RWawePR | Time Dependent Method to Solve the quantum reactive Scattering equations for triatomic systems | Academic |
| GROMACS | GROningen MAchine for Chemical Simulations | Academic |
| SCIVR | Semiclassical initial value representation method | Academic |
| **Framework** | **Description** | |
| GriF | Grid Framework enabling efficient and user-friendly scientific massive calculations | Free |
| Gcres | Quality of Users (QoU), Quality of Services (QoS) evaluation Framework | Free |
| ggamess | Front-end script for submitting multiple GAMESS-US jobs | Free |

**Table 1 – A list of the CMMST packages offered by COMPCHEM as tools, applications or Services**

Progress of active users to the class of *software providers* (which are users structuring their programs (or suites of programs) for usage by other users in a workflow) with the purpose of offering the possibility of building more complex applications of general interest out of the software provided by other people.
In addition to the creation of an interoperable distributed library of application software whose components will be maintained by the various members of the community other goals will be pursued by the VRC. For example, the ability of GriF of facilitating the selection of the computing platform more suitable for the intended calculations and enabling the composition of collaborative applications as services (some of which may be in competition among them) will be enhanced. This is addressed to provide other services of higher level and monitor the grid to the end of evaluating the QoS and the QoU useful to award credits to the members of the community depending on their commitment to the community goals (for this purpose COMCPHEM has developed a tool called GcreS). This is aimed at creating a community economy based on the award and the redemption of credits which are used as terms of exchange (toex) for buying better services or getting a larger share of the resources (hardware or financial) owned by the community. This is the driving force that COMPCHEM is planning to use to motivate people to contribute to the cooperative goals and to ensure its sustainability.

ISSN: 2279-8773