A MODERN APPROACH TO AB INITIO COMPUTING IN CHEMISTRY, MOLECULAR AND MATERIALS SCIENCE AND TECHNOLOGIES

ANTONIO LAGANA',
DEPARTMENT OF CHEMISTRY, UNIVERSITY OF PERUGIA, PERUGIA (IT)*

ABSTRACT

In this document we examine the present situation of Ab initio computing in Chemistry and Molecular and Materials Science and Technologies applications. To this end we give a short survey of the most popular quantum chemistry and quantum (as well as classical and semiclassical) molecular dynamics programs and packages. We then examine the need to move to higher complexity multiscale computational applications and the related need to adopt for them on the platform side cloud and grid computing. On this ground we examine also the need for reorganizing. The design of a possible roadmap to establishing a Chemistry Virtual Research Community is then sketched and some examples of Chemistry and Molecular and Materials Science and Technologies prototype applications exploiting the synergy between competences and distributed platforms are illustrated
for these applications the middleware and work habits into cooperative schemes and virtual research communities (part of the first draft of this paper has been incorporated in the white paper issued by the Computational Chemistry Division of EUCHEMS in August 2012)

INTRODUCTION

The computational chemistry (CC) community is made of individuals (academics, affiliated to research institutions and operators of chemistry related companies) carrying out computational research in Chemistry, Molecular and Materials Science and Technology (CMMST). It is to a large extent registered into the CC division (DCC) of the European Chemistry and Molecular Science (EUCHEMS) Society and is connected to other chemistry related organizations operating in Chemical Engineering, Biochemistry, Chemometrics, Omics-sciences, Medicinal chemistry, Forensic chemistry, Food chemistry, etc. for research, innovation and education. Such community operates also for Collaboration in Chemistry, Molecular and Materials Science and Technology (CMMST) within the homonymous domain of the European Initiative COST (Collaboration in Science and Technology) and for research based education within the ECTN (the European Chemistry Thematic Network) Association ECTNA. The latter is developing for that purpose a Virtual Education Community (VEC) to promote a shared European frame to harmonize and accredit Chemistry curricula (including CC syllabus and labels), to exploit modern computing technologies in education (electronic self evaluations tests, distributed repositories for learning objects etc.) and to promoting Euromaster and Eurodoctorate courses in Theoretical Chemistry and Computational Modelling (TCCM).

The present document is meant to offer a contribution to the redefinition of a coordinated CC policy of access and use of shared computing resources, to the fostering of the development of service oriented approaches and to the promotion of innovative research in CMMST. To this end reference is made to the European Grid Infrastructure (http://www.egi.eu/) and to its national partners (e.g. the Italian http://www.italiangrid.it/, the Polish http://www.plgrid.pl and the Spanish http://www.es-ngi.es/ ones) for ensuring the grounding of the CMMST calculations and methodologies on distributed computing (or High Throughput Computing, HTC) platforms in addition to High Performance Computing (HPC) and group (or Departmental and University) mid and small size computer clusters which are more popular within the CC community because traditionally offered by national or local computer infrastructures. Both HTC and HPC facilities (as well as clusters) are presently undergoing a radical change of computing paradigms (that is paralleled also by an analogous change in governmental policies in Europe).

ISSN: 2279-8773

Such radical changes imply, among others, modifications in the way research is carried out and computing applications are built. The document focuses first on **interoperability and modularity** and their enhancement through a continuous iteration stream of designing, implementing, running and validating data and software components by aggregating codes and data of different origins in workflows (as well as workflows of workflows). The second focus of the document is the analysis of how **Virtual Research Communities** can established and operate as a **credit based Service Oriented system** in which computing applications (and research itself) are structured as modular services grafted on an economy based on Quality of Service (QoS) and Quality of Users (QoU) mechanisms enabling the objective recognition of the contributions of the various members (with the consequent award of credits) as part of a collaborative/competitive endeavour.

COMPUTATIONAL CHEMISTRY PROGRAMS

The most popular CMMST ab initio programs are the quantum chemistry ones (see for example http://en.wikipedia.org/wiki/List_of_quantum_chemistry_and_solid_state_physics_software from which Table 1 is taken). Most of them are based on Hartree-Fock (HF) and some post Hartree-Fock methods. They may also include density functional theory (DFT), Molecular Mechanics (MM) or semiempirical quantum chemistry methods. The programs include both open source and commercial software. Most of them are large packages, often containing several separate programs, and have developed over many years.

Package	License[†]	Lang.	Basis	Periodic[‡]	Mol. mech.	Semi-emp.	HF	Post-HF	DFT
ABINIT	GPL	Fortran	PW	3d	Yes	No	No	No	Yes
ACES II	Academic	Fortran	GTO	No	No	No	Yes	Yes	Yes
ACES II MAB	Academic	Fortran	GTO	No	No	No	Yes	Yes	No
ACES III	GPL	Fortran/C++	GTO	No	No	No	Yes	Yes	No
ADF	Commercial	Fortran	STO	Any	Yes	Yes[4]	Yes	No	Yes
Atomistix ToolKit (ATK)	Commercial	C++/Python	NAO/EHT	3d[9]	Yes	Yes	No	No	Yes
BigDFT	GPL	Fortran	Wavelet	Any	Yes	No	Yes	No	Yes
CADPAC	Academic	Fortran	GTO	No	No	No	Yes	Yes	Yes
CASINO (QMC)	Academic	Fortran 95	GTO / PW / Spline / Grid / STO	Any	No	No	Yes	Yes	No
CASTEP	Academic (UK) / Commercial	Fortran	PW	3d	Yes	No	Yes[5]	Yes	Yes
CFOUR	Academic	Fortran	GTO	No	No	No	Yes	Yes	No
COLUMBUS	Academic	Fortran	GTO	No	No	No	Yes	Yes	No
CONQUEST	Academic	Fortran 90	NAO/Spline	3D	Yes	No	Yes[5]	No	Yes
COSMOS	Commercial	Unknown	Unknown	Unknown	Yes	Yes	No	No	No
CP2K	GPL	Fortran 95	Hybrid GTO / PW	Any	Yes	Yes	Yes	No	Yes
CPMD	Academic	Fortran	PW	Any	Yes	No	Yes	No	Yes
CRYSTAL	Academic (UK) / Commercial	Fortran	GTO	Any	Yes	No	Yes	Yes[10]	Yes
DACAPO	GPL ?[1]	Fortran	PW	3d	Yes	No	No	No	Yes
DALTON	Academic	Fortran	GTO	No	No	No	Yes	Yes	Yes
DFTB+	Academic / Commercial	Fortran 95	NAO	Any	Yes	Yes	No	No	No
DFT++	GPL	C++	PW / Wavelet	3d	Yes	No	No	No	Yes
DIRAC	Academic	Fortran 77,	GTO	No	No	No	Yes	Yes	Yes

Package	License[†]	Lang.	Basis	Periodic[‡]	Mol. mech.	Semi-emp.	HF	Post-HF	DFT
		Fortran 90, C		🟥	🟥	🟥	🟩	🟩	🟩
DMol3	Commercial	Unknown	Numeric AOs	3d	No	No	No	No	Yes
ELK	GPL	Fortran 95	FP-LAPW	3d	Unknown	Unknown	Yes	Unknown	Yes
ErgoSCF	GPL	C++	GTO	No	No	No	Yes	Yes	Yes
EXCITING	GPL	Fortran 95	FP-LAPW	3d	Unknown	Unknown	Yes	Unknown	Yes
FLEUR	Academic	Fortran 95	FP-(L)APW+lo	3d, 2d, 1d	No	No	Yes	Yes	Yes
FHI-aims	Commercial	Fortran	NAO	Any	Yes	No	Yes	Yes	Yes
FreeON	GPL	Fortran 95	GTO	Any	Yes	No	Yes	Yes	Yes
Firefly / PC GAMESS	Academic	Unknown	GTO	No	Yes[3]	Yes	Yes	Yes	Yes
GAMESS (UK)	Academic (UK) / Commercial	Fortran	GTO	No	No	Yes	Yes	Yes	Yes
GAMESS (US)	Academic	Fortran	GTO	No	Yes[2]	Yes	Yes	Yes	Yes
GAUSSIAN	Commercial	Fortran	GTO	Any	Yes	Yes	Yes	Yes	Yes
GPAW	GPL	Python / C	Grid / NAO / PW	Any	Yes	Unknown	Yes[5]	No	Yes
hBar Lab[7]	Commercial	Unknown	GTO	No	No	No	Yes	Yes	Yes
HiLAPW	Unknown	Unknown	FLAPW	3d	No	No	No	No	Yes
JAGUAR	Commercial	Unknown	GTO	Unknown	Yes	No[11]	Yes	Yes	Yes
MADNESS	GPL	C++	Wavelet	Unknown	No	No	Yes	No	Yes
MISSTEP	GPL	C++	PW	No	No	No	No	No	Yes
MOLCAS	Commercial	Fortran	GTO	No	Yes	Yes	Yes	Yes	Yes
MOLPRO	Commercial	Fortran	GTO	No	No	No	Yes	Yes	Yes
MOPAC	Academic / Commercial	Fortran	Unknown	Unknown	Unknown	Yes	No	No	No
MPQC	LGPL	C++	GTO	No	No	No	Yes	Yes	Yes

NWChem	ECL v2	Fortran 77 / C	GTO, PW	Yes(PW)
				No(GTO)
					Yes	Yes	Yes	Yes	Yes

Octopus	GPL	Fortran 95, C, OpenCL	Grid	Any	Yes	No	Yes	No	Yes
ONETEP	Academic (UK) / Commercial	Fortran	PW	Any	Yes	No	Yes[5]	No	Yes
OpenAtom	Academic	Charm++ (C++)	DVR	Unknown	Yes	No	No	No	Yes
OpenMX	GPL	C	NAO	3d	Yes	No	No	No	Yes
ORCA	Academic	C++	GTO	No	Yes	Yes	Yes	Yes	Yes
PLATO	Academic	Unknown	NAO	Any	Yes	No	No	No	Yes
PQS	Commercial	Unknown	Unknown	Unknown	Yes	Yes	Yes	Yes	Yes
Priroda-06	Academic	Unknown	GTO	No	No	No	Yes	Yes	Yes
PSI	GPL	C / C++	GTO	No	No	No	Yes	Yes	Yes
PWscf[6]	GPL	Fortran	PW	3d	No	No	Yes	No	Yes
PyQuante	BSD	Python	GTO	No	No	Yes	Yes	Yes	Yes
Q-Chem	Commercial	Fortran / C++	GTO	No	Yes	Yes	Yes	Yes	Yes
Quantemol-N	Academic / Commercial	Fortran	GTO	No	Yes	Yes	Yes	Yes	No
Quantum ESPRESSO	GPL	Fortran	PW	3d	Yes	No	Yes	No	Yes
RSPt	Academic	Fortran / C	FP-LMTO	3d	No	No	No	No	Yes
SPARTAN	Commercial	Fortran / C /	GTO	No	Yes	Yes	Yes	Yes	Yes

Package	License[†]	Lang.	Basis	Periodic[‡]	Mol. mech.	Semi-emp.	HF	Post-HF	DFT
		C++		🟥	🟩	🟩	🟩	🟩	🟩
SIESTA	Academic	Fortran	NAO	3d	Yes	No	No	No	Yes
TB-LMTO	Academic	Fortran	LMTO	3d	No	No	No	No	Yes
TERACHEM [8]	Commercial	C/CUDA	GTO	No	Yes	No	Yes	No	Yes
TURBOMOLE	Commercial	Fortran	GTO	No	Yes	No	Yes	Yes	Yes
VASP	Academic(AT)/ Commercial	Fortran	PW	Any	Yes	No	Yes	Yes	Yes
WIEN2k	Commercial	Fortran / C	FP-(L)APW+lo	3d	Yes	No	No	No	Yes
Yambo Code	GPL	Fortran	PW	3d	No	No	Yes	Yes	No

[†] “Academic”: academic (no cost) license possible upon request; “Commercial”: commercially distributed.

[‡] Support for periodic systems (3d-crystals, 2d-slabs, 1d-rods and isolated molecules): 3d-periodic codes always allow the simulation of systems with lower dimensionality within a supercell. Specified here is the capability for actual simulation within lower periodicity.

[1] The CAMPOS project (which includes Dacapo) states that all code is GPL. The Dacapo distribution itself does not contain any license information.

[2] Through interface to TINKER

[3] Through Ascalaph

[4] Through interface to MOPAC

[5] Using exact exchange DFT

[6] Distributed with Quantum ESPRESSO

[7] Web service integrating MPQC.

[8] TeraChem is the first fully GPU-accelerated quantum chemistry software.

[9] Atomistix ToolKit also contains finite-bias NEGF electron transport calculations with open boundary conditions.

[10] Through CRYSCOR program.

[11] However, available in the Schrödinger Suite.

Table 1 – Popular quantum chemistry packages.

The second less developed largest class of established CMMST programs are the MM (Molecular Mechanics) or the Molecular Dynamics (MD) ones in which the atoms and molecules are allowed to interact and move according to the equations of motion. Some of them have been already quoted in Table 1 (among which the most popular ones are CPMD, NWChem, CP2K, etc.). Other packages of this class are:

AMBER (a set of both MM force fields for the simulation of biomolecules “Ambertools12” (which are in the public domain, and are used in a variety of simulation programs) and a package of molecular simulation programs “Amber 12”; see for an overview [D.A. Case, T.E. Cheatham, III, T. Darden, H. Gohlke, R. Luo, K.M. Merz, Jr., A. Onufriev, C. Simmerling, B. Wang and R. Woods. The Amber biomolecular simulation programs. J. Computat. Chem. 26, 1668-1688 (2005)]).

CHARMM (a widely used set of force fields for MD and a simulation and analysis package. The CHARMM Development Project involves a network of developers throughout the world working to develop and maintain the package. [Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983). "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". J Comp Chem 4 (2): 187–217. DOI:10.1002/jcc.540040211]

DL_POLY (a general purpose MD simulation package continually developed at Daresbury Laboratory by W. Smith and I.T. Todorov under the auspices of EPSRC and NERC in support of CCP5. It can be used to simulate a wide variety of molecular systems including simple liquids, ionic liquids and solids, small polar and non-polar molecular systems, bio- and synthetic polymers, ionic polymers and glasses solutions, simple metals and alloys; see [Molecular Simulation, 28 (2002), pp 385]).

GROMACS (a MD simulator primarily designed for biochemical molecules, like proteins and lipids that have a lot of complicated bonded interactions, that is extremely fast at calculating the nonbonded interactions (that usually dominate simulations) and is equipped with tools for input assemblage and output analysis. [Berk Hess, Carsten Kutzner, David van der Spoel and Erik Lindahl, GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. (eprint) J. Chem. Theory Comput. 4(3):435-447 (2008); see also http://www.gromacs.org/content/view/12/176/]).

LAMMPS (a MD simulator mainly designed for studies in soft materials (biomolecules, polymers), solid-state materials (metals, semiconductors) and coarse-grained or mesoscopic systems. It can be used to model atoms or, more generically, as a parallel particle simulator at the atomic, meso, or continuum scale and is particularly esy to modify. [S. Plimpton, Fast Parallel Algorithms for Short-Range Molecular Dynamics. (1995)]; see also http://lammps.sandia.gov/)

MOLDY (a short-ranged MD program that makes use of Link Cells and Neighbour Lists, to fully exploit the short range of the potentials used, and the slow diffusion expected for solid systems. The code allows for a wide variety of boundary conditions, including constant pressure, temperature and strain rate. It also incorporates molecular statistics via the conjugate gradients minimisation of the enthalpy).

TINKER (a general package for MM and MD, with some special features for biopolymers making use of any of several common parameter sets, such as Amber (ff94, ff96, ff98, ff99, ff99SB), CHARMM (19, 22, 22/CMAP), Allinger MM (MM2-1991 and MM3-2000), OPLS (OPLS-UA, OPLS-AA), Merck Molecular Force Field (MMFF), Liam Dang's polarizable model, and the AMOEBA (2004, 2009) polarizable atomic multipole force field. Parameter sets for other widely-used force fields are under consideration for future releases).

YASARA (a molecular-graphics, -modeling and -simulation program for Windows, Linux and Mac OS X developed since 1993, that finally makes it really easy to answer your questions. With an intuitive user interface supports several YASARA create high interaction with artificial reality that allows you to focus on the problem more than on technicalities).

Though mainly based on in-house developed codes also quantum mechanics approaches (and semiclassical as well) are available for carrying out molecular dynamics calculations (though to a great extent restricted to small sytems).

MULTISCALE SIMULATIONS AND VIRTUAL EXPERIMENTS

The above mentioned programs and packages (and/or a proper combination of them) are structured in a way that allows the evaluation of some properties of theoretical interest related to given physical observables rationalized via a model treatment. A typical case of this is the evaluation of the macroscopic thermodynamic properties of realistic systems based on the ergodic hypothesis (statistical ensemble averages are equal to time averages of the system) for which an averaging over unobserved elementary variables is so extended to not allow a direct evaluation of observables from first principles.

The most straightforward correspondence between theoretical properties and physical observables is, indeed, obtainable for rarefied gas systems experiments. A typical case of that are the Crossed Molecular Beam (CMB) experiments in which the cross section of single collision elementary (state to state) processes can be individually studied. In particular, in recent times, a cooperative application called Grid Empowered Molecular Simulator (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)] has been assembled by combining as building blocks of a workflow programs devoted to quantum chemistry calculations of the electronic structure of the molecular system considered (INTERACTION), to the fitting (FITTING) of the calculated potential energy values to a suitable functional form (after corrections and quality controls), to the quantum or classical (whatever is more appropriate) dynamical calculations (DYNAMICS), to the statistical averaging over the unobserved variables to derive physical observables (OBSERVABLES). In this way it has been possible to work out in a complete ab initio fashion the measured CMB product beam intensity (virtual beam) thanks to the exact knowledge of the geometry of the experimental apparatus [A. Lagana', 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) -> H + CO2 reaction, Faraday Discussion of Chem. Soc. 157, 415-436 (2012)]. Moreover, though still in the experimental stage, GEMS is modular and usable as a service by both theorists and experimentalists (here service means provide as working tool to embody inside a grid based workflow any software that from a well defined input produces a specific well determined output). The usability of GEMS, however, obviously depends on the availability on the grid platform of the appropriate computers not only because different classes of programs may pose different requests in terms of computer resources (and this is even more so when multi-scale simulations are considered) but also because the same package in different conditions may drastically vary its requests depending either on the level of theory adopted or on the values (and related ranges) of input parameters used. The choice of the level of theory to adopt, in fact, largely impacts the size of memory and the length of calculations.

As a matter of fact, the computer platforms most used for the CMMST quantum chemistry calculations of the INTRACTION block are the best HPC machines available at the time. HPC platforms have always been the natural haven for CMMST memory hungry - cpu time greedy calculations whose most emblematic prototypes are the quantum chemistry computer programs. In this type of calculations high performances are obtained only by implementing concurrency (simultaneous execution of computational tasks) at very fine grain tightly coupled level of computations and then, possibly, by exploiting also the highest possible parallelism of loosely coupled tasks.

The corresponding more popular packages are listed in Table 1 and are usually installed on the supercomputers of large scale facilities (especially the electronic structure ones). Their maintenance is taken care by specialized ICT experts who are usually in charge also of answering to the tickets open the users when utilizing the package. Typical performances of this type of codes are described below by making reference to the package GAMESS-US (information reported by us in our most recent PRACE application http://www.prace-project.eu/ quoted in https://www3.compchem.unipg.it/virtl-comm/?q=ict) implemented on an IBM Power 6 system. In Fig. 1 elapsed times (lhs panel) and speedups (rhs panel) are plotted as a function of the number of nodes used.

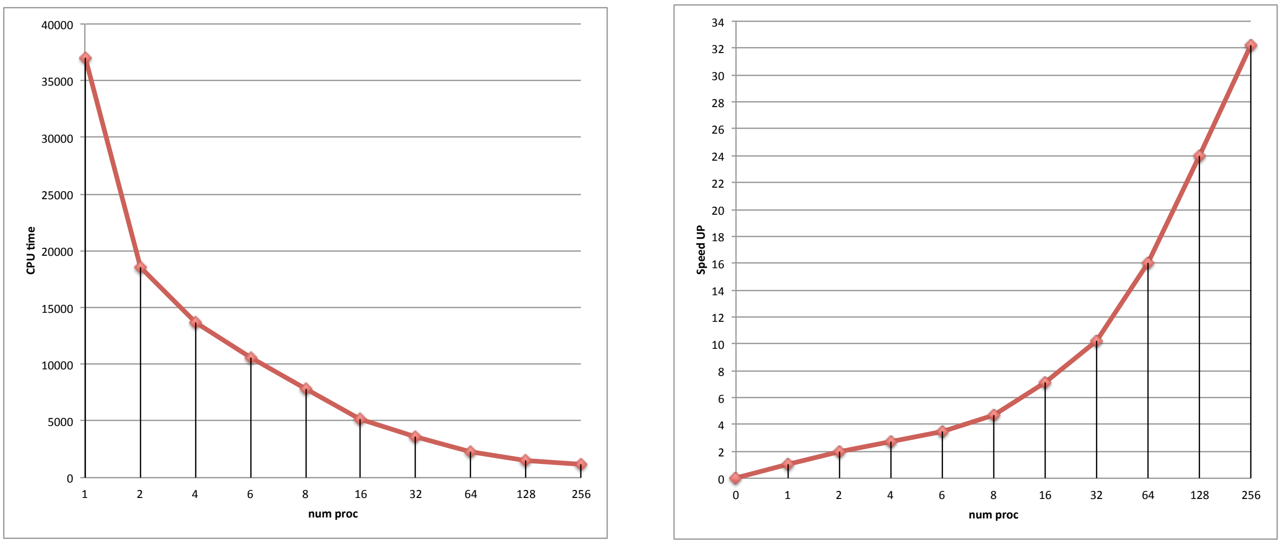

Figure 2 - CPU Time/s (lhs panel) and Speed up (rhs panel) for a GAMESS-US typical run on the IBM Power6 575 machine.

Yet, a warning should be issued about the suitability for quantum chemistry calculations of more modern HCP architectures (like Blue Gene or GPUs which rely on a huge increase in the number of processors (and/or cores) at the expense of core memory. This is, however, less a problem for dynamical calculations (especially the MM and the MD ones) unless a huge number of atoms or molecules is considered. The integration of classical equations of motions is, in fact, high in time consumption yet low in memory occupation (essentially only positions and momenta of the considered particles need to be stored). Less a problem is also the running of the FITTING and OBSERVABLES blocks of GEMS because they do not pose usually particularly heavy demand of memory or time. However, variants of GEMS involving for its OBSERVABLES block the integration of kinetics and fluid dynamics sets of equations might do so. For this reason, the management of GEMS has to be equipped with a tool able to carry out a proper selection of the resources to be used including Workflow Management System (WfMS), Quality of Service (QoS) and Quality of User (QoU) instruments.

GRID AND CLOUD COMPUTING

Yet, the mentioned computer technology evolution is not the only (and not even the major) problem for CMMST users. In fact, the management policies adopted in general by the computer centres are highly unsatisfactory for such community because centred on the idea of keeping the machines as busy as possible and granting computer time (on the ground of applications evaluated by a panel of experts) by limiting the total amount of core-hours assigned and pre-determining the period of accessibility. Other limitations are the maximum and minimum number of cores used (in parallel) per run, the number of planned runs of the whole project, the maximum (and minimum) number of cores per run, the total size of RAM per run with maximum (and minimum) number of cores, the maximum (minimum) job wall clock time, the inclusion in the code of check points/restart features, the total storage/hard disk needed for the whole duration of the project, the total storage/hard disk needed for one production run, what other applications and libraries the package requires, how is the application parallelized (MPI, OpenMP, hybrid, etc.), the intensity of I/O, previous benchmarks and tests of the code, middleware used, etc.. Moreover, the assignment of computer time and storage is subject to an ex ante evaluation of the quality (from the point of view of the evaluator rather than from the success of the research line to which it is inspired) of the proposed research and this could be quite different from the actual impact of the proposed research in its science field. After all Computer centres most often aim at excellence (meant often as "record breaking" and "first time" criteria), at new codes implementation and benchmarking and at full occupation of the machines without much awareness of the new science development heavily affecting, so far, on the long range, the impact and the sustainability of the proposed research.

Although widely adopted, the above mentioned policy is not adequate to meet the needs of the CMMST community (and is clearly not appreciated by the community). The perception of such inadequacy became particularly acute for the Physics community around the turn of the millennium when planning the computational campaigns for dealing with the massive amount of data going to be produced [Robert J. Wilson (22 October 2001). "The European DataGrid Project". Snowmass 2001 (American Physical Society). Retrieved 2 October 2011] by the Large Hadron Collider (LHC) project of CERN (Centre Europeenne pour la Recherce Nucleaire). To face such problem a new collaborative model to produce and allocate computer cycles was developed that was funded within the European Framework Programmes for Research and Technological Development ["EGEE-III Project Publishable Summary". Enabling Grids for E-sciencE. 8 June 2010. Retrieved 2 October 2011; "The DataGrid Project". CERN. March 2004. Retrieved 2 October 2011] through the international "Research and Technological Development for an International Data Grid" ["Research and Technological Development for an International Data Grid". Project IST-2000-25182 description. European Community Research and Development Information Service. Retrieved 7 October 2011] three years project with a budget of about 12 millions of Euros. Along this strategic line on April 2004 the Enabling Grids for E-Science in Europe (EGEE) project (led by the information technology division of CERN[5]) was funded by the European Commission through the Directorate-General for Information Society and Media. This 24-month project of the Sixth Framework Programme had a cost of over 46 million Euro and used high-capacity computing to model complex systems and to process experimental results. The consortium included 70 institutions in 27 countries ["Enabling grids For E-science". Project description. European Community Research and Development Information Service. Retrieved 2 October 2011]. The LHC Computing Grid continued to be a major application of the EGEE technology [LHC Computing Grid: Technical Design Report. The LCG TDR Editorial Board. 20 June 2005. ISBN 92-9083-253-3. Retrieved 3 October 2011]. By April 2006 the "in Europe" specification of the EGEE acronym was changed into Enabling Grids for E-sciencE in EGEE-II when the project was renewed for two years. This two-year further phase cost about 52.6 million Euro ["Enabling grids for E-sciencE-II". Project description. European Community Research and Development Information Service. Retrieved 2 October 2011]. The new name reflected the more universal mission of the project. As a matter of fact at that point e-Science (thanks mainly to a middleware software package known as gLite developed for EGEE ["Enabling grids for E-sciencE-II". Project description. European Community Research and Development Information Service. Retrieved 2 October 2011].) became increasingly based on open collaboration between researchers across the world and made Grid computing popular for scientific disciplines such as high-energy physics, bioinformatics to share and combine the power of computers and sophisticated, often unique, scientific instruments ["History of EGI". EGI.eu website. Retrieved 4 October 2011]. In addition to their scientific value, on May 2008 the EU Competitiveness Council promoted "the essential role of e-infrastructures as an integrating mechanism between Member States, regions as well as different scientific disciplines, also contributing to overcoming digital divides" [European Competitiveness Council, Conclusions on European Research Infrastructures and their regional dimension, Brussels, 3 June 2008]. A third two-year project phase called EGEE-III ran from 2008 to 2010. On 30 April 2010 the EGEE project ended ["EGEE Portal: Enabling Grids for E-sciencE". Official web site. Archived from the original on 19 June 2010. Retrieved 2 October 2011].
In September 2007 the EGI Design Study (EGI_DS) project was launched to evaluate requirements and use cases, identify processes and mechanisms for establishment, define the structure, and initiate the organization ["European Grid Initiative Design Study". Official website. 2009. Retrieved 2 October 2011; "Europeans connecting through the Grid". European Research Headlines (European Commission). 14 September 2007. Retrieved 3 October 2011]. By the year 2009 the governance model evolved towards the European Grid Initiative (EGI), building upon National Grid Initiatives (NGIs) ["ICT Infrastructures for e-Science". Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee of the Regions. Brussels. 5 March 2009. Retrieved 2 October 20]. At the same time other projects like Diligent ["Diligent Project". Defunct website. Archived from the original on 4 November 2008. Retrieved 3 October 2011], D4SCIENCE ["DIstributed colLaboratories infrastructure on Grid enabled technology 4 science". Project description. European Community Research and Development Information Service. Retrieved 3 October 2011], BEinGRID ["Business experiments in GRID". Project 034702 fact sheet. European Community Research and Development Information Service. Retrieved 9 October 2011] were funded. In March 2009, the policy board of EGI announced that its headquarters would be hosted in Amsterdam, the Netherlands, at the Science Park Amsterdam ["Business Experiments in Grid: BEinGRID". Project web site. European Community Research and Development Information Service. Archived

from the original on 23 July 2011. Retrieved 9 October 2011]. The EGI.eu foundation was officially formed on 8 February 2010 in Amsterdam ["Scientists join forces to create European grid computing infrastructure". News release (European Community Research and Development Information Service). 25 September 2007. Retrieved 2 October 2011]. with the final "I" of the acronym now meaning "infrastructure" so as to reflect the transition from a series of short-term research projects to a more sustainable service ["Establishing EGI.eu". Projects (British Publishers): p. 26. Retrieved 3 October 2011] and the NGIs being the stake holders and the supporters of operations, users, dissemination within individual countries. EGI is governed by a Council of representatives of the member NGIs and manages their collaboration allowing researchers to share and compose computing resources in international cooperation. A still undergoing 32 million Euro project, named the EGI-Integrated Sustainable Pan-European Infrastructure for Research in Europe (EGI-InSPIRE), was funded in September 2010 to EGI.eu. A 1.5 million Euro project called e-ScienceTalk was also funded in 2010 for 33 months to support websites and publications covering the EGI ["e-ScienceTalk: Supporting Grid and High Performance Computing reporting across Europe". Project 260733 summary. European Community Research and Development Information Service. 1 September 2010. Retrieved 4 October 2011] to continue the action of an earlier programme known as GridTalk that was funded from 2008 to 2010 ["GridTalk: co-ordinating grid reporting across Europe". Project 223534 summary. European Community Research and Development Information Service. 1 May 2010. Retrieved 4 October 2011; "About GridTalk". Official website. 2010. Archived from the original on 4 October 2011. Retrieved 4 October 2011].
In the EGI model, its publicly networked (Internet) infrastructure provides geographically distributed users with uniform and secure access to the available computing and storage resources (even if structured in a series of independent administrative domains) allowing their efficient, scalable and robust usage. Moreover, while the access and usage of the EGI resources through the standard protocols and middleware is independent of the technical specificities and varieties adopted by the various centres, the central services and tools adopted by the Grid allow a global planning and accounting of the use of the resources by the National infrastructures (NGI) and Virtual Organizations (VO) or Research Communities (VRC). VRCs are groups of like-minded individuals organised by discipline or computational model. A VRC can establish a support relationship, formalised through a Memorandums of Understanding (MoU), with EGI. EGI VRCs [http://www.egi.eu/community/vrcs/] typically have an established presence in their field and represent well-defined scientific research communities. Multi-national scientific communities can draw various benefits from having a VRC partnership with EGI like benefitting from the resources and support of the NGIs, from the workshops and forums organised by EGI, from EGI services to resolv specific technical issues, from becoming involved in the user-focussed evolution of EGI's production infrastructure.

More recently the development of the commercial Cloud Computing has widened the domain of the Grid distributed computing through a new paradigm of virtualization of the resources and of their web management. Thanks to the Cloud services physical resources are converted into a variety of virtual computing and storage environments which can be flexibly and straightforwardly provided on demand. This approach has allowed companies like Google and Amazon to offer on the net commercial services for institutions and companies. The resulting integration has led to the creation of highly important digital infrastructures for variety of users.

GRID SUPPORT TO VRCS

For VRCs the advent of grid computing met not only their needs for a more open (and possibly user friendly) access to additional computing and storage resources but also their needs (in various stages of their research like design, development, compute and interpretation which most often have to be iterated several times up to self-consistency of the solutions) for a more flexible composition of applications out of existing heterogeneous codes, data and tools. Currently, the main EGI VRCs are (http://www.egi.eu/community/vrcs/):

- WeNMR: Structural biology
- LSGC: Life sciences
- HMRC: Hydro-meteorology
- WLCG: High energy physics

including CLARIN for Humanities (that is still at level of Letter of Intent). Other communities [https://indico.egi.eu/indico/sessionDisplay.py?sessionId=40&confId=679#20120329] are struggling to become VRCs like Earth Science (H. Schwichtenberg, Fraunhofer), Biomed (F. Michel, CNRS), Digital Cultural Heritage (S. Andreozzi, INFN), Cherenkov Telescope Array Computing Grid (G. Lamanna, CNRS), GAIA/ASTRA (N. Walton, Cambridge), AUGER (J. Chudoba, Cesnet), Astronomy & Astrophysics (C. Vuerli, INAF), CHAIN (F. Ruggeri, INFN). These efforts are presently concentrated on implementing a Europe-wide coordination and interaction among research communities and national resource infrastructure providers. This means also finding a way of managing maintenance, operation and delivery of an open uniform Europe-wide federated production infrastructure, developing and promoting technologies for gathering new resources and supporting integration of scalable interdisciplinary Virtual Research Environments personalised to the VRC. To this end a variety of services have been already implemented in EGI like:

DASHBOARDS

- EDMS (Experiment Dashboard Monitoring System), is a software monitoring, transferring data and site commissioning that provides also assistance and VO management that was originally designed to support LHC, CMS, ATLAS and ALICE experiments but can operate on several Grid middlewares to cover the full range of the needed computational activities

USER INTERFACES AND FRAMEWORKS

- GANGA is an easy to use front end for job definition and management that offers a uniform environment across multiple distributed computer systems; DIANE is a lightweight task processing framework utilizing an application aware scheduler allowing an efficient and robust execution of large number of computational tasks in unreliable and 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 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 (as an example SOMA2 for the molecular science community)

DATA MANAGEMENT

- GREIC (Grid Relational Catalog) provides a set of advanced data grid services aimed at transparently, efficiently and securely managing databases on the Grid, HYDRA is 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**,** DPM (Disk Pool Manager), LFC (LCG File Catalog), FTS (File Transfer Service) is a lightweight but fully functional set of services supporting data management.

More information about the EGI products to use can be found for

- Catalogues
    - Catalogue of existing solutions: http://go.egi.eu/sciencegateways
- Gateway components repository
    - SCI-BUS portlet repository http://*www.**sci**-**bus**.eu*
- "How-to" documentation
    - Develop an EGI science gateway primer (through a new EGI Virtual Team project and portal-community@mailman.egi.eu)
- Portal & gateway workshops
    - Portal workshops by NGIs
    - Community workshops on portals
- Collaboration with workflow system developers
    - SHIWA project; ER-Flow proposal (http://*www.youtube.com/user/**ERFLOW)***, VOs and VRCs
    - Requirements, workshops (e.g. http://go.egi.eu/workflowworkshops)
    - Middleware API table: https://wiki.egi.eu/wiki/Service_APIs
- Applications Database (EGI software catalogue)
    - Store information
    - Organise software into clusters

  - Subscription, notification
  - Integrate the catalogue with the EGI platform (e.g. Info system), integrate the catalogue with other catalogues
- Training Marketplace
  - Advertise and search for events, materials, online courses
  - Web gadgets.

A VRC FOR CMMST

The CMMST community is also involved in adopting the VRC model. At European level such effort is grounded on the existing VOs (COMPCHEM, GAUSSIAN and chem.vo.ibergrid.eu) which already recognise the advantages of being a VRC of EGI in terms of access and use of national computing resources that are federated in EGI. To this end there is a need to:

i. design a plan aimed at assembling a structure offering to CMMST researcher VRC out of the existing Chemistry, Molecular & Materials Science and Technology oriented EGI VOs and from the applications, tools and other resources and services that NGIs and projects of EGI provide. ORGANIZZAZIONE LOCALE

ii. identify 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. SERVIZI

iii. develop a governance scheme proposal to establish a new CMMST VRC in EGI. Besides the technical aspects, the proposal will define the organisational and funding models for the VRC. LIVELLO GENERALE

The strong bias to user needs of the Grid infrastructure and organization finds its natural implementation in the fact that grid access is completely free (within the limits of the availability of the resources and the respect of the network etiquette). Accordingly users, once registered after being accredited both by national (of their own country) and a thematic (like the COMPCHEM VO [A. Lagana', A. Riganelli, O. Gervasi, On the structuring of the computational chemistry virtual organization COMPCHEM, Lecture Notes in Computer Science 3980, 665-674 (2006)]) authorities, may choose to run programs and packages already implemented by others on the Grid (*passive users*) or can implement their own programs and utilities (*active users*). In this case the users, which represent a significant fraction of the community members, will have the option either of writing their own scripts or, as an easier choice, of accessing portals, WfMS (like 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) DOI 10.1016/j.future.2010.08.006] that is a 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 more biased towards MMST. Some members of the community may choose to use, for example, SOMA2 that is an Application Oriented Molecular Modelling Workflow (http://soma-server.sf.net/) or simply fish in the application Data Bases (AppDB) of EGI (http://*appdb.egi.eu/*). COMPCHEM is still lacking of a specific portal and will have to evaluate whether to adopt an existing product or assemble its own. In any case, however, the applications listed in Table 2 (together with some of those listed in Table 1) will be made available to the users.

Application	Description	License	Porting in EGI	Service [1]
ABC	Solving the Schrodinger equation for triatomic systems using time independent method	Academic	✔	GriF, Gcres
MCTDH	MultiConfigurational Time Dependent method Hartree	Academic	✔	

FLUSS	Lanczos iterative diagonalization	Academic	✔	
VENUS96	Quasi-classical dynamics of reactive collisions		✔	
DL_POLY	Classical Molecular Dynamics		✔	GriF, Gcres
NAMD	Classical Molecular Dynamics	Academic	✔	
GAMESS-US	General Atomic and Molecular Electronic Structure System	Academic	✔	GriF, Gcres ggamess
RWawePR	Solving the Schrodinger equation for triatomic systems using time dependent method	Academic	✔	GriF, Gcres
GROMACS	GROningen MAchine for Chemical Simulations	Academic	✔	
SCIVR	Semiclassical initial value representation methods	Academic	✔	
CRYSTAL	General-purpose program for the study of crystalline solids	Academic (UK) / Commercial	Work in progress	Work in progress
Framework	**Description**			
GriF	Grid Framework enabling efficient and user-friendly scientific massive calculations	Free		
Gcres	Quality of Users (QoU), Quaily of Services (QoS) evaluation Framework	Free		
ggamess	Front-end script for submitting multiple GAMESS-US jobs	Free		

[1] The Application has been integrated in the listed Frameworks and made available for the CC Community as a Service (Application as a Service -- AaaS)

Table 2 – A list of the MMST packages offered by COMPCHEM as tools, applications or Services

Thanks to the layered structure of COMPCHEM, an even larger mass of in-house or public software produced either by the EGI as such or by its VOs and VRCs (including COMPCHEM itself) will be added to the already mentioned tools and packages made available by the *software providers* (which are users structuring their programs (or suites of programs) for usage by other users) with the purpose of offering the possibility of building more complex applications of general interest out of the software provided by other people and coordinated in common workflows for cooperative usage.

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. As already mentioned, all this is managed by GriF. GriF facilitates the selection of the computing platform more suitable for the intended calculations and enables the composition of collaborative applications as services (some of which may be in competition among them) 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. Among these cooperative goals is also research and long term development and innovation.

TOWARDS A HIGH PERFORMANCE GRID: A PROTOTYPE USE CASE

Fundamental to the implementation of cooperative accurate realistic multiscale computational chemistry applications of higher level of complexity is the ability of Grif of redirecting the jobs to the most appropriate HPC and HTC platforms. This is meant to enable the overcoming of the present highly unsatisfactory situation in which neither HPC nor HTC are completely fit alone to meet the requests of complex MMST applications and this is the way to provide COMPCHEM with higher level services. After all, also on the resource providers side (and not only on the user one) there are good reasons for coordinating the use of HPC and HTC e-infrastructures to the end of interoperating large computational applications. This in fact allows an optimization of the usage of both HTC and HPC computing resources because it not infrequent the case in which a user utilizes HPC platforms not as such but as a bunch of loosely coupled processors underutilizing their fast dedicated network. At the same time HTC users may utilize massively distributed HTC platforms to solve tightly coupled computational tasks ending up by wasting a large amount of time in transferring data on the net. A coordination of the two types of platforms to interoperate via a single workflow (or workflow of workflows) and properly manage the various components on the most appropriate hardware, would instead allow a clever composition of complex applications optimizing the use of the various computing resources and providing the users with the best level of performance.

We have already mentioned the application of GEMS for the simulation of the virtual signal of a crossed molecular beam experiment of interest for combustion [A. Lagana , 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) -> H + CO2 reaction, Faraday Discussion of Chem. Soc. in press 2012; doi: 0.1039/c2fd20046e]. Yet GEMS once implemented on a proper MMST portal can apply equally well to spectroscopy, spectrometry or any other molecular and material based simulation). Quantum detailed simulations of this type require, in fact, nodes (or clusters of nodes) equipped with large (of the order of many GB) memories and processors performing at the level of several Gigaflops because the whole Potential Energy Surface (hereinafter PES) governing the nuclear motion needs to be worked out first. Due to the accuracy required by this type of investigations, in fact, we need to employ high level ab initio quantum chemistry methods and large basis-sets which make each single molecular geometry calculation extremely CPU-time and memory hungry. Moreover, in order to work out a global PES, we need to repeat each single point calculation by varying many times the geometry of the system. Then, once a PES is generated, the kinetic and dynamical data needs to be calculated for a large number of initial conditions. Dynamical calculations (if based on classical mechanics approaches commonly consisting of the integration of a large set of Hamilton (or equivalent) equations) usually do not require large memories and can be run sequentially on off the shelf machines. Yet, they need statistical treatments based on a large number (millions especially in the threshold region of scarcely efficient reactions) of trajectories. Accordingly ab initio and dynamics calculations bear opposite (HPC and HTC) computing requests and the assemblage of the related computational workflow enabling cooperation between HPC and HTC platforms needs to be a key component of the CC community.

An example of a simple HTC – HPC application of GEMS is given in Figure 2 where a skeleton (HTPC1) based on a HTC-HPC scheme that distributes a large quantity of independent tasks on a HTC platform whose outcomes are individually passed as input to an HPC one. As shown by the figure, in the first section of HTPC1 (that is of the HTC type) an emitter (triangle) generates a (large) number of independent events (circles) each of which provides the input for a HPC highly coupled calculation (square). The outcomes of the distributed HPC tasks are returned (lower layer of arrows) and gathered together by a collector (inverted triangle). In case the information collected is insufficient the sequence is further iterated a certain number of times.

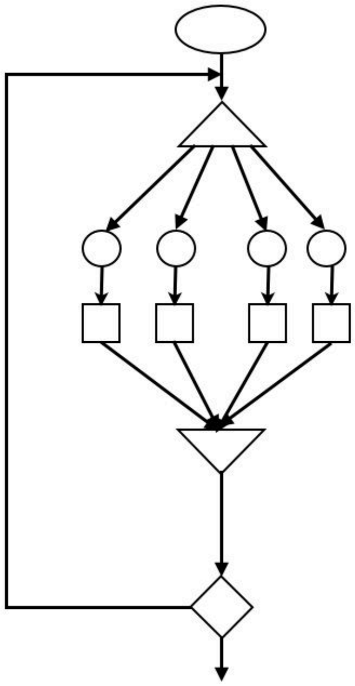

Figure 2 – Skeleton HTPC1: a HPC computation following a HTC one

An example of a strongly coupled treatment implemented on a HPC platform followed by the distribution of a large quantity of independent tasks to be calculated on a HTC platform (skeleton HTPC2) is given instead in Figure 3. Also in this case the sequence of the two sections is checked against convergence and further iterations are performed (with the associated switch between the two platforms) until either convergence or a maximum number of iterations has been reached.

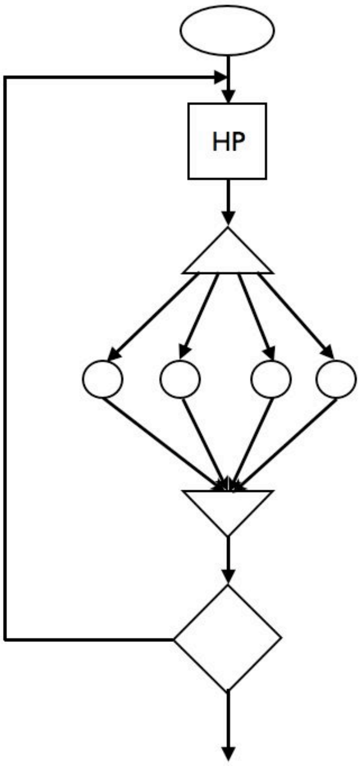

Figure 3 - Skeleton HTPC2: a HPC computation preceding a HTC one

The systematic sampling and construction of a full dimensional PES (and the evaluation of some related properties) using a high level ab initio electronic structure package (see in Table 1 Dalton, Gamess, Gaussian, Molcas, Molpro, Nwchem, etc.) for complex molecular systems encompasses such scheme [L. Storchi, F. Tarantelli, A. Lagana', Computing Molecular energy surfaces on the grid, Lecture Notes in Computer Science 3980, 675-683 (2006)]: generation on a HTC platform the molecular geometries to be considered and then launch of the ab initio package for each of them on a HPC one. On the contrary HTPC2 scheme for calculating kinetic coefficients by quantum mechanical flux correlation functions is amenable to packages using a Multi Configuration Time Dependent Hartree scheme.

TOWARDS A HIGH PERFORMANCE GRID: METODOLOGY

To implement the above mentioned use-cases the following steps need to be undertaken [https://indico.egi.eu/indico/contributionDisplay.py?contribId=157&confId=1019]:
- allow access to the involved platforms by the users, integrate the HTC and the HPC workflows,
- integrate the application into the workflows,
- design and implement a bridging tool for the combined HTPC platform,
- develop and implement services on the HTPC platform,
- optimize the applications for the HTPC platform.

This can be performed by implementing the above mentioned steps through two levels of intervention on the overall computing framework having distinctive characteristics (see fig.4):

- a research community or a VO layer in which the case study will be prepared, implemented and run.
- an infrastructure layer to be managed by the resource providers to allow access to computing and storing resources by enabling different resources of HTC and HPC type belonging to different middleware to interoperate.

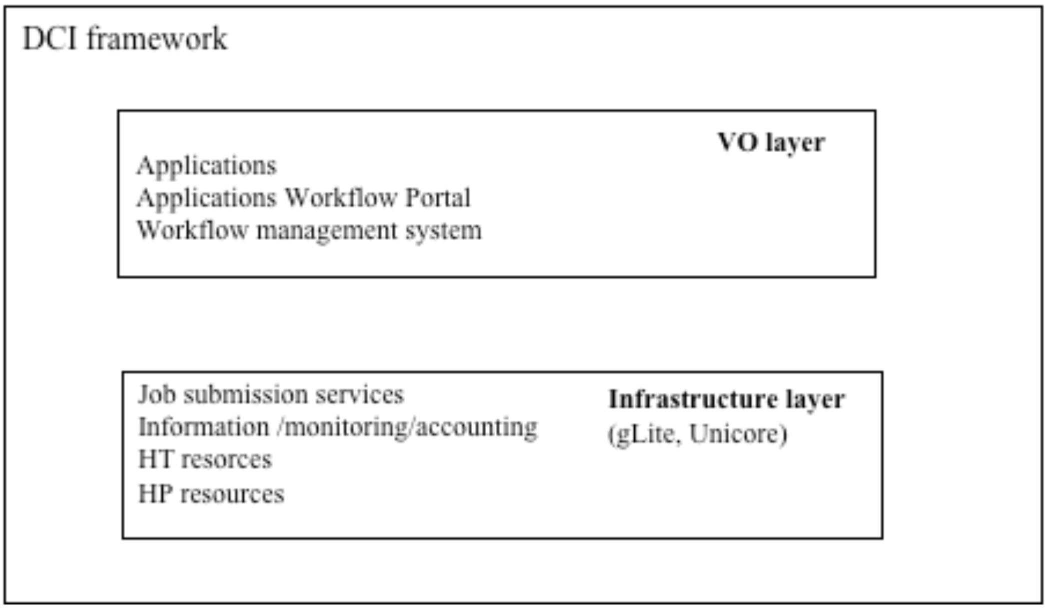

Figure 4 – Synoptic view of the service and research activities

This is a perspective that the CC community is aiming at. To this end COMPCHEM and other CC VOs will be encouraged to work in closer synergy to form a Virtual team of EGI and to gather together the necessary effort to assemble a proposal for a EU project. To this end an advanced Workflow Management System (WfMS) capable to represent and manage mixed HTC and HPC computational applications will be assembled or borrowed from existing EGI middleware (say the most recent EMI releases (http://www.eu-emi.eu/middleware) for example). Within the project the user, through a specific portal, will be enabled to input the workflow structure to the WfMS, that, by interoperating with a job submission service (like the Workload Management System of gLite), will allow the submission of HTC and HPC computing tasks to the most suitable computing resource available in the various distributed computing infrastructures. In the middleware of the platforms the abstract interfaces through which computing resources announce themselves will be made interoperable using the potentialities of the framework GriF.