# VHERLA: A VIRTUAL MOLECULAR SCIENCE DATA CENTER ALLOCATED ON THE GARR CLOUD

Giuseppe Vitillaro, CNR ISTM - UOS Perugia

Antonio Laganà, Dipartimento di Chimica, Biologia e Biotecnologia, University of Perugia

## Abstract (introduction)

VHERLA, the images of HERLA (an infrastructure assembled in Perugia for the activities of the CMS$^2$ consortium aimed to carry out research on Molecular Sciences and training students on relevant distributed and parallel computing) generated in June 2017 for an Open Science Cloud School and running on the OpenStack platform is presented. VHERLA, has proven to work properly in a Cloud environment as a foundation infrastructure of the local Molecular Science Community, in order to develop applications for Open Molecular Science Cloud services. In this paper the upgrading of the activities of the **MOSEX (MOlecular Simulator Enabled Cloud Services)** project, previously proposed as a **European Open Science Cloud Pilot** to the end of generating and collecting as a service validated Molecular Science data on VHERLA, is discussed. In the project, some already implemented applications (like **GEMS,** a distributed workflow of programs simulating molecular scattering processes) are illustrated and its possible extensions to other Molecular Science packages proposed during the last meeting of the Computational and Theoretical Chemistry Division of EUChemS (August 2018) as possible additional components of MOSEX, are considered to strengthen the collaborative support to the high level calculation of molecular properties and to the validation of data (both experimental and theoretical) to be made available in an Open Science scheme on the cloud.

## 1-Introduction: meta- and grid-computing

In the recent past, the members of the Molecular Science community started being active in combining, through collaborative initiatives, research activities and ICT technologies within the stream of grid computing European projects. In particular, significant efforts have been spent to the end of assembling concurrent high level ab initio calculations enhancing the "realism" of Molecular Science simulations. Relevant European activities in this field started at the beginning of the present century thanks to COST (www.cost.eu/), ECTN (www.ectn.eu/) and EuChemS (https://www.euchems.eu/).

As to COST, Action D23 (METACHEM, www.cost.eu/COST_Actions/cmst/D23 launched by the University of Perugia in the year 2000) initiated the development of tools connecting the activities of different Molecular Science research laboratories operating on a shared computing platform made of a geographically distributed cluster of heterogeneous computers connected on a network (LAN, MAN or WAN) through a software co-ordinating them as a single virtual parallel machine [1]. This established a network of European Meta Laboratories fostering innovative solutions for chemical applications and a new paradigm for collaborative research, making it feasible to develop new a priori realistic

ISSN: 2279-8773

simulations for several scientific, technological and environmental applications through the working groups on Multi-reference quantum chemical methods, 4-component relativistic quantum chemical calculations, a priori simulation of crossed molecular beam experiments, Quantum Mechanical studies of structure, dynamics and spectroscopy of systems relevant to environment, materials, energy, education, etc. in Chemistry. The COST Action METACHEM ended in the year 2005.

In the year 2006 a new COST Action (D37 | Grid Computing in Chemistry: GRIDCHEM), aimed to develop grid solutions and paradigms for molecular science research, sprouted out of D23. GRIDCHEM leveraged the creation and use of distributed computing infrastructures ("Grids") and drove collaborative computer modelling and simulation in chemistry towards new frontiers in complexity and a new regime of time-to-solution [2]. The areas of application covered traditional chemistry, materials science, molecular biology and environmental chemistry through the working groups Photochemistry and photobiology, Dynamics engines for Grid molecular simulators (for this see Fig.1), distributed e-learning

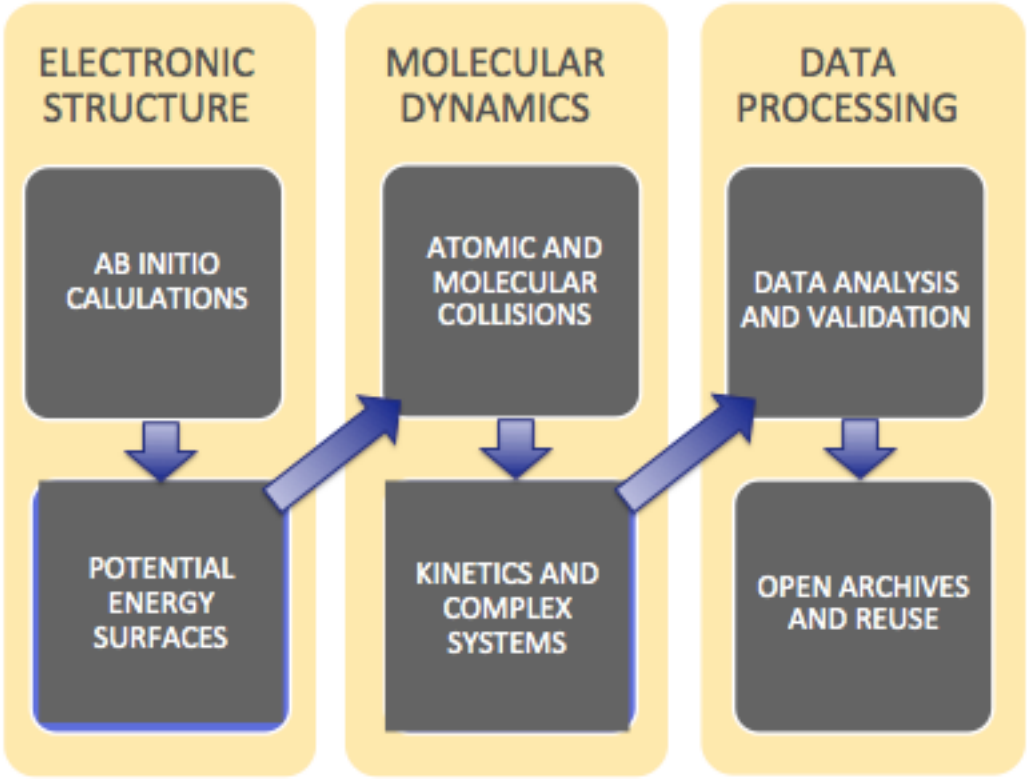


**Fig. 1** – The Molecular Dynamics Engine of the GRIDCHEM

in chemical knowledge transfer and assessment, codes interoperability in quantum chemistry, and workflows in chemical data production and validation. In other words, the GRIDCHEM action enabled and accelerated the transition of

Molecular Science to the "infrastructure of computational science of the 21st century".

## 2-Open Molecular science for competence dissemination

As to ECTN, its VEC (Virtual Education Community) Committee promoted the social use of Chemistry knowledge via Open collaborative Educational initiatives among its members. The main VEC initiatives are the **EChemTest®** and **GLOREP** cloud services leveraging an ECTN international network of both National Test Centres (NTC)s (and Accredited Test Sites (ATS)s) [3] and repositories. In more detail **EChemTest®** offers computer based Self Evaluation Sessions (SES)s at NTCs and ATSs computer laboratories under controlled conditions. NTCs and ATSs located at the Higher Education Institutions (HEI)s having signed with ECTN an ad hoc agreement within which the HEIs act both as users (utilizing relevant services for their academic activities) and as producers (contributing to the development of relevant tools) [3] using the Open Source software EOL (Exams On Line) http://echemtest.libreeol.org/doc/ECHEMTestGuide.pdf).
Levels of competence in Chemistry are articulated as follows: **Pre-University Level-1** for persons at the end of compulsory education (*General Chemistry 1 (GC1)*), **Pre-University Level-2** for persons at the beginning of Chemistry related University studies (*General Chemistry 2 (GC2)*), **University Bachelor Level-3** for persons at the end of the Core Chemistry Syllabus at the University Level as defined in the «**Chemistry Eurobachelor®**» (*Analytical Chemistry 3 (AC3)*; *Biological Chemistry 3 (BC3)*; *Chemical Engineering 3 (CE3)*; *Inorganic Chemistry 3 (IC3)*; *Organic Chemistry 3 (OC3)*; *Physical Chemistry 3 (PC3)*) and **University Master Level-4** for persons at the end of a Master degree in one of the specialized chemistry areas in agreement with the «**Chemistry Euromaster®**» requirement (*Computational Chemistry 4 (CC4)*; *Conservation Science 4 (CS4)*; *Advanced Organic Chemistry 4 (AOC4)*). On its side **GLOREP** offers a distributed repository for storing and handling Learning Objects (LO)s, audio-visuals, virtual reality applications and MOOCs (Massive Open Online Courses) designed to enhance Chemistry knowledge among students, citizens and professional workers. Support to the above mentioned initiatives has been given also by EuChems (the European Society of Chemical Societies) through its Computational Chemistry inter-divisional group that, in order to gain greater authority in carrying out these tasks, was promoted to the level of Division of Computational and Theoretical Chemistry (CTC).
…..The Open collaborative user/producer model adopted for the above mentioned Molecular Science knowledge processes (Educational in the specific case) is often referred to as Prosumer [3,4]. In the Open collaborative scheme of the Prosumer model the activities of the members continuously feed new information into the system (as producers) improving so far its quality by leveraging the feedbacks

obtained from usage and fostering its progressive evolution from tacit to explicit knowledge. The main features of such Open knowledge scheme are: self-consistency, modularity, availability, reusability, interoperability with no locality and no (or extremely limited) ownership constraints as is typical of cloud services. These features are highly desirable as they give utmost thoroughness to information allowing them to be flexibly aggregated and readily used (and re-used) thanks also to the association with metadata which allow to verify the existence and the properties of the information themselves.
The possibility of connecting and handling the various components of the Molecular Science resources for an adequate and full fruition has called (as is also the case of other disciplines) for the adoption of a powerful research environment bearing the proper cloud functionalities.

## 3-Towards a suitable Molecular Science Collaborative Virtual Research Environment

The answer to the request for a suitable Virtual Research Environment (VRE) was provided by the EGEE (Enabling Grids for E-sciencE) first and the EGI (European Grid Infrastructure) projects. These projects were, indeed, successful in:
-proving that a globally distributed computing Grid plays an essential role for large-scale, data intensive science in many fields of research;
-consolidating the operations and middleware of the European Grid for use by a wide range of scientific communities (such as astrophysics, computational chemistry, earth and life sciences, fusion and particle physics, etc.).
As a matter of fact, EGEE (in the first decade of the years 2000) became a unique and powerful resource for several European sciences by allowing researchers to have a plaform for scientific collaboration on common challenges for several disciplines. As to Molecular Science, during the EGEE-III project, a pilot computational application was assembled (see Fig. 1) to prove that accurate calculations of cross sections and rate coefficients of reactive and non reactive molecular processes can be efficiently and cooperatively performed by distributing on the GRID high level ab initio calculations of structured Potential Energy Surfaces (PES)s for a very large number of molecular geometries [5] and of trajectories [6]. This prompted the establishing of the COMPCHEM Virtual Organization (VO) [7] and to the generalization of the distributed procedures computing the cross sections and the rate coefficients of reactive and non reactive molecular processes into the Grid Empowered Molecular Simulator **GEMS** [8] by composing:

**3a1.** the production and/or collection of high level ab initio information on the electronic structure of the involved molecular system;

**3a2.** the fitting of available data using appropriate short and long range formulation of the PES;

**3a3.** the checking, correcting and coding of the PES into a high performing routine;

**3a4.** the calculating of an extended set of detailed dynamical quantities and their averaging to evaluate the desired observable

Next step was the establishing of the EGI CMMST (Chemistry, Molecular and Materials Science a Technologies) VRC (Virtual Research Community) [9] within the activities of the IGI (http://www.italiangrid.it) Joint Research Unit (JRU) (made of more than 20 Italian academic and research institutions) of the EGI-INSPIRE [10] project (https://www.egi.eu/about/egi-inspire/). The purpose of this was to graft the activities of the Molecular Science community on an operational environment providing a proper set of services. In this spirit, the CMMST VRC, in spite of being mainly a loosely coupled aggregate of independent Chemistry laboratories, was able to develop and exploit a further set of collaborative services. In particular, the CMMST VRC was enabled to:

**3b1.** orchestrate the activities of the e-infrastructure experts and of the members of the involved communities so as to enable an effective intra- and trans-community networked implementation and coordination of a collaborative/competitive (synergistic) research environment by:
- allowing a selection of the compute resources based on quality parameters and a composition of higher level of complexity chained applications through the coordinated usage of distributed hardware and software,
- fostering the use of specialized web portals and workflows facilitating the production of data and know how in science and innovation, the direct re-use of the produced data and knowledge in education, training and further research
- rewarding the work done by its proactive members on behalf of the community.

**3b2.** produce and provide computational services useful to molecular and materials disciplines when carrying out multi-scale treatments necessary to reproduce the observables of realistic systems in the area of energy, environment, materials, pharmacology, chemistry, biology, biotechnologies, medicine, etc. by means of:
- state-of-the-art first principles electronic structure and nuclei dynamics computations,
- high level of accuracy multiscale design of complex molecular systems,
- knowledge management for training and education in sciences and technologies.

**3b3.** turn (in collaboration with partner SMEs) the versatility of the adopted e-

infrastructure tools, the richness of the developed CMMST knowledge and the credit mechanism supporting the synergistic operating into a business model enabling an efficient transfer of the activities to the market ensuring business sustainability Accordingly, the CMMST VRC would collect the requirements, validate the developed procedures, disseminate them, carry out related integration, user support and knowledge transfer, adopt existing scientific gateways, workflows, data management and commons and associate them with the synergistic model and a credit based economy.

An important outcome of the COMPCHEM VO and CMMST VRC activities were:
**3c1.** the First Training Workshop on Grid porting of computational chemistry applications. The event, held in Rome (January, 2014) was hosted by the GARR consortium (http://www.garr.it/b/eng) and was devoted to the discussion of the porting on the grid of some Molecular Science applications by exploiting the services and resources of the Italian National Grid Infrastructure available within the gLite middleware. Three real life use cases were considered for that purpose:
**-VENUS**, a classical trajectory direct dynamics code developed at the Texas Tech University (https://cdssim.chem.ttu.edu/nav/htmlpages/licensemenu.jsp),
**-CRYSTAL**, a quantum chemistry program for solid state physics and chemistry (http://www.crystal.unito.it/index.php) developped at the University of Torino, Italy and
**-QUANTUM ESPRESSO**, (http://www.quantumespresso.org/).an open-source software for electronic structure calculations and materials modeling at the nanoscale developed by the CNR-IOM DEMOCRITOS National Simulation Center in Trieste (Italy).
**3c2.** The launch (under the coordination of the Universidad Autonoma of Madrid) of the Theoretical Chemistry and Computational Modeling (TCCM) European Master and of the homonymous European Doctorate (https://tccm.qui.uam.es/). These two initiatives have played in the last 15 years a key role in forming a new generation of young researchers tackling the development of modern chemistry, biochemistry, chemical biology and material science applications (i.e. state-of-the-art first principles electronic structure and nuclei dynamics codes, to the implementing of accurate multi-scale simulators of smart energy carriers in combustion, energy storage, space missions, bioinorganic chemistry, using both ab initio and empirically parameterized kinetic data, designing materials and modelling supra-molecular phenomenology, handling extended information systems for the investigation of the structure and processes of complex molecular systems relevant to pharmacology, medicinal and biological systems, and managing distributed knowledge processes, etc.) with particular focus on Open Molecular Science Cloud approaches. This has resulted into the School of Open Science Cloud, (SOSC) (see https://indico.cern.ch/event/605204/overview and

http://services.chm.unipg.it/ojs/index.php/virtlcomm/issue/view/25) held in Perugia on June 2017 and into specific sessions of the one month year long Intensive Course (IC) of the last edition of the already mentioned TCCM Erasmus Mundus Master (September 2018). Work is also in progress to assemble a H2020c Open Molecular Science Cloud project (to be illustrated at the European Conference of the CTC division of EuChemS to be held in Perugia from the 1st to 5th September 2019 and at the following satellite workshop to be held at the Accademia delle Scienze (dei 40) in Rome).

## 4- The Implementation of the Open Molecular Science Cloud

….. According to the scheme *Hypothesis → Data collection → Processing → Storing data and results → Long term preservation → Publication and distribution → Reuse* (https://www.fosteropenscience.eu/content/what-open-science-introduction) published by Gema Bueno de la Fuente, Open Science Cloud does indeed represent a **new approach to scientific process**. The Open Science approach is based on cooperative work and new ways of diffusing knowledge by using digital technologies and new collaborative tools. This implies that *the primary outputs of publicly funded research results, publications and research data should be publicly accessible in digital format with no or minimal restriction* so as to extend the principles of openness to the whole research cycle. This fosters knowledge sharing and collaboration as early as possible and entails a systemic change to the way science and research are done.

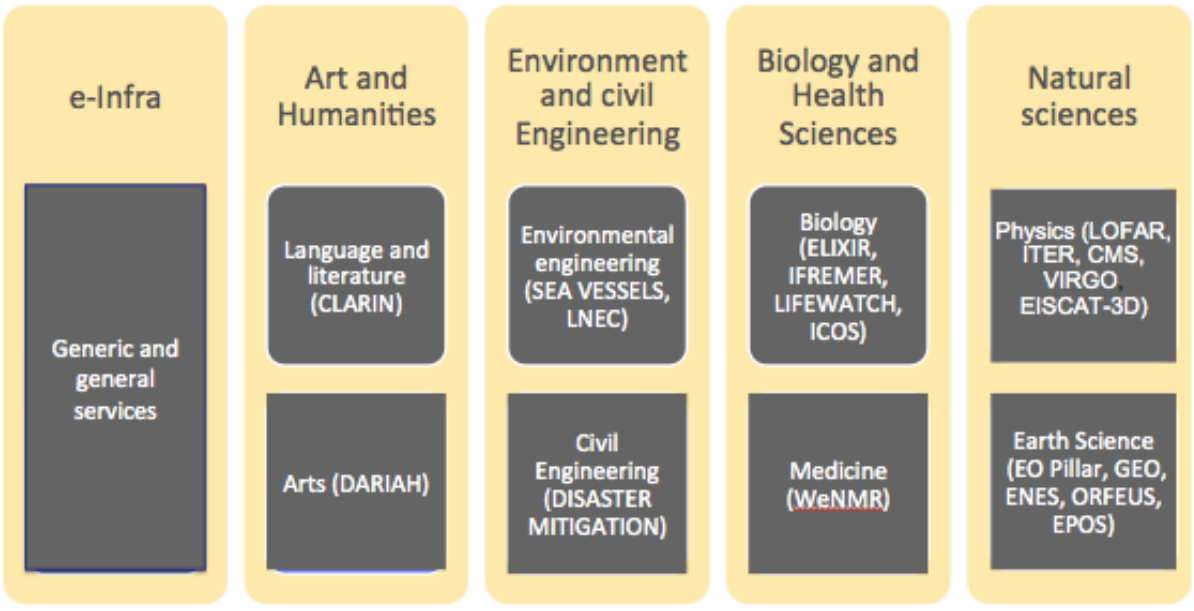


**Fig. 2** – EOSC-Hub service providers

The European Open Science Cloud (EOSC) H2020 project (EOSC-Hub) launched

in January 2018 with a budget of 33 Meur is indeed an initiative designed to provide European researchers with a virtual environment supporting the storing, managing, analysing and reusing of data for research, innovation and education. This will happen by creating common interfaces and standard as well as maintenance, interoperability and sustainability structures for data, protocols and methodologies. Unfortunately, as shown in Fig. 2, in the Natural science block (the rightmost one) of the project the Chemistry and Molecular Science sub-fields, in which both the CMMST VRC and the COMPCHEM VO operate, were not included. This has forced the Molecular Science Community to find its own way and search for proper partners to develop the above mentioned virtual environment.
For this reason, an embryonic cloud platform, named VHerla, was first established. at the end of the year 2013 at the Dipartimento di Chimica, Biologia e Biotecnologia (DCBB) of the University of Perugia and then extended two years later into that of the **Consortium of the University of Perugia** (DCBB, Dipartimento di Matematica ed Informatica (DMI), Dipartimento di Fisica e Geologia (DFG)) and **INFN Perugia,** CNR **ISTM - UOS Perugia** plus **Master-UP srl** and **Molecular Horizon srl**. Herla has been built as a "Beowulf Model", a paradigm locally developed from the end of the '90s, aimed to solve HPC problems for research in Molecular Science and training in parallel computing. The platform consisted in a couple of HPC clusters, (CG/training) and (FE/research), running Scientific Linux 6.x, with two distinct access nodes, one physical (cgcw/CG/32-bit) and one virtual (fecw/FE/64-bit/VMWare). The "training (CG) cluster" made available a small set of resources (10 nodes, 22Gb of memory, 11 cores, 2Tb of storage) to students, while the "research (FE) cluster" made available a larger set of resources (13 nodes, 52 cores, 2GPU, 5Tb of storage) to scientists, for research projects. The clusters were connected using NIS, in a "single image system".

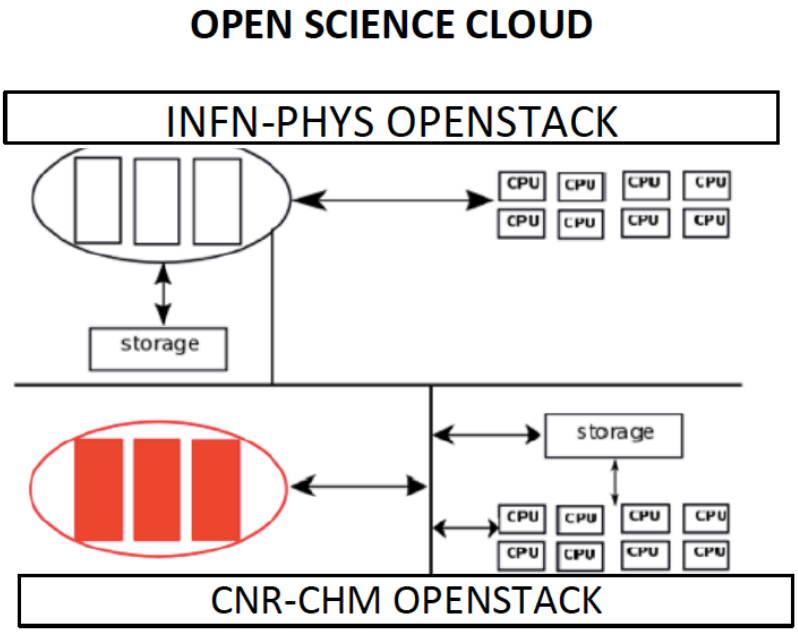


**Fig. 3** – The Herla (INFN-PHYS and CNR-CHM) platform

As a first step towards the creation of a cloud image for the Perugia Consortium (named at this point CMS[2]) infrastructure, in collaboration with the Department of Physics and geology **(DFG)** and **INFN Perugia,** the **VHERLA** images were deployed on a CEPH storage (locate at DCBB), on which the activities of the School Open Science Cloud (SOSC17) held in Perugia on June 2017 running under the INFN OpenStack platform, were carried out (see Fig. 3). The next step of the process consisted in allocating early this year, **a Virtual Data Center** on the **GARR Cloud** to the end of generating a virtual cluster for Molecular Sciences, with the support of A. Barchiesi (GARR CSD) and G. Attardi (GARR Cloud). Next, during July 2018, the "cnr-istm" project was allocated on the GARR Palermo node (see Fig. 4), with the assignment of the following resources: 8 istances, 128 VCPU, 384Gb RAM, 1Tb Storage and 10 Volumes

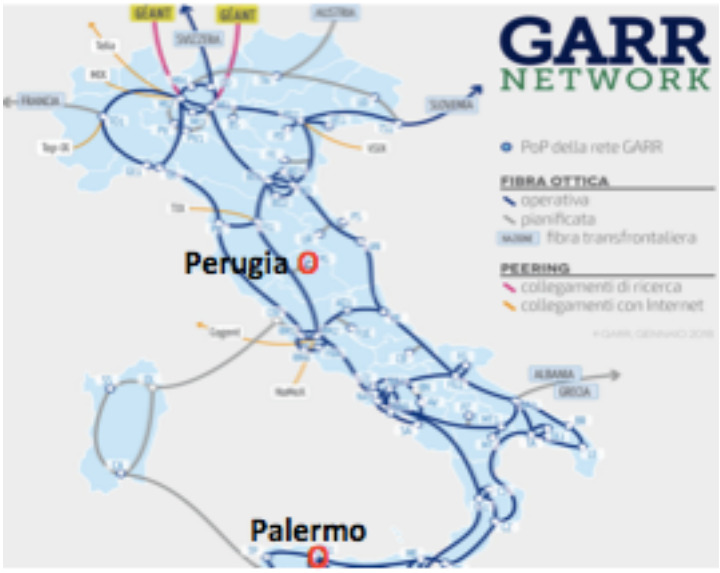


**Fig. 4** – GARR Network (Perugia and Palermo are marked in red)

Later, the images of VHERLA (previously located at DFG and INFN Perugia) were transferred on the OpenStack GARR Cloud platform within the project "cnr-istm" and were used to install the version VHERLA(GARR-CLOUD) hscw (http://hscw.herla.unipg.it/ganglia/?p=2&c=FrontEnd) bearing the following features:

Access Node hscw (2 core, 4Gb RAM, 200Gb storage)
Cluster (7 nodes, 96 cores, 360Gb RAM, 700Gb storage).

The access node, hscw, can be reached at the IP address [90.147.189.20] via SSH and the following 7 nodes, 96core, ~380Gb, 512Gb scratch are defined at Torque(PBS)/MAUI as [Intel Xeon E3-12xx v2(Ivy Bridge)/2.6Ghz]

```
hs01 np=16  t100 x86_64
hs02 np=16  t100 x86_64
hs03 np=16  t100 x86_64
hs04 np=16  t100 x86_64
hs05 np=16  t100 x86_64
hs06 np=8   t100 x86_64
hs07 np=8   t100 x86_64
```

with the first 5 nodes (01-05) having 64Gb/16core, hs06 having 32Gb/8core and hs07 having 16Gb/8core (see Fig. 5). It is important to point out here that the above mentioned nodes are not Infiniband connected (300Mb/sec bandwidth, 80 μs latency) implying a computing performance degradation when inter-node communication is required.

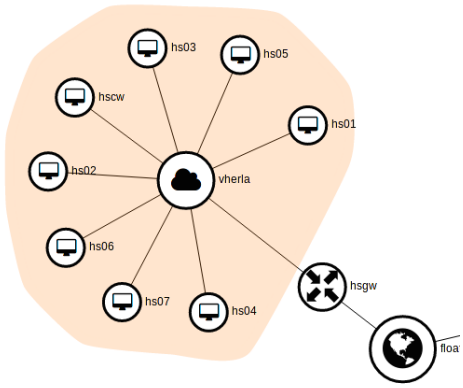


**Fig. 5** – Scheme of the seven nodes of the EOSC-Hub service provider of VHERLA

## 5. VHERLA MOLECULAR SCIENCE APPLICATIONS

During August 2018 VHERLA(GARR-CLOUD) passed base tests and benchmarks of the following Herla Molecular Science applications:

- **QUANTUM ESPRESSO**/39.62
- **GAUSSIAN**/09-c01-omp
- **NWCHEMS**/6.5.26243
- **GAMESS-US**/050113R1
- **MOLPRO**/2010p-omp
- **SIESTA**/3.2-pl-4

with satisfactory results for single node calculations (yet with unsatisfactory MPI parallel performances for multiple concurrent multiple nodes applications). As already pointed out this is due to the limited bandwidth and high latency of the GARR-CLOUD "cnr-istm" nodes.

Further tests were carried out in the period September 3-20, 2018 when VHERLA

was mainly devoted to the activities of the 13th EM (European Master) TCCM (Theoretical Chemistry and Computational Modelling) intensive course (http://www-old.chm.unipg.it/chimgen/mb/theo2/TCCM2018/EM-TCCM2018/EM-TCCM/Welcome.html), with the reference web page being at http://hscw.herla.unipg.it URL.

During the tests, production calculations have been planned (for the period ranging up to 13/01/2019) for the software of **MOSEX (MOlecular Simulator Enabled Cloud Services)** the previously proposed **European Open Science Cloud Pilot** designed as a follow-up of the EGI COMPCHEM Virtual Organization (VO) activities. MOSEX aims to gather together and offer as a service on the cloud for the Molecular Science (as reported at the last meeting of the CTC Division Board of EuChemS at the end of August 2018 in Liverpool and will be more in detail planned during the forthcoming OMSC (Open Molecular Science Cloud) meeting next September at the "Accademia Nazionale delle Scienze (dei XL)" in Rome) by implementing the Prosumer Model for the following suggested applications:

**5a1. Molecular electronic structure and dynamical properties programs**
**GEMS** (small molecules chemical processes (including ab initio determination of analytical Potentials), their Quantum/Classical dynamics efficiency and properties obtainable from the analysis of the wavefunction), **SHARC** (Molecular Dynamics on excited states), **ORCA** (electronic structure), **TURBOMOLE** (electronic structure), **VMS:** (Virtual Spectrometer), **GROMACS MD** (Molecular dynamics), **VENUS** (Quasiclassical Molecular dynamics), **QCL** (Quantum-classical dynamics)
**5a2. Drug design programs and cloud services** (QSPR and 3D QSPR models)
**5a3. Distributed repository of molecular science data**
**CHEMCONNECT (**organization, publication and storage of molecular information on combustion), **IOCHEM** (organization, publication and storage of molecular information on materials)
**5a4. Dissemination and evaluation of molecular knowledge**
**MolSSI** (Molecular science software tools development and training), **GLOREP** (Distributed repository of shared Learning Objects and educational tools), **LIBREEOL (**EChemTest distributed software for assessing, evaluating and certifying Chemical knowledge using e-tests under controlled conditions), **VIRT&L-COMM** (e-magazine for virtual community activities dissemination).

MOSEX is also committed to improve the services offered on the cloud for the Molecular Science members by implementing the following extensions and enhancements of the project

**5b1. Extension of the project "cnr-istm" on GARR-CLOUD to the end of the**

ISSN: 2279-8773

**year 2019 (or replacement with a new project)** with associated re-definition of hardware (mainly storage and flavors)
**5b2. Availability of Infiniband** (or at least Ethernet 40/100 Gbe) for "cnr-istm" in order to exploit parallelism in production runs
**5b3. Availability of GPUs** on the platform
**5b4. Support by GARR-CLOUD** to implement on DCBB/ISTM/PG a local OpenStack/CEPH for federation with GARR-CLOUD and extension to DFG and DMI
**5b5. Transfer of VHERLA** from GARR-CLOUD to the DCBB/ISTM/PG cluster with the termination of "cnr-istm" or definition of continuation condition for the "cnr-istm" project.

## 6. CONCLUSIONS

By leveraging the VHERLA infrastructure assembled in Perugia for training the students on distributed Molecular Science computing, an Open Science Cloud has been established to the end of activating relevant services and applications. A specific focus has been devoted to the Grid Empowered Molecular Simulator (GEMS) a distributed workflow of programs simulating molecular scattering processes. The extension of the Cloud services to the components of MOSEX has also been considered in order to enhance the quality of support to the high level calculation of molecular properties and to the validation of related data (both experimental and theoretical) made available in repository for use in large simulations (e.g. combustion, energy storage, material science, astrochemistry, etc.. Future connections of the initiative with a possible initiative of the Italian GARR infrastructure is being considered.

## REFERENCES

1] I. Foster, C. Kesselman Eds., The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann Publ., San Francisco (1999).
2] I. Foster, C. Kesselman, S. Tuecke, The anatomy of the Grid: Enabling Scalable Virtual Organisations, International J. Spercomputer Applications, 15(3), 200-222 (2001).
3] A. Laganà, O. Gervasi, S. Tasso, D. Perri, F. Franciosa, The ECTN Virtual Education Community prosumer model for promoting and assessing chemical knowledge, Lecture notes comput science 10964, 533-548 (2018).
4] I. Nonaka, H. Takeuchi, The Knowledge-Creating Company, Oxford University Press (1995).
5] L. Storchi, F. Tarantelli, A. Lagana', Computing Molecular energy surfaces on the grid, Lecture Notes in Computer Science 3980, 675-683 (2006).
6] O. Gervasi, A. Lagana', SIMBEX: a portal for the a priori simulation of crossed beam experiments, Future Generation Computer Systems, 20(5), 703-716 (2004).
7] A. Laganà, A. Costantini, O. Gervasi, N. Faginas Lago, C. Manuali, S. Rampino, COMPCHEM: progress towards GEMS a Grid Empowered Molecular Simulator and beyond, Journal of Grid Computing, 8(4), 571-586 (2010).
8] S. Rampino, F. Pirani, A. Lagana', E. Garcia, A study of the impact of long range interactions on the N + $N_2$ reactivity using GEMS, Int. J. Web and Grid Services 6, 196-212 (2010).
9] https://wiki.egi.eu/wiki/Towards_a_CMMST_VRC
10] https://wiki.egi.eu/wiki/EGI-InSPIRE:Main_Page