PROCEEDINGS OF SCIENCE

Porting workflows based on small and medium parallelism applications to the Italian Grid Infrastructure

Daniele Cesini[1]
INFN-CNAF
V. B. Pichat 6/2; Bologna,Italy
E-mail: daniele.cesini@cnaf.infn.it

Marco Bencivenni
INFN-CNAF
V. B. Pichat 6/2; Bologna,Italy
E-mail: marco.bencivenni@cnaf.infn.it

Alessandro Costantini
INFN and University of Perugia, Italy
E-mail: alessandro.costantini@pg.infn.it

Emidio Giorgio
INFN-Catania
Via Santa Sofia 64, Catania, Italy
E-mail: emidio.giorgio@ct.infn.it

Giuseppe La Rocca
INFN-Catania
Via Santa Sofia 64, Catania, Italy
E-mail: giuseppe.larocca@ct.infn.it

Vania Boccia
INFN-Napoli
Via Cintia, I , Napoli, Italy
E-mail: vania.boccia@unina.it

Roberto Alfieri
INFN and University of Parma
Viale G.P. Usberti n.7/A (Parco Area delle Scienze) - 43124 Parma, Italy
E-mail: roberto.alfieri@unipr.it

[1] Speaker

PoS(ISGC 2013)024

PROCEEDINGS OF SCIENCE

Roberto De Pietri
INFN and University of Parma
Viale G.P. Usberti n.7/A (Parco Area delle Scienze) - 43124 Parma, Italy
E-mail: roberto.depietri@unipr.it

Luciano Gaido
INFN-Torino
Via Pietro Giuria, 1, Torino, Italy
E-mail: luciano.gaido@to.infn.it

Alessandro Venturini
CNR-ISOF Bologna
Via P. Gobetti 101, Bologna - Italy
E-mail: alessandro.venturini@isof.cnr.it

Stefano Ottani
CNR-ISOF Bologna
Via P. Gobetti 101, Bologna - Italy
E-mail: stefano.ottani@isof.cnr.it

Andrea Buzzi
CNR-ISAC Bologna
Via P. Gobetti 101, Bologna - Italy
E-mail: A.Buzzi@isac.cnr.it

Piero Malguzzi
CNR-ISAC Bologna
Via P. Gobetti 101, Bologna - Italy
E-mail: P.Malguzzi@isac.cnr.it

Daniele Mastrangelo
CNR-ISAC Bologna
Via P. Gobetti 101, Bologna - Italy
E-mail: D.Mastrangelo@isac.cnr.it

The Italian Grid Infrastructure (IGI) is one of the National Grid Initiatives (NGIs) composing the European Grid Infrastructure (EGI) and provides computational and storage resources to scientific communities belonging to various domains and disciplines. Starting from the early 2000s, the infrastructure was originally shaped on the Large Hadron Collider community needs, mainly addressed by the paradigm of the High Throughput Computing (HTC) which is able to fully exploit the geographically distributed resources running sequential applications. The support for parallel codes was therefore initially limited. In the last decade, however, facilitated by European projects such as the EGEE series (Enabling Grid for eScience in Europe) and EGI-InSPIRE, the usage of the infrastructure was significantly extended to other scientific communities, requiring new applications and new types of workflows to be ported to the Grid.

PoS(ISGC 2013)024

PROCEEDINGS OF SCIENCE

As a consequence the request to support non-embarrassingly parallel applications increased. Moreover, the latest hardware enhancements, in particular the wide spread of multicore processors, boosted the request for High Performance Computing (HPC) support by the Grid. This support is now greatly improved also taking advantage from new Grid middleware capabilities fostered by dedicated working groups within both EGI and the NGIs, including IGI.

In this paper we will present the achievements obtained by recent collaborations between IGI (its Training and User Support unit in particular) and several user communities in exploiting the IGI resources to run a set of case studies provided by the user communities, whose workflows require the execution of parallel applications.

Within such collaborations, IGI provided not only the necessary resources but also the know-how to appropriately modify the applications and make them suitable for an effective and efficient use of the distributed computing environment.

The work made and the adopted solutions are of great interest across the various scientific domains, ranging from Computational Chemistry, to Geophysics and Bioinformatics. Moreover the parallel applications have been selected in order to cover a wide range of technological challenges for what concerns resource usage and software requirements.

The paper shows to what extent some types of HPC is now feasible on the Italian Grid comparing the speed-up factors obtainable with parallel runs. Since not all the Grid sites support parallel jobs, for some use cases effort has been spent to combine HTC (or scalar) and HPC (or parallel) runs of the same application to maximize the exploitation of the infrastructure.

Focus is also given to the issues encountered during the porting process, in particular those concerning the resource discovery and the tuning of the executables to the underlying hardware infrastructure. Possible further improvements from an user point of view, such as better support to accelerators (manycores processors and GPUs) and interoperability with supercomputing, as in the case of Computational Chemistry domain, is also discussed.

The International Symposium on Grids and Clouds (ISGC) 2013
March 17-22, 2013
Academia Sinica, Taipei, Taiwan

PoS(ISGC 2013)024

1. Introduction

The wide spread of Grid technologies for research communities in Europe started in the early 2000s pushed by big experiments and projects, in particular in the field of high energy physics (HEP). To address the computational and storage needs of the Large Hadron Collider (LHC) at CERN a worldwide infrastructure was created (the Worldwide LHC Computing Grid, WLCG) granting thousands of scientists access to distributed resources and exploiting the paradigm of High Throughput Computing (HTC).

In the last decade, fostered by many projects supported by European Commission , such as the EGEE project series [1] and EGI-InSPIRE [2], new user communities started adopting the same Grid infrastructure to address their computational problems. New users communities means also new requirements. Among these requirements, the support needed by strongly coupled parallel applications was probably the one with the greatest impact on the infrastructure, on its operations and on the middleware used to run its services. Only recently, thanks to the availability of multicores processors, to the introduction of new middleware features and to the efforts of various working groups in this field, including a Virtual Team within the EGI-InSPIRE project [3], parallel computing on the Grid is becoming attractive.

The Italian Grid Infrastructure (IGI or NGI_IT) [4], which is one of the biggest National Grid Initiatives (NGI) of the European Grid Infrastructure (EGI), in the last couple of years improved the High Performance Computing (HPC) support to address the needs of communities belonging to various disciplines. In this paper we will present real life use cases based either on parallel applications or on workflows containing parallel applications that successfully run on the IGI infrastructure. The obtained performance and the main issues encountered during the submission phase will be analyzed.

2. The IGI infrastructure within EGI

IGI, the Italian National Grid Infrastructure (NGI), is one of the major partners of EGI. Currently it comprises 55 distributed sites (Figure 1) spread all over the country, and offers a total of about 32000 CPU cores and 30PB of storage capacity split between disk and tape. IGI supports more that 50 Virtual Organizations (VOs) belonging to more than 10 scientific domains. While some of this VOs are international and are supported by many NGIs within EGI, some others have a national scope and are devoted to new national user communities letting them to try and test the services offered by the Grid infrastructure. One of this national VOs is called GRIDIT and was extensively used to assess the use cases presented in this work.

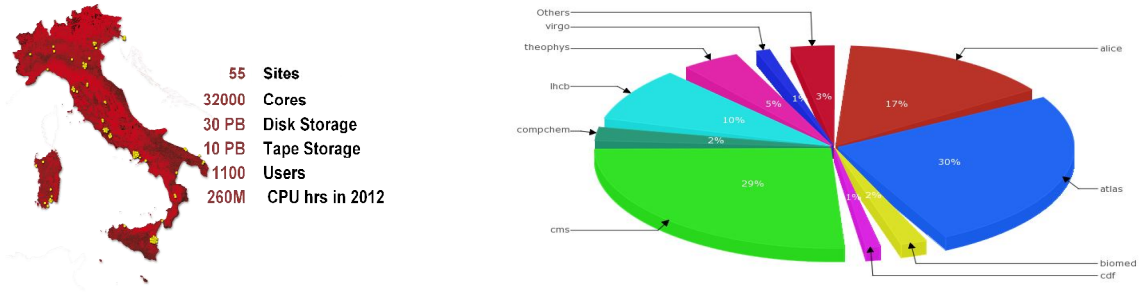

Figure 1- Left: Sites location and main numbers of the IGI infrastructure. Right:Usage of the infrastructure by Virtual Organisations (% of CPU time) in the last 18 months(source is the EGI Accountign Portal). CMS, ATLAS, LHCB, ALICE are the LHC experiments VOs that account for about the 85% of the IGI usage.

PoS(ISGC 2013)024

IGI is managed within 3 operative units (Operations, Middleware Development, User Support and Training) that collaborate with other NGIs and also with EGI. During 2012, IGI resources have been used by more than 1100 users, for a total of 260 millions CPU hours.
The middleware used on the infrastructure is a national release (called IGI release) based on the gLite middleware stack produced by the European Middleware Initiative (EMI).

3.The need for HPC and the related work

The number of EGI users belonging to disciplines outside the High Energy Physics has increased in recent years; HEP still represents the largest fraction, but new communities now count a significant number of users. Analyzing the use cases of the Italian communities belonging to such disciplines, in particular from Computational Chemistry, Earth Science and Bioinformatics (all of them were interested in exploiting the IGI computing capabilities) it was evident that better support of parallel applications was needed. In mid 2011 a working group has been set up to analyze the issues connected with parallel applications on Grid and to improve their support within the IGI infrastructure [5]. The working group took advantage of the new capabilities of the gLite middleware in supporting MPI and parallel jobs (§3.1) and of the effort that the EGI project put in the same topic through a dedicated MPI Virtual Team [3].
The IGI working group was created as a strong collaboration between Grid managers, site managers and various users communities and its main outcome can be summarized in the following points:

- The increased number of sites supporting MPI and parallel jobs for the national VOs (GRIDIT in particular) – from 3 to 10 - two of them counting thousands of cores and low latency connection.
- A series of recipes and solutions to better manages the coexistence of scalar and parallel jobs on the same site resources
- Ad hoc solutions to support checkpointing of long parallel jobs acting on huge data sets
- Recipes and standard semantics to publish into the grid information system specific software optimized for the sites resources and available to users as pre-installed binaries

Thanks to this effort it is now possible to smoothly run small and medium-sized HPC jobs (up to 128 computing nodes per run in some sites) on the Italian Grid Infrastructure. Reaching a higher number of computing nodes is possible in theory for sites enabled with low latency network connections but that number is limited by the interference with scalar jobs that results in too high waiting times in batch systems queues to have the needed resources available. The most interesting HPC use cases that were run on IGI will be described in the following sections.

3.1 HPC support in gLite

In the last couple of years the support to parallel jobs in gLite, which is the middleware on which the Italian Grid Infrastructure is based on, greatly improved thanks to new granularity attributes that allow to describe in details the resources needed by the parallel jobs, e.g. new Job Description Language statements (Table1) and on the reengineered mpi-start framework [6]. The mpi-start tool is an abstraction layer, located between the middleware and the underlying Local Resource Management System (i.e. Torque or LSF) that offers a unique interface to start parallel jobs with different implementations and takes care of the whole data distribution among nodes and execution of the jobs.

PoS(ISGC 2013)024

Attribute	Meaning
CPUNumber=P	Total number of required CPUs
SMPGranularity=C	Minimum number of cores per node
HostNumber=N	Total number of required nodes
WholeNodes=true	Reserve the whole node (all cores)

Table 1 - New gLite granularity attributes allowing to describe in the JDL requirements for parallel jobs

4. Selected use cases of HPC-based workflows

In this section we present a selection of the real life use cases and workflows based on parallel applications ported recently to the Italian Grid Infrastructure that have been provided by various Italian user communities belonging to different scientific domains.

4.1 Molecular Dynamics: NAMD

NAMD [7] is a powerful parallel Molecular Mechanics(MM)/Molecular Dynamics(MD) code particularly suited for the study of large biomolecules. It is compatible with different force fields, making possible the simulation of systems of quite different characteristics. NAMD can be efficiently used on large multi-core platforms and clusters. The NAMD real life use case was provided by the CNR-ISOF [8] group located in Bologna, it consists of the simulation of a lipidic bilayer in a water box made up of about 36,000 atoms to be run for 70ns of simulated time. Using 16 cores on a dual AMD 6238 machine the simulation requires 85 days of wall clock time. The application was ported to exploiting two sites, one located in Naples and the other one in Bologna. Naples site is equipped with InfiniBand network connection while the Bologna site has gigabit interconnection. NAMD was ported to Grid by rebuilding the binaries on Scientific Linux 5 and linking OpenMPI v4.3 libraries. We performed a series of preliminary scalability tests to choose the number of computational nodes to run the full simulation. Results of these preliminary tests are shown in Figure2.

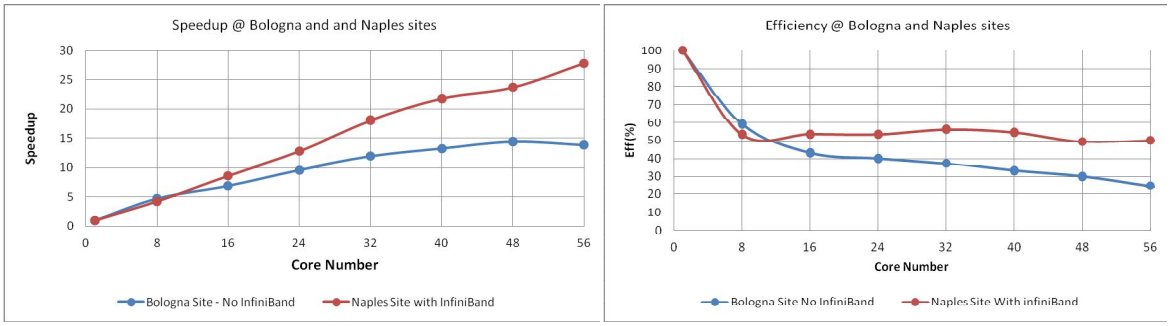

Figure2 – Speedup and efficiency tests for the CNR-ISOF lipidic bilayer simulation based on NAMD_v2-9

Given the results of these tests, the number of computation cores for the simulation was fixed to 48 in order to have a reasonable total computational time minimizing the waiting on queue times needed to obtain enough free CPUs on the sites. However, even with this number of computational nodes, the run time is much longer than the maximum wallclock and CPU times available on Grid local resource manager queues, so a Grid-level checkpointing was needed: the simulation was split in 70 segments of 1ns of simulated time each and a submission system was developed in python. The submission system took care of the data management and

PoS(ISGC 2013)024

execution of all the simulation segments, feeding the output of a segment as input to the next one.The whole submission resulted in the workflow represented by a linear Direct Acyclic Graph (or DAG) as depicted in Figure3.

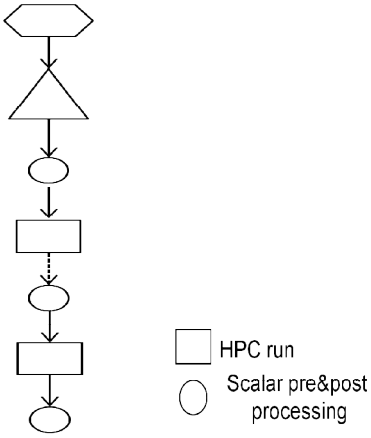

Figure 3 – The workflow used to run the CNR-ISOF simulation on the Grid infrastructure. To overcome time limits on Grid queues the simulation was split in 70 segments. Grid data management was needed to feed each segment with the output data of the previous one. The entire workflow was handled by an ad-hoc newly developed submission system.

The submission system was based on an algorithm that submitted the segments as Grid MPI jobs to the faster (InfiniBand enabled) site if enough resources were available otherwise the job was submitted to the slower site. The time evolution of the entire simulation is reported in Figure4 comparing the ideal evolution on the two sites with the real evolution using balanced submissions on both. The whole simulation took 32 days to complete.

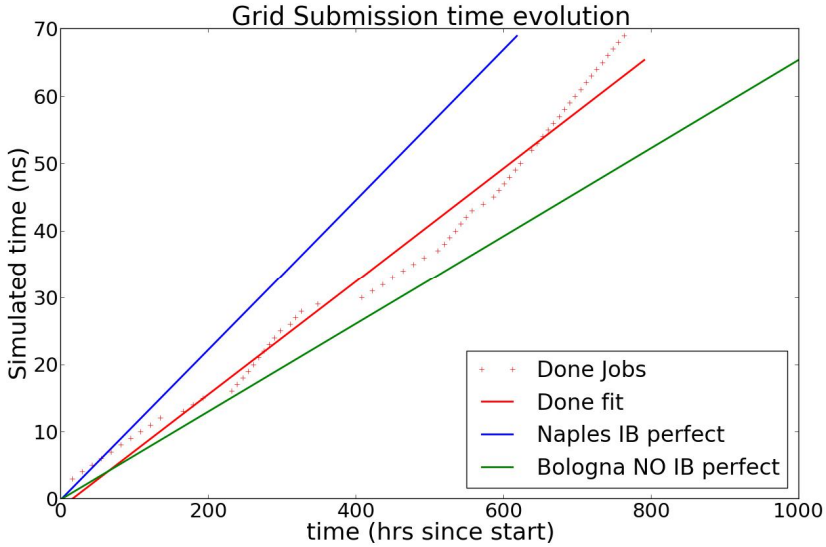

Figure 4 – The CNR-ISOF simulation actual time evolution (red dots) compared with the ideal evolution that it would have had on the two sites used for the Grid submissions (Naples sites is faster due to low latency network connections). The whole simulation took 32 days to complete.

PoS(ISGC 2013)024

4.2 Atmospheric Science: the GLOBO general circulation model

GLOBO is a hydrostatic, grid point, Atmospheric Global Circulation Model (AGCM) developed by the Institute of Atmospheric Science of the Italian National Research Council (CNR-ISAC) [9], Bologna Department. The Grid infrastructure was exploited by the community in order to perform a calibration [10] and fine tuning of the model through reforcasts of a 30-year period (from 1981 to 2010). A total of 2190 jobs were needed to cover the period, each job simulating 35 days and starting every 5 days in the 30-year period. The application was ported to the Grid environment using OpenMPI and 16 computational nodes per run. Each job required from 6 to 8 hours on the Grid sites depending on the network interconnection available.

Grid data management was needed for pre and post processing of each jobs. The workflow representing this use case is depicted in Figure5 and it can be seen as a combination of High Throughput Computing (HTC) and small sized High Performance Computing (HPC) calculations. The use case was successfully run on the IGI and the whole simulation took about two months with the time evolution represented in Figure 6.

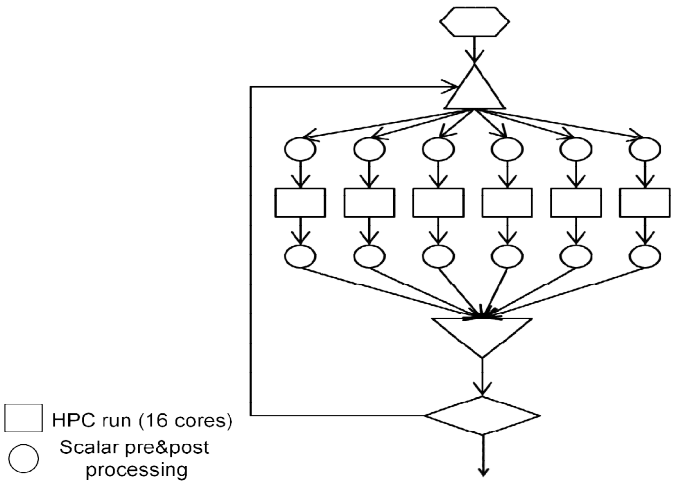

Figure 5 – The workflow representing the CNR-ISAC use case. It is a combination of HPC and HTC paradigms which requires 2190 parallel jobs, 16 cores each, with data management needed for pre and post processing.

PoS(ISGC 2013)024

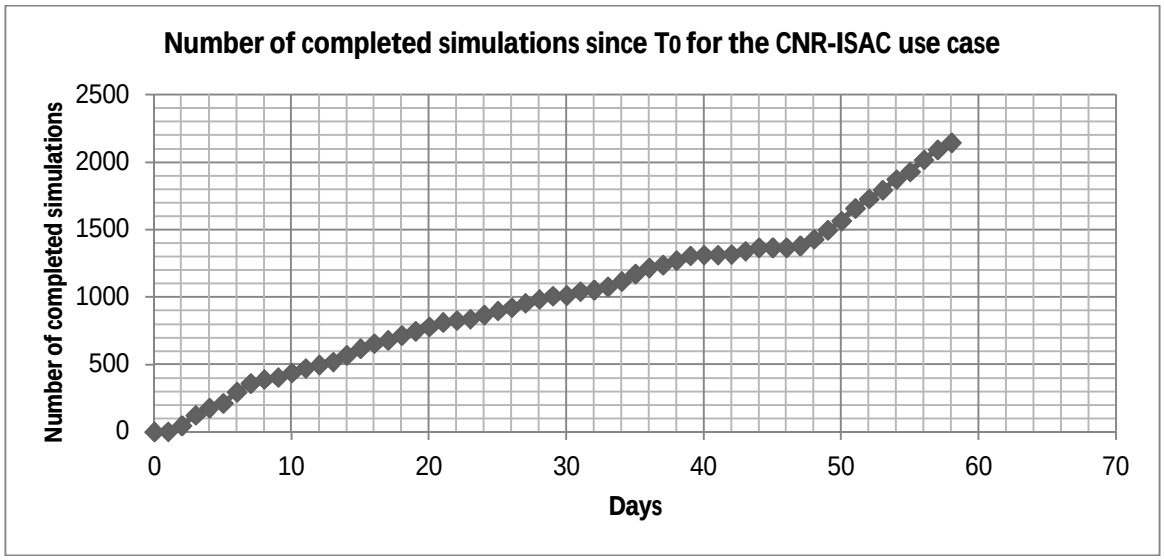

Figure 6 - Time evolution of the CNR-ISAC simulation. About 2200 HPC jobs were run in 2 months on IGI.

4.3 Theoretical Physics: the Einstein Toolkit

The Einstein Toolkit [11] is an open software that provides the core computational tools needed by relativistic astrophysics, to solve the Einstein's equations coupled to matter and magnetic fields. It solves time-dependent partial differential equations on mesh refined three-dimensional grids. The application has been ported to the grid infrastructure using an hybrid MPI and OpenMP code and required rebuilding the binaries executables on Scientific Linux 5. It is currently used in production with simulation involving up to 256 cores on the INFN-PISA site. The use case provided by the University of Parma and by INFN-Parma researchers involved long (of the order of 2 weeks using 128 cores) HPC jobs to simulate the evolution of a stable general relativistic TOV-Star model on a cubic multi-grid mesh with five levels of refinement (each of local size 40x40x40). Given the required execution time checkpointing was needed as in the NAMD case discussed in §4.1, but since this application also acts on big datasets (of the order of tens of GB) to ease the data management jobs were run on a single, InfiniBand enabled site (the INFN-PISA site). At this site a dedicated, permanent checkpoint storage area (data in this area outlives the job that created them) was prepared. Preliminary scalability tests were run to test the parallelized code (Figure 7) and finally the production runs were performed.

The workflow representing this simulation is very similar to one of Figure 3, but in this case the data management steps (circles) are much easier thanks to the local checkpointing area that allow POSIX calls to access the data in read/write mode.
The whole simulation required 6 restarts of 48 hours each using 128 cores of the site.

PoS(ISGC 2013)024

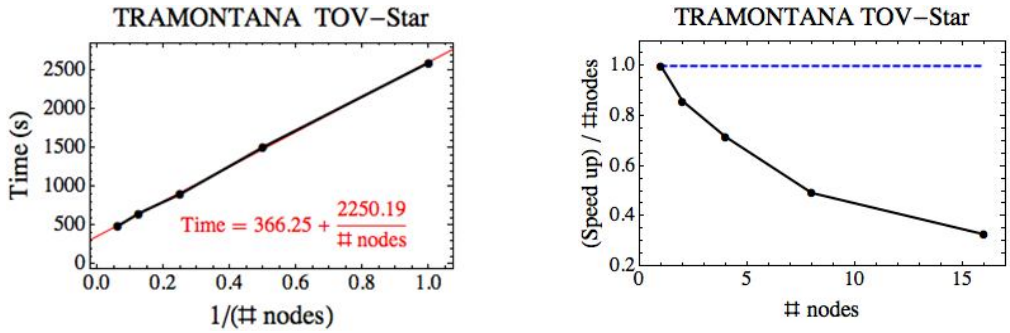

Figure 7 – Preliminary efficiency and scalability tests for the University of Parma/INFN-Parma use case about the simulation of a relativistic star. Real run used 128 cores on the INFN-Pisa site (InfiniBand enabled) and required 6 restarts of 48 hours each.

4.4 Computational Chemistry enhanced workflows

In the previous sections three of the HPC-based workflows that were recently ported to the Grid by users in strong collaboration with the Italian NGI user-support unit have been presented.

Other use cases, in particular from the Chemistry and Molecular and Materials Science and Technology (CMMST) community, at present operatively supported in EGI by the COMPCHEM VO [12] are based on different types of workflows, but still implying a combination of HTC (many scalar runs) and HPC (parallel runs). Examples of these workflows are provided in Figure 8: (a) an HPC run, possibly requiring many computational nodes (up to 256), followed by several scalar runs performing post processing and independent analysis on the data produced; (b) many HTC scalar jobs preparing the inputs used by a single HPC simulation.

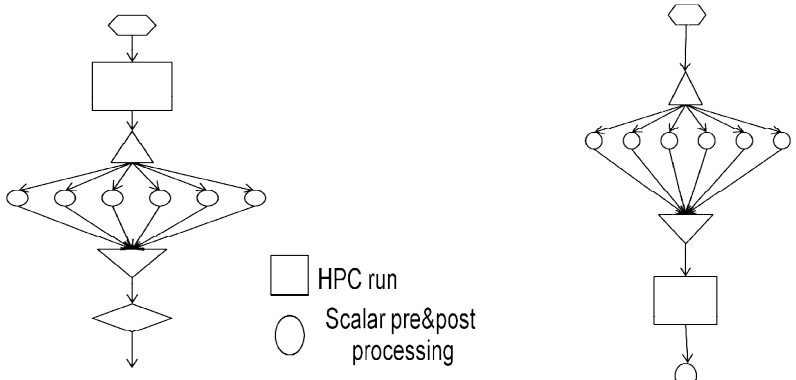

Figure 8 – HPC/HTC workflows schemas (a) an HPC run, possibly requiring many computational nodes, followed by several, scalar runs that performs post processing and independent analysis on the data produced (b) many HTC scalar jobs that prepare the input for a big HPC simulation

To support more effectively these workflows the Italian NGI and the COMPCHEM VO and the CMMST members are collaborating in a series of initiatives aimed at realizing basic interoperability between different infrastructures (HTC and HPC) in order to execute big parallel jobs on dedicated HPC machines not available on the IGI infrastructure. A number of

use cases that will benefit from such an interoperability have been collected by the CMMST community and submitted to the EGI-XSEDE [13] collaborative use examples call that was recently launched by EGI.
The activity on HTC-HPC workflows interoperation using in different infrastructure is still in progress and will be presented in future papers.

5. Conclusion and future work

In this paper the improvements obtained in the last couple of years by the Italian Grid Infrastructure in supporting MPI and parallel applications have been presented showing how small and medium size HPC can be now achieved using this infrastructure thanks to the new multicore CPU architectures, to new middleware capabilities and to the effort put in this field by working groups that acted at nation and European level.

It has also been shown how the infrastructure was successfully exploited to run complex real life use cases based on workflows requiring both parallel (HPC) and scalar (HTC) jobs. The use cases were provided by various user communities belonging to different scientific domains, three of this use cases were presented in details.

The activity of porting new applications and real life use cases based on parallel jobs is still ongoing as a collaboration between the IGI user-support unit and various Italian user communities belonging to different scientific disciplines. From the user point of view, one of the main issues encountered during the porting process is the steep learning curve needed to acquire the know-how necessary to submit complex workflows into the Grid environment. Long consultancy and training periods with Grid experts were needed in order to put in place an efficient workflows submission system; to reach this goal dedicated web interfaces are being built for some user communities. These interfaces will be added, as ad-hoc portlets, to the modular IGI submission and data management web portal [14] which is already an NGI production service based on WS-P-Grade [15] and Liferay technology [16].

From the site and grid managers point of view the main issues to be addressed when the number of parallel jobs increases is the fair coexistence of these kind of jobs with the scalar ones on the same sites. If the sites are partitioned to have a fraction of them made up of resources more suitable to run parallel jobs (i.e. they are connected with low latency network) it is important to route HPC jobs on these resources. However when there are no HPC jobs running the parallel resources should not be wasted and scalar jobs should be submitted as well. On the contrary if the site is not partitioned it is important to set local policies to guarantee that the resources needed by parallel jobs are made available in a reasonable time. Solutions to this kind of problems have been discussed in [5].

Various communities, such as the CMMST one and others, raised the requirement to run the HPC sections of the workflows, at least the most computationally complex, on HPC dedicated infrastructures. To address this requirement at least a basic interoperability among different infrastructures is needed and work is ongoing to achieve this.
A still open issue is the infrastructure support to application requiring GPUs and accelerators such as the Intel Xeon Phi: how to publish GPU resources in the information system and how to efficiently use them are important issues to be addressed.

PoS(ISGC 2013)024

References

[1] T. Ferrari, L. Gaido, *Resources and Services of the EGEE Production Infrastructure*, *Journal of Grid computing, Volume 9, Number 2, 119-133 (2011)*

[2] EGI, *"European Grid Infrastructure - An Integrated Sustainable Pan-European Infrastructure for Researchers in Europe (EGI-InSPIRE)", White Paper, 18 April 2011 - Document Link: https://documents.egi.eu/document/201*

[3] EGI, *EGI MPI Virtual Team – Reference Link https://wiki.egi.eu/wiki/VT_MPI_within_EGI*

[4] IGI Home Page: http://www.italiangrid.it

[5] R. Alfieri et al., *"The HPC testbed of the Italian Grid Infrastructure", In Proceedings of the 21st Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP'13), Belfast, Northern Ireland*

[6] K. Dichev, S. Stork, R. Keller, E. Fernández, *"MPI Support on the Grid", Computing and Informatics, vol 27, No 2 (2008)*

[7] NAMD homepage - http://www.ks.uiuc.edu/Research/namd/

[8] CNR-ISOF homepage: http://www.isof.cnr.it/

[9] CNR-ISAC home page - http://www.isac.cnr.it/

[10] Mastrangelo, D., Malguzzi, P., Rendina, C., Drofa, O., and Buzzi, A. *"First outcomes from the CNR-ISAC monthly forecasting system"*, Adv. Sci. Res., 8, 77-82, 2012

[11] F. Loffler, et al. *"The Einstein Toolkit: A Community Computational Infrastructure for Relativistic Astrophysics". Classical and Quantum Gravity, 29(11), (2012).*

[12] COMPCHEM Virtual Organization home page - https://www3.compchem.unipg.it/compchem/

[13] EGI-XSEDE webpage: http://www.egi.eu/news-and-media/newsfeed/news_2013_0003.html

[14] IGI submission and data management portal: https://portal.italiangrid.it

[15] P. Kacsuk et al., *"WS-PGRADE/gUSE Generic DCI Gateway Framework for a Large Variety of User Communities", Journal of Grid Computing, Vol 10, No 4, pp 601-630, DOI: .1007/s10723-012-9240-5*

[16] Liferay homepage: http://www.liferay.com/

PoS(ISGC 2013)024