THE COMETA E-INFRASTRUCTURE
A Platform for Business Applications in Sicily

Marcello Iacono Manno, Pietro Di Primo, Gianluca Passaro, Emanuele Leggio
COMETA Consortium, Via Santa Sofia 64, Catania, Italy

Roberto Barbera
Physics & Astrophysics Department, Catania University, Via Santa Sofia 64, Catania, Italy

Giuseppe Andronico, Riccardo Bruno, Emidio Giorgio, Marco Fargetta, Giuseppe La Rocca
Salvo Monforte, Diego Scardaci, Fabio Scibilia
National Institute for Nuclear Physics, Catania Section, Via Santa Sofia 64, Catania, Italy

Keywords: Grid, e-Infrastructure, Business, Parallel Computing, Security.

Abstract: The COMETA e-Infrastructure running in Sicily is compliant with the EGEE middleware and specifications offering a great computing power and huge storage capacity. Since its beginning one of the main goals has been to extend the adoption of Grid paradigm from the academic to business world. Several software and hardware extensions have been implemented in order to enhance the infrastructure performances; they include a new low-latency net layer reserved to heavy parallel applications; some modifications to the parallel job submission and execution procedures for a better support of MPI-based applications; new tools for job monitoring and file catalogue interaction; a scheduling policy tailored on the requirements of a complex environment hosting heterogeneous jobs; the GridLM license server able to grant access to commercial software only to authorised users; finally, the Secure Storage Service defends from insider abuse and completes a very high security level environment. The proposed business model includes a wide range of services collectively defined as Infrastructure-as-a-Service. All the above components result in a powerful and flexible platform, easy to use for any applications and open to further developments.

1 INTRODUCTION

The COMETA (www.consorzio-cometa.it) Grid infrastructure is a distributed computing platform located in Sicily both for academic and business computing. It is the first in Italy scaled to involve a whole region. Since its preliminary design, a primary target has been the implementation of business applications. For this reason, the environment of Grid computing has been adapted to the demanding requirements of such applications.

Grids can be useful as cheap host platforms for several kinds of business applications ranging from High Performance Computing (HPC) to Cloud Computing (CC), from data intensive to applications requiring dedicated user-friendly interfaces. Particularly, Small and Medium Enterprise can reduce both their cost of ownership and time to market by outsourcing their design and/or production activities. Even Public Institutions (such as hospitals and/or cultural institutions) may take advantage from this solution.

The business model is Software-as-a-Service (SaaS) that evolves into an Infrastructure-as-a-Service (IaaS) as the whole infrastructure is transparently used as a facility providing on-demand services for heterogeneous applications.

The effort to really and effectively support such a scenario required deep adjustments both to hardware and software compared to the standard EGEE project (www.eu-egee.org/) gLite middleware

Iacono Manno M., Di Primo P., Passaro G., Leggio E., Barbera R., Andronico G., Bruno R., Giorgio E., Fargetta M., La Rocca G., Monforte S., Scardaci D. and Scibilia F. (2010).
THE COMETA E-INFRASTRUCTURE - A Platform for Business Applications in Sicily.
In *Proceedings of the International Conference on e-Business*, pages 256-261
DOI: 10.5220/0003025802560261

(www.glite.org) that only provides generic tools to access the bare resources. The following sections describe the efforts about hardware design, policy tailoring and software support focusing on each specific issue. Section 2 introduces the COMETA e-Infrastructure. Section 3 describes the adoption of the InfiniBand net layer. Section 4 is about the scheduling policy, a critical aspect of the management of such infrastructures. Section 5 describes the modifications to the standard official middleware for a better support of MPI-based parallel applications. Section 6 focuses on GridLM, the newly-developed license server allowing only entitled users to run licensed software on the Grid under the license terms and the accounting system. Section 7 describes the Secure Storage Service implemented to solve the insider abuse problem. Section 8 is about the so-called "watchdog" tool to monitor long lasting jobs and recursive commands for massive interaction with the data catalogue. Section 9 illustrates the most common porting procedures and related computing schemas. Finally, some conclusions are drawn in Section 10.

2 THE COMETA CONSORTIUM AND THE SICILIAN E-INFRASTRUCTURE

The COMETA Consortium gathers the Universities of Catania, Messina Palermo , the National Research Institutes for Nuclear Physics (INFN), Astro-Physics (INAF), Geo-Physics and Volcanogy (INGV), and the SCIRE Consortium. These institutions have been involved at different levels in the EGEE project and other European projects deploying international e-Infrastructures. Nevertheless, the idea behind the development of a new regional e-Infrastructure is to foster the adoption of Grid computing for massive computation among Sicilian researchers both from academia and business worlds. For this reason the Sicilian e-Infrastructure has been built on grants coming form the Italian Ministry of University and Research (PI2S2 Project, www.pi2s2.it/) and the Sicilian Regional Government (TriGrid Project, www.trigrid.it). The infrastructure, which adopts the gLite middleware, is fully compliant with the international standards on interoperability. Figure 1 shows the location of the seven sites of the infrastructure in the cities of Catania, Messina and Palermo; they collect an overall amount of ~2000 cores and a storage capacity of >250 TB.

Figure 1: The sites of COMETA e-Infrastructure.

3 INFINIBAND NET LAYER

Usually, local network connection among the nodes of a Grid site is provided by Ethernet links at 1 Gb/s, since the communication among nodes is not a bottleneck for many applications performing on the Grid. Currently, the majority of the Grid applications are trivially parallel so they can be split in different not communicating chunks. As a result, the bandwidth provided by the GigaEthernet link is sufficient for most types of applications in order to run these jobs efficiently. However, an efficient execution of not trivial parallel jobs requires much faster exchanges of short messages instead of sustained communications. So, a low latency is required. This is the reason why each processor in the Sicilian e-Infrastructure is equipped with two network cards, one for the usual Ethernet connection devoted to Grid services and normal (non-parallel) jobs, and the other one for the InfiniBand connection, reserved to parallel computing only.

The impact of InfiniBand on the programs is noticeable. The latency of communication drops from the Ethernet value of about 50 μs down to 1-2 μs. Optimized parallel programs exploit this feature resulting in a much shorter execution time. The advantage of using the low-latency communication protocol is more sensitive the more nodes are allocated for the computation. For parallel computations involving several tens of cores and more, the adoption of the InfiniBand net layer is highly recommendable. In many cases, the optimization of net communication is a challenging problem, particularly for those parallel codes that have been adapted from different architectures, such as shared memory, to a distributed computing environment. In such cases, the InfiniBand net layer

is very effective too (see also the discussion about MPIGranularity in Section 5).

4 SCHEDULING POLICY

The coexistence of several heterogeneous jobs running on the same infrastructure is a major difference between Grids and dedicated clusters. Particularly, the execution of long lasting and multi-core parallel codes, also called HPC applications, requires an efficient resource assignment. So, the scheduling policy has to implement the best achievable trade-off among the different requirements. Each site has different queues dedicated to short, long and infinite jobs with respective longest durations of 15 minutes, 12 hours and 21 days (lifetime of the longest proxies). The priority is given to the short jobs in order to keep the turnaround time a reasonably small fraction of the execution time. However, this usual scheduling policy is not flexible enough as parallel jobs requiring hundreds of cores (sometimes a significant fraction of all the available resources of a site) would remain scheduled forever. Therefore, the pre-emption and reservations policies have been carefully studied and implemented in order to solve this problem.

Pre-emption puts the incumbent job in a suspended state and assigns the resource to the incoming job. This policy is acceptable only if the latter is short, so it has been implemented for emergency jobs (related to volcanic surveillance and other civil protection applications).

Common parallel jobs reserve the needed resources as soon as they become available. Provided that the amount of short jobs is higher compared to the others, their turnover is fast enough to keep this policy pretty valid. Figure 2 summarizes the scheduling policies. Scheduling is largely dependent on the amount of jobs on each queue and the adopted policy must be updated when significant changes occur in the job distribution.

This issue also witnesses the importance of the overall infrastructure monitoring that has to collect periodic statistic information about job distribution in the various queues.

Gustav is a CPU accounting tool developed by INFN, COMETA and KISTI institute in South Korea. Gustav collects accounting records from resources and publishes them to a centralised relational database that can be queried through a web interface (gustav.consorzio-cometa.it).

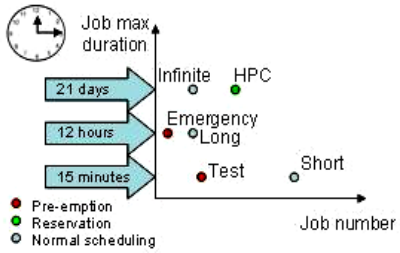

Figure 2: Main queues durations and populations.

5 UPGRADES OF THE GLITE MIDDLEWARE

The gLite middleware supports the Message Passing Interface (MPI) as its only libraries for parallel computing. The complexity of the porting procedure is a major factor against the adoption of the Grid paradigm for HPC. For this reason new wrappers have been developed to support each available combination of compilers, net layers and library versions. The MPIType tag added to the gLite Job Description Language (JDL) allows users to simply select the proper value to run their parallel code in the proper HW/SW environment.

MPIGranularity tag reserves the desired number of cores on the same physical processor for the incoming job. Thus. the communication over the net, a usual bottleneck in such situations, is reduced. Figure 3 shows how the impact of this technique on the execution time is higher compared to that of the InfiniBand net layer.

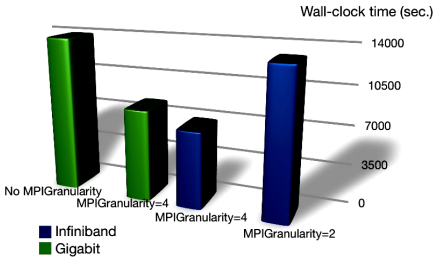

Figure 3: The impact of MPIGranularity and InfiniBand on the wall-clock execution time of a parallel application.

Although gLite has recently extended the MPI support including the so-called “mpi-start” approach, based on a set of scripts able to identify the library requested for each application, the advantages of the developed solution do not allow to move to the official support and some of the ideas were submitted to the MPI Working Group of EGEE

who is evaluating their integration in a future release of the gLite middleware.

6 THE COMETA LICENSE SERVER AND ROBOT CERTIFICATES

The delivery of license files to authorized users in a distributed environment is a non-trivial task. For security reasons, the license server is a single machine for the whole infrastructure, only locally redundant for the sake of service continuity. Thus, the license file has to be delivered to the remote infrastructure sites travelling on a public network. Moreover, solely entitled users must be granted.

The license system developed by COMETA has been designed to be used with the FLEXlm (www.globes.com/support/flexlm_index.htm) free license server that is the most used license management system for commercial software. The "floating" license mechanism has been adapted to the Grid distributed environment, evolving into GridLM. The secured communication channel created from the remote Computing Element (CE) to the license server allows to first identify the user by his/her certificate, check the number of granted licenses and then issue the license file.

Figure 4 illustrates the schema of the license server mechanism available on the COMETA e-Infrastructure.

The team of Grid developers is currently extending the license server mechanism to cope with "robot" certificates as well. They are a novel feature that allows running a program without any personal certificate. The user can store this special certificate on a smart card and insert it into a personal computer running the user interface (usually a web portal). This is enough to authenticate and launch a program on the Grid. This solution fits the requests from large user communities that want to simplify the approach to Grid computing. The accounting system is another key element for a sustainable use of the Grid infrastructures especially when they are open to commercial exploitation. gLite offers various tools for resource metering. PI2S2 project developers extended the measuring to the disk storage with SAGE (https://forge.eu-eela.eu/forum/forum.php?forum_id=31).

This tool computes the amount of disk energy used, i.e. the amount of space integrated over the time of occupation.

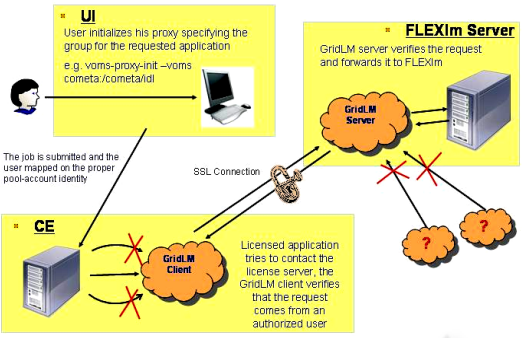

Figure 4: The Schema of the COMETA License Server Mechanism.

After a first version, tailored for DPM (www.gridpp.ac.uk/wiki/Disk_Pool_Manager) based devices, a second version was developed that is adaptable to a generic disk storage device. SAGE implements a good trade-off between precision and computing demand of the metering tool. A dedicated web portal called Sight-on-SAGE has been developed as well to allow an easy access to the accounting data both for users and VO managers.

7 SECURE STORAGE

One of the main benefits of Grid infrastructures is the possibility to use distributed storage space. A community might like to use Storage Elements (SE) owned by an external organization to delegate the management of these machines and to avoid to buy specialized hardware and to hire specialized personnel.

In this way the community could rent the storage space as needed and minimize both human and hardware costs.

In the case of confidential data, this scenario is not feasible. Indeed, the community should satisfy strong privacy requirements, for example when it has to manage medical or financial data. A mechanism to prevent the administrator of the machine accessing the data is required to store the confidential data in a storage element managed by an external organization.

The gLite middleware provides the same security infrastructure for all its services (using X.509 infrastructure to authenticate the users and the VOMS attributes to authorize the users).

However, data are stored in a clear format. The storage element administrator can in principle access them bypassing the Grid security infrastructure. This is known as the insider abuse problem.

The Secure Storage Service developed for the gLite middleware provides users with a set of tools to store in a secure way and in an encrypted format confidential data on SEs solving the insider abuse problem. Data stored trough the tools provided are accessible and readable only by authorized users.

The Secure Storage Service (www.ias07.org/) has been designed to be integrated in the gLite middleware; it is made up by the following components:

- Command Line Applications: commands integrated in the gLite User Interface to encrypt & upload and decrypt & download files on the Storage Elements;
- Application Program Interface (API): allowing the developer to write programs able to manage confidential data using the Secure Storage Service;
- Keystore: a new Grid element used to store and retrieve users' keys in a secure way;
- Secure Storage Framework: a service component, internally used by the other components; it provides encryption & decryption functions and other utility functions; it takes care of interaction with the gLite Data Management System.

8 WATCHDOG, VISUAL GRID AND RECURSIVE CATALOGUE INTERACTION

Many complex jobs often need long execution times. Such a long period increases the probability to have errors due to either infrastructure or program faults. Even a brief network interruption may cause the job to fail. On the other hand, if the job is performing an erroneous computation due to wrong input data or application bugs (many applications are undergoing a continuous development and many executions are needed to produce bug-free codes) the user can verify the output only after the execution. This leads to a consequent waste of time and resources, even though many problems could have been revealed a few moments after they have occurred during the run. Due to the intrinsic complexity of parallel jobs and their long durations, they require constant job monitoring and checkpoint features.

Although current gLite middleware offers the perusal job technique for job inspection, COMETA e-Infrastructure offers two more sophisticated tools for job monitoring: Watchdog and VisualGrid.

As the evolution of the job is detectable by inspecting the files produced by the job in the working directory, it is useful to have them copied on the Grid file catalogue at regular intervals, where the user can access them. The Watchdog utility performs such full or incremental backup by a customizable script also allowing the storage on the AMGA (amga.web.cern.ch/amga) metadata catalogue.

The VisualGrid tool (Andronico, 2010) allows the encoding of images produced by the job into a video that is immediately streamed to a public IP address. The result is a powerful tool for demonstrations but also for a direct, visual, control of the job work flow.

The tool has been tested with the FLUENT (www.fluent.com/) commercial application and its simulation of the Marmore falls in Italy (see Figure 5).

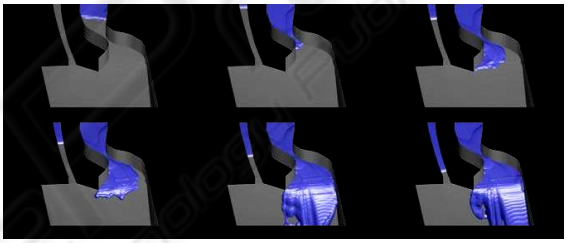

Figure 5: Some frames of the sequence produced by FLUENT and streamed by the VisualGrid tool.

Other tools developed by COMETA help users in their interaction with storage resources. Sometimes, MPI jobs write many files organised in (sub-) directories. The current gLite middleware forces the user to download them one-by-one unless he/she produces an automatic script. Similarly, there is no command for bulk upload or to delete all the files in a directory of the data catalogue. As these operations become more frequent with MPI jobs and the users may not want to write their own scripts for such standard operations, some new tools have been developed for recursive interaction with the Grid data catalogue. Their explanation is available on the web (https://grid.ct.infn.it/twiki/bin/view/PI2S2/WikiConsorzioCometa). The three basic commands are listed in the following: `lcg-rec-cr:` uploads all the files of a local directory on a Storage Element (SE) and registers the entry on the LFC catalogue, copying the structure of the sub-directories and its content as well; `lcg-rec-cp:` performs the recursive download of an entire directory together with its sub-directories; and `lcg-rec-del:`

removes all the entries of a directory from the catalogue and the corresponding files form the SE.

9 PORTING PROCEDURES AND EXECUTION PERFORMANCES

The porting procedure consists in bringing a program already running on a common dedicated platform to run in a Grid environment. The procedure is usually straightforward for simple jobs but it may become rather complicated for complex parallel jobs. This stresses the importance of the computational schema, i.e., the strategy followed to exploit the computing opportunities offered by the Grid. Nevertheless, as the knowledge about the Grid mechanisms and tools is greatly useful for job optimisation, a Grid expert usually helps the user during the porting process of his/her application.

This strategy was adopted in several cases. For instance, the FLUENT parallel implementation required an ad-hoc computing schema as the package is given as a "close box", with no access to the code. Provided that the COMETA HW was supported, FLUENT has been "embedded" into the infrastructure.

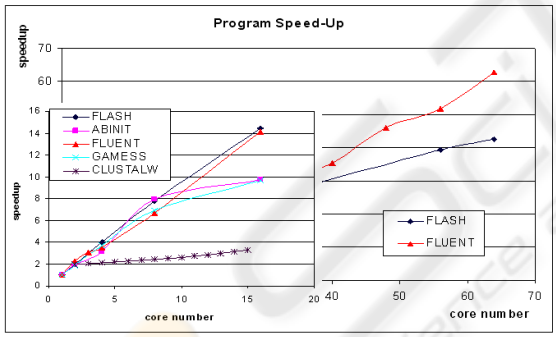

Figure 6: Speed-up of some parallel applications on the COMETA e-Infrastructure.

As shown in Figure 6, the result has been a pretty sensitive speed-up in the execution time comparable with similar dedicated clusters and confirmed by some other parallel applications like FLASH (Orlando, 2007), CLUSTALW (Lombardo, 2010), ABINIT and GAMESS (www.pi2s2.it/applications/), belonging to other scientific domains.

10 CONCLUSIONS

The Sicilian e-Infrastructure has been fully adapted to real business applications. Future developments concern the development of further tools supporting the exploitation of the infrastructure for business purposes. The targeted audiences range from small/medium enterprises to large public bodies such as health care institutions. Interested readers can contact the first author at marcello.iacono@ct.infn.it to get detailed information about the COMETA activities.

ACKNOWLEDGEMENTS

This work makes use of results produced by the PI2S2 Project managed by the Consorzio COMETA, a project co-funded by the Italian Ministry of University and Research (MIUR) within the Programma Operativo Nazionale "Ricerca Scientifica, Sviluppo Tecnologico, Alta Formazione" (PON 2000-2006). More information is available at www.pi2s2.it and www.consorzio-cometa.it.

REFERENCES

Andronico, G., Barbera, R., Iacono Manno, M., La Rocca, G., 2010, VisualGrid: On-Line Video Streaming for Application Control and Demonstration, *Proceedings of the Final Workshop of Grid Projects "PON Ricerca 2000-2006 – Avviso 1575"*, ISBN 978-88-95892-00-3, pp. 128-131, published in Catania by the COMETA Consortium

Orlando, S., Peres, G., Reale, F., Bocchino, F., Sacco, G.G., Miceli, M., Bonito, R., Pagano, P., Argiroffi, C., Yelenina, T., 2007, High Performance Computing on the COMETA Grid Infrastructure, *Proceedings of the Grid Open Days at the University of Palermo*, ISBN 978-88-95892-00-9, pp. 181-188, published in Catania by the COMETA Consortium

Lombardo, A., Lanzalone, G., Muoio, A., Barbera, R., Iacono Manno, M., 2007, Performance of ViralPack applications on PI2S2 Grid infrastructure, *Proceedings of the Final Workshop of Grid Projects "PON Ricerca 2000-2006 – Avviso 1575"*, ISBN 978-88-95892-00-3, pp. 446-451, published in Catania by the COMETA Consortium