Mem. S.A.It. Suppl. Vol. 13, 10

Memorie della
Supplementi

The Cometa Consortium and the PI2S2 project

U. Becciani

Istituto Nazionale di Astrofisica – Osservatorio Astrofisico di Catania, Via S. Sofia 78, I-95125 Catania, Italy, e-mail: ugo.becciani@oact.inaf.it

Abstract. The new grid e-Infrastructure in Sicily is offering new perspectives and important resources and starts to give new great opportunity for research using the HPC resources. We will show the infrastructure of the Cometa Consortium, the main activities of the PI2S2 project and the new challenges, mainly in the HPC area, that the project is carrying out. A simple but useful procedure for running HPC is also described.

Key words. Grid Computing, HPC

1. Introduction

This paper presents the new infrastructure of the PI2S2 project and the main activities of the Cometa consortium carried out in the last three years (U. Becciani 2005). It drives the user with a sequence of steps to use the Cometa computational grid for hpc jobs.

The computational resources are more and more a resource for addressing the major challenges of science and the astrophysical community has important computational tasks. Theoretical astrophysic produce huge amount of data and require significant computational resources for simulations and data analysis. Moreover the study and the interpretation of observational data represent a fundamental task of research activities. Therefore, the international astronomical community devotes a great effort to the management of data and the general distribution of scientific data-sets, like source catalogues and images, is a very common astronomical research activity.

E-Infrastructures, based on the *Grid paradigm*, are implementing, since a decade and in several parts of the world, the so-called e-Science vision for scientific collaboration. Computing and storage centres, hosting huge new generation detectors or unique very sophisticated instruments or databases of medical and biological data which grow at incredible speeds, are interconnected by large bandwidth research networks and glued together with a so-called *middleware*, a special software which is mid-way between the hardware of the resources and the software of scientific applications. The middleware allows all centres connected through it to behave as a huge single distributed computer and permits multi-disciplinary Virtual Organizations (VOrgs) of researchers, distributed across several geographic and institutional domains, to ubiquitously access, through virtual services, the data they need in order to produce important scientific results. The grid infrastructures start to be important resources for the astrophysical community, and the Cometa Consortium, with its own grid, gives

Send offprint requests to: U. Becciani

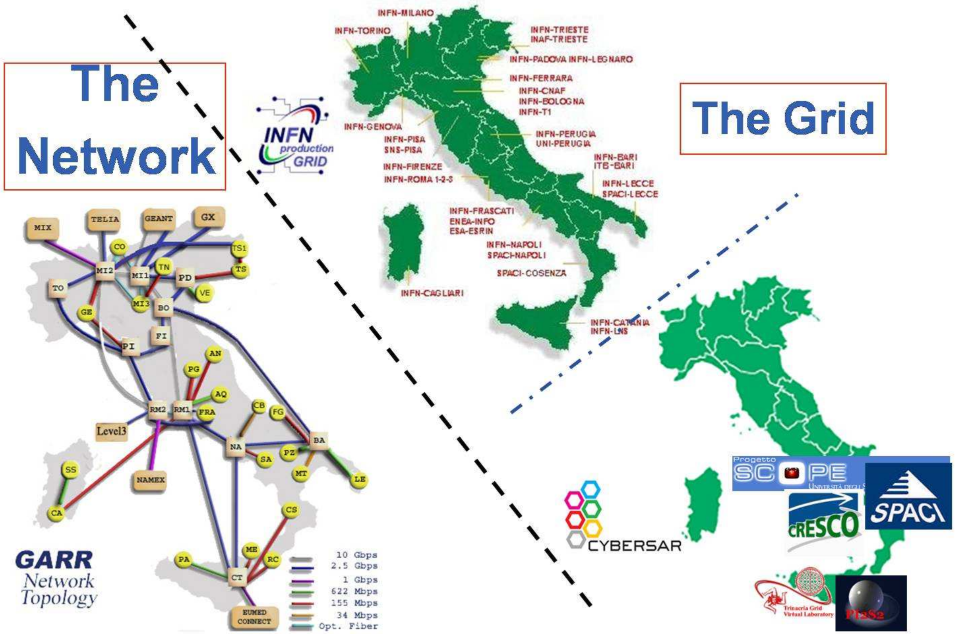

Fig. 1. The Garr Research Network in Italy

a contribution in many fields and in the HPC area.

2. The Cometa consortium and the PI2S2 project

The Cometa consortium (COnsorzio Multi Ente per la promozione e l'adozione di Tecnologie di calcolo Avanzato) is formed by seven partners in Sicily: the Universities of Catania, Messina and Palermo, the National Institute for Astrophysics (INAF), the National Institute for Nuclear Physics (INFN), the National Institute for Geophysics and Volcanology (INGV). The last partner is the SCIRE consortium formed by the University of Roma Tor Vergata and Elasis a spin-off enterprise of FIAT (the most popular italian Industry for commercial vehicles). Since 2005 the consortium has been mainly involved in the development of the PI2S2 (Progetto per l/Implementazione e lo Sviluppo di una e-Infrastruttura in Sicilia basata sul paradigma della grid) project, co-funded by the Italian Ministry of Research, using EU funds for Objective 1 regions. The project mainly aims to build up a large grid infrastructure in Sicily, for both research and industrial applications, using the grid paradigm. The PI2S2 project is one of the four projects funded by the Italian Ministry of Research for regional grids in Southern Italy.

The pillars of the Italian e-infrastructure for research are based on a National Network linking all the main regional research sites, with high speed links of more than 2.5 Gbps and 10 Gbps high speed interconnection. The National research network is managed by the consortium GARR supervised by the Ministry of Research. Sicily is connected in Catania with Rome and Naples with two 2.5 Gbps links (Fig 1) Before the funds of the regional projects, the Italian National grid was mainly composed by computational nodes made available by the INFN. New projects in Southern Italy are building new relevant grid infrastructures: in Naples SCOPE and CRESCO, in Sicily PI2S2 and Trigrid VL (Trigrid VL 2000), and in Sardinia CYBERSAR are changing the national scenario of grid resources distribution, making available new grid resources for science and industrial application.

3. PI2S2 main activities

The main goals of the e-Infrastructure in Sicily can be summarized as follow:

- Create a Virtual Laboratory in Sicily, for both scientific and industrial applications, built on top of a grid infrastructure
- Connect the Sicilian e-Infrastructure to those already existing in Italy, Europe, and the rest of the world improving the scientific collaboration and increasing the competitiveness of e-Science and e-Industry in Sicily
- Disseminate the grid paradigm through the organization of dedicated events and training courses
- Trigger/foster the creation of spin-offs in the ICT area in order to reduce the brain drain of brilliant young people to other parts of Italy and beyond

In the following we will describe the main activities of the project considering the four workpackages of the project plan.

3.1. The infrastructure

The Sicilian e-Infrastructure started in 2005 with a first project called Trigrid VL (Trinacria Grid Virtual Laboratory) funded by the regional government using funds from EC for Objective 1 regions and with the PI2S2 project. The infrastructure is distributed in seven sites of the three main towns of Sicily: Catania, Messina and Palermo (Fig. 2) as listed below and five out of the seven total sites are located inside the campus of the University of Catania.

- Site Name: COMETA-INAF-CATANIA - The site is hosted by the Astrophysical Observatory of the Italian National Institute for Astrophysics (INAF);
- Site Name: COMETA-INFN-CATANIA The site is physically located inside the Department of Physics and Astronomy of the University of Catania and it is operated by the staff of the "Sezione di Catania" of the Italian National Institute of Nuclear Physics (INFN);
- Site Name: COMETA-INFNLNS-CATANIA The site is hosted by the Laboratori Nazionali del Sud of the Italian National Institute of Nuclear Physics (INFN);
- Site Name: COMETA-UNICT-DIIT-CATANIA The site is hosted by the Faculty of Engineering of the University of Catania;
- Site Name: COMETA-UNICT-DMI-CATANIA The site is hosted by the Department of Mathematics and Informatics of the University of Catania.
- Site Name: COMETA-INGEGNERIA-MESSINA The site is hosted by the Faculty of Engineering of the University of Messina.
- Site Name: COMETA-UNIPA-PALERMO The site is hosted by the Department of Physical and Astrophysical Sciences of the University of Palermo.

The most important characteristic is that all sites have the same hardware and software configuration allowing high interoperability and realizing a homogeneous environment that is a fundamental requirement for HPC jobs. The computing infrastructure, based on the IBM Blade Centre each containing up to 14 IBM LS21 blades, interconnected with the low latency Infiniband-4X network, mainly provides for High Performance Computing (HPC) functionalities on the grid. Each blade is equipped with 2 AMD Opteron 2218 rev. F dual-core processors with a clock rate of 2,6 GHz able to natively execute x86 32 and 64 bits binary code. Each processor has 2 GB of DDR2 RAM (8 GB in total per blade). The storage infrastructure is based on IBM DS 4200 Storage Systems that provide high features of redundancy, management and reliability. Overall, about 2000 CPU cores and more than 200 TB of disk storage space are currently available on the Sicilian e- Infrastructure.

The grid middleware is based on gLite V.30 (gLite 2000), the V3.1 deployment is expected in the next few months.

Several commercial softwares are currently installedand licensed on the Grid. The main

Fig. 2. The PI2S2 infrastructure in Sicily

important softwares are: ABAQUS, a tool for Finite Element Analysis for fluidodynamical studies. Industrial applications and many research fields use this suite; Fluent, used mainly to develop and to drive computation fluidodynamics. IDL the most popular software in Astrophysics for data analysis and visualization, and MATLAB the popular software for scientific simulations.

3.2. Middleware integration

The Middleware integration activity is to design and implement new grid services together with their test procedures and benchmarks. Several activities are carried out: the interoperability with other grid infrastructures and middleware, the creation of frameworks for digital repositories, the support for HPC and MPI-2 enabled applications.

The porting of gLite services on Microsoft Windows platforms represents one of the most important activities in this field. Up to now the grid users access the resources with a Linux-based User Interface (UI) with a command line interface and all the gLite resources are Linux-based. This implies that only Linux-based applications can be deployed onto the Grid and often grid users need to be trained. The gLite porting has two main targets: to implement the graphical UI (we have already a preliminary and stable version) and the worker nodes on Windows. This activity will give new perspectives on the usage of the Grid, opening the infrastructure to new applications and new users.

3.3. Dissemination and training

The dissemination activities have produced several specialized events and more than thirty general articles have appeared on general and dedicated press. All of them are available on the COMETA Digital Library at the page (http://documents.consorzio-cometa.it/collection/Press%20Cuts.).

Since the beginning of the project, five tutorials have been carried out for both users and system administrators. More than 120 scientists and technical personnel have been trained on how to access and use the Sicilian e-Infrastructure. Five equipped class-rooms for Grid training and induction have been realized in the three main sites of the Cometa consortium, and are made available to the users of the Sicilian e-Infrastructure.

3.4. Applications on the grid

Since the beginning of the grid operability several scientific and industrial applications have used the infrastructure. More than 45 applications were running in the first three months of the new infrastructure, and in several fields of science: Astrophysics, Bioinformatics, Biomedicine, Chemistry, Computer Science, Nuclear Physics, Volcanology and so one (Pi2S2 applications 2000). Many of these applications were executed on the grid, with effort on porting of existing applications. Some of these applications, mainly in the Astrophysical field, are cosmological simulations that generally require HPC resources. The Astrophysical Italian computational community is using mainframes and clusters for HPC but some applications are now available on the grid. FLASH code form the University of Chicago is already ported on the grid, Gadget, Zeus-MP and other simulation codes will be soon running on the new grid infrastructure.

In the Astrophysical field, the Cometa consortium directly supports many research activities approved in the PI2S2 project: *Large Scale Structure of the Universe and IGM* (V. Antonuccio, J. Silk , U. Becciani et al.). High-resolution simulations of the interaction of Black Holes jet with the InterGalactic Medium. *MHD modeling of Coronal Mass Ejections* (P. Pagano, F. Reale, J. Raymond, S. Orlando, G. Peres) Propagation of shock waves in the solar corona generated during Coronal Mass Ejections by means of a numerical multi-dimensional MHD model. *Global Magneto-hydrodynamic simulations in Galaxyes* (A. Bonanno et al.). Understanding the non-linear evolution of the Magneto Rotation Instability and Tayler instability in the presence of Dark Matter. *Supernova remnants* (M. Miceli, F. Bocchino et al.). Simulations of the evolution of the stellar fragments ("shrapnels") ejected in a supernova explosion. *Virtual Observatory* (U. Becciani, A. Costa et al). Development of a Theoretical Virtual Observatory (ITVO) archive integrated with grid-services (execution of jobs in the grid). *X-ray emission from protostellar jets* (R. Bonito, S. Orland et al.). Studies on continuous supersonic protostellar jet propagating through an uniform medium. *3D MHD modelling of accretion processes in young stars* (G.Sacco, A. Mignone et al.) Interaction between the accreting material and the star atmosphere.

Other researches are carried out using the infrastructure or will soon use it. *COROT Mission* (A.F. Lanza, A. Bonomo). Search for extrasolar earth-like planets using the method of the transits. *GAIA Mission* Sphere reconstruction problem Status: code revision. *Model of the current Stellar Perturbations on the Oort Cloud* (G. Leto et al.). Evaluation of the overall effect of stellar perturbation on Oort Cloud objects, for the development of a complete evolution model of the Solar System including gravitational perturbations. *Radio Emission from Stellar Magnetosphere* (P. Leto et al.). Study and modeling of stellar magnetosphere from simulation of their radio emission, in comparison with radio measures with VLA. *Faint Galaxies in Astronomical Images* (R. Scaramella, S. Sabatini INAF OARM). Faint galaxies: galaxy formation processes and studies of the baryons and of the "dark matter" in the Universe. *Computational Cosmology* (S. Borgani - INAF-OATS). Detailed description of a number of physical and astrophysical effects (star formation, stellar and AGN feedback, metal enrichment, thermal conduction, gas viscosity).

4. PI2S2: new challenges

One of the most important challenges of the PI2S2 project consists in the opportunity to use the grid for HPC jobs. The PI2S2 infrastructure wants to be a gLite based grid with the main target of HPC to run parallel and big jobs on the grid. The second target is to create a new shared large scale infrastructure, bridging the new Southern Italy grid infrastructures. I also want to mention the Virtual Observatory activity: a TVObs compliant database will be published on the grid, with associated services for data exploration (VisIVO server).

4.1. HPC on the grid

Each PI2S2 site has the Infiniband high-speed low-latency network for HPC jobs. The most important feature is that the infrastructure provides the UI *on board* the grid node: the user interface shares the same hw and sw configuration with the worker nodes of the grid. It is also available a dedicated ui-cluster, with Platform LSF manager, for users that want to compile and test the run of the parallel code, before submitting the job in the grid. Anyway the production run will be submitted in the grid using the ui.

Using the LSF preemption characteristics the policy of the usage of the grid resource is to foster the hpc usage of the grid. A special hpc queue is defined on all nodes and users, that are specific credentials for hpc jobs, can use this queue. The preemption allows the hpc jobs to run anyway even if the grid node is fully busy with non-hpc jobs. The currently running *normal* jobs are stopped in favour of the hpc run having the highest priority. The hpc job can continuously run for 21 days and more than 200 CPU cores can be used.

Several MPI implementations with MPICH and MPICH2 compiled with different compilers are available. Both the GigaBit Ethernet and the InfiniBand networks are supported. The native protocol is used with the MPI library MVAPICH and MVAPICH2. Currently, MPI parallel jobs can run inside a single node (computing element - CE) only, but several projects are involved in studies concerning the possibility of executing parallel jobs on Worker Nodes belonging to different CEs.

However there are some constraints that an MPI job must consider. Some parallel applications require a shared home directory among the WNs: for performance reasons the home directory is not shared among the WNs, but a shared area (managed with a gpfs parallel filesystem) is available for applications that produce data of huge dimensions.

Another important feature is related to the monitor of a job execution. We have prepared a watchdog utility that consists of just a shell script file to be executed in background on the WN before starting the parallel job. The watchdog will initialize its environment and will start to monitor and report files to be watched.

5. Running HPC jobs

This section underlnes the main points the user needs to consider to start an HPC job on the grid. The purpose of this description is to show the main points a user must take into account when using the PI2S2 (EGEE based) grid. The user needs a valid login in a valid User Interface and must initialize a proxy certificate for the period when the job runs. Then the following steps must be considered.

5.1. Preparing scripts, jdl command file and submitting a job

The most important limit before starting a job submission, consists in the dimension of the input sandbox: data are directly sent to the grid node when a job is submitted. The input sandbox cannot be larger than 10 MBytes. This limit often does not permit you to send input data files and executables. The user must copy data in the catalogue or can create (and copy) a tarball file containing all data needed by the run. The user can ask to store data, when possible, in the storage element near the grid node where the job will run. The user must prepare the *mpi.pre.sh* scripting file that will run before a parallel execution on the assigned master node, mainly for setting environment variables and starting the watchdog procedure. It can also be used to extract data from the catalogue. The *watchdog.sh* scripting file will be used to monitor the run. The *mpi.post.sh* scripting file will be used to stop the watchodg and to collect output data in the grid catalogue. The last file the user must prepare is the jdl command file where he must specify the file list to be sent and retrived from the grid node, the HPC requirements and eventually the grid node he wants to use for the run. An example of a jdl file is here reported:

```
Type = "Job";
JobType = "MVAPICH_GCC4";
NodeNumber = 8;
Executable = "P-Gadget2";
StdOutput = "P-Gadget2.out";
StdError = "P-Gadget2.err";
Arguments= "input_param.txt";
InputSandbox = {"P-Gadget2",
"input_param.txt","mpi.pre.sh",
"mpi.post.sh","watchdog.sh"};

OutputSandbox = {"P-Gadget2.out",
"P-Gadget2.err"};

Requirements = RegExp("hpc",
other.GlueCEUniqueID);
```

The job submission will give information on the grid node where the job will be effectively executed.

5.2. Input data file management and job status monitoring

The Input/Output data file should be managed with the gfal (Grid File Access Library 2000) libraries, a file access mechanism to access files from the storage system (using the grid catalogue) on the worker node. However to use this library you need to modify the I/O procedure of the code. The *mpi.pre.sh* scripting file is executed by the master node when the job starts. It can extract the tarball file from the catalogue and prepares all the files the code needs to run. An hpc reserved area is provided on each grid node that must be used to download very large datasets. Home directories are not shared among computing nodes, but all of them mount the hpc area with the gpfs parallel filesystem. During the rung the watchdog procedure can copy the stdout and stderr file in the grid catalogue and all other files needed by the user to check the status of the run.

5.3. Retrieving output

At the end of the run the *mpi.post.sh* scripting file will be executed by the master node. It will be used to stop the watchdog and can be also used to create a tarball file of data output that will be copied into the grid catalogue and that can be directly downloaded by the user using the glite commands in the user interface.

6. Future perspectives: connectivity and interoperability

The interoperability among the projects of Southern Italy, represents a fundamental challenge to the development of the grid in Italy. The interoperability is coordinated by the Italian Ministry of Research to share the grid infrastructures and to create a National Large extensive computational grid. Some important steps are done and the agreement on the following points is already set:

– Usage of a common communication standard based on gLite
– Deployment of services based on gLite and LCG
– Certification Authority

The following basic services are also determined: VOMS (Virtual Organization Membership Service), Resource Brokers, Information Index (BDII), Computing Element, Worker Nodes, Storage Elements and User Interface.

The following second level services will be soon determined:

– Virtual file catalog (LFC, AMGA, etc.)
– Portal for resources access (Genius, linemode, etc.)
– Monitoring (GridIce, Service Availability Montoring, etc.)
– Tickets (Xoops/Xhelp, Service Level Agreement, etc.)

- Accounting (DGAS, APEL, etc.)
- Tags (resources description, etc.)

The result will be a testbed platform ready in the next few months to verify the interoperability; at the end, the extensive grid facility will allow Virtual Organizations (VOrgs) partners to run everywhere with a top BDII registration of remote resources,reserved queues for VOrgs partners, reserved area on Storage Elements for VOrgs partners.

7. Summary and conclusions

The new e-Infrastructure of the Consorzio COMETA, created in the context of PI2S2, represents a new opportunity for research and industrial applications in Italy. Huge computing and storage resources are distributed in seven main sites. More than 120 researchers and engineers have been trained so far to access and actually use the Sicilian Grid infrastructure, and a large number of applications have already been gridified, deployed, and are actually running.

Grid and specific HPC facilities have still a gap, but some areas begin to be overlapped and PI2S2 is going in this direction. The Cometa grid will also offer new services for data exploration and data analysis, with new power for Web Services and new opportunities for large datasets.

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 (MUR) within the Programma Operativo Nazionale Ricerca Scientifica, Sviluppo Tecnologico, Alta Formazione (PON 2000-2006). Further information is available at http://www.pi2s2.it and http://www.consorzio-cometa.it. The author thanks Mrs. Luigia Santagati for the English revision of the text.

References

www.trigrid.it

www.glite.org

Becciani, U. *New Grid Infrastructure in Sicily* A Computational Grids for Italian Astrophysics: Status and Perspectives Rome, November 2005, Ed. L. Benacchio, F. Pasian Polimetrica International Scientific Publisher, ISBN 978-88-7699-057-1, 2007, 179-189

Comparato M. et al. 2007 The Publications of the Astronomical Society of the Pacific, 119, 858, 898

www.pi2s2.it/application.

http://grid-deployment.web.cern.ch/grid-deployment/gis/GFAL/gfal.3.html