IGI grid services for the bioinformatics community

L. Gaido[1,2]✉, **M. Bencivenni**[2,3], **D. Cesini**[2,3], **G. Donvito**[2,4], **P. Veronesi**[2,3]

[1]INFN Torino, Italy
[2]IGI, Italy
[3]INFN CNAF, Italy
[4]INFN Bari

Motivations

In the last decade many projects related to grids have been carried out in Europe at both national and international levels with an important economic contribution by the European Commission.The grid middleware developed within these projects has been deployed into the European grid infrastructure (EGI) made by more than 350 sites all over Europe. The Italian Grid Infrastructure (IGI) is part of EGI and is one of the most important and widest national grid infrastructures in Europe since it provides about 33000 CPU cores, 17 PB of disk space and 9 PB of tape capacity spread over more than 50 sites. Although the grid infrastructure has been initially built according to the needs of a few scientific communities (high energy physics, earth observation and Bioinformatics among the others), it has been gradually evolving in order to provide grid services to a wider and wider user base. Several scientific communities are observing that as the instruments become more and more powerful the need for storage and computing is increasing day by day. This will also increase the number of users that could benefit from a geographical distributed computing grid infrastructure.

Methods

In order to support new communities (users and resource providers), various activities have been started. Training has been considered very important, so tutorial for users and grid administrators are regularly organized. In addition great effort has been devoted to understanding the user needs, by defining appropriate use cases, and to supporting the user communities to port their applications on the grid environment (the so-called "gridification"). An important effort is also spent in developing new tools that could make the interaction between the final users and the grid as easy as possible. In particular web tools to submit jobs to the grid infrastructure have been deployed and used by some bioinformatics communities. In order to address the needs of users relying on high-level tools like Workflow managers (e.g. Taverna and other similar tools), a front-end web service has been developed. This web interface could be used as a bridge towards the EGI/IGI grid infrastructure. The IGI community is also providing a service that allows the exploitation of Relational Databases over the grid infrastructure, assuring a high level of security and privacy.

Results

The usage of the standard EGI/IGI resources and services, together with the high level services that IGI is providing on top of the grid, has provided the end users with the capability of carrying out their high demanding computing activities in an easy and reliable way. In the past years, indeed, IGI has supported several bioinformatics communities to "gridify" many different applications such as: ASPic, PAML, MrBayes, CSTGrid, DNAfan, BLAST, BayeSSC, FT-Comar, Muscle, Gene Ontology DB analysis, ABCtoolbox, EMBOSS, Bowtie, SAMtools, Illumina Solexa data processing, AmpliconNoise, BioPython, HMMER. As a result the CPU consumed by various Italian bioinformatics groups on the IGI grid infrastructure has exceeded the 10 years in a few days of activity, thus hugely reducing the overall time needed for the execution of the jobs. Thanks to IGI the bioinformatics users have carried out their analysis in an easy and transparent way, both through simple web interfaces and through complex WorkFlow managers, reducing the time needed to get the results of about 2-3 orders of magnitude. The IGI infrastructure has also been exploited by the Computational Biology group of the Bologna University, in the frame of DUCK (Distributed Unified Computing for Knowledge, a collaboration between multidisciplinary Academic and Research Institutions located in the Emilia Romagna region), to run protein annotation application based on the Bologna Annotation Resource (BAR) method, and to perform massively parallel genome sequencing including about 18 millions of protein sequences. More than 150 computing nodes in the IGI grid infrastructure have been used, successfully dropping the computational times if compared to the computing resources of the local cluster available to the group. The protein annotation application reached a dropping factor of 120, i. e. the computation was performed in few weeks instead of years. Data management facilities offered by the Grid were also exploited to easily handle input and output files. To provide an easy-to-use service to the user communities IGI is developing a web portal that will hide the complexity of the authentication/authorization mechanisms and will also integrate the computing frameworks needed by the different user communities. A prototype of this portal can be easily set up for the bioinformatics community.

Availability

http://www.italiangrid.it/