RUSSIAN JOURNAL OF EARTH SCIENCES, VOL. 11, ES3006, doi:10.2205/2009ES000378, 2010

PROCEEDINGS OF THE INTERNATIONAL CONFERENCE
Electronic Geophysical Year: State of the Art and Results
3–6 June 2009 • Pereslavl-Zalessky, Russia

Grid in Earth Science and its future

M. Petitdidier[1] and H. Schwichtenberg[2]

Received 11 November 2009; accepted 20 November 2009; published 29 January 2010.

The term Grid emerged in the nineties, when the rapid increase of network speeds facilitated the time-efficient integration of distributed computers and storage resources [*Foster and Kesselmann,* 1998]. In the meantime, the resources shared through Grids have been extended to personal computers, servers, compute clusters, supercomputers, storage systems and services like databases and so forth. In 2000 the first European project, DataGrid, to deploy a Grid over Europe was launched, followed up to now by other European projects, called EGEEI, EGEEII and EGEEIII. Since the beginning, 2000, the Earth Science community has started using the Grid. The experience acquired via academic and R&D applications has demonstrated that Grid infrastructure could respond to the ES requirements. However, the interface between the ES software environment and Grid middleware is not simple for many applications. After nearly a decade of European Grid projects like EGEE, that have developed the infrastructure and put it in production for various scientific communities, the next challenge is to find a sustainable model for a continued operational phase. For this purpose, an organizational structure with a central office, called EGI (European Grid Initiative), was founded. In this structure, the provisioning of infrastructure and related services is based on the National Grid Initiatives (NGI) of the participating countries. The user communities are organized according to their scientific domain in so called Grid Virtual Research Communities. *KEYWORDS: Grid, DEGREE, EGEE, EGI, Earth Science.*

Citation: Petitdidier, M. and H. Schwichtenberg (2010), Grid in Earth Science and its future, *Russ. J. Earth. Sci., 11,* ES3006, doi:10.2205/2009ES000378.

1 Introduction

The need for more resources and services in the Earth Science (ES) community follows naturally the evolution and the vision of the research and operation in industry and academy. Fifty years ago during the International Geophysical Year, instruments were widely deployed throughout the world to gain a better understanding of the earth system. At that time each new deployment of instruments at selected locations led to new results and discoveries. Since the introduction of the satellite forty years ago, global synoptic observation has enhanced the vision of the earth system and opened new scientific fields. Then there has been the explosion in the size of databases, with data from a plethora of different sources, observations as well as simulations. Satellite observations have opened (and continue to open) new fields in informatics relative to the handling of large data volumes and number of files, the processing of large amounts of data and their storage, and the discovery of data. As a matter of fact, Petabytes of already acquired data, distributed in different locations, make it difficult for scientists to discover and access data needed for deriving knowledge about the Earth System. To face this data deluge the Earth Science (ES) community has started to use new technologies such as the Web and Grid. The Web services were rapidly adopted in particular to facilitate the access to data, their processing and visualization. However more computing resources are required to exploit those large set of data and to combine them with simulations. Grid and Web-service approaches have much in common as a result of their underlying Internet technology. Grid with its virtual organizations is a perfect example of this new vision of collaboration among virtual teams who work across space, time and organizational boundaries with links strengthened by Internet communication technologies. The first part of this paper is

[1]LATMOS/IPSL, Velizy, France
[2]SCAI/FhG, Sankt Augustin, Germany

focussed on the earth science with its requirements and applications. After a decade of Grid project the European grid goes into an operational phase that it is described into part 2 with the impact on the user community using the Grid, especially on the Earth Science one.

2 Earth Science Domain

2.1 Requirements

The Earth Science scientific domain is a large mosaic of disciplines that are very heterogeneous as it concerns the complexity of informatics problems they have to solve, the need and characteristics of computing and storage resources. Furthermore, according to the stage of the research and maturity of the applications, the computing needs may vary from a laptop to supercomputer via clusters. Grid infrastructure in Europe is built with clusters and farms of computing elements.

A survey of the requirements of the ES community relative to data management, job management and control, and portals was carried out in the EU DEGREE (Dissemination and Exploitation of Grids in Earth Science) project (see deliverables on http://eu-DEGREE.eu). Once the requirements identified, DEGREE project also evaluated the Grid tools and services developed by various Grid projects that may fulfil those requirements in order to identify the gaps between the available infrastructure and middleware, and the ES requirements. DEGREE has disseminated the key ES requirements to Grid developers. In order to convey requirements to the Grid community, test suite specifications were developed [*Som de Cerff et al.* 2009]; test suites providing real applications for testing functional and non-functional aspects of Grid, contrary to typical whiteboard tests or use cases.

Among all the requirements, some critical points have been identified and concern the access of data located in geographically-distributed data centres outside the Grid infrastructure, the weather and risk prediction that needs a resource reservation in order to provide the information in time, and standard access of MPI all over the Grid sites. The security, respect of data policies, and confidentiality are key conditions for adoption of Grid by the ES community. As for any tools, the requirements for Grid are reliability, robustness and easy use.

Even if there are critical requirements not fulfilled for Earth Science, EGEE (Enabling Grid for e-Science) and the other related projects have been able to satisfy the needs of thousands of users from a variety of scientific domains. The penetration of Grid in ES communities is pointed out by the number of countries, around 20 European and associated countries, in which ES applications are ported and the variety of applications, topic of the next paragraph.

2.2 Applications on Grid

Due to its intensive data processing and highly geographically distributed organizations, the multidisciplinary Earth Science (ES) community is uniquely positioned for the uptake and exploitation of Grid technologies. Members of the ES community have been participating in Grid projects for a long time with great success. Members of the community have been participating amongst others in EGEE[1], CYCLOPS[2], SEE-GRID[3], EELA[4], regional Grids and in European Space Agency Grid initiatives[5]. The Grid ES community is of substantial size and still growing, consisting of members from Europe, Latin America, China and the USA. In Europe, ES persons, using the Grid and belonging to one of the seven ES virtual organizations, are spread over around 20 countries. Scientists from Taiwan participate also to the ES activities in EGEE.

Applications cover several disciplines, such as atmospheric chemistry, climatology, geosciences, hydrology, meteorology, pollution, seismology, etc. [*Renard et al.,* 2009; *Iapaolo et al.,* 2007; *Fernadez-Quirelas et al.,* 2009; *Lecca et al.,* 2009; *Kussul et al.,* 2009; *Clévédé et al.,* 2009] and many services such as research, regional weather prediction, risk assessment, civil protection [*Mazzetti et al.,* 2009; *Raoult et al.,* 2009]

There are too many applications ported on Grid to describe each of them. Many applications are based on the availability of a large number of computing resources. A success story, related to the pesticide risk assessment in the framework of the European project, FOOTPRINT (http://www.footprint.org), consisted in processing in a relatively short time, few months, a millions of jobs, around 1hr each. Other success story concern the sharing of data on the Grid. The first ES application on the Grid was to retrieve ozone profiles from GOME satellite data and validate the 7-years of data with the 7 ozone lidar data [*Iapoalo et al.,* 2007]. In this example, data are available for the authorized scientists on the Grid without any need to download them in the concerned institutes. Another interesting feature of the Grid is to deploy a software and make it available for end-users. The Compagnie Générale de Géophysique Veritas has achieved the implementation of its generic seismic platform software, based on Geocluster commercial software, and then support academic teams using it [*Delecluse et al.,* 2008]. It permits to those academic teams to have the last version of the software without having to download and install it on their machine and may use it at a larger scale due to computing resources available. There are also other examples in seismology and modelling.

To fulfil ES requirements some services and tools have been developed. The CYCLOPS project intended to point out the interest for the European Civil Protection to use the

[1]Enabling Grids for E-sciencE http://eu-egee.org

[2]Cyber-Infrastructure for CiviL protection Operative ProcedureS http://www.cyclops-project.eu

[3]South Eastern European GRID-enabled eInfrastructure Development http://www.see-grid.eu

[4]E-science grid facility for Europe http://www.eu-eela.eu

[5]Earth Observation GPOD (Grid Processing On Demand) http://gpod.eo.esa.int

Grid infrastructure, especially during critical events, such as fire and flood [*Mazzetti et al.,* 2009]. A Grid platform on the EGEE gLite-middleware was developed and included Open Geospatial consortium services to fulfill the need of Geographical Information Systems. Those OGC components in connection with Grid computing are also well suited to accomplish high processing performance in geo Sciences [*Lanig and Zipf,* 2009]. Another critical point is the search and discovery of data via metadata, especially when they are distributed in different locations. Climate-G, is a research effort devoted to the Climate Change community. It is a distributed testbed among several European centres for climate change addressing challenging data and metadata management issues at a very large scale. The testbed is an interdisciplinary effort joining expertise in the field of climate change and computational science. The main goal of Climate-G is to allow scientists carrying out geographical and cross-institutional data discovery, access, visualization and sharing of climate data by using the Grid Relational Catalog (GRelC) services [*Fiore et al.,* 2009].

A large number of applications that heavily rely on the use of Earth remote sensing data from space have been deployed. This includes rapid flood mapping from satellite radar imagery for the United Nations Platform for Space-based Information for Disaster Management and Emergency Response (UN-SPIDER) and International Federation of Red Cross [*Kussul et al.,* 2009]. Activities in this area provide contribution to the development of the Global Earth Observation System of Systems (GEOSS) and Global Monitoring for Environment and Security (GMES).

The next development will be the building of scientific gateways, as already exist in TeraGrid (http://www.teragrid.org) in the USA, to integrate Grid services (job submission, compute and storage resource discovery, catalog, etc.), generic tools such as workflow, visualization softwaresand specific tools relevant to specific user communities, such as algorithm, access to some data centers, knowledge database.

3 European Grid Initiative (www.eu-egi.org)

Since 2000 the pan-European grid infrastructure has been developed and operated through a series of e-Infrastructure projects such as DataGrid, EGEE, SEEGrid, DEISA and so forth. EGEE, with CERN for coordinator, is one of the largest Grid infrastructures in Europe and connects more than 260 sites in around 55 countries, in and out Europe, with an approximate number of 150.000 processing cores (available to users 24/7).

EGEE and the other related projects were able to satisfy the needs of thousands of users from a variety of scientific domains. Besides compute and data storage demands, the Grid infrastructure also supports new capabilities for e-collaboration and multi-scientific domain interaction. After a decade of European Grid projects, the operational phase consists of an operating large-scale production grid infrastructure to serve various scientific communities. For this purpose, EGI-DS (Design Study for a European Grid Infrastructure – EGI) provided the conceptual setup and operation of a new organizational model of a sustainable pan-European Grid infrastructure. EGI will establish a sustainable, pan-European grid infrastructure, available 24 hours-a-day, to support leading edge collaborative e-Science in Europe. This model is thought to be capable of fulfilling the vision of a sustainable European Grid infrastructure for e-Science.

3.1 European Grid Infrastructure

The deployment of Grid sites over Europe has led to the development of national organizations, called National Grid Initiative (NGI), in order to operate the Grid infrastructure and coordinate the Grid activities within their country. The foundation of EGI, that will take place in 2010, are the National Grid Initiatives (NGIs). EGI will interconnect existing NGIs and will actively support the setup and initiation of new NGIs. EGI is also based on the NGIs for scientific and technical expertise as well as funding; a part of the EGI funding coming also from the European Commission.

The EGI council, decisional board of EGI, gathers the representatives of all the NGIs that signed the Memorandum of Understanding (MoU) and provided funding, associated and nonvoting members. The first meeting of the EGI council was held in July 2009. The associated members sign also the MoU with EGI.eu and provide funding to EGI. Their number of votes depends on the amount of funding. CERN has already signed, ESA is ongoing. Non-voting representatives concern the non-European Grid partners.

The EGI infrastructure will be based on the existing Grid middlewares like gLite, Unicore and ARC, used by the largest infrastructures in operation EGEE, DEISA and NorduGrid, respectively. Future developments and interoperability actions will be done by the European Middleware Initiative (EMI) that will not be part of EGI.eu but will be strongly linked to EGI. From the first European Grid project, the user community has been organized by scientific domains such as High Energy Physics (related to LHC experiment at CERN), Life Sciences, Earth Science, Astronomy & Astrophysics, Computational Chemistry, Fusion, Grid Observatory. Those Grid virtual communities have been building through different Grid projects. The core of the Grid Virtual Research community (VRC) constitutes the Specialized Support Centre of the VRC.

3.2 Earth Science Grid Virtual Research Community

The main goal of a given scientific VRC is to give support to user groups of its scientific community in their effort of effectively access and exploit the Grid-based DCI (Distributed Computing Infrastructure) operated by EGI.eu and by NGIs for their scientific, commercial, industrial or social objectives. VRCs should have a visibility in regard to the community they represent, a medium and long-term lifetime (several years) and answer the requirements of the commu-

nity in regard of the Grid use. They have close links with EGI, the other VRCs and the other Grid related projects.

A questionnaire about the interest in an ES Grid VRC and their aspired involvement was widely disseminated to ES Grid partners and groups. Until today we have a considerable number of replies from more than twenty institutes, universities and laboratories from Albania, Armenia, Bielorus, Bulgaria, France, Germany, Greece, Italy, Netherlands, Romania, Russia, Slovakia, Spain, Switzerland, Turkey, UK and Ukraine. They all expressed their intention to participate in the ES Grid VRC at different degrees of involvement. Partners from the various projects have shown a strong scientific and technical interest to work together in the same structure in order to disseminate the Grid usage in the ES community. The Grid VRC with all the Grid end-users constitutes the ES Grid Virtual Research Community. Different ES disciplines are involved in dedicated thematic physical infrastructures for data storage, access and exchange as proposed by the European Strategy Forum on Research Infrastructures (ESFRI). Relevant data (observation and model) and information are mainly outside of the Grid infrastructure, thus a major challenge is providing support for deploying the application and accessing data located in various locations. The ES specialized support center must address this data Grid issue with qualified support from ES and computer science experts. The data access to the ES related data centers is not only a technical question to solve, where the ES VRC has to provide interfaces and to consult about standards, esp. the OpenGIS Web Services of the Open Geospatial Consortium (OGC). It also concerns management related questions about establishing relationships with data and information providers and managers. Other challenges are the implementation of usual ES environment and of complex applications on Grid infrastructure.

The proposal to ask for funding the European Commission will be submitted by the end of November 2009.

4 Conclusion

This short overview points out the penetration of Grid within the European ES community and ES disciplines. The future organization of the Grid user communities into a Virtual Research Community with a Specialized Support Centre, as an expertise centre, will permit to address critical requirements that are not yet fulfilled.

References

Clévédé, E., D. Weissenbach, B. Gotab (2009), Distributed jobs on EGEE Grid infrastructure for an Earth Science application: Moment Tensor computation at the centroid of an Earthquake, *Earth Science Informatics, 2*(1–2), 97-106. doi:10.1007/s12145-009-0029-4

Delescluse, M., N. Chamot-Rooke (2008), Serpentinization pulse in the actively deforming Central Indian Basin, *Earth Planet. Sci. Lett., 276,* 140-151.

Fernández-Quirelas, V., J. Fernández, C. Baeza, A. S. Cofino, J. M. Gutierrez (2009), Execution management in the GRID, for sensitivity studies of global climate simulations, *Earth Science Informatics, 2*(1–2), 75-82. doi:10.1007/s12145-008-0018-z

Fiore, S., S. Vadacca, A. Negro, G. Aloisio (2009), Data issues at the Euro-Mediterranean Centre for Climate Change, *Earth Science Informatics, 2*(1–2), 23–35. doi:10.1007/s12145-009-0023-x

Iapaolo, M., et al. (2007), Gome Ozone Profiles Retrieved By Neural Network Techniques: A Global Validation With Lidar Measurements, *Journal of Quantitative Spectroscopy and Radiative Transfer, 107,* 105–119.

Kussul, N., A. Shelestov, S. Skakun (2009), Grid and sensor web technologies for environmental monitoring, *Earth Science Informatics, 2*(1–2), 37–51. doi:10.1007/s12145-009-0024-9

Lanig, S., A. Zipf (2009), Interoperable processing of digital elevation models in Grid infrastructure, *Earth Science Informatics, 2*(1–2), doi:10.1007/s12145-009-0030-y

Lecca, G., C. Lai, F. Murgia, R. Biddau, L. Fanfani, P. Maggi (2009), AQUAGRID: an extensible platform for collaborative problem solving in groundwater protection,, *Earth Science Informatics, 2*(1–2), 83–95. doi:10.1007/s12145-009-0019-6

Mazzetti, P., S. Nativi, V. Angelini, M. Verlato, P. Fiorucci (2009), FiorucciA Grid platform for the European Civil Protection e-Infrastructure: the Forest Fires use scenario, *Earth Science Informatics, 2*(1–2), 53–62. doi:10.1007/s12145-009-0025-8

Raoult, B., G. Aubert, M. Gutiérrez, C. Arciniegas-Lopez, R. Correa (2009), Virtual organisation in the SIMDAT meteorological activity: a decentralised access control mechanism for distributed data, *Earth Science Informatics, 2*(1–2), 63–74. doi:10.1007/s12145-009-0026-7

Renard, P., V. Badoux, M. Petitdidier, R. Cossu (2009), Grid computing for Earth Sciences, *Eos, AGU transactions, 90*(14), 117–119.

Som de Cerff, W., M. Petitdidier, A. Gemünd, L. Horstink, H. Schwichtenberg (2009), Earth science test suites to evaluate grid tools and middleware-examples for grid data access tools, *Earth Science Informatics, 2*(1–2), 117–131. doi:10.1007/s12145-009-0022-y

M. Petitdidier, Laboratoire d'Atmosphere, Milieux, Observations Spatiales, IPSL-10-12 Avenue de l'Europe, 78140 Velizy, France (monique.petitdidier@latmos.ipsl.fr)

H. Schwichtenberg, SCAI/FhG, Sankt Augustin, Germany (horst.schwichtenberg@scai.fraunhofer.de)