Geoinformation for European-wide Integration, Benes (ed.)

SpaceGRID: An international programme to ease access and dissemination of Earth Observation data/products – How new technologies can support Earth Observation Users Community

F. Lamberti & S. Beco
DATAMAT S.p.A., Rome, Italy
fiorella.lamberti@datamat.it

Keywords: Earth Observation, GRID, collaboration, distribution, computation, application

ABSTRACT: Dozens of satellites are constantly collecting data about our planetary system - the Earth in particular - 24 hours a day 365 days a year. Satellite data is used for many purposes, for example for telecommunications, navigation systems and environmental monitoring. However, even with the most powerful computers processing all this data is time-consuming and expensive. Distributing these tasks over a number of low-cost internet-connected platforms would provide enormous potential, at a relatively low cost, for many space applications. To look into this exciting possibility ESA started in September 2001 to work on its SpaceGRID study financed by the Agency's General Study Program. The main aim of this project is to analyze and assess how GRID technology can serve requirements across a large variety of space disciplines, in particular Earth Observation (EO), Space research (e.g. Space Weather Simulation, S/C Plasma interaction, Radiation Transport simulation), Solar System Research, Spacecraft (Mechanical) Engineering. The basic idea behind the GRID technology is to harness users' computers through the Internet in order to exploit the full capacity of such distributed computing power to solve major computational and distribution problems, instead of using just one big and expensive computer. This paper provides a summary of the first results about the analysis of the benefits that can be derived from the application of the GRID Technology to current EO applications.

1 INTRODUCTION

At present several thousand ESA users world-wide have online access to EO missions related meta-data (10 million references), data and derived information products acquired, processed and archived by more than 30 world-wide stations. ESA-ESRIN is the core management facility of the European EO infrastructure, which operates, since more than 20 years, EO payload data from numerous Earth Observation Satellites (ERS, Landsat, JERS, AVHRR, SeaWiFS are the presently active missions) owned by ESA, European and other International Space Agencies. Currently ESA managed EO missions download about 100GBytes of data per day. This number is currently estimated in the range of 500Gbytes after the launch of the ENVISAT satellite in March 2002. The products archived by ESA facilities are estimated at present in 800 Terabytes and will grow in the future with a rate of at least 500 Terabytes per year.

The advent of the GRID technology stimulates the analysis of existing EO applications, with the major aim to verify whether the EO Users Community can benefit from the GRID. This means to analyze how to improve archive access, products generation and distribution, and in general to enhance the exploitation of the EO data and products.

Figure 1. The SpaceGRID Project Logo

In the frame of the SpaceGRID project, the EO domain has been investigate for identifying users requirements to be taken into account should EO applications be ported in a GRID environment. These specific requirements have been derived by crossing GRID technical features with the users' needs expressed in the answers to a questionnaire, mailed to many ESA/ENVISAT Principal Investigators.

This paper will present the results achieved so far in the course of the project. First of all, an introduction of the SpaceGRID project will be provided along with an overview of the GRID technology. Then, the analysis of requirements will be presented with the major conclusions.

2 THE PROJECT

The major goals of the SpaceGRID project are:

- to assess how GRID technology can serve requirements across a large variety of space disciplines (spacecraft mechanical engineering, space weather, space science, earth observation
- to foster collaboration and enable shared efforts across space applications
- to sketch the design of an ESA-wide (and common) infrastructure
- to demonstrate proof of concept through prototyping
- to involve both industry and research centers
- keep Europe up with GRID efforts!!

Both industries and research centers are involved in this study to keep Europe up with GRID efforts. An international consortium of industry and research centers led by Datamat S.p.A. (Italy) constitutes the project team. Other members include Alcatel Space (France), CS Systemes d'Information (France), Science Systems Plc (UK), QinetiQ (UK) and the Rutherford Appleton Laboratory of the UK Council for the Central Laboratory of the Research Councils.
. The results of the study should be available within 18 months, but meanwhile ESA will be kept informed of the progress being made and project activities will be synchronized with the ESA internal GRID initiative.

Two other aspects of this project are of particular importance for ESA: finding a way to ensure that the data processed by the SpaceGRID can be made available to public and educational establishments, and ensuring that SpaceGRID activities are coordinated with other major international initiatives.

The paper will present the results of the study activities on EO applications up to the conference date, mainly addressing the results of requirements analysis of EO data-intensive applications requiring high volumes of processing resources. This analysis will be the basis for the creation of set of requirements for "GRID-enabling" standards, which the developers of GRID-aware EO applications have to comply with. Another important activity will address the definition of a EO user community collaborative environment, e.g. a support environment, which helps the collaboration between scientists and provides a common e-space where people can interactively and in real-time process and visualize results. This should allow user community an immediate sharing of experiences towards a common goal (e.g. analyze a complex event to provide an objective response in the shorted delay).

3 THE GRID TECHNOLOGY

This section provides a summary description of the GRID Technology and of the expected benefits that can result from its application to space domain.

3.1 *What is the GRID*

"The GRID is an emerging infrastructure that will fundamentally change the way we think about - and use - computing. The GRID will connect multiple regional and national computational grids to create a universal source of computing power."
("The GRID: Blueprint for a New Computing Infrastructure", Foster and Kesselman Eds., Morgan Kaufmann Publishers)

The idea behind is simple. As very few people or organizations use the full capacity of their computers' processing power, the idea is to harness this unused resource through the Internet to solve major computational problems that require far more 'memory' and computing power than that available at any one site.
Extending this concept, a GRID can be seen as an infrastructure that couples:

- Heterogeneous computers (PCs, workstations, clusters, traditional, supercomputers)
- Software
- Databases
- Special instruments (e.g. radio-telescope, like in SETI@home, see after)
- People

across the Internet and presents them as a unified integrated single resource.

A scaled-down version of this idea was first put into practice a year ago by the Search for Extra Terrestrial Intelligence (SETI) project, with SETI@home. By enlisting the help of volunteers, they are able to 'plug into' the processing power of millions of home computers to sift through radio signals to search for messages from space. The enormous potential of this idea gave rise to different GRID projects, which have now been set up in many areas of the world.

Within this context, a computational GRID is a collection of distributed, possibly heterogeneous resources, which can be used as an ensemble to execute large-scale applications; a computational GRID is also called metacomputer. The term "computational GRID" comes from an analogy with the "electric power grid":

- Electric power is ubiquitous
- Doesn't need to know the source (transformer, generator) of the power or the power company that serves it

Therefore, in the same way we plug into the electricity grid to run an electrical appliance whenever needed, one can plug into a computational GRID to run a computing power-demanding application whenever needed.

Let us analyze now what is the difference between GRID computing, Cluster computing and the Web.

- Cluster computing focuses on platforms often consisting of homogeneous (PCs or workstations) nodes in a single administrative domain, interconnected using relatively fast networks. In case of Cluster computing, application focus is on cycle-stealing computations, high-throughput computations, distributed computations.
- Web focuses on platforms consisting of any combination of resources and networks which support naming services, protocols, search engines, etc. The Web consists of very diverse set of computational, storage, communication, and other resources shared by an immense number of users: in this case the application focus is on access to information, electronic commerce, etc.
- GRID computing focus on ensembles of distributed heterogeneous resources used as a single virtual platform for high performance computing: in this case, some resources may be shared, other may be dedicated or reserved. GRID applications focus is on high-performance, resource-intensive applications.

3.2 *The GRID Opportunity*

GRID is an emerging global infrastructure that combines geographically distributed and heterogeneous computing capabilities into a single, powerful computing resource. The vision is to associate advanced computer networks, databases, sensors and people to a universal GRID infrastructure, which can be accessed by users in an inexpensive and consistent way. Thus, everybody could have access to an infinite amount of computing power, and new classes of high-complex applications could emerge (see [Foster, Kesselmann, 1999]).

The GRID technology integrates both, the coordinated use of networks, end-system computers, data archives, various sensors and advanced human computer interfaces, and the provision of a set of enhanced "middleware" services for obtaining information about available GRID components, locating and scheduling resources, accessing data resources, communicating, measuring performance, authenticating users and resources, ensuring the privacy of communications, etc. These services constitute a generic "virtual machine" that simplify and support GRID-aware application development.

In [Foster, Kesselmann, 1999], five major application classes have been identified, which could best use computational GRIDs. The first class is represented by ***distributed supercomputing applications***. They could use the GRID infrastructure to aggregate substantial computational resources (e.g. CPU, memory, etc.) in order to tackle problems that cannot be solved on a single system.

The second class includes ***high-throughput computing applications***, which could use the GRID to schedule large numbers of loosely coupled or independent tasks, with the goal of putting unused process cycles (often from idle workstations) to work.

The third class however describes ***on-demand computing*** applications.

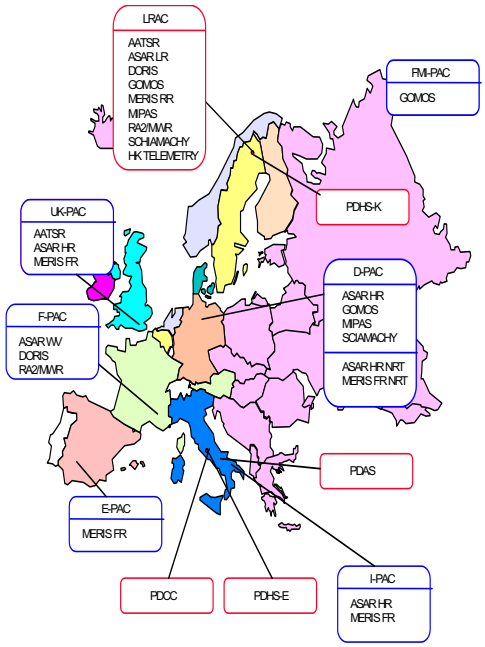

Figure 2 ENVISAT Distributed Payload Data Segment

They could use the GRID capabilities to meet short-term requirements for resources that cannot be cost-effectively or conveniently located locally.

The fourth class consists of ***data-intensive computing*** applications, which focus on synthesizing

new information from data that is maintained in geographically distributed repositories, digital libraries and databases by means of the GRID.

The fifth and last class includes ***collaborative computing*** applications, which are concerned primarily with enabling and enhancing human-to-human interactions. Here, the GRID infrastructure could support, for instance, the shared use of computational resources such as data archives and simulations.

At the moment several investigations are carried out to investigate about the benefits of this advanced technology in many different research areas.

Within the Earth Observation domain, there are several promising features. In fact, EO application systems are characterized by their intensive use of large EO data files, which are spread out over geographically distributed archives. At present, most of the data is accessed off-line and the service provided to the user community is far from being optimal, due to the complexity of product format, algorithms and processing required for meeting the specific user needs.

The GRID technology can be applied to an EO application framework as supporting infrastructure for the provision of access to this large amount of distributed EO data and computing resources. It could support the execution of EO application services as well as the processing of thematic application products by a vast set of collaborating users.

Besides, it could provide a positive impact to the near-real time delivery of thematic EO application products. Finally, the GRID infrastructure inherently could meet distribution and security, as well as performance and reliability requirements of any EO application.

4 THE ANALYSIS OF THE EO DOMAIN

The identification of EO application users requirements towards GRID technology has been carried out through the following steps:

1 Analysis of the Earth Observation domain, carried out through available documentation in order to provide a basic layout of the domain in terms of:

- The system, e.g. the space and the ground segments, with major emphasis on the archives where EO data are maintained.
- A general description of the basic level of product generation
- An overview of EO applications and related critical points

2 The EO User Community: in parallel to the collection of the suitable information for the above steps, a questionnaire campaign has been used for refining the knowledge about the users community. In addition some EO applications looking suitable for porting to a GRID environment, have been analyzed and specific interviews have been carried out with the involved personnel. Based on the analysis of this input information, EO related requirements have been identified and classified.

A summary of these tasks is herein reported in the following.

4.1 *The EO DOMAIN*

Many studies and workshops of several space agencies and industrial companies on the exploitation of EO data coming from spacecraft instruments identified a vast number of potential EO applications.

To give an idea of EO data volumes some examples are given in the following:

ESA/ESRIN in the next few years will handle 400 TB of EO data per year. ESA archives already hold 800 TB of data that has to be maintained, including 86,000 high-density tapes, and 42,000 tapes have been transcribed and recycled more than once. Such a large volume of data introduces a great deal of complexity with different data formats and metadata information, with ensuing complications for providing data integration and accessibility. Yet this complexity is expected to increase, along with the data volumes, with future planned ESA missions, particularly after the launch of Envisat carrying ten new EO instruments.

At the end of 1999 there were around 7.5 million SPOT scenes in the SPOT Image central catalogue and at different sites distributed around the world (Toulouse, Kiruna, etc.). SPOT data in Toulouse and Kiruna alone were approximately 130 TB.

NOAA has an archive of approximately 1Pb of data from its major satellite programs. In the next ten years this will grow to 9Pb and will reach 14 Pb in 2014 with the ingestion of data from the Earth Observation System program.

The most promising applications and where the scientific work is well advanced can be mainly identified in the following areas:
meteorology, ozone and climate modeling, agriculture, forestry, flood monitoring, fire detection and monitoring, hydrology, geology, sea state climatology, oil spill detection and monitoring, sea ice mapping, and coastal zone monitoring.

Until recently, EO applications were mainly limited to scientific use. However, in order to justify investments in space technologies, which still are largely public funded, new emphasis has been given in recent years to the development of ground segments with focus on exploiting data streams received from the satellites for commercial and operational applications.

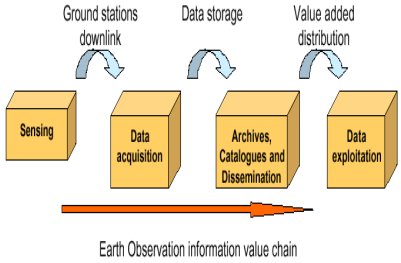

Figure 3 The Earth Observation information value chain

EO application systems are characterized by the provision of thematic, value-added services that are based on the intensive use of EO data from space. Raw EO data products, which typically include digital units of several hundred megabytes, are elaborated into increasingly complex and specialized higher-level products using a series of processing steps. This processing converts digital units to geophysical parameters and may involve many different levels of complexity, depending on the type of instrument, the data being analyzed and the thematic application.

Consequently, the thematic processing of EO data products typically requires large computational power, and access to a high amount of large data files.

EO application systems usually use and are built on top of a set of generic functions, which occur in most EO applications such as image processing or product cataloguing functions.

Due to hardware dependencies and data policies, most of these functions are restricted to specific locations and, thus, can only be invoked remotely. Furthermore, EO application systems involve collaborating actors, which typically are spread out across different geographical locations. Although these locations sometimes are connected by low-performance and unreliable communication networks, thematic EO application services are supposed to be executed as fast as possible, in order to eventually guarantee near-real-time processing during for instance emergency phases.

EO application systems are still mainly used in scientific environments. Processing routines are thus in most cases experimental and subject of constantly changing requirements. Besides, EO application systems strongly depend on data sources, whose availability and quality can unexpectedly vary (e.g. as in the case on an unexpected damage of a spacecraft instrument).

These constraints lead to a restricted use of EO data/products with respect to the huge level of information that is available in the EO data. In fact the lack of consolidated performance, reliability and operational requirements may affect the end users acceptance of EO applications as data provision elements within their operational context. In fact the integration of EO data/products within an operational context strongly depends on the value and quality of the thematic information extracted within the EO application system.

4.2 *The EO Users Community*

EO application users interested in grid metacomputing facilities may belong to the following categories

- Commercial and scientific end-users
- Operational users
  - National agencies
  - Science institutions
- Intermediate users
- Scientific and research end users
- Data providers
- Value Adding Companies and service companies

Many users focus their investigations on clearly delimited areas of geographical or topical interest; they may subscribe to delivery of products only for a specific time period or may require specific data related to the study of particular natural events.

In order to improve the knowledge of the EO User Community and to the interest towards GRID technology, a questionnaire has been prepared and distributed through an e-mail campaign.

The questionnaire had a twofold purpose:

- On the one hand, identify the types of EO applications for which a great interest and related activity exist in the EO User Community
- On the other hand, to analyze the type of interest of the EO User Community towards a GRID technology and the benefits that could be derived when porting EO applications within a GRID environment.

The following aspects were addressed in the questionnaire:

- ***Organization,*** to analyze whether the use of EO data/products is for a research institute/ private company
- ***Application***, to check whether a common trend exists about the use of EO data/products
- ***Type of EO Data/Products,*** also addressing the physical support means to receive them.
- ***Data utilisation,*** to understand to which extent the use of EO data/products is within research purpose or if already operational applications are available.
- ***Collaboration,*** to have a glance about the current ways of exchanging results
- ***Awareness of the GRID technology,*** to perceive how scientists and in general EO User Community feel that his/her own application can benefit with the GRID Technology.

The analysis of the collected answers led to the definition of general requirements and specific constraints able to describe if and how Earth Observation applications are suitable to be ported in a GRID environment.

The requirements were generated by a combination of:

- EO expert experiences
- Systems specification documents from EO applications
- Feedback from users (collected through questionnaires and interviews)

Additional requirements of a more general nature have also been added as a reflection of the broad and long-term vision of grids expressed by users.

These EO application requirements have been classified basically into the following main groups:

- Functional requirements
- Constraints requirements
- Performances
- Security
- Information service requirements
- Scheduling requirements
- Remote data access
- User services

These sets of requirements will represent the basis for the definition of a GRID-aware infrastructure suitable for the needs and features of Earth Observation applications.

5 CONCLUSIONS

In general the use of GRID in specific application processing is still considered by users in a very embryonic phase. However, positive feedbacks have been collected as GRID opens up the possibility of projects handling much larger amounts of data, and this is considered as a general trend in user needs.

Ocean and sea ice applications appear particularly promising for the development of GRID projects, because of the need to handle large amounts of data (often from several different sensors) to demonstrate behaviour over regional scales and over potentially long times. However, in some cases the potential is limited by data availability - e.g. there are not yet sufficient regular, repeat SAR observations over sea ice to demonstrate the full potential of SAR to monitor the evolution of sea ice on long time scales. In particular there are some problems concerning the evaluation of some SAR products from ERS-2 with respect to the ENVISAT ASAR due to the different features of the two sensors.

The issue of data volume is generally more important than processing power for many EO applications. For instance, shipping data around is a big job, but the real issue is to stimulate data usage by supporting research using EO data. This creates a data user community and product development. Data fusion and image classification could also represent interesting fields to explore.

Also reported from users, there is the need for "commercial in confidence". Then security of any application running on such GRID is essential. In particular, many processing software tools are of specialist nature and require knowledgeable operators.

EO data products and users are usually spread out across countries around the world. Moreover, in order to cope with the complexity of data analysis and processing, users often share experiences and simultaneously cooperate with one another.

An EO GRID infrastructure may support cooperative work in distributed environments, and provides a common work and information space where scientists can share in real time and interactively process results. Since remote sensing high-resolution products are generally very expensive and related data policies very restrictive, an EO GRID infrastructure could also provide security against illegal and unauthorized access and ensures dissemination of sensitive data in accordance to data policies.

Distribution is an indispensable characteristic of today's computing systems. However, it has the inherent property that data is transferred over public networks and thus it has to be protected against external hacker attacks.

In addition, EO applications demand for high quality and performance as support to near-real-time processing and provision of application services. Since EO high-resolution data is generally very cumbersome to handle (e.g. 140 MB for an ERS-2 PRI image) and a large amount of users might simultaneously access the computing system, this feature is a significant challenge for any technological solution based on the GRID Technology.

ACKNOWLEDGMENTS

Special thanks to the ESA staff, in particular to Mr. Pier Giorgio Marchetti and Luigi Fusco for the support provided during the study.

The authors would also like to thank all Principal Investigators that provided a valuable contribution to the SpaceGRID project.

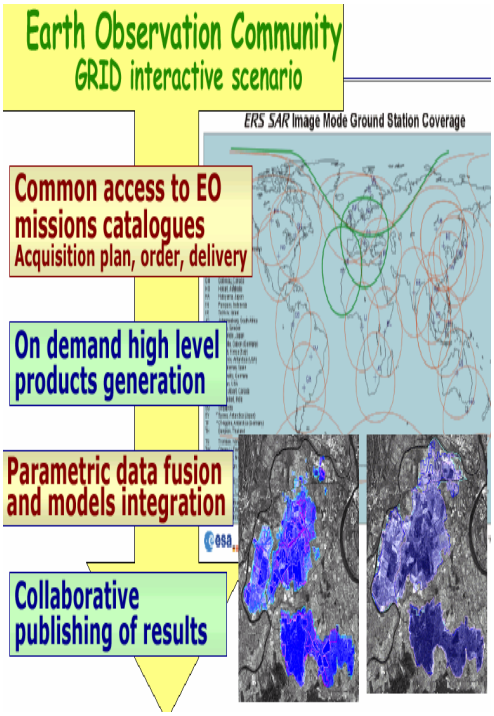

Figure 4. EO Communities: GRID Interactive Scenario

REFERENCES

I. Foster and C. Kesselman, Eds., *The GRID Blueprint for a New Computing Infrastructure,* Morgan Kaufmann, 1999

G. Megle, C.J. Readings, *The Earth Explorer Missions - Current Status*, Earth Observation Quarterly, No. 66, July 2000

European Space Agency: Opportunities for Science and Applications. Envisat Mission. ESA Publications Division.1998.

Data Grid Requirements Specification, DataGRID–09-D9.1-0101-1_2-Requirements, date 27/08/2001

EU DataGRID Project, http://www.eu-datagrid.org, http://tempest.esrin.esa.it/~datagrid

Globus Project, http://www.globus.org

Foster, I., Kesselman, C. and Tuecke, S. *The Anatomy of the Grid: Enabling Scalable Virtual Organizations.* International Journal of High Performance Computing Applications, 15 (3). 2001

http://www.globus.org/research/papers/anatomy.pdf.

Ian Foster, Cark Kesselman, Jeffrey M. Nick, Steven Tuecke, 'The Physiology of the GRID, an Open Grid Services Architecture for Distributed Systems Integration,

http://www.globus.org/research/papers/ogsa.pdf

Additional information about the SpaceGRID project can be found at the following web page:

http://esagrid.esa.int/spacegrid/index.html