Published in EOS, Transactions American Geophysical Union 90, issue 14, 117-119, 2009
which should be used for any reference to this work

Grid Computing for Earth Science

By P. Renard, V. Badoux, M. Petitdidier, and R. Cossu

The fundamental challenges facing humankind at the beginning of the 21st century require an effective response to the massive changes that are putting increasing pressure on the environment and society. The worldwide Earth science community, with its mosaic of disciplines and players (academia, industry, national surveys, international organizations, and so forth), provides a scientific basis for addressing issues such as the development of new energy resources; a secure water supply; safe storage of nuclear waste; the analysis, modeling, and mitigation of climate changes; and the assessment of natural and industrial risks. In addition, the Earth science community provides short- and medium-term prediction of weather and natural hazards in real time, and model simulations of a host of phenomena relating to the Earth and its space environment. These capabilities require that the Earth science community utilize, both in real and remote time, massive amounts of data, which are usually distributed among many different organizations and data centers.

The Earth science community can benefit greatly from technology that can provide ready access to computing resources and services, easily managed data and metadata storage in distributed systems or in data centers, clearly defined data policy, authentication, confidentiality, and electronic collaboration. Grid infrastructure and systems meet these requirements as a distributed resource system. Grid computing permits the sharing of resources between institutions and allows for scaling up computing power and storage capacity in a way that is impossible for a single institution to do. Also, grid computing offers a transparent collaborative platform for users, allowing them to have access to more resources at a given time. This access is especially important for exploiting large data sets scattered in several locations, for running large statistical jobs, and for sharing data and algorithms among many partners without the need for conversions.

Earth science computing and data management needs traditionally have been provided for by local and national institutions. The limitations of cost and the number of computer central processing units (CPUs) available at any one site can be overcome by geographically distributed systems for accessing data, computing resources, and Web services. Geographically distributed computing began by catering to specialized purposes or particular user groups (e.g., the Distributed European Infrastructure for Supercomputing Applications). The Berkeley Open Infrastructure for Network Computing (BOINC), introduced in 1999, pioneered the use, on a volunteer basis, of the enormous processing power of personal computers (PCs) around the world. Earth scientists adopted BOINC only for specific climate and hydrology applications because it is not general enough and does not handle the problems of PC heterogeneity and confidentiality requirements.

Grid Computing

Grid computing emerged more than 10 years ago [*Foster and Kesselman*, 1998] as one type of distributed resource system. Grid computing consists of a network infrastructure comprising loosely coupled heterogeneous data storage and computing resources connected via the Internet and controlled for management and access by software (middleware) such as gLite, UNICORE, Globus Toolkit, and GRIA. A grid system is based on long-term and dynamic collaboration among grid partners (resource providers and user communities) with a trust agreement to guarantee security and confidentiality. A user must be authorized by a certification authority and must belong to a recognized virtual organization: a user community providing the rights to access to grid resources (computing, storage, data, software, services). The user can then execute simple tasks (jobs) or complex computation workflow operations by specifying only the characteristics of the computing resources needed and a logical name for data to be accessed via the grid storage.

Because of its architecture, a grid can efficiently tackle a large ensemble of computations running independently. A grid is also ideally suited for analyzing and producing large data sets and for sharing data within large teams. Several grid infrastructures have already been deployed around the world, for example, in North and South America, Asia, Australia, North Africa, and in 2008 in Senegal.

The largest grid deployment to date, Enabling Grids for E-Science (EGEE; http://www.eu-egee.org/), is designed for analyzing petabytes of data that will be produced by the European Organization for Nuclear Research's (CERN) Large Hadron Collider experiment in Geneva, Switzerland. Access to EGEE is not restricted to high-energy physics and is currently used by other scientific communities including bioinformatics, Earth sciences, and astronomy. As of March 2009,

EGEE is deployed at more than 300 sites. EGEE provides more than 80,000 CPUs and more than 20 petabytes of storage, and it is capable of running up to 100,000 jobs concurrently.

Grid computing has become a basic tool for particle physics and biotechnology researchers, but it still is not used routinely by Earth scientists.

Earth Science Community Needs

Since the International Geophysical Year in 1957–1958, the Earth science community has been deluged with data from the worldwide deployment of instruments. All observations need to be archived, but synoptic (time series) observations are particularly important because they cannot be repeated. For the past 4 decades, satellite observations have driven developments in computer sciences for handling, storing, and processing large volumes of data. The total data archive of the Earth observation program, managed by the European Space Agency, now accounts for some 5 petabytes, distributed geographically among several European data centers. Plans for the next 10–15 years foresee the accumulation of 10–15 times more data per year than at present. As models and simulations become increasingly sophisticated, they also generate large amounts of data. Model simulations place heavy demands on computing and storage resources, and often require supercomputers and/or distributed CPU capabilities that cannot be met at the institutional level. As a consequence, observational and simulation data are underutilized. This is particularly apparent in the understanding and prediction of climate change, where much unexplored data await exploitation.

The DEGREE Project

To identify the technical obstacles facing the adoption of grid methods in the Earth sciences (the complexity of setting up the hardware and installing and configuring the grid software, the use of grid middleware, and the implementation of Earth science tools), a consortium of Earth science academic and private institutions, space agencies, and computer science institutes launched the Dissemination and Exploitation of Grids in Earth Science European project (DEGREE; http://www.eu-degree.eu/) in 2006. Over a period of 2 years, 2006–2008, the consortium conducted a survey of the tools, software, and protocols commonly used by the Earth science community and of those provided by grid projects. DEGREE addressed three main topics: data management, job management, and portals to integrate Earth science and grid tools. Survey results, which are published on the DEGREE Web site, were discussed with grid project developers, and applications were proposed to test critical functionalities. An outcome is a grid road map for the Earth sciences that proposes a series of steps leading to the adoption of grid computing on a large scale (see Earth science white paper on grids, at http://www.eu-degree.eu/DEGREE/internal-section/wp6/DEGREE-D6.1.2_v2.8.pdf). The first step is to build an Earth science grid community by disseminating grid technologies and to support the deployment of new applications. The 5-year target is for a dedicated Earth science grid platform on the grid infrastructure to share computing and storage resources, data, knowledge, algorithms, and services over a continuum of time and over a variety of geographical scales.

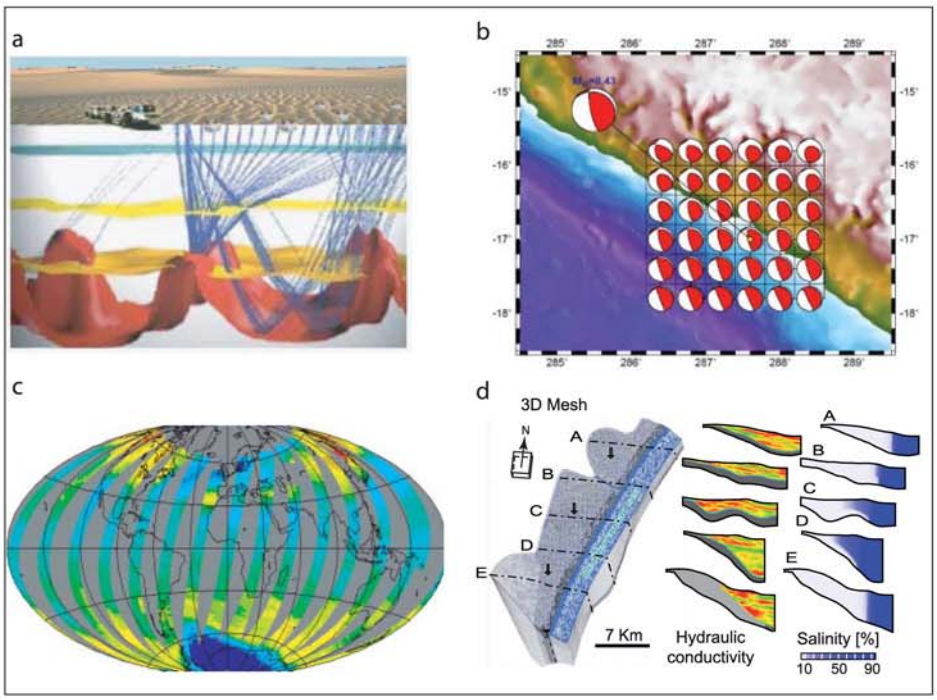

Fig. 1. Illustrations of the four Earth science grid applications described in the text. (a) Schematic view illustrating the three-dimensional computation of seismic wave propagation in a reservoir model for seismic inversion by GeoCluster. Image courtesy of CGGVeritas. (b) Interpretation of near–real time seismograms, with the "beach balls" being the common graphic representation of the source mechanism. Image courtesy of E. Clévédé, Institut de Physique du Globe de Paris, France. (c) Visualization of atmospheric ozone content. Image courtesy of the European Space Agency. (d) Forecasting seawater intrusion in a coastal aquifer. Image courtesy of J. Kerrou, University of Neuchâtel, Switzerland.

Grid Applications in Earth Science

The following four examples illustrate current applications of grid computing in the Earth sciences.

1. GeoCluster (http://www.cggveritas.com/default.aspx?cid=4-13-1925) is a software platform developed by CGGVeritas for seismic data processing, imaging, and underground reservoir characterization. It is operated in a grid environment (http://www.egeode.org; Figure 1a) using EGEE middleware and different infrastructures to accommodate academic and business needs. This first commercial grid application, which began operating in 2007, has demonstrated grid computing in a field where data security and confidentiality are extremely important. The application also has demonstrated the enormous benefit to users of ready access to computing resources and the updated software, and the ease of collaboration with distant colleagues.

2. The Institut de Physique du Globe de Paris has developed and deployed on EGEE an application to interpret, in near real time, earthquake data from the worldwide Geoscope seismometer network (http://geoscope.ipgp.jussieu.fr). Results are delivered as space-time location of a seismic rupture, the seismic energy released, the source mechanism, and earthquake duration (Figure 1b). Grid computing allows hundreds of simulations to be treated concurrently and then to be combined for final results, reducing delivery times from 1 week to less than 6 hours, suitable for seismic early warning systems.

3. In April 1995, the Global Ozone Monitoring Experiment was launched on board the European Remote Sensing (ERS) satellite. A neural network algorithm was used to retrieve atmospheric ozone profiles, which were then validated using ground-based light detection and ranging (lidar) observations (Figure 1c). Grid computing was used to produce and validate 7 years of ERS satellite data, allowing scientists from different institutions to easily manage about 70,000 files, handle the metadata for geospatial queries collocating the satellite's lidar observations, and share raw and analyzed data [*Iapaolo et al.*, 2007].

4. Grid computing can help emerging countries access high-performance computing and solve severe environmental problems such as seawater intrusions in coastal aquifers. The finite element code known as Coupled Variable Density and Saturation 3-D (CODESA-3D) has been

used to compute probabilistic maps of seawater intrusion in Tunisia's Korba aquifer (Figure 1d) by using a Monte Carlo method. The simulations include flow and density-dependent transport processes in a 3-D heterogeneous coastal aquifer [*Kerrou et al.*, 2007]. In this example, grid computing maintains two key advantages over classical distributed computing. First, a very large number of Monte Carlo simulations, numbering at least in the hundreds, can be run in parallel with substantial gains in time and accuracy. Second, the grid analysis can be controlled simply from a Web browser (e.g., http://www.eumedgrid.org) by collaborating scientists located in Europe and northern Africa.

A Vision for the Future

The above examples show that grid computing can fulfill most of the computing requirements of Earth scientists and offers new ways for efficient collaboration. Grids such as EGEE—consisting of clusters and farms of CPUs—cannot handle massive computations requiring parallel computing, shared memory, and intense communication between the processors. Other grids, such as TeraGrid (http://www.teragrid.org), can fulfill these needs.

Grid computing permits the sharing of resources between institutions and for scaling up the computing power and storage capacity in a way that is impossible for a single institution. Also, grid computing offers a transparent collaborative platform for users, allowing them to access more resources at a given time. This is especially important for the exploitation of large data sets scattered in several locations; for running large statistical jobs; and for sharing data and algorithms, without the need for conversion, among large numbers of partners.

Grid computing is currently available and can meet most of the technical requirements for the Earth sciences. Although some technical gaps still exist, grid developers are aware of them and are working to meet Earth science needs in the next generation of grid development. Significantly more effort will be required before transparent grid usage will be widespread in the Earth sciences. The vision is that the grid infrastructure must provide, within the next 5 years, a dedicated platform for sharing knowledge, algorithms, data, and services over a wide range of time and spatial scales. Such a platform will help provide efficient and timely answers to many fundamental challenges facing mankind.

Acknowledgments

The work presented in this article was funded by the European Community (contract DEGREE-IST-2005-034619). We gratefully acknowledge Charles Barton for polishing our text and all the partners of the project: G. Lecca (Center for Advanced Studies, Research and Development in Sardinia, Italy), G. Vetois (CGGVeritas, France), W. Som de Cerff (Royal Netherlands Meteorological Institute (KNMI)), L. Fusco and J. Linford (European Space Agency), L. Hluchy and V. Tran (Institute of Informatics, Slovak Academy of Sciences, Slovakia), C. Plevier (Dutchspace, Netherlands), H. Schwichtenberg (Fraunhofer Institute for Algorithms and Scientific Computing, Germany), and M. Zhizhin (Institution of the Russian Academy of Sciences Geophysical Center (RAS)), without whom this work could not have been done.

References

Foster, I., and C. Kesselman (Eds.) (1998), *The Grid: Blueprint for a New Computing Infrastructure*, 677 pp., Morgan Kaufmann, San Francisco, Calif.

Iapaolo, M., et al. (2007), GOME ozone profiles retrieved by neural network techniques: A global validation with lidar measurements, *J. Quant. Spectrosc. Radiat. Transfer*, *107*, 105–119.

Kerrou, J., G. Lecca, F. Murgia, and P. Renard (2007), Grid-enabled simulation of the impact of exploitation uncertainty on the seawater intrusion of the Korba aquifer (Tunisia), in *IST-Africa 2007 Conference Proceedings*, edited by P. Cunningham and M. Cunningham, Int. Inf. Manage. Corp., Dublin.

Author Information

Philippe Renard and Vincent Badoux, University of Neuchâtel, Neuchâtel, Switzerland; E-mail: philippe.renard@unine.ch; Monique Petitdidier, Laboratoire Atmosphères, Milieux, Observations Spatiales, Paris, France; and Roberto Cossu, European Space Agency, Frascati, Italy