J3.3 The Earth System Grid II: Turning Climate Datasets Into Community Resources

Ian Foster[1], Ethan Alpert[2], Ann Chervenak[4], Bob Drach[3], Carl Kesselman[4], Veronika Nefedova[1], Don Middleton[2], Arie Shoshani[5], Alex Sim[5], and Dean Williams[3]

[1]Argonne National Laboratory, Argonne, IL., [2]NCAR, Boulder, Co., [3]Lawrence Livermore National Laboratory, Livermore, CA., [4]USC/ISI, Marina del Ray, CA., [5]Lawrence Berkeley National Laboratory, Berkeley, CA.

1. Introduction

Global coupled Earth System models are vital tools for understanding potential future changes in our climate. As we move towards mid-decade, we will see new model realizations with higher grid resolution and the integration of many additional complex processes. The U.S. Department of Energy (DOE) is supporting an advanced climate simulation program that is aimed at accelerating the execution of climate models one hundred-fold by 2005 relative to the execution rate of today.

This program, and other similar modeling and observational programs, are producing terabytes of data today and will produce petabytes in the future. This tremendous volume of data has the potential to revolutionize our understanding of our global Earth System. In order for this potential to be realized, geographically distributed teams of researchers must be able to effectively and rapidly develop new knowledge from these massive, distributed data holdings and share the results with a broad community of other researchers, assessment groups, policy makers, and educators.

The Earth System Grid II (ESG-II) is aimed at addressing this important challenge. The broad goal is to define, develop, and deploy a next generation environment that harnesses the combined potential of massive distributed data resources, remote computation, and high-bandwidth wide-area networks as an integrated resource for the research scientist. We envision ESG-II as a foundation for next-generation analysis applications, web-based data portals, and collaborative problem-solving environments, and thus as important enabling infrastructure for sustaining and advancing climate and other environmental research.

Corresponding Author Address: Ian Foster, Math & Computer Science Div.; Argonne National Laboratory; Argonne, IL 60439; email: foster@mcs.anl.gov.

In an earlier phase of this work, ESG-I [2], we developed and demonstrated promising new core technology. Building upon these foundations, ESG-II will be focused upon integrating and extending a range of Grid and collaboratory technologies, researching promising intersections of the Grid and the DODS framework for remote access to scientific data, extending Globus Toolkit™ technologies for authentication, resource discovery, and resource access, and leveraging Data Grid technologies developed in other projects. We also intend to develop new technologies for creating and operating "filtering servers" capable of performing sophisticated analysis steps and delivering the results to users.

2. Grid and The Globus Project™

ESG-II builds upon a substantial foundation of work in distributed computing. The Globus Project™, a joint effort of Argonne National Laboratory, the Information Sciences Institute of the University of Southern California, and the University of Chicago, has been working for the past 5 years to solve precisely the problems faced in the ESG effort - facilitating scientific collaboration within flexible "virtual organizations" by connecting globally dispersed collaborators to complex and large-scale instrumentation, data, computing, and visualization resources [5]. The results of the Globus Project™ are being studied, developed, and enhanced at institutions worldwide to create new grids and services, and to conduct computing research. Globus Toolkit™ services were used, in particular, in ESG-I, as we discuss below.

Components of the Globus Toolkit™ provide the infrastructure needed to create "grids" of computing resources and users; track the capabilities of resources within a grid; specify the resource needs of user's computing tasks; mutually authenticate both users and resources;

and deliver data to and from remotely executed computing tasks. The so-called Data Grid community has started to address some of the requirements associated with distributed management and analysis of large-scale data. Over the past two years, project participants have established a broad national—and indeed international—consensus on the importance of Data Grid concepts [8] and on the specifics of a Data Grid architecture [4]. A tightly coordinated set of projects has been established that together are developing and applying Data Grid concepts to problems of tremendous scientific importance, in such areas as high energy physics and astronomy as well as (in the context of ESG-I) climate research. For example, the DOE-funded PPDG project (http://www.ppdg.net) is focused on the application of Data Grid concepts to the needs of a number of U.S.-based high energy and nuclear physics experiments, the NSF-funded GriPhyN (http://www.griphyn.org) project is focused on the automatic generation and management of derived data, and the EU-funded European Data Grid (EDG) project (http://www.eu-datagrid.org), is focused on the development of an operational Data Grid infrastructure. All three of these projects have adopted a common Globus Tookit™ based infrastructure.

3. ESG-I

In ESG-I, we took the first steps towards the realization of the Earth System Grid concept. Specifically, we developed techniques for the high-speed movement of data between centers and users, replica catalogs for keeping track of data location, request managers for coordinating multiple transfers, and a Grid-enabled version of the data analysis package produced by the Program for Climate Model Diagnosis and Intercomparison (PCMDI). We demonstrated our ability to manage the location and movement of large datasets from the user's desktop. We also learned a lot about user requirements: in particular, the emerging importance of thin clients and standard data access protocols as a means of delivering ESG capabilities to the largest possible audience, and the emerging importance of moving analysis processes to the data so that movement of terascale data holdings is minimized. New capabilities developed in ESG-I are now in production use in climate research.

4. ESG-II

Many components are necessary to create the long-term goal of the ESG-II distributed collaborative infrastructure, including:

- Teraflop computers to run highly sophisticated atmospheric, ocean, and coupled atmospheric ocean models;
- Large-scale data processing and analysis engines;
- Data storage facilities that can hold petabytes of raw and processed climate data with fast access times;
- Networking hardware that supports both the transmission of terascale data and collaborative audio and visual communication, white board interaction, visualization, and animation; and
- Powerful, cost-effective local and desktop facilities for visualization and data processing.

Tying these diverse components together in an easily interconnected and comprehensible way is the integrating middleware software: that is, the services such as security and resource discovery that sit between application and network, and that help researchers navigate and manage a secure system.

The ESG-II project will both leverage a range of existing technologies and develop new technologies in several key areas. Key building blocks are:

- High-speed data movement that uses data movers with extended FTP (GridFTP) [1] to make effective use of the grid-wide physical networking topology.
- High-level replica management that moves data objects between storages systems such as home repositories and network enabled disk caches located throughout the ESG at points where they can improve throughput and response time.
- A sophisticated grid-wide integrated security model that gives each user a secure, authentic identity on the ESG and permits management of grid resources according to various project-level goals and priorities.

- Remote data access and processing capabilities that reflect an integration of GridFTP and the DODS framework.
- Enhanced existing user tools for analysis of climate data such as PCMDI tools [9], NCAR's NCL, other DODS-enabled applications, and web-portal based data browsers.

As part of this project, we will also develop two important new technologies:

- Intelligent request management that uses sophisticated algorithms and heuristics to solve the complex problem of what data to move to what location at what time, based on knowledge of past, current, and future end-user activities and requests.
- Filtering servers that permit data to be processed and reduced closer to its point of residence, reducing the amount of data shipped over wide area networks.

5. Summary

The availability of a usable Earth System Grid has the potential to fundamentally change and enhance the way climate research scientists work together and address major challenges in climate simulation. Scientists will be able to request complex data products via convenient "Grid-enabled" desktop tools, including "thin client" Web interfaces. In its ultimate instantiation, the Grid will determine where these data products should be computed, stored, post-processed and visualized, using the most effective combinations of local and remote resources.

The ESG-II project will leverage, test, and showcase these new developments in the crucible of production operation of distributed global Earth System models, analysis of the resulting simulation data, and in impacts and assessment studies.

6. Website

http://www.earthsystemgrid.org

7. Acknowledgements

This work is supported by the U.S. Dept. of Energy Scientific Discovery Through Advanced Computing, contract #2001-184.

8. Literature Cited

1. Allcock, B., Bester, J., Bresnahan, J., Chervenak, A.L., Foster, I., Kesselman, C., Meder, S., Nefedova, V., Quesnel, D. and Tuecke, S., Secure, Efficient Data Transport and Replica Management for High-Performance Data-Intensive Computing. in *Mass Storage Conference*, (2001).
2. Allcock, B., Foster I., Nefedova, V., Chervenak, A., Deelman, E., Kesselman, C., Lee, J., Sim, A., Shoshani, A., Drach, B., Williams, D. High-Performance Remote Acces to Climate Simulation Data: A Challenge Problem for Data Grid Technologies, SC2001.
3. Catlett, C. and Smarr, L. Metacomputing. *Communications of the ACM*, *35* (6). 44--52.
4. Chervenak, A., Foster, I., Kesselman, C., Salisbury, C. and Tuecke, S. The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Data Sets. *J. Network and Computer Applications*.
5. Foster, I., Insley, J., Laszewski, G.v., Kesselman, C. and Thiebaux, M. Distance Visualization: Data Exploration on the Grid. *IEEE Computer*, *32* (12). 36-43.
6. Johnston, W. Realtime Widely Distributed Instrumentation Systems. in Foster, I. and Kesselman, C. eds. *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann, 1999, 75-103.
7. Messina, P. Distributed Supercomputing Applications. in Foster, I. and Kesselman, C. eds. *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann, 1999, 55-73.
8. Moore, R., Baru, C., Marciano, R., Rajasekar, A. and Wan, M. Data-Intensive Computing. in Foster, I. and Kesselman, C. eds. *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann, 1999, 105-129.
9. Williams, D.N. and Mobley, R.L. The PCMDI Visualization and Computation System (VCS): A Workbench for Climate Data Display and Analysis, Lawrence Livermore National Laboratory, 1994.