The Earth System Grid II: Turning Climate Model Datasets into Community Resources

David E. Bernholdt*, Kasidit Chanchio, Mei Li Chen, Line Pouchard, Oak Ridge National Laboratory; Ian T. Foster, Veronika Nefedova, Argonne National Laboratory; Alex Sim, Arie Shoshani, Lawrence Berkeley National Laboratory; Bob Drach, Dean N. Williams, Lawrence Livermore National Laboratory; David Brown, Luca Cinquni, Peter Fox, Jose Garcia, Don E. Middleton, Gary Strand, National Center for Atmospheric Research; Shishir S. Bharathi, Ann Chervenak, Carl Kesselman; University of Southern California

Summary

In pursuit of DOE Climate Change research goals, global climate simulations are being run on supercomputers across several DOE sites and at NCAR. The resulting data archive, distributed over several sites, currently contains upwards of one hundred terabytes of simulation data. Looking towards mid-decade and beyond, we must prepare for distributed climate data holdings of many petabytes. The Earth System Grid (ESG) is a collaborative interdisciplinary project aimed at addressing the challenge of enabling management, discovery, access, and analysis of these enormous and extremely important data assets. As part of the ESG team, ORNL researchers are focusing on issues of metadata to support the management of climate data from diverse sources, and working to deploy a robust, secure distributed environment.

The goal of the Earth System Grid (ESG, http://www.earthsystemgrid.org) project is to develop an environment that enables the management, discovery, distributed access, processing, and analysis of distributed terascale climate research data.

Global climate simulations, carried out at a small number of supercomputer sites around the country. Datasets from these simulations are of interest to numerous climate researchers for different purposes, and must also be archived. Tracking these datasets and moving them to where they are needed for archiving or analysis are, at present, tedious and time-consuming tasks. Much of the current modeling activity is focused on the upcoming Intergovernmental Panel on Climate Change (IPCC) assessment, and supporting the data management needs of these simulation efforts has become a major priority for the ESG team.

The Earth System Grid project brings together remote access protocols from the environmental science community, grid-based technologies for authentication, data discovery, and resource access, tools for storage management, and new tools, to create a distributed environment allowing researchers easy discovery and use of new and existing climate datasets. The result is a large, distributed software system that links most of the major US climate computational and archive sites with tools for high throughput data movement, metadata cataloging and searching.

Recent work at ORNL has focused on two areas: the importance of metadata, and security issues around deploying a complex

* (865) 574-3147, bernholdtde@ornl.gov

software system that is distributed across multiple sites.

Metadata

Descriptive information about the datasets registered with the ESG (their *metadata*) is of crucial importance in allowing researchers to locate datasets of interest to them. While data and metadata are frequently specific to a given field or even a particular project, many of the issues of representing, organizing, querying, and using metadata cross all boundaries.

ORNL researchers have been working actively with data management specialists across the country and around the world to help insure that ESG's strategies for dealing with metadata remain at the state of the art. This has included sharing of ESG schemas and prototype ontologies with other researchers, presentations and papers at national and international scientific meetings and workshops, and other activities. A highlight has been the publication of an invited paper "Data Grid Discovery and Semantic Web Technologies for the Earth Sciences," in the Journal of Digital Libraries, written by ORNL's Line Pouchard and David Bernholdt together with Andrew Woolf of the Rutherford Appleton Laboratory in the United Kingdom, describing work by both the ESG and the National Environment Research Council (NERC) data grid project in the UK.

Security for Distributed Systems

Wide-area distributed computing systems pose particular issues with regard to security because they link together resources at multiple sites with differing security policies and requirements. The situation is exacerbated when the distributed computing system is intended to provide access to unique and highly valued resources at the various sites. In the case of the ESG, the climate data sets of interest are stored in mass storage systems, along with other data from numerous other projects. The computer centers rightly view their mass storage systems and the data they contain as precious resources and are very careful about anything that might adversely affect them. ESG researchers at ORNL have been working together with ESG team members at other sites, the staff of the ORNL Center for Computational Sciences and the Computer (CCS) and Network Security team to integrate the Earth System Grid software environment into the CCS infrastructure in order to allow the ESG to serve data from the CCS High Performance Storage System (HPSS). This work will allow climate researchers to access many terabytes of climate simulation data already archives at ORNL as well as results from the on-going work towards the IPCC assessment.

For further information on this subject contact:
Mary Ann Scott, Program Manager
Mathematical, Information, and Computational Sciences Division
Office of Advanced Scientific Computing Research
Phone: 301-903-6368
scott@er.doe.gov