

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

PI's: Ian Foster (ANL), Don Middleton (NCAR),
& Dean Williams (LLNL/PCMDI)

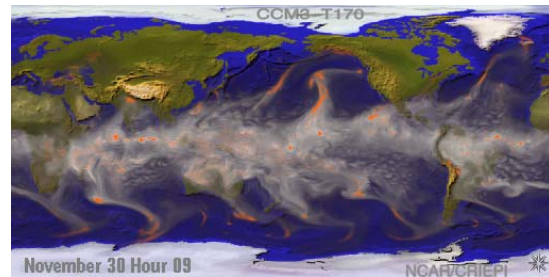
Senior Personnel: David Bernholdt (ORNL), David Brown (NCAR), Kasidit Chanchio (ORNL), Ann Chervenak (USC/ISI), Luca Cinquini (NCAR), Bob Drach (LLNL/PCMDI), Peter Fox (NCAR), Jose Garcia (NCAR), Carl Kesselman (USC/ISI), Veronika Nefedova (ANL), Line Pouchard (ORNL), Arie Shoshani (LBNL), Alex Sim (LBNL), and Gary Strand (NCAR) Additional Collaborators: William Allcock (ANL), Lawrence Buja (NCAR/CCSM), Joe Link (ANL), Laura Pearlman (USC/ISI), Von Welch (ANL), John Drake (ORNL)

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 volumes of data.

In 2003, DOE-sponsored climate change research has produced at least 72 terabytes of scientific data that is stored across several of the DOE sites and at NCAR. Related data (e.g. ocean simulation) exists as well. For the modeling teams, the daily management and tracking of their data is already proving to be a significant problem. The primary customers for the data are climate researchers who reside at various centers and universities across the U.S. but their ability to discover and use the data is not nearly what it should be. That's today, and the problem is rapidly escalating.

In the future, the computers on which we run the models will be much faster and the models themselves will become increasingly complex. Furthermore, the geographic resolution that is studied will be much finer. The image below depicts experimental work where the resolution of the simulation is several times that of present-day operational climate models and provides much greater regional detail.



All of this adds up to an enormous increase in the volume and complexity of data that will be produced. Moving these data will become increasingly costly and we will often need to leave the data where they are computed.

ESG's mission is to chart a viable course into the future that allows us to manage and use the climate simulation data. To this end, our goals include researching new strategies, developing new technologies and tools, and building an operational environment. Scientific needs define and drive the project while Grid technology serves as the underlying foundation for federating distributed computing systems. ESG works

in between these two areas, developing the information systems and applications for the end user.

One of our goals is to deliver a simple but powerful web portal for climate data. At the SC'02 supercomputing conference in Baltimore, ESG demonstrated a prototype that addressed security, management, discovery, access, and visualization of climate data that was distributed across multiple centers. The portal reflected a massive integration of new and emerging Grid technologies, and, invisibly to the user, harnessed computing, disk and archival storage resources at ORNL, LLNL, LBNL, USC/ISI, ANL, and NCAR. The portal incorporated strong security and the ability to flexibly allow access for various groups. These core requirements for building ESG's collaborative work environment were addressed in cooperation with the SciDAC *Security and Policy for Group Collaboration* project.

Our data management efforts center upon moving terascale data across the network and a close collaborative effort with the SciDAC *Scientific Data Management ISIC* is a cornerstone of the project. We have already demonstrated large-scale data transfers managed via our web portal and reached an important milestone: deep interoperability between DOE and NCAR archival storage systems.

ESG is developing new strategies and technologies for capturing metadata ("data about data") so that as data stream out of the supercomputing models, detailed metadata are captured in catalogs. These catalogs are crucial for managing the data and locating them later on.

One of ESG's strategies is to reduce dramatically the amount of data that must be

moved over the network, and we are engaged in groundbreaking work in developing generalized remote data access capabilities. This involves heavy collaboration with the SciDAC *High-Performance Datagrid Toolkit* project, as well as joint work with the community *OPeNDAP* project. We demonstrated partial functionality at SC'02 and in 2003 we'll show new analysis applications capable of operating on remote data over the network.

The ESG project makes heavy use of collaboration technology, especially the AccessGrid for project meetings and interactions with scientists and other collaborators. ESG will be an early adopter of new AccessGrid technology releases from the *Middleware to Support Group to Group Collaboration* SciDAC project.

ESG's collaborations and interactions are extensive, and span several SciDAC projects, the Globus™ Project, the U.K. e-Science program, the NSF National Science Digital Libraries program, multi-agency efforts to develop data portal technology, and NASA-sponsored projects in the area of distributed data access.

Our next step in 2003 will be a phased deployment, evaluation, and refinement process for new technology, tools, and environments. The first of these will be in the area of data management, followed by a public version of the ESG web portal for search, access, and analysis.

For further information on this subject:

<http://www.earthsystemgrid.org/>

Contact the ESG PI Team at:

esg-manage@earthsystemgrid.org

Or contact

Don Middleton

NCAR Scientific Computing Division

don@ucar.edu, 303-497-1250