DOE SciDAC's Earth System Grid Center for Enabling Technologies

Final Report for

University of Southern California Information Sciences Institute

October 1, 2006 through September 30, 2011

Principal Investigator:
Ann Chervenak

Table of Contents

1. Overview of Earth System Grid: Executive Summary

The mission of the Earth System Grid Federation (ESGF) is to provide the worldwide climate-research community with access to the data, information, model codes, analysis tools, and intercomparison capabilities required to make sense of enormous climate data sets. Its specific goals are to (1) provide an easy-to-use and secure web-based data access environment for data sets; (2) add value to individual data sets by presenting them in the context of other data sets and tools for comparative analysis; (3) address the specific requirements of participating organizations with respect to bandwidth, access restrictions, and replication; (4) ensure that the data are readily accessible through the analysis and visualization tools used by the climate research community; and (5) transfer infrastructure advances to other domain areas.

For the ESGF, the U.S. Department of Energy's (DOE's) Earth System Grid Center for Enabling Technologies (ESG-CET) team has led international development and delivered a production environment for managing and accessing ultra-scale climate data. This production environment includes multiple national and international climate projects (such as the Community Earth System Model and the Coupled Model Intercomparison Project), ocean model data (such as the Parallel Ocean Program), observation data (Atmospheric Radiation Measurement Best Estimate, Carbon Dioxide Information and Analysis Center, Atmospheric Infrared Sounder, etc.), and analysis and visualization tools, all serving a diverse user community. These data holdings and services are distributed across multiple ESG-CET sites (such as ANL, LANL, LBNL/NERSC, LLNL/PCMDI, NCAR, and ORNL) and at unfunded partner sites, such as the Australian National University National Computational Infrastructure, the British Atmospheric Data Centre, the National Oceanic and Atmospheric Administration Geophysical Fluid Dynamics Laboratory, the Max Planck Institute for Meteorology, the German Climate Computing Centre, the National Aeronautics and Space Administration Jet Propulsion Laboratory, and the National Oceanic and Atmospheric Administration.

The ESGF software is distinguished from other collaborative knowledge systems in the climate community by its widespread adoption, federation capabilities, and broad developer base. It is the leading source for present climate data holdings, including the most important and largest data sets in the global-climate community, and—assuming its development continues—we expect it to be the leading source for future climate data holdings as well.

Recently, ESG-CET extended its services beyond data-file access and delivery to include more detailed information products (scientific graphics, animations, etc.), secure binary data-access services (based upon the OPeNDAP Data Access Protocol), and server-side analysis. The latter capabilities allow users to request data subsets transformed through commonly used analysis and intercomparison procedures. As we transition from development activities to production and operations, the ESG-CET team is tasked with making data available to all users seeking to understand, process, extract value from, visualize, and/or communicate it to others—this is of course if funding continues at some level. This ongoing effort, though daunting in scope and complexity, would greatly magnify the value of numerical climate model outputs and climate observations for future national and international climate-assessment reports. The ESG-CET team also faces substantial technical challenges due to the rapidly increasing scale of climate simulation and observational data, which will grow, for example, from less than 50 terabytes for the last Intergovernmental Panel on Climate Change (IPCC) assessment to multiple Petabytes for the next IPCC assessment. In a world of exponential technological change and rapidly growing sophistication in climate data analysis, an infrastructure such as ESGF must constantly evolve if it is to remain relevant and useful.

Regretfully, we submit our final report at the end of project funding. To continue to serve the climate-science community, we are currently seeking and must identify additional funding. Such funding would allow us to maintain and enhance ESGF production and operation of this vital endeavor of cataloging, serving, and analyzing ultra-scale climate-science data.

At this time, the entire ESG-CET team would like to take this opportunity to sincerely thank our funding agencies in the DOE Scientific Discovery through Advanced Computing (SciDAC) program and the Office of Biological and Environmental Research (OBER)—as well as our national and international collaborators, stakeholders, and partners—for allowing us to work with you and serve the community these past several years.

2. ISI Accomplishments in the Earth System Grid

Next, we summarize the specific contributions of the University of Southern California Information Sciences Institute team in the Earth System Grid project.

2.1. Replication in Earth System Grid

ESG researchers in the U.S. and international climate researchers formed the Earth System Grid Federation (ESGF), which deploys ESG software at sites around the world to create an international federation of climate data nodes and gateways. Some member institutions of the ESGF act as *mirror sites* for climate data sets originally published by other institutions. A mirror site replicates existing data sets on its local data node and publishes metadata about the replicated data sets on its gateway so that those replicas may be discovered and accessed by researchers.

Widespread replication of climate data sets was long considered impractical because of their large size. However, with the growing importance of IPCC data sets, several sites around the world decided to host a replica or mirror of a key subset of climate data sets. The goals of these mirror sites are to provide reliable access to these data sets for local scientists; to reduce wide area data access latencies; and to improve fault tolerance of the distributed system by making data sets available at multiple sites. The first two ESGF mirror sites are the British Atmospheric Data Centre (BADC) and German Climate Computing Center (DKRZ). Additional mirror sites are being deployed at the Australian National University and the University of Tokyo.

The ISI team developed a replication client that enables ESG users to publish and replicate data sets within the ESG architecture. In this section, we describe the functionality of the replication client software.

2.1.1. The Replication Client Software

The replication client is implemented as a Python egg archive. It requires the existing ESG Publisher Python library developed at Lawrence Livermore National Laboratory. Installed using the *easy_install* utility, the replication client becomes available in the command shell of the data node.

Replication in ESG proceeds in three phases. Phase one queries metadata about a data set from an ESG gateway and prepares a data movement request as either a control file or a script for a data movement agent to process. The second phase is the actual data transfer by a data transfer agent. The final phase prepares the replica's metadata catalog and publishes its availability to the gateway for the mirroring organization.

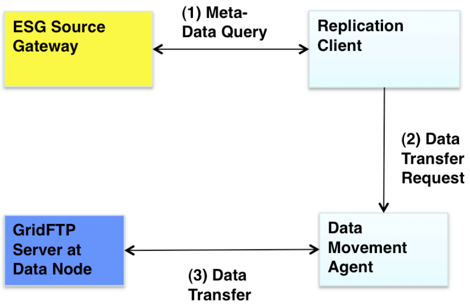

Figure 1: Metadata query and data transfer in ESG.

Metadata Query

The replication client does not directly transfer data. Instead, it queries a gateway metadata server and uses that metadata to prepare control files or scripts for use by one of the supported data movement agents. These steps are illustrated in Figure 1. To query the metadata, the replication client sends a data set identifier to the metadata server on the gateway. In response, the server returns an XML file describing the data set, including the names of the files, their sizes, what services are available for the file transfer, and the URL of the THREDDS metadata catalog on the data node hosting the data set. The replication client downloads this THREDDS metadata catalog from the data node and saves it for use in the third phase of replication. The client-side interaction with the gateway web service is managed using the Python implementation of the Hessian web services library found in the data node Python library.

Creating a Data Movement Request

ESG data nodes may allow data access using multiple transfer protocols. The replication client currently supports two data movement agents for performing data transfers. The client design is modular, and support for additional transfer agents can be added as necessary.

The default data movement agent is the Bulk Data Mover (BDM) developed at Lawrence Berkeley National Laboratory. BDM is a file transfer management system designed to maximize transfer throughput over networks of varying capacity. The replication client creates a transfer request for BDM by extracting relevant elements (file names, sizes, etc.) from the data set's metadata and writing appropriate commands to copy the files from their source location to the mirror site into an XML file that conforms to the BDM XML Schema. This file is used as the transfer control file for BDM.

An alternative data movement agent is the globus-url-copy utility from the Globus Toolkit. Globus-url-copy was developed at Argonne National Laboratory. The utility is designed for effective transfers of data from a GridFTP server or between GridFTP servers. When using globus-url-copy as its data movement agent, the replication client creates a plain text file of source/target pairs (one per line) that can be read by the command line globus-url-copy utility to perform the data transfers.

The replication client also supports the Globus Online (GO) service, a hosted data transfer service currently being developed at Argonne National Laboratory. The replication client can produce the GO command format, which is similar to that of globus-url-copy. The GO service is hosted in the Amazon Cloud. Requests submitted to GO are managed reliably by the GO service, which is responsible for coordinating transfers between GridFTP servers at ESG data nodes.

Once the command file for the data movement agent is constructed, an administrator at the mirror site initiates the replication of the data set by submitting the command file to that agent. The replication client does not automatically initiate the data transfers itself because mirror sites in the ESGF require the ability to control the timing of data- and network-intensive replication operations for large climate data sets.

Replica Publication

After the data movement agent transfers the files of the data set to the mirror site's data node, the replication client is used to prepare, check, and publish the data set as a replica on the mirror site's gateway node. These steps are illustrated in Figure 2.

To prepare the data set for publication as a replica on the mirror site, the replication client calls on the ESG publication library to first scan all the files in the replicated data set and then use the results of the scan to create a new THREDDS metadata catalog, which is stored in an XML file. This catalog file will be used by the mirror site's data node to publish the data set to the scientific community.

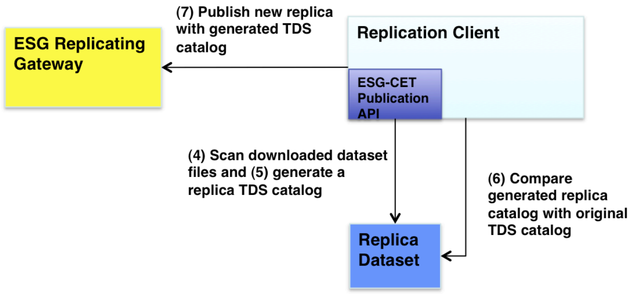
ESG Replicating Gateway
(7) Publish new replica with generated TDS catalog
Replication Client
ESG-CET Publication API
(4) Scan downloaded dataset files and (5) generate a replica TDS catalog
(6) Compare generated replica catalog with original TDS catalog
Replica Dataset

Figure 2: The preparation, checking, and replica publication process in ESG.

The replication client compares the newly created THREDDS metadata catalog for the replicated data set to the THREDDS metadata catalog downloaded from the publication data node after the initial metadata query. ESG data sets may have large numbers of files, which can create very large metadata catalogs. To limit the memory footprint required to compare two large XML files, the replication client uses a white list of properties to use for the comparison. The replication client parses each XML catalog file separately to create smaller Python objects. The class overrides the equals operator to provide flexibility and ease-of-use.

Properties of individual data set files are used for the comparison. The user is able to select which properties, from a predefined list, are to be checked and, of those properties, which ones are required to match and which ones are optional. "Optional" in terms of the replication client means that if one of the catalogs contains the property and the other does not, the comparison does not fail. If both catalogs have an optional property, those properties must match. If a property is mandatory, the comparison will fail if one catalog does not have that property.

The comparison of the THREDDS metadata catalog files is only intended as a check on the scan and replica catalog generation and is not a quality control check on the data files themselves. Users rely on the functionality of the transfer agents, such as the checksum matching in the Bulk Data Mover, to verify the correctness of data transfers. The original data publishers perform quality control for scientific purposes.

Once the replication client successfully compares the new THREDDS metadata catalog to the catalog from the site where the data set was originally published, the user may publish the replicated data set at the mirror site. Publishing makes the replicated data set's metadata visible on the mirror site's gateway node. After publication, this metadata will be shared with other gateway nodes in the distributed system. Climate scientists and other users around the world are then able to use any gateway to discover and access the newly replicated data set. The replication client uses the ESG data publication library to perform the publication operation in similar fashion to the existing ESG publication command line tools.

2.1.2. Performance of Replication Client

We measured the performance of the replication client on two data sets of size 5 GBytes and 50 Gbytes. We downloaded these test data sets from the PCMDI ESG gateway and data node at Lawrence Livermore National Laboratory. The performance of several sub-components of the replication client is shown in Table 1. The means are calculated for 5 or more runs of the replication client. These measurements include the metadata query done in phase 1 of the replication client's operation and the file scan, THREDDS metadata catalog generation, and THREDDS catalog comparison done in phase. The standard deviation of some of these measurements is high, resulting in, for example, the scan taking longer on average for the 5 GByte data set than for the 50 GByte data set.

The data transfer times are not included in the table. We downloaded each data set once using the globus-url-copy option for the replication client. The transfers took 827 seconds and 5854 seconds for the 5 GByte and 50GByte data sets, respectively.

Table 1: Performance of replication client components. All times shown are in seconds.

	Metadata 5GB	Metadata 50GB	Scan 5GB	Scan 50GB	Generate Catalog 5G	Generate Catalog 50G	Compare Catalogs 5GB	Compare Catalogs 50GB
Mean time (seconds)	0.98	1.95	1.01	0.58	2.08	3.96	0.52	0.55
Standard Deviation	0.08	1.32	1.06	0.01	2.42	6.89	0.00	0.01

Replication Client Testing

The ESG team supported the expanded use and testing of the ESG replication client by three sites: PCMDI at Lawrence Livermore National Laboratory, the British Atmospheric Data Centre (BADC) and the German Climate Computing Center (DKRZ). Personnel at these sites exercised and tested the features of the replication client and identified bugs and issues regarding the use of the client and how it functioned. The replication client was improved based on these bug reports and feature requests. Additional work includes packaging the software to make it easily deployable by participating sites and documenting the software to improve its ease of use by climate scientists.

2.2. Monitoring the Earth System Grid

As Grids for scientific applications become larger and more complex, the management of these environments becomes increasingly difficult. Commonly, these scientific Grids consist of a large number of heterogeneous components deployed across multiple administrative domains, including storage systems, compute clusters, Web portals, and services for data transfer, metadata management, and replica management. Monitoring these components to determine their current state and detect failures is essential to the smooth operation of Grid environments and to user satisfaction.

Monitoring systems collect, aggregate, and sometimes act upon data describing system state. This information can help users make resource selection decisions and help administrators detect problems. Monitoring systems can typically be queried and, in many cases, can take actions based on events. Grids present additional challenges for monitoring systems because of the frequency with which resources are added and removed and because of the distributed nature of the responsibility for administering resources in a Grid.

In the first several years of the Earth System Grid project, we used the Globus Toolkit Version 4 (GT4) Monitoring and Discovery System (MDS4) to monitor the ESG infrastructure. The Globus Toolkit provides middleware to support secure resource sharing among participants in a Grid. MDS4 defines and implements mechanisms for service and resource discovery and monitoring in distributed environments. MDS4 currently includes two higher-level services: the Index service, which collects and publishes aggregated information about Grid resources, and the Trigger service, which collects resource information from the Index Service and performs actions when certain conditions are met.

The services that were monitored in the ESG infrastructure include GridFTP data transfer services, the OPeNDAP service that filters requested information to reduce the amount of data transferred, the ESG Web portal, HTTP servers for data access, Replica Location Service catalogs at several sites, Storage Resource Managers at several sites, and three hierarchical mass storage systems. The Index and Trigger services for ESG run on the machine at NCAR but are maintained and updated by ISI staff.

Features of the monitoring infrastructure for ESG included the following. The ESG portal displays an overall picture of the current status of the ESG infrastructure based on our monitoring information, giving users and administrators an understanding at a glance of which resources and services are currently available. In addition, failure messages provided by the Trigger service help system administrators to identify and quickly address failed components and services. The monitoring system helps avoid system downtime by warning of the imminent expiration date of host certificates on services, so that they can be renewed without service interruptions. Finally, when a particular site has scheduled downtime for site maintenance, it is not necessary to send failure messages to system administrators regarding components and services at that site. We have developed a simple mechanism that disables particular triggers for the specified downtime period.

ISI Accomplishments in Monitoring

- The ISI and ANL teams developed the monitoring services infrastructure (including the MDS4 Index Service and Trigger Service) that is used by ESG to detect component failures. This development is ongoing and includes feature improvements and bug fixes. As new features are added to these services, they are incorporated into the ESG monitoring infrastructure.

- The ISI team deployed the ESG monitoring infrastructure in the distributed ESG environment and supports this deployment. As the ESG architecture evolves (for example, as the new gateway architecture is being deployed), the ISI team has modified the monitoring service deployment to follow this evolution.

- The ISI team has added new triggers as needed by ESG to detect additional failure conditions.

- When failure messages occur, the ISI team helps to identify the cause of failures in the ESG infrastructure. In one recent example, we assisted the ORNL and LBNL teams in tracking down problems with SRM servers that were failing on several sites: mainly at ORNL, but also at NCAR and LBNL.

- The ISI team helped the ORNL team set up the monitoring probes for the SRM, TRM, and RLS services running on the new ESGtg machine.

- The ISI monitoring team is also doing ongoing work to integrate the portal’s monitoring framework with the latest version of the Globus Toolkit Version code.

3. Publications

- SciDAC ’11 Conference Proceedings

D. N. Williams, J. Ahrens, R. Ananthakrishnan, M. Balman, G. Bell, S. Bharathi, D. Brown, M. Chen, A. L. Chervenak, L. Cinquini, R. Drach, I. T. Foster, P. Fox, S. Hankin, D. Harper, N. Hook, P. Jones, D. E. Middleton, N. Miller, E. Nienhouse, R. Schweitzer, G. Shipman, A. Shoshani, F. Siebenlist, A. Sim, W.

G. Strand, F. Wang, C. Ward, P. West, H. Wilcox, N. Wilhelmi, and S. Zednik, "Earth System Grid Center for Enabling Technologies: A Data Infrastructure for Data-Intensive Climate Research," Journal of Physics: Conference Series, SciDAC '11 Conference Proceedings.

- SciDAC '10 Conference Proceedings

D. N. Williams, J. Ahrens, R. Ananthakrishnan, M. Balman, G. Bell, S. Bharathi, D. Brown, M. Chen, A. L. Chervenak, L. Cinquini, R. Drach, I. T. Foster, P. Fox, S. Hankin, D. Harper, N. Hook, P. Jones, D. E. Middleton, N. Miller, E. Nienhouse, R. Schweitzer, G. Shipman, A. Shoshani, F. Siebenlist, A. Sim, W. G. Strand, F. Wang, C. Ward, P. West, H. Wilcox, N. Wilhelmi, and S. Zednik, "Earth System Grid Center for Enabling Technologies: Building a Global Infrastructure for Climate Change Research," Journal of Physics: Conference Series, SciDAC '10 Conference Proceedings.

- SciDAC Review Article

D. N. Williams, R. Ananthakrishnan, D. E. Bernholdt, S. Bharathi, D. Brown, M. Chen, A. L. Chervenak, L. Cinquini, R. Drach, I. T. Foster, P. Fox, S. Hankin, V. E. Henson, P. Jones, D. E. Middleton, J. Schwidder, R. Schweitzer, R. Schuler, A Shoshani, F. Siebenlist, A. Sim, W. G. Strand, N. Wilhelmi, M. Su. "The Planet at Their Fingertips: Climate Modeling Data Heats Up", Spring 2009.

The ESG-CET team completed the SciDAC Review Article entitled, "The Planet at Their Fingertips: Climate Modeling Data Heats Up". The article talks about the increasing importance of climate modeling and the tremendous need for the Earth System Grid to allow fast and accurate access to hundreds of petabytes.

(URL: http://www.scidacreview.org/0902/html/esg.html)

- Paper in the Bulletin of the American Meteorological Society (BAMS)

D N Williams, R Ananthakrishnan, D E Bernholdt, S Bharathi, D Brown, M Chen, A L Chervenak, L Cinquini, R Drach, I T Foster, P Fox, D Fraser, J Garcia, S Hankin, P Jones, D E Middleton, J Schwidder, R Schweitzer, R Schuler, A Shoshani, F Siebenlist, A Sim, W G Strand, M Su, N. Wilhelmi, "The Earth System Grid: Enabling Access to Multi-Model Climate Simulation Data", in the Bulletin of the American Meteorological Society, February 2009.

(URL: http://ams.allenpress.com/perlserv/?request=get-abstract&doi=10.1175/2008BAMS2459.1)

- Poster and Paper: SciDAC '07 Conference

The ESG team presented a peer-reviewed paper in the SciDAC 2007conference proceedings. The complete citation is: R Ananthakrishnan, D E Bernholdt, S Bharathi, D Brown, M Chen, A L Chervenak, L Cinquini, R Drach, I T Foster, P Fox, D Fraser, K Halliday, S Hankin, P Jones, C Kesselman, D E Middleton, J Schwidder, R Schweitzer, R Schuler, A Shoshani, F Siebenlist, A Sim, W G Strand, N. Wilhelmi, M Su, and D N Williams, "Building a Global Federation System for Climate Change Research: The Earth System Grid Center for Enabling Technologies (ESG-CET)", in the Journal of Physics: Conference Series, SciDAC '07 conference proceedings.

- Paper and Presentation: eScience'06

Ann Chervenak (ISI/USC) presented a peer-reviewed paper at eScience 2006 in Amsterdam, Netherlands. The complete citation is: Ann Chervenak, Jennifer M. Schopf, Laura Pearlman, Mei-Hui Su, Shishir Bharathi, Luca Cinquini, Mike D'Arcy, Neill Miller, and David Bernholdt, Monitoring the Earth System Grid with MDS4, in Second IEEE International Conference on e-Science and Grid Computing (e-Science'06), page 69, Los Alamitos, CA, USA, 2006, IEEE Computer Society.

4. Progress Reports

In this section, we include progress reports that show the contributions of the USC/ISI team throughout the course of the Earth System Grid project.

4.1. Semi-Annual Progress Report: October 1, 2010 through March 31, 2011

The data replication team at ISI deployed the ESG replication client to the British Atmospheric Data Centre (BADC) and the German Climate Computing Center (DKRZ) for testing. The issues identified by the expansion of testing were analyzed and solutions were added to the software.

The Globus Online service was added as an option for data transfers.

The replication client integrates functionality from several ESG providers. It queries the metadata catalog at the gateway; it pulls the THREDDS catalog from the original publication site's data node; it creates control files for either the Bulk Data Movement client or the GridFTP globus-url-copy client to transfer data; and it invokes the publication API from LLNL to scan the replicated files, generate a THREDDS catalog for those files, and publish the mirrored data set to the gateway. The replication client compares the THREDDS catalog generated for the newly copied files with that created at the original publication site; the data set is only published if this comparison verifies the correctness of the THREDDS catalog.

4.1.1. Data Replication

The data replication team at ISI supported the expanded use and testing of the ESG replication client by deploying the software to sites for two European ESG Federation partners: the British Atmospheric Data Centre (BADC) and the German Climate Computing Center (DKRZ). Personnel at these sites exercised and tested the features of the replication client and identified some issues regarding the use of the client and how it functioned. These issues were recorded and solutions tracked in the ESG Federation bug tracker. Testing continued at the US sites PCMDI, USC/ISI and NCAR.

DKRZ successfully used the ESG replication client to mirror datasets from BADC.

The replication client typically runs on a data node that will store a replica of an existing data set (the mirror data node). The replication client is given the name of a data set to be replicated and begins by querying the metadata catalog on a gateway to determine the list of files in that data set. The replication client also queries the gateway for a pointer to the THREDDS catalog on the data node where the data was originally published (the source data node). Next, the replication client prepares data transfer requests to copy the data from the source data node to the mirror data node. These transfer requests may be prepared for use by the Bulk Data Movement client, the GridFTP globus-url-copy client, or the newly added Globus Online service. The replication client then copies the THREDDS catalog from the source data node to the mirror data node. After the data set has been successfully copied, the replication client invokes the publication API developed at LLNL to scan the copied files and create a THREDDS catalog for the replicated data set. Next, the replication client software compares the newly generated THREDDS catalog for the replicated data set to the THREDDS catalog from the source data node; the client verifies that essential elements of the THREDDS catalog are consistent in both catalogs. If this check succeeds, then the replication client invokes the publication API to publish the replicated data set to a gateway. The replicated data set can then be discovered and accessed by ESG users.

4.1.2. Specific accomplishments of this period include the following:

1. Packaged the software to make it easily deployable by participating sites.
2. Enhanced documentation to clarify the correct use of the software.
3. Added enhancements to the software in response to end-user needs.

a. New command line switch to better integrate with the ESG Publication API

b. Added Globus Online as a data movement agent option.

4. Implemented bug fixes as needed.

5. Expanded the replication test plan as documented in the ESG Federation Wiki.

Plans for the coming 6-month period are to continue to provide support to our European partners to further deploy and test the replication client functionality; integrate the replication client source into the ESG-CET software package; hold follow-on discussions to determine additional requirements for replication; and to implement these as appropriate for the ESG and Go-ESSP collaborations.

4.2. Semi-Annual Progress Report: April 1, 2010 through September 30, 2010

The data replication team at ISI successfully completed the implementation of the ESG replication client during this period and performed end-to-end testing of this client to mirror data sets from PCMDI to two sites: USC/ISI and NCAR. The replication client integrates functionality from several ESG providers. It queries the metadata catalog at the gateway; it pulls the THREDDS catalog from the original publication site's data node; it creates control files for either the Bulk Data Movement client or the GridFTP globus-url-copy client to transfer data; and it invokes the publication API from LLNL to scan the replicated files, generate a THREDDS catalog for those files, and publish the mirrored data set to the gateway. The replication client compares the THREDDS catalog generated for the newly copied files with that created at the original publication site; the data set is only published if this comparison verifies the correctness of the THREDDS catalog. The completed replication client has been tested successfully to provide end-to-den data replication of a 5 GByte and 50 GByte data sets from PCMDI to data nodes at USC/ISI and NCAR.

4.2.1. Data Replication

The data replication team at ISI successfully completed the implementation of the ESG replication client during this period and performed end-to-end testing of this client to mirror data sets from PCMDI to two sites: USC/ISI and NCAR.

The replication client typically runs on a data node that will store a replica of an existing data set (the mirror data node). The replication client is given the name of a data set to be replicated and begins by querying the metadata catalog on a gateway to determine the list of files in that data set. The replication client also queries the gateway for a pointer to the THREDDS catalog on the data node where the data was originally published (the source data node). Next, the replication client prepares data transfer requests to copy the data from the source data node to the mirror data node. This data transfer can be performed using either the Bulk Data Movement client or the GridFTP globus-url-copy client. The replication client then copies the THREDDS catalog from the source data node to the mirror data node. After the data set has been successfully copied, the replication client invokes the publication API developed at LLNL to scan the copied files and create a THREDDS catalog for the replicated data set. Next, the replication client software compares the newly generated THREDDS catalog for the replicated data set to the THREDDS catalog from the source data node; the client verifies that essential elements of the THREDDS catalog are consistent in both catalogs. If this check succeeds, then the replication client invokes the publication API to publish the replicated data set to a gateway. The replicated data set can then be discovered and accessed by ESG users.

4.2.2. Specific accomplishments of this period include the following:

(1) Implemented, tested, and documented the replication client functionality that queries the gateway metadata server and prepares a control file (XML for BDM) or an executable shell script (for globus-url-copy) to download the data files for a requested data set. Created a simple database to manage the originating data node THREDDS catalog so that it could be used later for comparison to the newly generated THREDDS catalog for the replicated data set.

(2) Implemented, tested, and documented functionality to invoke the ESG-CET publication API developed at LLNL to scan the downloaded files, create a THREDDS catalog, and publish the replicated data set if the catalog comparison is successful.

(3) Designed, implemented, tested, and documented code to compare two THREDDS catalogs: the catalog from the source data node and the newly generated catalog from the mirror data node (see (2)). A candidate replica data set will not be published to the gateway unless the comparison succeeds, indicating that the required information is present in the THREDDS catalog for the replicated data set.

(4) Packaged the replication client and documented its dependencies for easy installation at NCAR.

(5) Ran end-to-end replication tests of data sets less than a gigabyte, 1 GByte, 5 GByte, and 50 GByte in size. These tests mirrored data sets from PCMDI to a mirror data node at USC/ISI. ISI personnel also worked closely with NCAR personnel, providing support for them to install the replication client at their site; test different portions of the replication logic; and finally perform end-to-end replication of the test data sets using the globus-url-copy data transfer client.

Plans for the coming 6-month period are to complete testing at NCAR using BDM for data transfer; provide support to our European partners to deploy and test the replication client functionality; hold follow-on discussions to determine additional requirements for replication; and to implement these as appropriate for the ESG and Go-ESSP collaborations.

4.3. Annual Report June 2010

During the last year, the main accomplishment of the ISI team under the ESG-CET project was the design and development of a data Replication client for ESG. The goal of this client is to support mirroring of large climate modeling data sets among institutions within ESG as well as at key international climate data centers, including the British Atmospheric Data Center and the Max Plank Institute. This replication client is described in detail below.

Our first tasks were to identify the requirements for the Replication client, to understand how existing ESG components could be integrated to provide this replication capability, and to identify missing pieces of functionality that needed to be developed. To identify replication requirements, we held a series of teleconferences with the ESG team and with our international collaborators. Dr. Chervenak attended the replication meeting held in Hamburg, Germany, in October 2009 in connection with the Go-ESSP meeting. This replication design meeting identified data mirroring and replication requirements for our European, Australian and Japanese collaborators. In particular, researchers at the British Atmospheric Data Center and the Max Plank Institute had strong opinions about the functionality needed for successfully mirroring key ESG data sets at their institutions, and this input was incorporated into our design. In the past year, we have attended weekly phone calls with these international collaborators, and we have given frequent updates on the status of data replication functionality.

The design of the replication client began with the ISI team proposing a draft API specification for the replication client, which was then refined and modified during extensive discussions with both the ESG team and our collaborators in the Go-ESSP team.

We held a series of design meetings to understand the impact of replication and mirroring of ESG datasets on the design of gateways and data nodes. Based on these discussions, the NCAR team developed a scheme for sharing metadata about replicated data sets among the gateways so that replicas could be discovered and accessed. They later modified the gateway to include additional metadata that pointed to the THREDDS metadata catalog on the data node where the data set was originally published; this source THREDDS catalog is fetched by our replication client and compared to the THREDDS catalog generated for the newly replicated data set.

We worked with the LLNL team on data publication issues for replicated data sets, including identifying necessary functionality that needed to be added to the publication client. The LLNL team responded by modifying the publication client to add metadata to identify published data sets as replicas. They also developed an API for our use that includes the ability to separately scan a mirrored directory, create a THREDDS catalog for it, and to publish the mirrored data set to a designated gateway node.

We also worked with the LLNL team to discuss operational issues related to data replication. These include the requirement to run a BDM-enabled GridFTP server at LLNL for us to download the data in our testing. We identified issues related to this GridFTP server, which needs to run on a dedicated machine because of the resources it consumes. We also agreed on a security model for this GridFTP server, since it will use the role-based authorization being developed by the security team rather than the older token-based authorization scheme. This BDM-enabled GridFTP server was deployed at LLNL in May 2010 and is currently being tested by the ISI team.

The ISI team has also done extensive testing of our replication client and the publication client developed by LLNL. We have tested the documentation and procedures for installation of a data node, both manually using the instructions provided by the LLNL team and using the recently developed installation scripts. The ISI team has worked to identify problems and bugs in these instructions and scripts and relayed this information to the LLNL team so that they can improve the documentation and scripts.

Finally, the ISI team succeeded in downloading data from the ESG web portal to the local data node, publishing that data to the PCMDI gateway using the ESG publication client, and then discovering and downloading that newly published data.

4.3.1. Replication Client

The development of the Replication client requires the integration of many components, some developed at ISI and some at LLNL, LBNL and NCAR. The steps required to perform data replication and the current status of each of these steps in the development of the Replication client include the following:

(1) The Replication client requests the list of files in the data set that will be mirrored from an ESG gateway.

Status: The replication client running at ISI data node successfully queries the LLNL gateway to retrieve the list of files in a data set.

(2) The Replication client creates a transfer request to move the files from the source data node to the mirror site. This request is a command file for the Bulk Data Movement client (developed at LBNL) that includes the files names retrieved in the previous step for BDM. This request is formatted in XML according to the schema provided by LBNL.

Status: Successfully creates BDM client commands consistent with the required schema.

(3) The Replication client initiates the transfer of the files to the mirror site by issuing BDM commands via the BDM client running at the ISI data node to the BDM-enabled GridFTP server at LLNL

Status: Currently testing BDM transfers initiated from ISI on the BDM-enabled GridFTP server deployed at LLNL.

(4) Once the BDM data transfers are complete, the Replication client generates a local THREDDS metadata catalog for the newly replicated files. This is generated by calling two commands from the replication API developed by LLNL:

- scanDirectory: Scan one or more directories and generate a mapfile
- generateReplicaThreddsCatalog: Scan a collection of files as defined in a mapfile, and generate THREDDS catalogs of the resulting datasets

Status: Successfully called scanDirectory; exception when calling generateThreddsCatalog. Currently investigating the source of the problem.

(5) The Replication client next retrieves the original THREDDS metadata catalog at the source data node where the data set was originally published. This requires a change in the metadata stored at the gateway to include a pointer to the source THREDDS catalog.

Status: The NCAR gateway team modified the gateway metadata to include a placeholder for this pointer to the source THREDDS catalog. We have successfully retrieved the source THREDDS catalog using pointers retrieved from the gateway metadata catalog.

(6) The Replication client compares the THREDDS catalog generated for the mirrored data set with the original THREDDS catalog retrieved from the source data node where the data set was published. The replication client must verify that all required elements of the two THREDDS catalogs match.

Status: Have designed a scheme for comparing these threads catalogs. Have begun to identify which parameters are required in both catalogs. Still to be done: implementation of the comparison.

(7) Replication client uses the replication API developed at LLNL to publish the replicated data set to the mirror gateway node

Status: Will test the publication function of the replication API after the catalog comparison is complete. We have successfully published data sets from the ISI data node to the LLNL gateway using the old publisher client but not with the new replication API.

In summary, most of the components of the Replication client have been developed. We are working with other members of the ESG-CET collaboration to resolve the remaining issues, and we expect to complete the implementation of the Replication client in Summer 2010.

4.3.2. Other Accomplishments:

The ISI team has used the ESG project to assist in the training of graduate and undergraduate students through research projects and summer internships. In June 2009, Shishir Bharathi received his PhD from the University of Southern California. Shishir worked on several aspects of ESG during the years of his MS and PhD study, including replica management and monitoring.

In summer 2009, Erin Brady did a summer internship on the ESG project. Erin was participating in the Computing Research Association's Distributed Research Experience for Undergraduates (DREU) program, which provides summer research opportunities for undergraduate women and underrepresented minorities in Computer Science. Erin worked on an early version of the data replication client. Her work was presented in a poster at the Supercomputing Conference in November 2009. (Brady, E., Chervenak, A. L., "Mirroring Earth System Grid Data Sets", Supercomputing (SC09) Conference, ACM Student Poster Session, Portland, Oregon, November 2009.)

4.3.3. Monitoring and Replica Catalogs

Finally, the ISI team continued this year to support monitoring infrastructure and replica catalogs for the Earth System Grid project. This monitoring software identifies when services fail or when certificates are going to expire, thus reducing the downtime of the ESG infrastructure. The replica catalogs are used to discover the locations of ESG data files so that those files may be downloaded by ESG users. The monitoring and replica catalog services provided by ISI have been used for several years by the ESG team. However, the use of these services is being phased out in the new, federated version of the Earth System Grid.

4.3.4. Plans for the Next Funding Period

The Replication client will be a key component of the new federated ESG infrastructure. In the remainder of the ESG-CET project, our team will continue development, testing, documentation and deployment of the Replication client.

We plan to complete the initial development of the client in the summer of 2010. Next, we will perform end-to-end testing of data mirroring from LLNL/PCMDI to ISI. Once this works reliably, we will provide the Replication client to our European partners for testing at their sites in Summer and Fall 2010. We will provide support to them in the installation and testing of this functionality in their environments. We will use their feedback to drive improvements to the software, providing periodic releases and extensive documentation of our tools.

We will continue to add functionality to enrich the ESG replication capability. In particular, we plan to integrate the Replication client with the versioning system and the notification system being developed for ESG. It is essential that users performing data mirroring be able to determine whether they have the latest version of a data set, and it is desirable to provide access to earlier versions of data sets at mirror sites. It is also important that mirror sites be notified when data sets change or when new data for an existing data set becomes available. We plan to add these capabilities in the remainder of the ESG-CET project.

Finally, the UK team has recently expressed a strong desire to for the capability to revoke replicated data sets after they are published. We will work to develop a revocation capability.

Our goal in the remainder of the ESG-CET project is to be agile and responsive to the needs of the climate community, tailoring the functionality of the Replication client according to meet new requirements that arise during large-scale, production operation of the ESG infrastructure. By the end of the ESG-CET project, we will have implemented a stable, robust, scalable, high performance, well-documented and thoroughly tested data replication client.

4.4. Renewal Report, June 2009

This report summarizes ESG-CET's progress and accomplishments since its inception under SciDAC-2 support. In this section, we summarize the specific contributions of the team at USC Information Sciences Institute. The ISI team contributes to ESG-CET in two main areas: data replication and system monitoring. Ann Chervenak is the lead investigator at ISI and manages these efforts on the ISI team. Robert Schuler is the lead architect and developer of software related to data replication for ESG. Mei-Hui Su has led the efforts on monitoring the ESG infrastructure. Shishir Bharathi, a PhD student, has contributed to ESG in the areas of data replication services and monitoring infrastructure. The ISI team participates regularly in ESG teleconferences and face-to-face meetings and works to integrate our software with the larger ESG architecture. The overall nature of ESG-CET is highly collaborative, and most activities are accomplished as a team. Here, we highlight the areas that are led by ISI personnel.

In the area of data replication management, ISI has led two efforts over the funded period of ESC-CET. First, ISI has provided a highly scalable and reliable tool called the Replica Location Service (RLS) that is used in the current ESG deployment to register the locations of all files stored at the distributed ESG sites. Robert Schuler is the lead architect and developer of the RLS and has been responsible for development of new features, documentation, maintenance, bug fixes, and technical support of ESG users.

For the last year, the ISI team has led a second effort at defining and implementing data mirroring functionality for the Earth System Grid. This work is described in detail in Section 5.10. Replication of climate data sets was not initially a goal of the Earth System Grid project. Such replication was thought to be impractical because of the large size of these data sets. However, in the last year, there has been increasing interest in mirroring key subsets of ESG data sets at sites around the world to provide easier, faster and more reliable access for scientists. In response to this new requirement, Dr. Chervenak has led a series of design discussions to define the data mirroring/replication use case and to define the functional requirements for a data mirroring tool that integrates with the rest of the ESG infrastructure. The implementation of this design is currently underway.

Another major contribution to ESG by the ISI team has been in the area of system monitoring. The ISI team has deployed and maintained the monitoring infrastructure for ESG. This work is described in detail in Section 5.7 of this report. Mei-Hui Su on the ISI team has led the effort to install the Globus Monitoring and Discovery System (MDS) on the ESG infrastructure. This monitoring infrastructure has significantly increased the reliability of ESG by quickly identifying components that have failed in the distributed system. This has resulted in much faster recovery of failed services in the distributed ESG infrastructure, resulting in less overall downtime. The ISI team will continue to lead monitoring efforts, which will be more complicated and sophisticated in the next-generation federated ESG deployment.

ISI closely manages our expenditures of ESG-CET funding to make sure that these expenditures directly serve project requirements. We monitor and do projections for this funding to maintain a steady and sustainable rate of expenditures. We are currently on track to maintain the current level of funding through the remainder of this funding period, and we plan to maintain a similar level of funding through the end of the project.

Ann Chervenak was lead author on a peer-reviewed paper on the ESG monitoring infrastructure at the eScience 2006 Conference in Amsterdam, Netherlands. The complete citation is: Ann Chervenak, Jennifer M. Schopf, Laura Pearlman, Mei-Hui Su, Shishir Bharathi, Luca Cinquini, Mike D'Arcy, Neill Miller, and David Bernholdt, Monitoring the Earth System Grid with MDS4, in Second IEEE International Conference on e-Science and Grid Computing (e-Science'06), page 69, Los Alamitos, CA, USA, 2006, IEEE Computer Society.

Dr. Chervenak and other members of the ISI team have also contributed to several papers listed in Section A.7.

ESG-CET funding was used for partial support of one PhD student, Shishir Bharathi. Currently, we are using ESG funding to support an undergraduate summer intern, Erin Brady, who is working on the ESG data mirroring use case in summer 2009.

4.4.1. Data Replication

Replication of climate data sets was not initially a goal of the Earth System Grid project. Because of the large size of climate data sets, replicating them to multiple sites was considered to be impractical for the first several years of the project. Over time, as the importance of the IPCC data sets stored at PCMDI has increased, several international sites expressed interest in replicating or “mirroring” key portions of the IPCC data sets. Replicating these key data sets has several advantages. Scientists in a particular geographical region would have access to a Figure nearby copy of the data, reducing wide area latencies for data access. Having multiple copies of data sets also provides an increased level of fault tolerance, since data sets are available at other sites even if one site fails or becomes unavailable.

In the past year, the ESG project has worked toward defining the use case for data replication and mirroring with the help of our collaborators at potential mirror sites in the UK and Germany. Initially, we expect that a key subset of the data (the “common core”), which represent approximately 10% of the total data, will be extracted and made available from PCMDI. Mirror sites can replicate this common core or they may construct their own subset of the data for mirroring.

Replication of data sets to a mirror site requires copying the relevant data sets to a Data Node at the mirror site, copying the necessary metadata associated with those data sets to the mirror site's Data Node, and publishing the replicated data sets by making them visible to users of the mirror site's Gateway.

Implementation of the data replication functionality has begun. This work involves the integration of several key ESG components. Once the data sets to be replicated have been identified, the replication service invokes the Bulk Data Movement component to move the data sets reliably. The replication service uses existing ESG metadata API operations to query and replicate the relevant metadata information to the mirroring Data Node. The replication service will use a modified version of the ESG publication client to publish the newly replicated data sets at the mirror site's Gateway, where publication entails making these data sets visible and accessible through the Gateway. Updates to data sets will be identified and propagated to mirror sites by making use of the Versioning functionality. In addition to integrating these existing ESG components, the replication service is responsible for choosing among available source replicas for the data and metadata and eventually for subscription and notification operations.

Initially, we will use manual notifications to inform mirror sites when new data sets are available or when already-mirrored data sets have failed. In the future, we plan to deploy an automated notification/subscription system. Once a mirror site has replicated a data set, it can register a subscription to that data set. If the Data Node that originally published a data set updates the data set, it issues a notification that the data set has changed, which will automatically trigger notifications to any sites that have subscribed to information about that data set. We hope to use an available open source notification/subscription system, but may implement this functionality ourselves.

Our plans for the remainder of the project with respect to data replication include completing the implementation of the data replication service, deploying it and testing it within ESG. We will also work with collaborating institutions in the UK and Germany, making the data replication software available to them and providing support as they mirror data sets. One of our design goals for the replication service is to provide well-defined interfaces to functionality such as bulk data movement, metadata queries, publishing, etc., that allow our collaborators to use the portions of the ESG replication functionality that are most useful to them without requiring them to deploy the entire ESG software stack. We hope to make it possible for our collaborators to use the ESG replication tools along with their own data movement services or metadata catalogs. We expect that additional mirror sites will also join the collaboration, possibly including sites in Asia and Australia.

4.4.2. Monitoring

One technology that has contributed significantly to the robustness of the ESG infrastructure is MDS, the Globus Monitoring and Discovery System. MDS monitors the status of components in the distributed system, including GridFTP services, SRMs, the NCAR portal, http data services, the OpenDAP service, and replica catalogs at all the ESG sites. The MDS consists of two components. One is the Index Service collecting status information from providers at each ESG component, including whether a particular service is currently working correctly. The second is the Trigger Service, which takes action based on monitored conditions. In particular, the Trigger Service sends emails to the ESG administrators' mailing list when components fail. This has resulted in much faster recovery of failed services in the distributed ESG infrastructure, resulting in less overall downtime. Prior to the deployment of the monitoring system, it was not uncommon for failures in the infrastructure to be first detected by ESG users, resulting in longer unavailability of services or data, as well as frustration on the part of our users. With the help of the automated monitoring provided by MDS, the ESG team is quickly informed when components fail, allowing the team to quickly restart failed services.

Monitoring operation of the various components of the distributed ESG environment is central to keeping the system operating effectively and ensuring a good user experience. The ESG's monitoring capability is primarily focused on the needs of the ESG staff that maintain the ESG systems, but it also provides useful information to users.

Analysis and experience show that the ESG user community can generally tolerate service outages of reasonable duration, as long as they have adequate information about the situation. Therefore, the ESG team makes an effort to announce planned outages (e.g. preventative maintenance on an ESG system) in advance, and the user-facing gateways include live status information on the various components of the ESG at each participating site, fed by the monitoring system. The ESG team receives more detailed information from the monitoring system via automatic email messages.

The production monitoring system is based on the Globus Monitoring and Discovery System (MDS), which collect information from probes for each of the ESG's component services. When a service is unresponsive, MDS triggers notifications to the ESG team, and appropriate staff undertake diagnostic and corrective action. The systems provide controls to tune the threshold for sending notifications (for example, brief network interruptions are frequent, and can be perceived as service outages unless the monitoring service checks for multiple successive failures of the probe), and frequency of notifications.

For the next-generation ESG environment, we anticipate that the same basic MDS-based infrastructure will provide us the capabilities we need. Instead, we are focused primarily on the need to extend the breadth and depth of the monitoring applied to each component of the system. An example of broadening the monitoring of the ESG enterprise is to include not just liveness of the various services, but also to monitor resource utilization of the underlying systems. This capability, which we anticipate providing by integrating local monitoring systems like Ganglia or Nagios with MDS, will give managers of ESG sites and the ESG-CET team a better handle on the overall health and stability of the system and help plan for additional resources that may be required to support increasing usage over time. Examples of extending the depth of monitoring would include more sophisticated probes of individual services and multi-site monitoring (executing remote probes from multiple locations) in order to provide more specific and reliable information to the ESG operations staff.

We are currently collaborating with the Metrics working group to understand the extent to which Monitoring and Metrics can share some of the same data collection, aggregation, and reporting infrastructure. While it is not universally the case, for many types of information, the difference is largely the fact that Metrics provides an historical record while Monitoring provides real-time information.

4.5. Annual Progress Report June 2008

4.5.1. Overview of Earth System Grid Project

The Earth System Grid (ESG) project has developed and delivered a production environment that serves a worldwide community with climate data from multiple climate model sources (e.g., CCSM, PCM, IPCC), ocean model data (e.g., POP), CCSM source code distribution, and analysis and visualization tools. Data holdings are distributed across multiple sites including LANL, LBNL, LLNL, NCAR, and ORNL. ESG also operates a dedicated portal that supports the IPCC community in the development of its 4th Assessment Report (AR4). ESG now supports over 6,000 registered users from around the globe; manages over 180 TB of data, models, and tools; and has delivered more than 250 TB of data to its users. It is estimated that over 300 scientific publications have been authored focused upon the analysis of the IPCC data alone.

In 2006, we launched the current phase of the ESG effort, the ESG Center for Enabling Technologies (ESG-CET). The primary goal of this stage of the project is to broaden and generalize the ESG system to support a more broadly distributed, more international, and more diverse collection of archive sites and types of data. An additional goal is to extend the services provided by ESG beyond access to raw data by developing "server-side analysis" capabilities that will allow users to request the output from commonly used analysis and intercomparison procedures.

ESG is a large, production, distributed system – a Data Grid – with primary access points via three web portals: one for general climate research data; another dedicated to the IPCC activity; and a third for the Community Climate System Model (CCSM) Biogeochemistry (BGC) Working Group, which is just going into production at ORNL. The deployment of these three separate portals is driven by international data requirements, restrictions, and timelines. However, they are all based on the same underlying software system. Our goal in ESG-CET is to achieve complete integration of these focused archives, while providing the tailored access and other controls required by the various data owners. In this way, we will provide ESG users with coherent access to ever-growing and increasingly diverse collections of global community climate data.

USC Information Sciences has contributed to the Earth System Grid effort in two main areas: providing the monitoring infrastructure that checks the status of Earth System Grid components and reports errors and working with other ESG partners on the distributed ESG architecture, particularly in the areas of metadata federation and management.

4.5.2. Monitoring the Earth System Grid

As Grids for scientific applications become larger and more complex, the management of these environments becomes increasingly difficult. Commonly, these scientific Grids consist of a large number of heterogeneous components deployed across multiple administrative domains, including storage systems, compute clusters, Web portals, and services for data transfer, metadata management, and replica management. Monitoring these components to determine their current state and detect failures is essential to the smooth operation of Grid environments and to user satisfaction.

Monitoring systems collect, aggregate, and sometimes act upon data describing system state. This information can help users make resource selection decisions and help administrators detect problems. Monitoring systems can typically be queried and, in many cases, can take actions based on events. Grids present additional challenges for monitoring systems because of the frequency with which resources are added and removed and because of the distributed nature of the responsibility for administering resources in a Grid.

To monitor this infrastructure, we use the Globus Toolkit Version 4 (GT4) Monitoring and Discovery System (MDS4). The Globus Toolkit provides middleware to support secure resource sharing among participants in a Grid. MDS4 defines and implements mechanisms for service and resource discovery and monitoring in distributed environments. MDS4 currently includes two higher-level services: the Index service, which collects and publishes aggregated information about Grid resources, and the Trigger service, which collects resource information from the Index Service and performs actions when certain conditions are met.

The services that are monitored in the ESG infrastructure include GridFTP data transfer services, the OPeNDAP service that filters requested information to reduce the amount of data transferred, the ESG Web portal, HTTP servers for data access, Replica Location Service catalogs at several sites, Storage Resource Managers at several sites, and three hierarchical mass storage systems. The Index and Trigger services for ESG run on the machine at NCAR but are maintained and updated by ISI staff.

Features of the monitoring infrastructure for ESG include the following. The ESG portal displays an overall picture of the current status of the ESG infrastructure based on our monitoring information, giving users and administrators an understanding at a glance of which resources and services are currently available. In addition, failure messages provided by the Trigger service help system administrators to identify and quickly address failed components and services. The monitoring system helps avoid system downtime by warning of the imminent expiration date of host certificates on services, so that they can be renewed without service interruptions. Finally, when a particular site has scheduled downtime for site maintenance, it is not necessary to send failure messages to system administrators regarding components and services at that site. We have developed a simple mechanism that disables particular triggers for the specified downtime period.

ISI Accomplishments in Monitoring

- The ISI and ANL teams developed the monitoring services infrastructure (including the MDS4 Index Service and Trigger Service) that is used by ESG to detect component failures. This development is ongoing and includes feature improvements and bug fixes. As new features are added to these services, they are incorporated into the ESG monitoring infrastructure.
- The ISI team deployed the ESG monitoring infrastructure in the distributed ESG environment and supports this deployment. As the ESG architecture evolves (for example, as the new gateway architecture is being deployed), the ISI team has modified the monitoring service deployment to follow this evolution.
- The ISI team has added new triggers as needed by ESG to detect additional failure conditions.
- When failure messages occur, the ISI team helps to identify the cause of failures in the ESG infrastructure. In one recent example, we assisted the ORNL and LBNL teams in tracking down problems with SRM servers that were failing on several sites: mainly at ORNL, but also at NCAR and LBNL.
- The ISI team helped the ORNL team set up the monitoring probes for the SRM, TRM, and RLS services running on the new ESGtg machine.
- The ISI monitoring team is also doing ongoing work to integrate the portal's monitoring framework with the latest version of the Globus Toolkit Version code.

4.5.3. Data Architecture Contribution

Federation of Metadata Catalogs

One of the most important issues in the current generation of the Earth System Grid is the question of how best to federate an increasingly large and diverse set of participants. ISI has been fully engaged in architecture discussions related to federation. The federated ESG architecture consists of a set of ESG gateways and data nodes. There will be at least three Gateway located at NCAR, LLNL and ORNL. Each gateway will provide rich ESG services and sophisticated administrative support as well as access to a distributed and replicated metadata catalog and a query interface through which clients can request any ESG data set by name or by querying for metadata attributes. Gateways need to be fault tolerant, so that they continue to operate even if other gateways in the federated system fail.

The data nodes represent other institutions that contribute data sets to ESG, for example, institutions that publish IPCC data sets. Data nodes will produce data sets, will be responsible for authoring at least some of the metadata for those data sets, and they may use ESG services to assist with this authoring. Data nodes are considered the authoritative source of the data and the associated metadata. Users at a data node must coordinate with an ESG gateway to publish data sets, as described below.

ISI Accomplishments with respect to the federated ESG architecture include:

- The ISI team led a discussion on federated metadata catalogs that began at the February 2007 ESG Metadata Meeting, and followed up with an architecture document that was refined over several document versions based on input from the ESG collaboration. This document describes a variety of options for federated metadata catalogs, including a system that has a single master catalog and systems with multiple masters. Based on this discussion, the ESG participants initially decided on a single master architecture, although this architecture continued to evolve over time.

- The ISI team produced a document and led a discussion on Federation Use Cases in October 2007 on an ESG teleconference. The goal of this document was to discuss issues related to data publication and access and how these were handled by the proposed federated ESG architecture.

- The ISI team led a discussion on the use of the Globus Replica Location Service in the next generation of the Earth System Grid in December 2007 during an ESG teleconference. The discussion concerned how/whether the RLS fit into the evolving ESG distributed architecture.

- The ISI team led a discussion on Options for Accounting and Auditing in ESG in March 2008. This discussion looked at issues for doing accounting and auditing in the distributed ESG infrastructure and described what other projects, such as Teragrid and Open Science Grid, are doing in these areas. The goal of this discussion was to determine whether additional accounting and auditing mechanisms are needed in addition to the metrics work already being done at NCAR.

- The ISI team produced a document and led a discussion on Federated Metadata at the ESG All-Hands Meeting in April 2008. This discussion revisited the issues of how best to federate metadata catalogs given the current design and implementation of ESG gateways and data nodes. The discussion of these issues is ongoing.

The Replica Location Service in ESG

The existing deployed ESG infrastructure uses the Replica Location Service deployed at several sites. The Replica Location Service is a distributed catalog that keeps track of mappings from logical names to one or more physical locations where those files are stored. ESG currently uses the RLS to keep track of the locations of hundreds of thousands of files. During data discovery, the current ESG portal infrastructure queries both the metadata catalog and the RLS to find the location of the files of interest to a user. The RLS returns the location and size of each file, which are used to estimate data transfer times.

ISI accomplishments related to RLS include:

- The ISI team developed the RLS and provides ongoing support for the RLS catalog servers that are deployed in ESG. This support includes the development of new features, providing bug fixes for existing servers, and support for configuring RLS servers in the ESG environment. In the last year, that support has included debugging problems with RLS servers at LANL, LLNL and NCAR. In addition, ISI staff helped ORNL to troubleshoot a problem with the new RLS server deployed on the ESGtg machine and to synchronize the new RLS catalog with the older catalog on the sleepy machine.

- The ISI team recently completed a pure Java client for the RLS, a feature that was requested by the NCAR team to improve the ease of development and the reliability of the ESG portal.

- As already discussed, ISI led a discussion on the future role of RLS in the next-generation distributed ESG architecture.

- ISI has investigated extending the state sharing mechanisms used by RLS to support sharing of RDF triples that store ESG search metadata among ESG gateways. In particular, we are investigating providing to the ESG team a service that would extract RDF triples from a triple store on one gateway and then upload those triples into a triple store at another gateway.

4.5.4. Additional Information

Project Web Site

www.earthsystemgrid.org

Students funded

Students funded: 1 PhD student at ISI

Publications

- "Building a global federation system for climate change research: the earth system grid center for enabling technologies (ESG-CET)," R Ananthakrishnan, D E Bernholdt, S Bharathi, D Brown, M Chen, A L Chervenak, L Cinquini, R Drach, I T Foster, P Fox, D Fraser, K Halliday, S Hankin, P Jones, C Kesselman, D E Middleton, J Schwidder, R Schweitzer, R Schuler, A Shoshani, F Siebenlist, A Sim, W G Strand, N Wilhelmi, M Su and D N Williams, Proceedings of SciDAC 2007 Conference, 24.28 June 2007, Boston, MA. (Also appeared in Journal of Physics: Conference Series, Volume 78, 2007.)
- "The Earth System Grid: Enabling access to multi-model climate simulation data," D. N. Williams, R. Ananthakrishnan, D. E. Bernholdt, S. Bharathi, D. Brown, M. Chen, A. L. Chervenak, L. Cinquini, R. Drach, I. T. Foster, P. Fox, D. Fraser, J. Garcia, S. Hankin, P. Jones, C. Kesselman, D. E. Middleton, J. Schwidder, R. Schweitzer, R. Schuler, A. Shoshani, F. Siebenlist, A. Sim, W. G. Strand, and N. Wilhelmi, 2008: Bulletin of the American Meteorological Society (in review).

Internal ESG Documents and Talks for Architecture Discussions

- "ESG Federated Metadata Architecture", ESG Metadata Meeting, February 2007 (Talk).
- "Metadata Architecture for ESG," Ann Chervenak, February 2007 (Document).
- "Data Federation Use Cases," Ann Chervenak and Frank Siebenlist, October 2007 (Document).
- "The Replica Location Service in the Earth System Grid," Ann Chervenak and Robert Schuler, December 2007 (Talk).
- "Options for Accounting and Auditing in ESG," Ann Chervenak, March 2008 (Talk).
- "Search Metadata: Storage and Sharing Considerations," Robert Schuler and Ann Chervenak, April 2008 (Document).
- "Federated Metadata", Ann Chervenak, ESG All-Hands Meeting, April 2008 (Talk).

Talks

- The Earth System Grid: Turning Climate Datasets into Community Resources, Supercomputing '07, Reno, Nevada, November 2007.

4.6. Progress Report, March 2008

4.6.1. Monitoring

The ISI team continues to provide the monitoring services infrastructure that allows ESG to detect and repair component failures. These monitoring services are essential for the reliable operation of the ESG portals and services. When failure messages occur, the ISI team helps to identify the cause of failures in the ESG infrastructure.

In the last quarter, this support for the ESG monitoring infrastructure included several efforts. We helped the ORNL team set up the monitoring probes for the SRM, TRM, and RLS services running on the new ESGtg machine. We also assisted the ORNL and LBNL teams in tracking down problems with SRM servers that were failing on several sites: mainly at ORNL, but also at NCAR and LBNL. Finally, the monitoring team is also doing ongoing work to integrating dataportal's monitoring framework with the

Globus Toolkit Version 4.1 codebase. In particular, the goal of this work is to capture information on transitions of services from up to down and the reverse.

4.6.2. Support for Current RLS Deployments in ESG

The ISI team has provided ongoing support for RLS catalog servers that are deployed in the current ESG infrastructure. During this quarter, that support has involved tracking and debugging required for RLS servers at LANL, LLNL and NCAR. In addition, ISI staff helped ORNL to troubleshoot a problem with the new RLS server deployed on the ESGtg machine and to synchronize the new RLS catalog with the older catalog on the sleepy machine.

4.6.3. Role of RLS in New ESG Architecture

The ISI team led a detailed discussion and email thread about the future role of the RLS in the next generation ESG architecture. As a result of this discussion, we planned to investigate the utility of RLS catalogs for replicating replica location mappings among gateways. As part of this work, Rob Schuler of ISI worked with Luca and Nate at NCAR to extract mappings from their gateway database and prepare these for insertion in an RLS catalog. In the quarter ahead, we plan to demonstrate the utility of the RLS by replicating these mappings on gateway nodes at ORNL, NCAR and LLNL.

4.6.4. Extending RLS State Sharing Mechanisms for Sharing RDF Triples Among Gateways

One outcome of the discussion of the future role of RLS was the identification of the RLS state sharing mechanism, which uses soft state update techniques, as a technique that might be generally useful for sharing state among ESG gateways. In particular, the ESG team would like to have a service that can extract RDF triples from a triple store on one gateway and then upload those triples into a triple store at another gateway. The RLS team has begun to investigate the design of such a service. Important design questions include what the size of these updates would be and whether it would be possible to compress them. In the coming quarter, we plan to design this service and provide an initial prototype.

4.6.5. Investigation of Accounting and Auditing Technologies on Other Grid Projects

A final effort by the ISI team this quarter was to survey the mechanisms used for auditing and accounting in three major grid projects (OSG, TeraGrid and CEDPS). The results of this survey were reported to ESG, with the hope that some of these techniques may prove useful, in addition to the logging and accounting already performed by ESG. We also identified accounting and logging issues that ESG has not yet addressed, such as how accounting data gathered at data nodes will be communicated to gateways.

4.7. Progress Report, September 2007

ISI Monitoring, Data Catalogs, and Federation Highlights

The ISI team continues to provide the monitoring services infrastructure that allows ESG to detect and repair component failures. These monitoring services are essential for the reliable operation of the ESG portals and services. This work has involved incorporating new features into the ESG monitoring infrastructure, particularly related to the Trigger service that reacts to the failed state of services, as these features are provided by the Globus Monitoring and Discovery Service team. ISI staff also monitor these services to ensure they are operating correctly and to register scheduled downtime to avoid unnecessary failure messages. In addition, the ISI team maintains and improves the Replica Location Service (RLS) catalogs for the Earth System Grid Project. During this reporting period, the ISI team completed a pure Java client for the RLS, a feature that was requested by the NCAR team to improve the ease of development and the reliability of the ESG portal. Finally, the ISI team is working on the design of federated metadata catalogs and on design issues related to the federation of data sources and gateways in the ESG distributed architecture. Currently, the ISI team is working with ANL to develop use cases for federation.