Globus Toolkit Support for Distributed Data-Intensive Science

W. Allcock[1], A. Chervenak[2], I. Foster[1,3], L. Pearlman[2], V. Welch[1], M. Wilde[1]
[1] (Argonne National Laboratory, Argonne, IL USA)
[2] (USC Information Sciences Institute, Marina del Rey, CA USA)
[3] (University of Chicago, Chicago, IL USA)

Abstract

In high-energy physics, terabyte and soon petabyte-scale data collections are emerging as critical community resources. A new class of "Data Grid" infrastructure is required to support distributed access to and analysis of these datasets by potentially thousands of users. Data Grid technology is being deployed in numerous experiments through collaborationssuch as the EU DataGrid, the Grid Physics Network, and the Particle Physics Data Grid[1] . The Globus Toolkit is a widely used set of services designed to support the creation of these Grid infrastructures and applications. In this paper we survey the Globus technologies that will play a major role in the development and deployment of these Grids.

1 Introduction

Grid computing environments that enable "virtual organizations" to share their computing resources as they pursue common goals are characterized by heterogeneous resources, decentralized control, and the lack of existing trust relationships [2]. Future LHC experiments will require sophisticated multi-tiered data grids, some spanning over 140 institutes, 30 countries, and 1800 researchers[3]. In the sections below, we describe the Globus Toolkit services for managing data, resources, and security that support this new Grid computing paradigm and this scale of resource deployment.

2 Data Grid Architecture

In the tiered data grid of an international physics collaboration, reconstructions of raw and simulated data will propagate between computing centers, across high speed wide-area networks, from the point of origin to multiple points of analysis, and analysis results will flow among the centers in various patterns, not all of which are completely understood at this point. Two fundamental Globus facilities that support this paradigm are the fast, efficient, and secure transport of large and numerous data files, and the management and tracking of replicated (cached) files at multiple sites to reduce data access latency and the demands on network bandwidth.

One model for this collaboration is as follows: groups of experimenters focused on specific research topics will use grid information services (described below) to locate computing resources with sufficient secondary storage to hold a working set of the data files they intend to process. The researchers will locate logical files names of interest to them (from the experiments' metadata catalogs) and request, either directly or through their application frameworks, that replicas of these files be created in a repository from which their computation jobs can access them. In the background, grid-specific policy functions will decide where in the grid various data collections should be replicated, while space management utilities will be periodically scanning storage servers to locate replicas that, based on their usage history and other policies, are candidates for removal.

2.1 Transport

The GridFTP transport service is based on extensions to standard FTP protocol that create a universal grid-wide transport protocol. It provides secure, high throughput data transfer even on high-speed wide area networks, and third party transfers which allow the source, destination, or both to be striped, with arbitrary and potentially different topologies or even file access mechanisms (e.g., library API vs. kernel system calls). A plug-in interface supports the widely-ported MPI-IO interface, which allows GridFTP to access filesystems such as the Parallel Virtual File System, PVFS [4] and Sun's Parallel File System, PFS. It also enables applications to readily develop parallel I/O access methods to customized disk layouts and topologies. This capability is illustrated in figure 1, below.

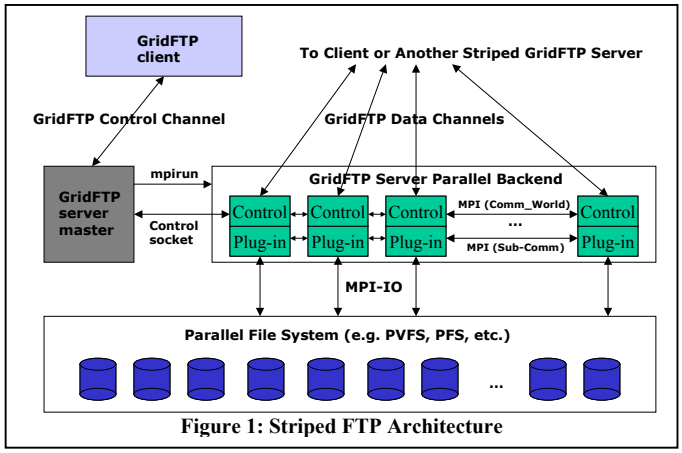

Figure 1: Striped FTP Architecture

Features that significantly increase GridFTP's speed are the ability to flexibly set TCP tuning parameters, and to create multiple TCP data channels between source and destination hosts (where necessary to achieve high WAN utilization). GridFTP supports 64-bit file sizes and offsets, and has the ability to transfer partial files in arbitrary non-contiguous segments. Other advanced features include checkpoint/restart of transfers, data channel reuse for consecutive file transfers, and user plug-ins that allow advanced applications to perform memory-to-memory transfer, server-side computation without additional data copies, and provide integrated instrumentation for the recording of logging and audit trails. All the mechanisms described here (except for striping, which is in testing) are part of the Globus Data Grid Alpha release.

The wide area network performance achieved by GridFTP is impressive and is a major focus of our ongoing research and development. Its can currently sustain 500 Mbit/sec for hours and achieve 5 second bursts over 1 Gbit/sec, continuously transferring 2-GByte files striped between 8 servers at each endpoint (Dallas to Berkeley), over a 2.5 Gbit/sec NTON OC-48 link [5].

2.2 Replication

The second major Globus component provided for constructing data grids is the *replica management service* which caches files in a distributed computing system so as to optimize the performance of the data analysis process. This service consists of a replica catalog where information about replicas is stored and a set of registration and query operations: register a file, create and delete a replica of a registered file, and locate a replica.

The replica management service can be used by higher-level services, for example, by a *replica selection service* that selects among replicas based on predicted data transfer time or by a *replica creation service* that automatically generates and registers new replicas in response to data access patterns and the current state of the grid.

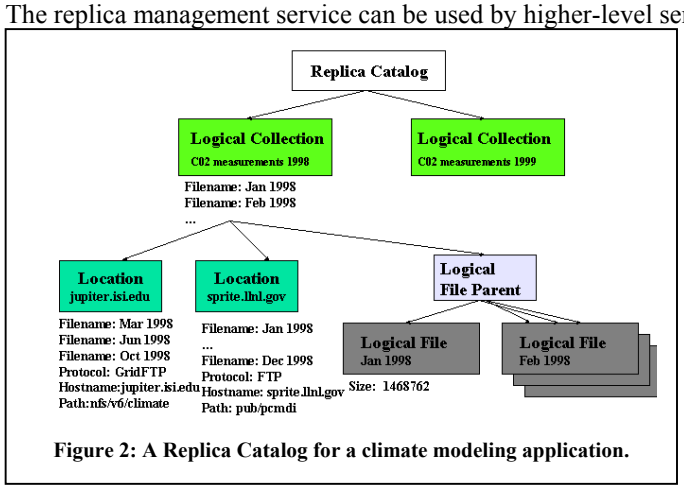

Figure 2: A Replica Catalog for a climate modeling application.

Figure 2 shows a replica catalog for a climate modeling application with two logical collections of CO_2 measurements for 1998 and 1999. The 1998 collection has two physical locations, a partial collection on the host jupiter.isi.edu and a complete collection on sprite.llnl.gov. The location entries contain attributes that list all files stored at a particular physical location, and the protocol, hostname, port, and path required to map from logical names for files to physical URLs. The example catalog also contains logical file attributes (such as file size).

3 Resource Discovery, Monitoring, and Management Architecture

A vital component of Globus for distributed virtual organizations is the ability to discover, share, and monitor the resources that make up the VO. Grids of the scale demanded by international physics experiments require a grid resource information service that is distributed, for both high performance and fault tolerance. The Globus MDS-2 architecture[6] implements a service that consists of Grid Resource Information Servers (GRISs) associated with resources and a Grid Information Index Server (GIIS) that aggregates information from multiple GRISs. These services are linked by a *Grid Information Protocol* (used by clients of the MDS to query the information service) and a *Grid Registration Protocol* (used by resource providers to register information about resources and to refresh that information periodically).

Performance monitoring is another essential service for data grids. Applications and resource brokers need to monitor the state of the grid, including network bandwidth, space available at storage servers, the load on computational resources, and the progress of data transfers. Based on this state information, the application or broker can make improved scheduling and resource selection decisions. Work is under way to enhance the monitoring capabilities of the Globus toolkit, and in particular, to link GridFTP servers into the resource monitoring architecture, to provide an information base from which to perform intelligent replica selection [7].

GRAM, the Globus Resource Allocation and Management service[8], provides the ability to schedule computations on hosts throughout the grid. This fundamental component is beyond our scope to discuss here, but we note that the Condor high-throughput computing environment has been adapted (via Condor-G[9]), using Globus, to harness resources across a Grid as if they all belonged to one large personal computing domain.

4 Security: Authentication, Authorization, and Communities

The Grid Security Infrastructure (GSI) provides facilities for the mutual authentication of services and their users, protection of data, and for the delegation of user credentials [10]. GSI is used for all resource access and sharing. Using GSI for all GridFTP control and data exchange enables GridFTP to provide secure data transport between all Grid storage servers.

Under development is a Community Authorization Service (CAS), which facilitates community-based access control. The administrator of a resource server grants permissions on a resource to the CAS server; the CAS server then grants fine-grained permissions on subsets of that resource to members of the community. For example, the administrator of a GridFTP server may grant permissions on a filesystem to a CAS server, which then be used to grant permissions to files and directories within that filesystem to community members. A CAS server provides these functions:

1. It acts as a community registration authority: community administrators establish trust relationships with users and resource providers and then register their credentials with the CAS server.

2. It provides an interface to an authorization database, allowing community members to grant permissions on groups of objects (possibly residing on different servers) to groups of people.

3. It grants capabilities (credentials that grant specific access rights -- based on the policies stored in the authorization database) to users, who then use those capabilities to authenticate to resource providers. The resource providers then verify that their own local policies allow the CAS server to grant capabilities for each request and that each request is authorized by the policy encoded in the capability.

4. It provides a scalable approach to managing information about multiple certificate authorities.

CAS will be used to provide community access control for file servers, replica catalogs, and MDS information servers, among other applications.

5 Experience

The NASA Information Power Grid is perhaps the most extensive deployment of the MDS-2 Grid Information Service, and has done extensive measurements of that service. Recent simulations of US

airspace traffic were conducted across the three sites of this grid [11]. The Earth Systems Grid [12] brings together shared resources and climate scientists from ANL, NCAR, ORNL, and LBL. It is progressing past the demonstration stage, developing software and infrastructure to tackle important problems in the modeling of climate processes.

Within the CMS experiment, an application called GDMP [13] has been modified to use the replica catalog and GSI-enabled FTP and is being used experimentally to move data files containing object-oriented databases between remote database federations. GDMP will soon be integrated with GridFTP.

6 Future Plans

Above GridFTP and the replication services we are designing new data grid layers to fully automate the process of transporting large numbers of large files reliably between sites, with automatic recovery from server crashes and space exhaustion, determination of striping configurations, and selection of the most efficient replica for the source of a transfer. To make the replica catalog service itself more scalable and fault-tolerant, we are exploring various architectures for distributing and replicating the catalog service itself. We are also investigating requirements, architectures, and collaborations to Grid enable HPSS and other tertiary storage systems (such as Enstore and Castor).

As part of the GriPhyN project, we are studying the integration of *virtual data* into the data grid architecture, which allows data to be either retrieved or recomputed to satisfy a request. Support for virtual data requires mechanisms that can generate the desired data if it is not available or if recomputation is more cost-effective than data transfer. We are also studying the problems of making grid-wide resource scheduling decisions in this more complex virtual data paradigm.

The ultimate goal for the data grid of the future is to give researchers a virtual computing environment in which to run their applications, treating their grids as resources that are as simple to use as their desktop computer systems are today. This will require the fully automated and globally optimized location and scheduling of computing, networking, and storage resources.

References

1. http://www.eu-datagrid.org, http://www.ppdg.net, http://www.griphyn.org
2. Foster, I., C. Kesselman, and S. Tuecke, *The Anatomy of the Grid: Enabling Scalable Virtual Organizations.* Intl. J. Supercomputer Applications, 2001. (to appear).
3. Bethke, S., et al., *Report of the Steering Group of the LHC Computing Review.* 2001.
4. Carns, P., et al., *PVFS: A Parallel File System For Linux Clusters.* Proceedings of the 4th Annual Linux Showcase and Conference, Atlanta, GA, October 2000, pp. 317-327, 2000.
5. Allcock, B., et al. *Secure, Efficient Data Transport and Replica Management for High-Performance Data-Intensive Computing*. in *Mass Storage Conference*. 2001.
6. Czajkowski, K., et al. *Grid Information Services for Distributed Resource Sharing*. in *IEEE International Symposium on High Performance Distributed Computing*. 2001: IEEE Press.
7. Vazhkudai, S., S. Tuecke, and I. Foster. *Replica Selection in the Globus Data Grid*. in *International Workshop on Data Models and Databases on Clusters and the Grid (DataGrid 2001)*. 2001: IEEE Press.
8. Czajkowski, K., et al., *A Resource Management Architecture for Metacomputing Systems*, in *The 4th Workshop on Job Scheduling Strategies for Parallel Processing*. 1998. p. 62--82.
9. Frey, J., et al. *Condor-G: A Computation Management Agent for Multi-Institutional Grids*. in *Tenth IEEE Symposium on High Performance Distributed Computing (HPDC10)*. 2001.
10. Butler, R., et al., *Design and Deployment of a National-Scale Authentication Infrastructure.* IEEE Computer, 2000. 33(12): p. 60-66.
11. NASA web site: http://www.nas.nasa.gov/Main/Features/2001/Summer/ipg_aviation.html
12. Allcock, B., et al., *High-Performance Remote Access to Climate Simulation Data: A Challenge Problem for Data Grid Technologies.* Supercomputing 2001 Proceedings, to appear, 2001.
13. Samar, A. and H. Stockinger, *Grid Data Management Pilot (GDMP): A Tool for Wide Area Replication.* IASTED International Conference Applied Informatics (AI 2001) February 19-22, 2001 Innsbruck, Austria, 2001.