

SciDAC DataGrid Middleware
A High-Performance Data Grid Toolkit:
Enabling Technology for Wide Area Data-Intensive Applications
Quarterly Report July 2003 thru September 2003

Accomplishments this Quarter:

XIO improvements based on use and feedback

The Globus eXtensible Input Output (XIO) system underwent some feature improvements to better facilitate layered protocols such as reliable UDP variants. There has also been debugging based on initial usage under the new GridFTP server implementation. We are confident that XIO will be the most stable, heaviest tested, and best documented V1.0 component in the history of the Globus Toolkit.

New GridFTP server is moving data, but still lacks many features

We achieved a significant milestone in this quarter with the first data movement by the new server. The server still lacks many features, but the design and infrastructure are in place and we are already beginning to test the server, while simultaneously adding to the feature set. Both the wuftp based server and the new server will appear in the 3.2 release of the Globus Toolkit.

Feature Enhancements to wuftp based GridFTP server

The wuftp based GridFTP server received several feature enhancements for the GT3.2 release. This will be the last feature enhancements made to the wuftp based server. Only bug fixes will be provided on this version from this point forward. The features included:

- **MLSx commands:** MLST and MLSD are RFC defined commands for FTP that provide consistent structured directory listings that are programmatically parseable.
- **File Globbing:** globus-url-copy now supports moving multiple files with a single command. It can accept a directory or a glob (such as *.dat) and will move all the files. All features of the protocol are functional and channel caching is used to improve performance.
- **Checksum Command:** Allows for end to end verification of the transfer by doing a checksum on the disk image on each end after a transfer. This can detect things that the TCP checksum might miss.
- **Switch for RFC1738 compliant URLs:** Most FTP servers, and our GridFTP server, do not follow the RFC when it come to interpreting FTP URLs. For most situations, this is a good thing. There are some communities for whom this is a problem. We now provide a switch that will interpret the URLs according to the RFC
- **chmod Support:** The GridFTP server now adds chmod to its list of filesystem operations it can perform. Note that this is not standard and will have to be introduced to the GGF for standardization.

RFT Performance improvements and Functionality Enhancements

The Reliable File Transfer (RFT) Service has been the flagship OGSi service since the release of GT3.0. The reliability aspects of it have been very good, but the XML deserialization was

limiting us to about 450 files and could take up to 30 minutes. We have made changes to the request schema that greatly reduces the overhead. We also allowed the specification of a directory, which makes the client simpler and also helps with the limit on the number of files and processing time. This work will be released in the 3.2 alpha.

Replica Location Service (RLS) improved

RLS was released in the Globus Toolkit GT3.0 release in late June 2003. This release has made the RLS part of an official Globus toolkit release for the first time. During this quarter, additional testing and some minor bug fixes have been incorporated into the toolkit. Also, some additional functionality was added related to bulk operations.

Hierarchical Replica Location Index (RLI) support added for GT3.2

A new version of the Replica Location Service that supports a hierarchical Replica Location Index was developed this quarter. It is currently undergoing testing and will be included in the Globus Toolkit 3.2 release scheduled for the next quarter.

Replica Manager Prototype completed

We have implemented a client tool that performs copy and registration operations, invoking gridFTP servers to copy data items and registering the resulting copies in the RLS. This tool is intended to provide the same functionality that was formerly provided by the replica management API in earlier versions of the Globus toolkit.

OGSI wrapper for RLS in progress

Finally, we have been developing a simple Grid service wrapper around the existing Replica Location Service implementation. An initial implementation is complete, and we expect to release this grid service through the GTR (Grid Technology Repository) in the first half of October 2003.

Ongoing work to maintain Compatibility between US and EU tools

An ongoing effort related to the RLS is to work with the European DataGrid project to overcome the problem of two implementations of RLS that are not interoperable. This situation arose last December after the EDG group diverged from our common design and implementation efforts and built an incompatible RLS. During the last quarter, we traveled to Italy for a meeting and devised a plan for achieving interoperability. While the general plan was agreed upon during that meeting, we are still waiting for resource commitments from the EDG/LCG projects before progress can be made on achieving interoperability. Discussions are actively ongoing.

Continued development of the Metadata Catalog Service based on OGSA DAI Service

We have continued development of the Metadata Catalog Service this quarter. After conducting an extensive performance study for the paper on MCS to be presented at the SC2003 conference, we have been conducting a re-implementation of the service as an extension of the OGSA Data Access and Integration Service, which provides a grid service front end to a variety of database back ends.

Continue to be a strong force in Standards Work

We continue to be extremely active in the Global Grid Forum OREP (OGSA Services for Data Replication) Working Group and the DAIS (Data Access and Integration) Working Group. We have attended multiple interim meetings and have hosted two of them.

Researched failure recovery in automated mass data transfers.

Development continues on Stork, which treats data allocation and transfer in the same manner as computation jobs. A primary focus this quarter was on failure recovery. When events can be identified as temporary failures, Stork is now capable of failing over to alternate protocols or services. When failures are deemed permanent, higher-level recovery can be performed at the workflow level through the use of “mitigation” jobs which attempt to un-do work that has already been accomplished. This is discussed in the “Run-Time” paper cited below.

<http://www.cs.wisc.edu/condor/stork>

NeST development and hardening.

The NeST storage appliance saw further development this quarter, refining the ability to bind several NeSTs together into a cooperative cache that can minimize wide area traffic by permitting input data to be discovered and accessed within the local area. In addition, significant effort was put into performance debugging, improving the latency of single operations by a factor of 2 to 3. <http://www.cs.wisc.edu/condor/nest>

Parrot User-Level Filesystem implemented and tested.

Following some initial design work the previous quarter, we completed the implementation of the Parrot user-level filesystem. This tool attaches to any arbitrary process (or process tree) through the debugger interface and converts ordinary POSIX I/O into operations on remote systems, currently including NeST, GSI-FTP, DCAP, and RFIO. It is now in use locally at Wisconsin, and has been deployed in an international test bed for performing BaBar monte carlo work over the wide area. A development version is now available to the public for testing.

<http://www.cs.wisc.edu/condor/parrot>

Researched coordinated access to CPU and I/O resources.

We continue to research efficient and expressive methods of integrating access to batch CPU and temporary disk space into order to execute large workloads with both CPU and I/O needs. Our current approach involves deploying a set of NeSTs alongside a Condor pool, which jobs access via Parrot. A global scheduler transfers data and submits jobs in a coordinated fashion, while dealing with failures. We call this a “batch aware” file system; our initial results are currently in review.

Papers Published or in Progress

“A Metadata Catalog Service for Data Intensive Applications”, Shishir Bharathi, Ann Chervenak, Ewa Deelman, Carl Kesselman, Mary Manohar, Sonal Patil, Laura Pearlman, Gurmeet Singh, will appear in Proceedings of SC2003, November 2003.

Douglas Thain and Miron Livny, “Parrot: Transparent User-Level Middleware for Data-Intensive Computing”, in Proc of Workshop on Adaptive Grid Middleware, New Orleans, 2003.

George Kola, Tefvik Kosar and Miron Livny, "Run-time Adaptation of Grid Data-placement Jobs", Proceedings of Workshop on Adaptive Grid Middleware, New Orleans, 2003.

John Bent, Douglas Thain, Andrea Arpaci-Dusseau, Remzi Arpaci-Dusseau, and Miron Livny, “Explicit Control in a Batch-Aware Distributed File System,” in review.

Presentations Given

02 Jul 2003: "GT3 Overview" GridPP Project Meeting, Oxford, England

July 18, 2003: “Data Management Services in GT2 and GT3” by Ann Chervenak at the International Grid Summer School in Vico Equense, Italy

19 Sep 2003: "Data Transport and OGSA" Data Mining Workshop, Minneapolis, MN

28 Sep 2003: “Parrot” lecture at AGridM 2003.

28 Sep 2003: “Run-Time” lecture at AGridM 2003.

29 Sep 2003: "Transition and Evolution: Moving to Grid Services" 7th European DataGrid Project Meeting, Heidelberg, Germany