

SciDAC DataGrid Middleware
A High-Performance Data Grid Toolkit:
Enabling Technology for Wide Area Data-Intensive Applications
Quarterly Report January 2005 thru March 2005

Accomplishments this Quarter:

GT3.9.5 (beta) released

With this release, the interfaces are frozen for the GT4.0 release. We will continue to performance tune and bug fix, but external developers can code to APIs, work with configuration files, etc, and be confident they will not change. This release includes several components supported by the SciDAC DataGrid Middleware project: GridFTP, RLS, RFT and XIO.

Statically linked GT3.9.5 GridFTP server released as optional component in VDT 1.3.3

This allows the new server to be a drop in replacement for older servers without having to worry about version mismatches with dynamically linked libraries.

Single GridFTP host supports 1800 clients

Graduate students at the University of Chicago employed a tool they have developed called DiPerf (Distributed Performance Testing facility). This employs the PlanetLab worldwide host base to launch clients against services. The DiPerf system failed, not the server. We believe the server would have continued to scale until it ran out of resources, probably file descriptors. It is interesting to note that the load on the machine remained manageable since either the disk or the network will become the bottleneck and after that additional clients require very little in the way of resources.

New GridFTP implementation continues to win praise

The new server implementation continues to demonstrate that it is faster, more stable, and more extensible than its predecessor. The Grid Physics Network (GriPhyN) project has been using the GT2.4.3 release and was adamant that they would not switch to newer versions until there was proof of stability. They have a large work flow they refer to as coadd. During a typical run, the 2.4.3 server would have on the order of 1000 failures during a run. We convinced them to install the 3.9.5 server from VDT on their testbed. They re-ran a coadd workflow with zero failures.

GridFTP Data Storage Interfaces (DSI) continue to be implemented

The Storage Resource Broker announced GridFTP access to SRB with their last release. This is achieved by replacing the standard POSIX file system DSI with the SRB DSI. The NeST DSI is complete but has not been distributed yet. The HPSS DSI is making rapid progress. We have a development meeting with HPSS in April, and expect the DSI to be complete next quarter.

Reliable File Transfer (RFT) Service moving the Sloan Digital Sky Survey (DR3) archive

With a single command, RFT began the movement of slightly under 1 million files for a total of three terabytes. We have intentionally killed the RFT server several times to test recoverability. Once we restarted the server, it recovered with no human intervention.

RFT performing well despite being outside its design space

RFT was designed to handle massive file transfers, i.e., relatively few submissions that would run for hours, days, and weeks, perhaps months. However, with GRAM using RFT for file staging, we are operating at the opposite end of the spectrum: many submissions per second, usually of one file that will transfer in a few seconds. This has proven to be a challenge. We have made several improvements and continue to work with GRAM to finish resolving problems.

The Extensible Input / Output (XIO) system gaining ground in the network research community

With its ability to plug-and-play drivers, compose functionality dynamically at run time, and to plug in under GridFTP with very little modification required, XIO is gaining increased interest amongst the network researchers.

Simulation framework for studying I/O intensive grid workloads

This quarter we continue the development of a discreet event based simulation framework to be used for studying I/O intensive grid workloads. This simulator consists of 15,000 lines of code and models disks, networks and memory buffer caches for a distributed compute cluster. Further, this simulator can bind together the distributed disks into a batch-aware distributed file system. The focus of the simulator is then to design and study scheduling algorithms by which the scheduler can best utilize this batch-aware distributed file system and coordinate the allocation of data with the placement of jobs.

Formalization of five distinct data allocations

Using the simulation framework, we then developed a taxonomy of allocation strategies and measured their effectiveness across a range of representative I/O workloads. We then identified eleven key factors which influence the relative performance of these allocations. Finally, using these eleven key factors, we have quantified the sensitivity of each allocation to each factor.

Developed analytic predictive models

Finally, we have developed simple analytic models by which a batch scheduler can dynamically determine which of the possible allocations will maximize throughput for a particular workload. We have then demonstrated the accuracy of these predictive models by obtaining results which closely mirror the simulation results.

Documentation

These findings are currently being codified into a dissertation document by John Bent. The goal of this document is to serve as a valuable reference about I/O intensive batch workloads. The design of the distributed file system framework, the different allocation strategies and the predictive models are all contained within this document. In the future, these ideas will be incorporated into production batch scheduling systems such as Condor.

Replica Location Service Development

We have continued to work on the code base of the Replica Location Service in preparation for the Globus Toolkit Version 4.0 release. This includes fixing bugs related to connection

management in the RLS and also identifying a file descriptor leak in the underlying XIO libraries. We are currently adding several features to the RLS, including a script that will provide service information about the RLS to the Monitoring and Discovery Service (MDS) and the ability to collect usage statistics from RLS services running in various Grids. We have also improved documentation for the service. We continue to support a growing number of applications that use the RLS in production, including the Laser Interferometer Gravitational Wave Observatory (LIGO) project, Earth System Grid, the US portions of the CMS and Atlas physics experiments, Nordugrid and others.

Design and Implementation of a WS-RF Data Replicator Service

One of the most important activities we have undertaken in the current quarter is the implementation of the Data Replicator Service (DRS), which will be included as a Technical Preview Component for the Globus Toolkit Version 4.0 release. The goal of this service is to generalize the publication component of the Lightweight Data Replicator system from the University of Wisconsin at Milwaukee. The DRS uses a pull-based replication model to bring a set of requested files to a local site. The DRS interacts with the RLS to determine whether files exist locally, and if not, to find the locations where those files exist in the Grid. Then the DRS creates file transfer requests using the Globus Reliable File Transfer (RFT) service. The Data Replicator Service (DRS) is implemented in Java and complies with the WS-RF specifications.

Integration of Replica Location Service with POOL

We continued to work with the US CMS project on the integration of the Globus RLS with the POOL environment, which provides persistent object management for particle physics applications. In the current quarter, we have worked with US CMS and the POOL project at CERN to package this integration for the next POOL release.

OGSA DAI Performance Evaluation

In this quarter, we continued our evaluation of the performance of the latest release of the OGSA DAI database service to determine whether it could be used to implement the Earth System Grid metadata catalog. On the positive side, we determined that the OGSA-DAI server did not crash, as it had done in earlier tests. We were able to successfully perform updates involving more than 10,000 SQL statements. On the negative side, the performance of OGSA-DAI is still slow for our update requirements.

Monitoring

We initially implemented a monitoring infrastructure for the Earth System Grid project to monitor the state of deployed ESG services, including the ESG portal, GridFTP servers, HTTP servers, mass storage systems and the OpenDAP-g server. This monitoring infrastructure was originally based on the Globus Toolkit Version 3 indexing and archive service.

During the current quarter, we have worked to generalize this monitoring infrastructure to make it suitable for use by other projects. In addition, we have begun work to migrate the monitoring infrastructure from Globus Toolkit Version 3 to Version 4, where we will be able to take advantage of ongoing development for the MDS Index Service and the WebMDS visualization component.

Open Science Grid

We are leading the effort on defining data management architecture in the Blueprint document of the Open Science Grid activity.

OGSA Data Services Group

This quarter, we began working with the OGSA Data Services Group on their data architecture document.

Plans for Next Quarter

1. Version 4.0 of the Globus Toolkit will be released as a stable release during the next quarter.
2. As a result of the Version 4.0 release, a significant portion of our efforts over the next quarter will revolve around bug fixes, responding to questions on the discuss email list, updating documentation based on feedback, and assisting various user communities in porting / integrating this new release. We feel this is critical to the success of this release.
3. Improvements to the Replica Location Service (RLS) to improve maintainability.
4. Continued development of the Data Replicator Service, including a new modular approach to the service architecture. Our goal is to have a well-defined reliable replication and replica registration components and a policy engine. This will allow us to implement a variety of replication services (push-based, pull-based, etc.) that re-use large amounts of code, with some differences associated with particular policies.
5. Complete HPSS DSI
6. Continue support for the RLS integration with the POOL infrastructure
7. Continue supporting existing monitoring infrastructure.
8. Porting the monitoring infrastructure to the GT4 environment to take advantage of development on Index Service and WebMDS visualization service
9. Continue interactions with Open Science Grid and OGSA Data Services groups.

Papers:

“The Earth System Grid: Supporting the Next Generation of Climate Modeling Research,”
D. Bernholdt, S. Bharathi, D. Brown, K. Chancio, M. Chen, A. Chervenak, L. Cinquini,
B. Drach, I. Foster, P. Fox, J. Garcia, C. Kesselman, R. Markel, D. Middleton, V. Nefedova,
L. Pouchard, A. Shoshani, A. Sim, G. Strand, D. Williams,
Proceedings of the IEEE, vol. 93, 3, pp. 485- 495, March 2005.

“Design and Implementation of a Data Replication Service Based on the Lightweight Data Replicator System,” A. Chervenak, C. Kesselman, S. Koranda, B. Moe, R. Schuler, submitted to 14th IEEE International Symposium on High Performance Distributed Computing (HPDC-14), 2005.

“The Globus Striped Framework and Server”, Bill Allcock, John Bresnahan, Raj Kettimuthu, Mike Link, Catalin Dumitrescu, Ioan Raicu, Ian Foster submitted to 14th IEEE International Symposium on High Performance Distributed Computing (HPDC-14), 2005.

Presentations:

“Design, Performance and Scalability of a Replica Location Service”, presented by Bill Allcock at GlobusWorld 2005, February, 2005.

“Overview of GT4 Data Services”, presented by Bill Allcock and Neil P. Chue Hong at GlobusWorld2005, February 2005.

“GridFTP for Users: The new server” a half day hands on tutorial taught at the National E-Science Center in Edinburgh, Scotland, presented by Bill Allcock, Jan 2005

“GridFTP for Admins: The new server” a half day hands on tutorial taught at the National E-Science Center in Edinburgh, Scotland, presented by Bill Allcock, Jan 2005

“GridFTP for Developers: The new server” a half day hands on tutorial taught at the National E-Science Center in Edinburgh, Scotland presented by Bill Allcock, Jan 2005

“GridFTP for Developers: The new server” a 90 minute tutorial taught at GlobusWORLD in Boston, presented by Bill Allcock, Feb2005

“An Overview of the Architecture and Performance of GT4” a 90 minute presentation to the EGEE project at CERN, presented by Bill Allcock, Jan 2005