

**SciDAC DataGrid Middleware**  
**A High-Performance Data Grid Toolkit:**  
**Enabling Technology for Wide Area Data-Intensive Applications**  
**Quarterly Report April 2004 thru September 2004**

**Accomplishments this Quarter:**

**Striping functionality added to GT3.9.2 Development Release**

The release of GT3.9.2 marks the first time striping functionality has been in an official release. Early performance testing on the TeraGrid yielded very promising results. Running memory to memory, we were able to achieve 27 Gbs on a 30 Gbs link (90% utilization) with only 32 nodes. This represents nearly perfect linear scaling. Disk to disk results achieved 17.5 Gbs with 64 nodes. We believe the current bottleneck is the disk sub system and intend to contact the sites and the GPFS technical staff to try and improve this performance.

**Reliable File Transfer (RFT) Service ported to Web Services Resource Framework (WSRF)**

The Reliable File Transfer (RFT) service has been successfully ported from the Open Grid Service Infrastructure (OGSI) to the Web Services Resource Framework (WSRF). We will continue to support the OGSI version at least until the release of 4.2 (late 2005). Initial testing indicates some scalability problems in the WSRF core libraries. We are working with the WSRF core developers to address these issues.

**Additional Drivers for the eXtensible Input Output (XIO) System**

We completed several new drivers to round out the functionality and usability of XIO. We added a MODE E driver, which enables Open/Close/Read/Write (OCRW) access to the GridFTP data channel (multiple TCP streams). Since MODE E can have out of order arrival of data, we implemented an ordering driver which buffers the blocks and presents them in order to the application. We also implemented a GridFTP driver which allows OCRW access to a file via a GridFTP server. Finally, we implemented a queuing driver to allow multiple outstanding writes. We are also working on a UDT (reliable UDP based protocol from Bob Grossman's group at UIC). Currently it is functional, but we are not getting the performance out of UDT that we should. We are currently working with the UIC crew to resolve these issues.

**Replica Location Service Development**

During the current quarter, most of our efforts have gone into improving the reliability and stability of the Globus Replica Location Service implementation. A number of applications are using RLS in production, including The Laser Interferometer Gravitational Wave Observatory (LIGO), Earth System Grid, the US portions of the CMS and Atlas physics experiments, Nordugrid and others. As the scale of RLS usage has increased to tens of millions of files and tens of sites, subtle bugs and operational problems were uncovered, including connection management issues, memory leaks and authorization errors. Most of our development effort in the last quarter has focused on tracking down and repairing these.

We also added some features requested by users in the current quarter. These include adding operations to *rename* existing entries with a different logical file name, as required by the Atlas physics application. We have also implemented a bulk or collective query to determine whether particular logical or physical names exist in the catalog, as required by the LIGO application.

Another focus of our work in the last quarter has been improving the documentation of RLS, including producing the first FAQ for RLS.

### **Integration of Replica Location Service with POOL**

For the last several years, there has been resistance on the part of some in the particle physics community to using the Globus RLS implementation. One reason was that the Globus RLS implementation did not provide some of the functionality provided by the European Data Grid project's RLS, such as generating Globally Unique Identifiers (GUIDs) and providing many-to-one mappings between logical file names and GUIDs. In the current quarter, we implemented this additional functionality to allow integration of the Globus RLS into the POOL environment, which provides persistent object management for particle physics applications. As part of Globus RLS/POOL integration, we implemented a mapping table from logical names to GUIDs, and we used the existing POOL logic for GUID generation. Our implementation has passed all unit testing by the US CMS application group, and it is currently being tested with CMS application codes.

### **Implementation and Evaluation of a Peer-to-Peer Version of the RLS**

Continuing our work from the previous quarter, we further developed and evaluated a peer-to-peer version of the Replica Location Service. This system uses a Chord peer-to-peer structured overlay network to distribute RLS mappings among a collection of peer-to-peer Replica Location Index nodes. (The Globus toolkit implementation of the Local Replica Catalog is unchanged.) A paper describing this work was accepted for publication at the SC2004 conference in Pittsburgh in November 2004.

### **Implementation of an OGSi RLS Grid Service**

We continued development and performance evaluation of an RLS Grid Service that was based on a OGSi Grid service standard defined in the Global Grid Forum OGSA Data Replication Services (OREP) working group. In the current quarter, we updated the implementation based on the latest version of the Globus Toolkit Version 3. We performed additional performance evaluation. A paper describing this service was accepted for the Grid 2004 workshop to be held in Pittsburgh in November 2004.

### **Evaluation and Integration of Lightweight Data Replicator Functionality**

We have spent time this quarter doing a careful study and evaluation of the Lightweight Data Replicator tool from Scott Koranda's group at the University of Wisconsin at Milwaukee. LDR is developed as part of the LIGO project. Our hope is that some of all of the functionality of LDR can be incorporated into the Globus toolkit. Based on our evaluation, we have decided that initially we will incorporate the LDR scheduled data transfer/RLS registration functionality into the GT4 release as a technology preview component.

### **Examination of potential information inaccuracy**

Wrote a technical report, TR1517, University of Wisconsin-Madison, "Coping with BAD-Users." This TR describes how the batch-aware distributed file system uses detailed user-supplied information about the I/O characteristics of workloads. Potential problems resulting from inaccurate information are then discussed and possible algorithms and designs for coping with these problems are presented.

### **Developed a batch-aware distributed file system and scheduler simulation**

Continuing to explore the above problem concerning scheduling with inaccuracy, I developed a simulation of a wide-area batch scheduling system. This simulation is highly realistic with detailed disk, network, and buffer cache models. I have begun using this simulation to present a batch scheduler with inaccurate information to more precisely quantify the adverse affects in regards to throughput and correctness which can result from bad information.

### **Plans for Next Quarter**

A major portion of work on RLS, GridFTP, RFT, and XIO over the next quarter will center on performance testing, hardening, and improvements in ease of use in preparation for the GT4 final release early next year.

We plan to work on moving the Sloan Digital Sky Survey archive from Fermi Lab to Starlight via RFT. This archive contains in excess of 1 million files.

We continue to work on custom data storage interfaces to the High Performance Storage System (HPSS), the Storage Resource Broker (SRB), and the University of Wisconsin Network Storage (NeST) system. This will allow GridFTP clients to access these systems via a GridFTP server.

A Beta release of GT4.0 is planned for December.

We will continue work on the interfaces and implementation for the Lightweight Data Replicator (LDR) scheduled transfer/registration functionality

The current centralized implementation integrates with POOL using a single RLS Local Replica Catalog. In a realistic grid environment, we need to support a distributed RLS deployment. We are currently designing the distributed configuration of the RLS/POOL integration, and we will complete this implementation in the next quarter

### **Papers Published or in Progress**

July, 2004: completed paper on "A Peer-to-Peer Replica Location Service Based on a Distributed Hash Table" by Min Cai, Ann Chervenak, Martin Frank, to appear in Proceedings of SC2004 Conference, November 2004.

August, 2004: completed paper on "Implementation and Evaluation of a ReplicaSet Grid Service" by Mary Manohar, Ann Chervenak, Ben Clifford, Carl Kesselman, to appear in Proceedings of Grid2004 Workshop, November 2004.

August, 2004: submitted final version of "The Earth System Grid: Supporting the Next Generation of Climate Modeling Research" by David Bernholdt, Shishir Bharathi, David Brown, Kasidit Chanchio, Meili Chen, Ann Chervenak, Luca Cinghini, Bob Drach, Ian Foster, Peter Fox, Jose Garcia, Carl Kesselman, Rob Markel, Don Middleton, Veronika Nefedova, Line Pouchard, Arie Shoshani, Alex Sim, Gary Strand, and Dean Williams, to appear in Proceedings of the IEEE.

Zhang, H., K. Keahey, and B. Allcock, Providing Data Transfer with QoS as Agreement-Based Service. IEEE International Conference on Services Computing (SCC 2004), Shanghai, China, September, 2004

### **Presentations Given**

Reliable Data Transport: A Critical Service for the Grid during the GGF11 Grid Services Workshop

August 13, 2004: Led a meeting on Replica Management at Argonne National Laboratories in Chicago with approximately ten participants from five institutions. Presented key concepts for replica management as well as status and plans for the Replica Location Service.

August 23-24, 2004: Participated in EGEE meeting at University of Wisconsin at Madison and presented status and plans for Replica Location Service.