

Data Management: Key Concepts

DRAFT

Data Management: Key Concepts

Overview of Data Management in GT4

The Globus Toolkit provides a number of components for doing data management. A very high level overview is presented here and then detailed information is given for the individual components by following the component links.

The components available for data management fall into two basic categories: data movement and data replication.

DRAFT

Table of Contents

1. Data movement	1
1. GridFTP	1
2. Reliable File Transfer (RFT) Service	2
2. Data replication	3
1. Replica Location Service (RLS)	3
2. Using RLS: An Example	3
3. For more information	4
3. Higher level data services	5
1. Data Replication Service (DRS)	5
4. Key Concepts for RLS	6
1. RLS Overview	6
2. Some Terminology	6
3. Using the RLS: An Example	7
4. Components of the Replica Location Service	7
5. Configuration Options for RLS Deployment (not in key)	8
6. The RLS and Replica Consistency Services	9
7. For more information:	9
Glossary	10

List of Figures

2.1. Example of the associations between a logical file name and three replicas on different storage sites (RLS).	3
4.1. Example of the associations between a logical file name and three replicas on different storage sites (RLS).	7
4.2. Example deployment of a Replica Location Service (RLS).	8

DRAFT

Chapter 1. Data movement

There are two components related to data movement in the Globus Toolkit: the Globus GridFTP tools and the Globus Reliable File Transfer (RFT) service.

1. GridFTP

GridFTP is a protocol defined by Global Grid Forum Recommendation GFD.020, RFC 959, RFC 2228, RFC 2389, and a draft before the IETF FTP working group. The GridFTP protocol provides for the secure, robust, fast and efficient transfer of (especially bulk) data. The Globus Toolkit provides the most commonly used implementation of that protocol, though others do exist (primarily tied to proprietary internal systems).

The Globus Toolkit provides:

- a server implementation called `globus-gridftp-server`,
- a scriptable command line client called `globus-url-copy`, and
- a set of development libraries for custom clients.

While the Globus Toolkit does not provide an interactive client, the [GridFTP User's Guide](#) does provide information on at least one interactive client developed by other projects.

If you wish to make data available to others, you need to install a server on a host that can access that data and make sure that there is an appropriate Data Storage Interface (DSI) available for the storage system holding the data. This typically means a standard POSIX file system, but DSIs do exist for the Storage Resource Broker (SRB), the High Performance Storage System (HPSS), and NeST from the Condor team at the University of Wisconsin – Madison. A complete list of DSIs is available [here]. If you need an interface to a storage system not listed here, please contact us. While we certainly cannot offer to write DSIs for every storage system, we can assist in the development, or if a broad enough community can be identified that uses the system, we may be able to obtain joint funding to develop the necessary interface.

If you simply wish to access data that others have made available, you need a GridFTP client. The Globus Toolkit provides a client called `globus-url-copy` for this purpose. This client is capable of accessing data via a range of protocols (`http`, `https`, `ftp`, `gsiftp`, and `file`). As noted above this is not an interactive client, but a command line interface, suitable for scripting. For example, the following command:

```
globus-url-copy gsiftp://remote.host.edu/path/to/file file:///path/on/local/host
```

would transfer a file from a remote host to the locally accessible path specified in the second URL.

Finally, if you wish to add access to files stored behind GridFTP servers, or you need custom client functionality, you can use our very powerful client library to develop custom client functionality.

For more information about GridFTP, see:

- the [documentation](#).
- [The Globus Striped GridFTP Framework and Server](#)¹

¹ http://www.globus.org/alliance/publications/papers/gridftp_final.pdf

2. Reliable File Transfer (RFT) Service

While globus-url-copy and GridFTP in general are a very powerful set of tools, there are characteristics which may not always be optimal. First, the GridFTP protocol is not a web service protocol (it does not employ SOAP, WSDL, etc). Second, GridFTP requires that the client maintain an open socket connection to the server throughout the transfer. For long transfers this may not be convenient, such as if running from your laptop. While globus-url-copy uses the robustness features of GridFTP to recover from remote failures (network outages, server failures, etc), a failure of the client or the client's host means that recovery is not possible since the information needed for recovery is held in the client's memory. What is needed to address these issues is a service interface based on web services protocols that persists the transfer state in reliable storage. We provide such a service and call it the Reliable File Transfer (RFT) service.

RFT is a Web Services Resource Framework (WSRF) compliant web service that provides "job scheduler"-like functionality for data movement. You simply provide a list of source and destination URLs (including directories or file globs) and then the service writes your job description into a database and then moves the files on your behalf. Once the service has taken your job request, interactions with it are similar to any job scheduler. Service methods are provided for querying the transfer status, or you may use standard WSRF tools (also provided in the Globus Toolkit) to subscribe for notifications of state change events. We provide the service implementation which is installed in a web services container (like all web services) and a very simple client. There are Java classes available for custom development, but due to lack of time and resources, work is still needed to make this easier.

For more information about RFT, see the [documentation](#).

Chapter 2. Data replication

The Replica Location Service (RLS) is one component of data management services for Grid environments. RLS is a tool that provides the ability keep track of one or more copies, or replicas, of files in a Grid environment. This tool, which is included in the Globus Toolkit, is especially helpful for users or applications that need to find where existing files are located in the Grid.

1. Replica Location Service (RLS)

RLS is a simple registry that keeps track of where replicas exist on physical storage systems. Users or services register files in RLS when the files are created. Later, users query RLS servers to find these replicas.

RLS is a distributed registry, meaning that it may consist of multiple servers at different sites. By distributing the RLS registry, we are able to increase the overall scale of the system and store more mappings than would be possible in a single, centralized catalog. We also avoid creating a single point of failure in the Grid data management system. If desired, RLS can also be deployed as a single, centralized server.

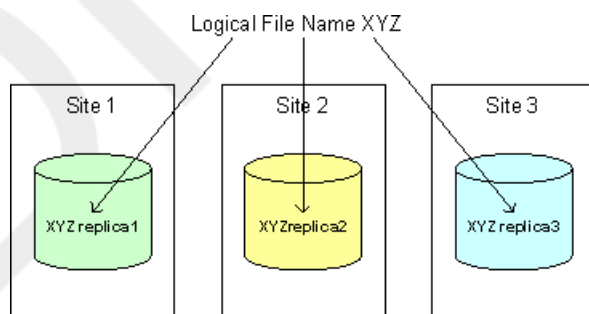
Before explaining RLS in detail, we need to define a few terms.

- A *logical file name* is a unique identifier for the contents of a file.
- A *physical file name* is the location of a copy of the file on a storage system.

These terms are illustrated in Figure 1 (below). The job of RLS is to maintain associations, or mappings, between logical file names and one or more physical file names of replicas. A user can provide a logical file name to an RLS server and ask for all the registered physical file names of replicas. The user can also query an RLS server to find the logical file name associated with a particular physical file location.

In addition, RLS allows users to associate attributes or descriptive information (such as size or checksum) with logical or physical file names that are registered in the catalog. Users can also query RLS based on these attributes.

Figure 2.1. Example of the associations between a logical file name and three replicas on different storage sites (RLS).



2. Using RLS: An Example

One example of a system that uses RLS as part of its data management infrastructure is the Laser Interferometer Gravitational Wave Observatory (LIGO) project. LIGO scientists have instruments at two sites that are designed to detect the existence of gravitational waves. During a run of scientific experiments each LIGO instrument site produces millions of data files. Scientists at eight other sites want to copy these large data sets to their local storage systems so that they can run scientific analysis on the data. Therefore, each LIGO data file may be replicated at up to ten physical

locations in the Grid. LIGO deploys RLS servers at each site to register local mappings and to collect information about mappings at other LIGO sites. To find a copy of a data file a scientist requests the file from LIGO's data management system, called the Lightweight Data Replicator (LDR). LDR queries the Replica Location Service to find out whether there is a local copy of the file; if not, RLS tells the data management system where the file exists in the Grid. Then the LDR system generates a request to copy the file to the local storage system and registers the new copy in the local RLS server.

LIGO currently uses the Replica Location Service in its production data management environment. The system registers mappings between more than 3 million logical file names and 30 million physical file locations.

3. For more information

For more detailed information about RLS, click [here](#).

For more information about RLS, see the [documentation](#).

DRAFT

Chapter 3. Higher level data services

GT 4.2.0 also provides a higher-level data management service that combines two existing data management components: RFT and RLS.

1. Data Replication Service (DRS)

For the Technical Preview of the Globus Toolkit 4.2.0 release we have designed and implemented a Data Replication Service (DRS) that provides a pull-based replication capability for Grid files. The DRS is a higher-level data management service that is built on top of two GT data management components: the Reliable File Transfer (RFT) Service and the Replica Location Service (RLS).

The function of the DRS is to ensure that a specified set of files exists on a storage site. The DRS begins by querying RLS to discover where the desired files exist in the Grid. After the files are located, the DRS creates a transfer request that is executed by RFT. After the transfers are completed, DRS registers the new replicas with RLS.

DRS is implemented as a Web service and complies with the Web Services Resource Framework (WSRF) specifications. When a DRS request is received, it creates a WS-Resource that is used to maintain state about each file being replicated, including which operations on the file have succeeded or failed.

1.1. For more information

For more information about DRS, go to [Data Replication Service \(DRS\)](#).

Chapter 4. Data Management: Key Concepts for RLS

The Replica Location Service (RLS) is a tool that provides the ability to keep track of one or more copies, or replicas, of files in a Grid environment. This tool, which is included in the Globus Toolkit, is especially helpful for users or applications that need to find where existing files are located in the Grid.

1. RLS Overview

The Replica Location Service is one component of data management services for Grid environments. It is a simple registry that keeps track of where replicas exist on physical storage systems. Users or services register files in the RLS when the files are created. Later, users query RLS servers to find these replicas.

The RLS is a distributed registry, meaning that it may consist of multiple servers at different sites. By distributing the RLS registry we are able to increase the overall scale of the system and store more mappings than would be possible in a single, centralized catalog. We also avoid creating a single point of failure in the Grid data management system. If desired, the RLS can also be deployed as a single, centralized server.

In addition, we are developing a higher-level WS-RF Data Replicator Service that uses the Globus Reliable File Transfer Service to copy files and then registers the new replicas in the RLS. This service is available in the Globus Toolkit Version 4.2.0 release as a Technical Preview component (for more information, click [here](#)).

2. Some Terminology

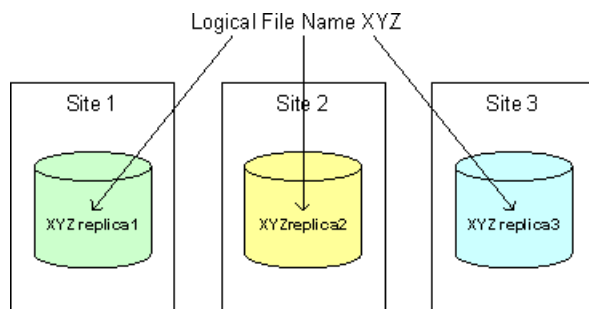
Before explaining the RLS in detail, we need to define a few terms.

- A *logical file name* is a unique identifier for the contents of a file.
- A *physical file name* is the location of a copy of the file on a storage system.

These terms are illustrated in Figure 1 (below). The job of the RLS is to maintain associations, or mappings, between logical file names and one or more physical file names of replicas. A user can provide a logical file name to an RLS server and ask for all the registered physical file names of replicas. The user can also query an RLS server to find the logical file name associated with a particular physical file location.

In addition, the RLS allows users to associate attributes or descriptive information (such as size or checksum) with logical or physical file names that are registered in the catalog. Users can also query the RLS based on these attributes.

Figure 4.1. Example of the associations between a logical file name and three replicas on different storage sites (RLS).



3. Using the RLS: An Example

One example of a system that uses the RLS as part of its data management infrastructure is the Laser Interferometer Gravitational Wave Observatory (LIGO) project. LIGO scientists have instruments at two sites that are designed to detect the existence of gravitational waves. During a run of scientific experiments, each LIGO instrument site produces millions of data files. Scientists at eight other sites want to copy these large data sets to their local storage systems so that they can run scientific analysis on the data. Therefore, each LIGO data file may be replicated at up to ten physical locations in the Grid. LIGO deploys RLS servers at each site to register local mappings and to collect information about mappings at other LIGO sites. To find a copy of a data file, a scientist requests the file from LIGO's data management system, called the Lightweight Data Replicator (LDR). LDR queries the Replica Location Service to find out whether there is a local copy of the file; if not, the RLS tells the data management system where the file exists in the Grid. Then the LDR system generates a request to copy the file to the local storage system and registers the new copy in the local RLS server.

LIGO currently uses the Replica Location Service in its production data management environment. The system registers mappings between more than 3 million logical file names and 30 million physical file locations.

4. Components of the Replica Location Service

The RLS design consists of two types of servers: the Local Replica Catalog and the Replica Location Index.

The Local Replica Catalog (LRC) stores mappings between logical names for data items and the physical locations of replicas of those items. Clients query the LRC to discover replicas associated with a logical name. The simplest RLS deployment consists of a single LRC that acts as a registry of mappings for one or more storage systems. Typically, when an RLS is deployed on a site, an administrator populates it to reflect the contents of a local file or storage system. If new data files are produced by a workflow manager or a data publishing service, these services typically register newly created files with the RLS as part of their publication process. Our performance studies for an LRC deployed on a moderately powerful workstation with a MySQL relational database back end show that the catalog can sustain rates of approximately 600 updates and 2,000 queries per second.

For a distributed RLS deployment, we also provide a higher-level Replica Location Index (RLI) server. Each RLI server collects information about the logical name mappings stored in one or more LRCs. An RLI also answers queries about those mappings. When a client wants to discover replicas that may exist at multiple locations, the client will pose that query to an RLI server rather than to an individual Local Replica Catalog. In response to a query, the RLI will return a list of all the LRCs it is aware of that contain mappings for the logical name contained in the query. The client then queries these LRCs to find the physical locations of replicas.

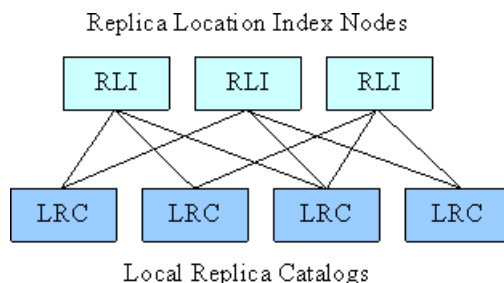
Figure 4.2. Example deployment of a Replica Location Service (RLS).

Figure 2 illustrates a distributed RLS deployment. Information is sent from the LRCs to the RLIs using soft-state update protocols. Each LRC periodically sends information about its logical name mappings to a set of RLIs. The RLIs collect this mapping information and respond to queries regarding the mappings. Information in RLIs times out and gets periodically refreshed by subsequent updates. An advantage of using such soft-state update protocols is that if an RLI fails and later resumes operation, its contents can be reconstructed using these updates.

Because each LRC may hold millions of logical file name mappings, updates from LRCs to RLIs can become large. Sending them across the network may be slow, especially in the wide area; when updates arrive at an RLI, they may consume considerable storage space there. One option for making these updates more efficient is to compress their contents. Various compression strategies are available. The one that we chose to implement for the RLS is based on Bloom filter compression. Each Local Replica Catalog periodically creates a bit map that summarizes its contents by applying a series of hash functions to each logical name registered in the LRC and setting the corresponding bits in the bit map.

5. Configuration Options for RLS Deployment (not in key)

We can determine the performance and reliability of a distributed RLS by the number of Local Replica Catalog and Replica Location Index servers we deploy and the way we configure updates among LRCs and RLIs. For example, we can improve reliability by configuring the system so that every LRC updates multiple RLI indexes.

In practice, current deployments of RLS systems are relatively small. For example, the LIGO physics collaboration deploys RLS servers at ten sites. For deployments of this scale the Replica Location Service is often deployed in a fully connected configuration, with a Local Replica Catalog and a Replica Location Index server deployed at each site, and all LRCs sending updates to all RLIs. Such a configuration has the advantage that every site has a complete picture of the replicas in the distributed system.

For larger deployments, however, this configuration is unlikely to scale well because it requires a large number of updates to be sent among servers and stored at each RLI. For example, in a system with hundreds of RLS sites, a fully connected deployment would require every site to send and receive hundreds of updates. In such large deployments it is likely that the system would be partitioned so that each RLI would receive updates from only a subset of the LRCs.

Ideally, the management of a distributed RLS should be automated and self-configuring, so that LRCs and RLIs discover one another and updates among them are redistributed automatically to balance the load when servers join or leave the system. While we are studying approaches for automated membership management, including peer-to-peer self-organization of RLS servers, the current RLS implementation uses a simple static configuration of LRC and RLI servers. An administrator must manually change the pattern of updates among servers.

6. The RLS and Replica Consistency Services

RLS is a fairly simple tool that provides registration and discovery of files. It is intended to be just one part of an overall system for managing data in Grids. Those considering whether to use an RLS should understand what functionality the system does and does not provide.

One of the most common assumptions new users make about RLS is that, when a new replica is registered in the RLS, the system checks to make sure that the registered entry is a valid replica of an existing file. Actually, the RLS does not perform such correctness or consistency checks on new entries. The RLS allows users to register mappings between logical names and physical locations without any verification that the physical files that are registered as replicas are actually copies of one another. Likewise, if registered replicas are modified so that they are no longer valid copies of one another, the RLS will not detect these changes or take action. Instead, the RLS relies on higher-level data management or consistency services to perform these functions.

There are several reasons for this design choice. One is the simplicity and efficiency of providing a registry that is not required to perform consistency checks on the registered files, which can be time-consuming. For example, a service that guarantees consistency might calculate checksums for all replicas and verify that they match. Providing consistency guarantees therefore creates additional overhead for the system that can limit its performance.

In practice, providing a high level of consistency checking during replica registration is not required for many applications because often only a small set of highly trusted users is allowed to publish data. These privileged users do not need to verify that registered files are actually replicas because the users control the entire publishing process. They typically need to publish a large number of files quickly and cannot tolerate extra overheads associated with performing consistency operations such as checksum calculations on every file registration.

Another reason we do not enforce replica consistency in the RLS is that users may have different definitions of what constitutes a “valid” replica. For some users, replicas are exact byte-for-byte copies of one another. For others, two files may be considered replicas if they are different versions of the same file or if the files contain the same content but in different formats, for example, compressed and uncompressed versions of the same data. We choose not to prescribe a particular definition for replicas in the RLS.

For these reasons we leave consistency-checking operations to higher-level services, which may perform these checks before or after replicas are registered in the RLS. If these services find that registered replicas are invalid, they will remove the replica mappings from the RLS.

7. For more information:

For more information about RLS, use the following links:

- [Replica Location Service \(RLS\)](#).
- “[Performance and Scalability of a Replica Location Service](#)¹,” Ann L. Chervenak, Naveen Palavalli, Shishir Bharathi, Carl Kesselman, and Robert Schwartzkopf, High Performance Distributed Computing (HPDC-13) Conference, June 2004.

¹ <http://www.globus.org/alliance/publications/papers/chervenakhpd13.pdf>

Glossary

L

logical file name A unique identifier for the contents of a file.

P

physical file name The address or the location of a copy of a file on a storage system.

DRAFT