

Globus Data Services

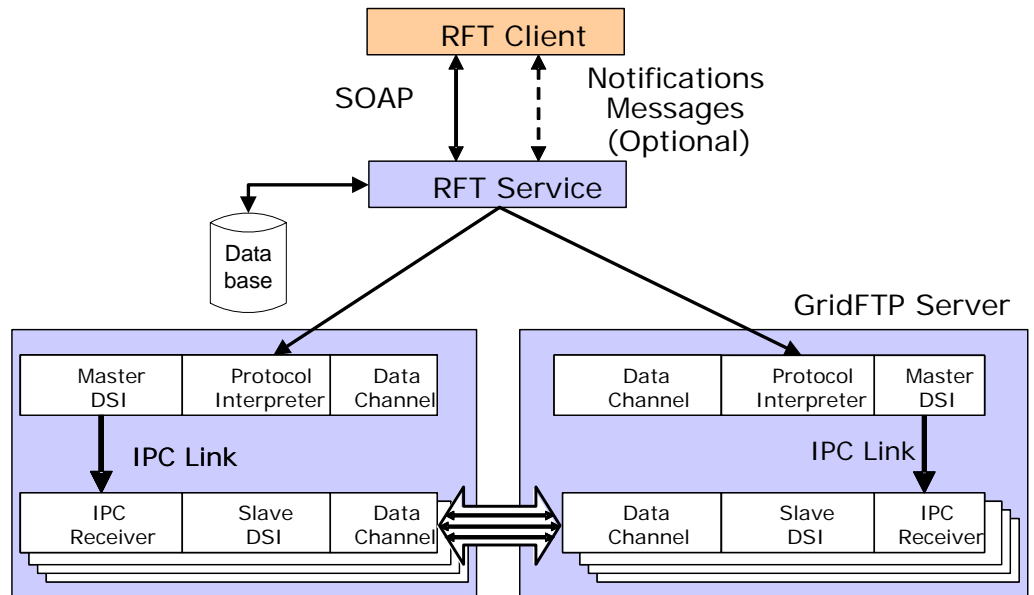
The Globus Toolkit includes a variety of tools used for data management, including data transport, reliable multi-file transfers, replication management and database access and integration.

Data Transport via the Globus GridFTP Family of Tools

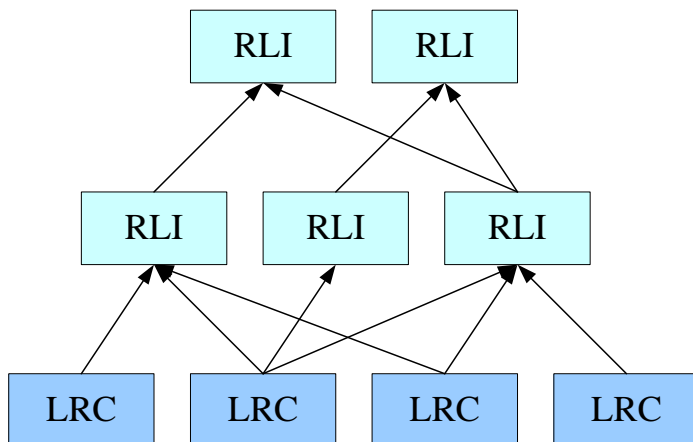
The GridFTP data transport protocol is a widely used standard for secure, efficient, and robust transportation of data. The Globus toolkit provides a reference implementation along with several related tools. At the heart is the Globus GridFTP server. It is built over the eXtensible IO (XIO) system, allowing a range of transport protocols (TCP, UDP, UDT, RBUDP, etc). We provide parallelism (multiple TCP connections between hosts) to work around TCP's well known limitations for high bandwidth, high latency transfers. As shown here, we also allow

striping, which allows multiple hosts on a parallel file system to efficiently utilize even 10 Gigabit networks. We provide two methods for invoking transfers. The first is a command line scriptable client called globus-url-copy. It has the benefits of the GridFTP restart mechanisms, but if it fails all state is lost, since it is held in memory.

To provide additional reliability, and to provide a service type interface (similar to a job scheduler), we provide the Reliable File Transfer (RFT) service. Much like a job submission, the client provides a description of the transfers it wants to take place, and the service runs the transfers on the client's behalf. Since the request and the restart markers are written to a database, transfers can be resumed with no more than 5 seconds of lost transmission should the RFT service fail.



Replica Location Service (RLS)



The Replica Location Service is a distributed 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. There are two types of RLS servers: a Local Replica Catalog (LRC) that contains mappings from logical names to physical locations, and a Replica Location Index (RLI) that aggregates state information about the contents of LRCs. RLS is used in a number of production Grid deployments. For example, the LIGO project currently deploys RLS servers on ten sites. The system registers mappings between more than 11 million logical file names and 120 million physical locations. RLS servers scale to support millions of replica mappings and up to 100 simultaneous clients.

Data Replication Service (DRS)

The function of the Globus Data Replication Service (DRS) is to replicate a specified set of files onto a local storage system and register the new files in catalogs. DRS builds on lower-level Grid data services, including the Globus Reliable File Transfer (RFT) service and Replica Location Service (RLS). DRS discovers where desired data files exist on the Grid by querying the Replica Location Service; transfers these files to the target storage system using the Reliable File Transfer Service; and registers these new files in the RLS so that other sites may discover them. Throughout DRS replication operations, the service maintains state about each file, including which operations on the file have succeeded or failed. DRS is implemented as a Web service compliant with the Web Services Resource Framework (WS-RF) specifications and is available in the Globus Toolkit Version 4.0.2 release.

Data Access and Integration Service (DAI)

OGSA-DAI provides a pure Java data service framework for accessing and integrating data resources, such as files, relational and XML databases, on to Grids. Towards this end, OGSA-DAI exposes intrinsic data resource capabilities – such as the ability to perform SQL queries on relational resources or evaluate XPath statements on XML collections – through web service-based interfaces, thus allowing data resources to be easily incorporated as first class citizens in Grids. OGSA-DAI also allows additional functionality to be implemented at the service – such as transformation of data coming out of a data resource – so as to avoid unnecessary data movement. In addition, OGSA-DAI provides a compact way of handling multiple potential interactions with a service within a single request via an XML document, called a perform document, where data is pipelined between different sets of activities that operate on a data stream coming out of, or going into, a data resource.

OGSA-DAI allows developers to easily add or extend the functionality supported at the service layer without requiring changes to the framework itself. OGSA-DAI also facilitates the provision of data integration capabilities, via OGSA-DAI services, from various sources to obtain the required information.

