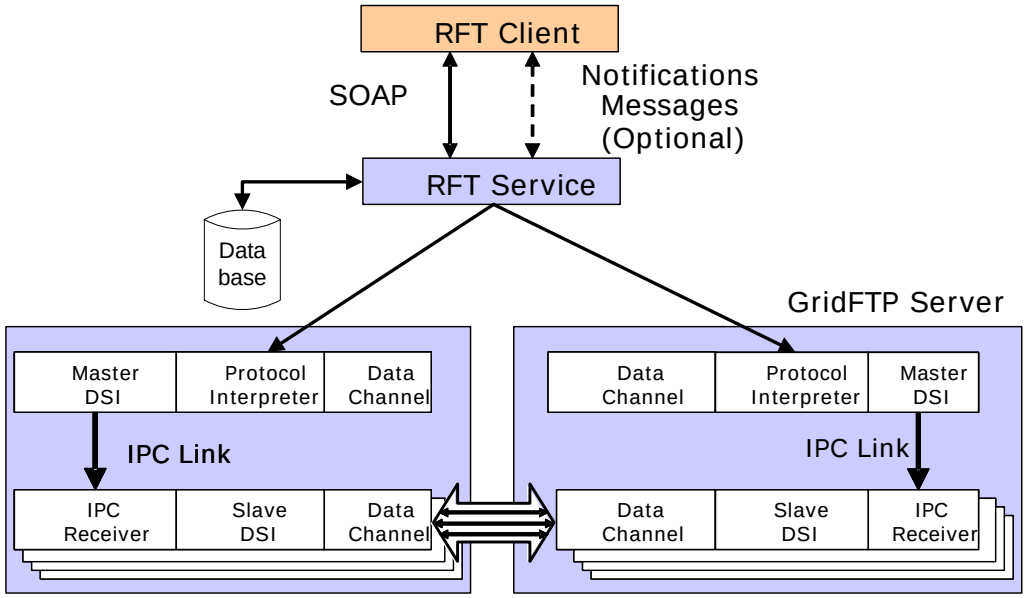
the globus alliance
www.globus.org

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.

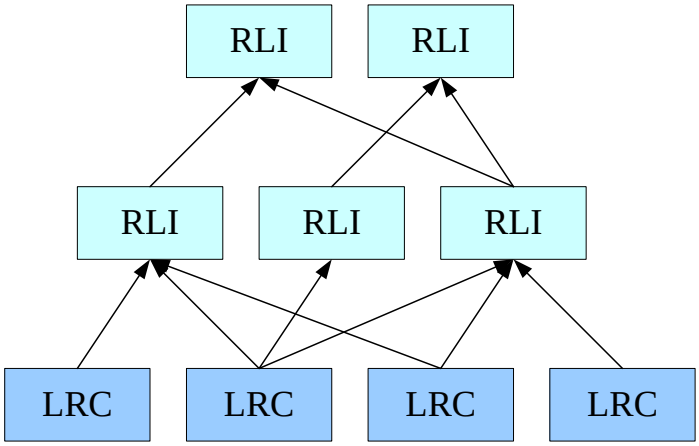
Data Transport via the Globus GridFTP Family of Tools

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 enables 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**, which has the benefits of the GridFTP restart mechanisms, but if this command fails, then all state is lost, since it is held in memory. The second method is the **Reliable File Transfer** (RFT) service, which offers additional reliability and provides a service-type interface (similar to a job scheduler). Much like a job submission, the client provides a description of the transfers it wants to take place, and the RFT 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



The Replica Location Service (RLS) 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.

the globus alliance
www.globus.org

Data Replication Service

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 service and Replica Location Service. By querying the Replica Location Service, DRS discovers where desired data files exist on the Grid. DRS then transfers these files to the target storage system using the Reliable File Transfer Service. Finally, DRS registers the new files in the RLS so that clients 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 specifications and is available in the Globus Toolkit Version 4.0.2 release.

Data Access and Integration Service

The Open Grid Services Architecture – Data Access and Integration (OGSA-DAI) provides a pure Java data service framework for accessing and integrating data resources, such as files or relational and XML databases, onto Grids. To 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. To avoid unnecessary data movement, OGSA-DAI also allows additional functionality to be implemented at the service, such as transformation of data coming out of a data resource. 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 enables 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.