WG Internal
OGSA-DMI

File Transfer Service Use-case

Gavin McCance, Paolo Badino
Version 1: 6 November 2006

Status of This Document

This document provides information to the Grid community regarding use cases for data transfer in the EGEE project. It does not define any standards or technical recommendations. Distribution is unlimited.

1. Abstract

This document provides information about the EGEE data transfer use-cases. It described briefly the EGEE model and VO topology. It then discusses the both the user-level and the service management-level use-cases.

Contents

2. EGEE and WLCG projects

The EGEE project, (EGEE Website 2006), is an EU funded project to develop a large scale eScience infrastructure for Europe. This project, now in its second phase, also has extensive links beyond Europe. One of its key goals is to develop a generic middleware stack, known as gLite (gLite Website 2006), to provide a high level of infrastructure for Virtual Organisations (VOs) using the grid to tackle their problems.

One of the key application areas for EGEE is high energy physics, with a particular orientation towards the Large Hadron Collider (LHC) experiment, currently under construction at CERN and due to begin data taking in 2007. The project to provide the computing infrastructure for the LHC is the Worldwide LHC Computing Grid (WLCG), which is a very important part of the EGEE project. The WLCG project also uses resources from the Open Science Grid (OSG Website 2006) in the United States and NorduGrid (NorduGrid Website 2006) in Northern Europe.

Reliable movement and storage of data is a cornerstone of distributed systems. For data grids, where the volumes of data to be moved are huge, data management must offer a high degree of control, to ensure that data is placed correctly for processing and safe keeping; but also offer a view at a sufficiently high level to ensure that data placement can be managed efficaciously. Robustness of data storage and replication is also paramount – the data management system must be able to cope with as many errors itself as possible, saving valuable human time for those events where intervention really is necessary.

2.1 Data management middleware components

The EGEE and WLCG grids rely on a wide variety of mass storage systems. However, all systems export the standard Storage Resource Manager (SRM) interface. This allows users to control the creation and copying of files into and out of the storage systems in a standardised way. The systems all support GridFTP as the de-facto standard WAN transfer protocol.

The EGEE middleware provides a number of client tools to work with files in the storage system, providing Posix access to the files and functions to copy the files between storage systems (and to update the associated file catalogs).

The middleware also provides an asynchronous file transfer service (FTS) for managing bulk file transfer requests between sites.

2.2 VO topology

We distinguish two type of VO: large VO and small VO. In the EGEE model, large VOs consist of many sites grouped into a number of clusters. Each cluster consists of one Regional Operations Centre whose purpose is to support the other sites within the cluster. The clusters tend to be defined geographically (or nationally), often with a large national computing centre playing the role of ROC for the cluster.

An example of a large VO is from High Enery Physics: the Worldwide LHC computing grid (WLCG), where we define multiple (order of 10) tier-1 centres (usually national physics labs, often the same site as that running the EGEE ROC), each looking after multiple tier-2 centres (for example universities, typically in the same country as the tier-1 centre).

The grid middleware services that do not need to be co-located with the resources (for example, job brokers and information system aggregators) tend to be run as managed services by the larger tier-1 centres. The tier-1 centres also have the responsibility for the safekeeping of the

physics experiment data (typically on mass storage tape systems) and much of its initial physics processing. For a data-grid of this sort, the data volumes and data transfer rates tend to follow the tiered model: we expect much more of tier-1 sites than of tier-2 sites, in terms of volume, performance and availability.

Small VOs are defined as VOs with a few sites, possibly (though not necessarily) supported by a single ROC. The data volumes and rates are less and do not generally follow any specific tiered structure.

3. Use-cases

The use-cases assume the following:

> The fundamental unit of transfer is always the file. VOs often prefer to manage datasets rather than files (a dataset consists of multiple files). However, the current EGEE grid middleware expects to be given files (i.e. it expects the production frameworks of the VOs to break to datasets into individual files for processing).

Before describing the transfer use-cases, we describe the basics of how files are created and accessed in the EGEE middleware.

3.1 Data generation

Analysis and production jobs write their results in to the local storage system. This may either be done directly (most of the SRMs export a Posix-like interface) or after the file is complete (upload into the local SRM). Once the file is committed to the mass storage system it is considered read-only.

Data is typically not written into remote SRMs, but after being written into the local SRM, it may be replicated to a different one using the file transfer tools.

3.2 Data access

Analysis and production job need to access files for processing. These are almost always accessed from the job via the local SRM (typically by a Posix-like interface). The grid middleware and VO production frameworks ensure that the correct data is on the given SRM before the job is sent to that site.

An analysis job may request other data (i.e. data it did not know that it needed before the job started) in which case it will be fetched by the file transfer tools and made available on the local SRM. For efficiency reasons, however, it is preferred to avoid this.

3.3 Internal site transfers

We do not perform internal site transfers using any standard middleware. Transfers of this sort may be necessary for operational reasons, but it is expected that the details will not be exposed to the Grid and will be controlled using mechanisms provided by the underlying storage layer.

3.4 Transfer use-cases

All use-cases come down to the same fundamental one: we want files transferred reliably between two sites. We define different views of transfer use-case.

The **end-user** wants to transfer files between storage systems on different sites reliably and as quickly as possible. They are interested in *their* files.

The **VO production manager** wants the same as the end-user, but additionally they need to be able to control more (prioritising some production datasets over others). They need more monitoring, so they can ensure that Service Level Agreements (SLA) are being met and so that they can address any problems that may occur or take some other action (e.g. redirect a stream to a different site if there is a problem). They are interested in *all* files for their VO.

The **site manger** wants to ensure that the site's resources are used sensibly and to their maximum. They are interested in policy (e.g. the relative share promised in the SLA to each VO) and interested in management (such that they can control what transfers are using what resources). The also require monitoring, so they can spot quickly any problems on their site that may be degrading the service they are providing to their VOs.

3.5 Production use-case 1: production export

The requirement is to transfer a known set of files from one storage system to several others, as quickly as possible. The required transfer rates are defined by SLAs and the file flux is fairly uniform in time (or at least well known). The files are copied directly from the online system into a local buffer on the local mass storage system and made available for immediate export. The local buffer is limited, so we require a very high level of service availability.

An example is the export of data from the CERN lab to the tier-1 sites of WLCG. A defined share of each experiment's RAW data is delivered to each tier-1 site for archiving and reprocessing using dedicated private network links.

3.6 Production use-case 2: file upload from tier-2 to tier-1

The tier-2 sites are producing files that need to be sent to the associated tier-1 site for archiving and analysis. The required volumes are defined by SLAs and the flux is fairly uniform in time. The files are typically written directly to the local tier-2 storage system and subsequently marked for upload by the VO production software framework. The service availability requirements are lower.

An example is the production of Monte Carlo physics data at tier-2 sites.

3.7 Production use-case 3: data broadcast

The requirement is to copy the same dataset (from a single source) to many sites. Typically this is a smaller sized dataset (for example a calibration or run conditions dataset). The production often depends on having the latest calibration dataset available, so there is usually a requirement to deliver it to all sites as quickly as possible. The data flux is irregular with a very high burst rate.

3.8 Production use-case 4: data-set collection: tier-1 to tier-1

Due to the large volume of data comprising the RAW data, it may need to be processed over multiple sites – consequently, the derived data will be spread over multiple sites. It is then

desirable to collect all data products belonging to the same dataset together. Typically, this will take place between tier-1 sites in WLCG. The data flux is likely to be quite bursty in time with high burst rates.

3.9 Analysis use-case 1: data placement for analysis

The VO production systems will attempt to place the analysis input data as sensibly as possible given the VO computing model and the geographical spread of the users. As much as possible will be planned as part of the production system, but there will always be cases where explicit datasets need to be replicated in response to user needs. The data volumes involved are typically less than production, but the requests are much more chaotic.

3.10 Transfer management use-cases

In addition to the user-level use-cases, we have a large number of requirements from the other users of the system (the production managers and site managers).

3.11 VO production manager use-case 1: control jobs in my VO

A VO production manager needs to be able to control the relative priority of jobs in his VO, as well as cancel jobs from users in his VO. The example case is a high-priority production set of 'exciting new data' that must be delivered as soon as possible to eager physicists.

3.12 VO production manager use-case 2: verify the delivered service level

A VO production manager requires a reasonable VO view of what is happening in order to verify the agreed SLA. What rates are being delivered by what sites? What is the current failure rate to a given site?

3.13 VO production manager use-case 3: feedback from monitoring system

The VO production manager may wish to have feedback from the file transfer to make decisions about what data goes where. In the case of a (fast-filling) data export buffer, it may be desirable to redirect a data stream to another site if the site originally scheduled to receive it is having problems.

3.14 Site manager use-case 1: selective control of transfers

Being able to control different streams of transfers is critical to stable operations. For example, running transfers to different sites across different network links may require distinct TCP tuning parameters and will certainly require different number of concurrently transferring files to obtain the optimal use of the different network links.

Being able to shut down a given stream while maintaining the others is useful is one of the sites your transfer to or from is undergoing scheduled (or unscheduled) maintenance. The jobs destined for the given site can be queued, and will resume when the site comes out of maintenance.

3.15 Site manager use-case 1: implementation of VO policy

The site manager should be able to balance the available network resources between the VOs that his site supports according the agreed SLAs. Policies should also be controllable per VO, allowing the site manager to apply, e.g. different retry strategies for different VOs.

3.16 Site manager use-case 2: monitoring (reactive and historical)

The transfer service knows about the status of every file transfer request going through it, including details of any failures at the network or storage layer. The site manager (both at the source and the destination of the transfer) can use this information to augment the existing monitoring provided their storage systems to discover problems as early as possible.

The historical monitoring information is also useful since it is typically a direct measurement of the metrics specified in the SLA; the site manager can use it to measure his site's compliance.

4. Intellectual Property Statement

5. Disclaimer

6. Full Copyright Notice