LHCb DATA REPLICATION DURING SC3

A. Smith*†, University of Edinburgh, Edinburgh, Scotland
A. Tsaregorodtsev, Centre de Physique des Particules de Marseille, Marseille, France

Abstract

LHCb's participation in LCG's Service Challenge 3 involves testing the bulk data transfer infrastructure developed to allow high bandwidth distribution of data across the grid in accordance with the Computing Model. To enable reliable bulk replication of data, LHCb's DIRAC system has been integrated with gLite's File Transfer Service middleware component to make use of dedicated network links between LHCb computing centres. DIRAC's Data Management tools previously allowed the replication, registration and deletion of files on the grid. For SC3 supplementary functionality has been added to allow bulk replication of data (using FTS) and efficient mass registration to the LFC replica catalog.

Provisional performance results have shown that the system developed can meet the expected data replication rate required by the computing Model in 2007. This paper details the experience and results of integration and utilisation of DIRAC with the SC3 transfer machinery.

INTRODUCTION TO DIRAC DATA MANAGEMENT ARCHITECTURE

The DIRAC architecture (as discussed in detail in [1]) is split into three main component types: **Services**, **Resources** and **Agents**. *Services* provide the various independent functionalities of DIRAC and are deployed and administered centrally on machines accessible by all other DIRAC components. *Resources* refer to GRID compute and storage resources at remote sites. *Agents* are lightweight software components who can request jobs from the central Services to carry out a specific purpose. The DIRAC Data Management System (DMS) is made up an assortment of these components as shown in Fig. 1.

The main components of the DIRAC DMS are: **Storage Element**, **Replica Manager** and **File Catalogs**. The *Storage Element* is an abstraction of Storage Resources on the GRID while the actual access is provided by specific plug-in modules for different underlying storage implementations. These plug-ins provide access to storage through any of the available transport protocols e.g. srm, gridftp, bbftp, sftp, http etc. The plug-in modules provide the ability to perform data management operations on the physical storage such as namespace management (creation and deletion of directories), uploading of files to the SE, file download from SE, deletion of files etc.

* a.smith@cern.ch
† Marie Curie Fellow at CERN, Geneva, Switzerland

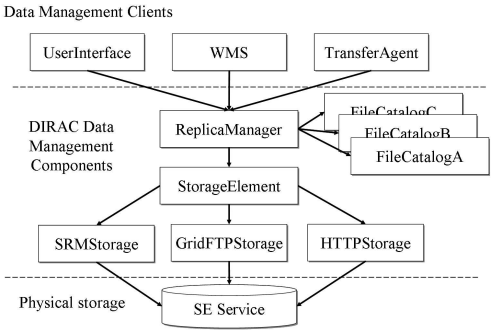

Figure 1: Schematic of DIRAC Data Management Architecture.

In Fig. 1 it can be seen that the Replica Manager can access a variety of available *File Catalogs*. A standard client API is exposed by the available catalogs to allow underlying operations to be performed in a transparent way. This allows the Replica Manager to access a variety of available catalogs and offers a certain level of redundancy of the information stored.

The *Replica Manager* is the point of contact for users of the DMS and provides an API for the available operations. These include uploading files from local cache to a GRID SE, copying files from GRID SEs to local system, file catalog registration, replication of files between GRID SEs and file removal. This API also allows users performing data management operations to be removed direct interaction with the File Catalogs or the physical Storage Elements.

LHCB TRANSFER AIMS DURING SC3

The structure of LCG's Service Challenge 3 (SC3)[2] can roughly be broken into two main sections: the Service Phase and the extended Service Phase. The goal of this initial Service Phase is to provide a platform of machinery for file transfers with proven reliability and quality of service. This platform can then be used by experiments to test their specific software and validate their computing Models in the extended phase.

LHCb's Data Replication goals during SC3 were determined from real use cases stated in LHCb's Computing TDR[3] to mimic eventual requirements. SC3 goals were set at roughly 40% of the 2007 needs and can be summarised as:

- Replication 1TB of stripped DST data from CERN to all Tier-1s.

- Replication of 8 TB of digitised data from CERN/Tier-0 to LHCb participating Tier1 centers in parallel.
- Removal of 50k replicas (via LFN) from all Tier-1 centres.
- Moving 4TB of data from Tier1 centres to Tier0 and to other participating Tier1 centers.

INTEGRATION OF DIRAC WITH FTS

The data replication tools deployed during SC3 utilised gLite's File Transfer Service (FTS)[4][5] to perform reliable bulk file transfers. FTS is the lowest-level data movement service defined in the gLite architecture and provides point-to-point movement of physical files (SURLs) between SRM compliant storage elements. The FTS takes a set of source-destination SURL pairs and assigns these file transfers to dedicated transfer channels linking two defined SRM endpoints. These dedicated channels are designed to take advantage of the high-bandwidth networking made available between CERN and LCG Tier1 sites. Routing of transfers is not provided by the FTS requiring a higher level service to resolve SURLs and determine to which sites, and therefore on which channel, a file should be transferred. The DIRAC DMS was employed to do this to allow integration with FTS. A schematic of the integration can be seen in Fig. 2.

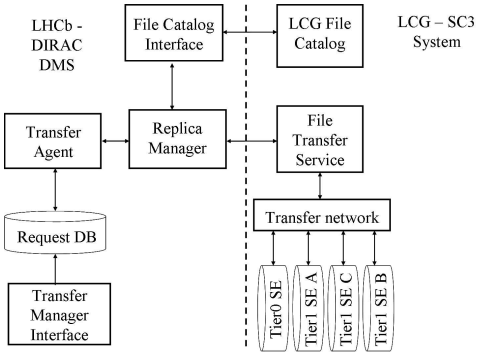

Figure 2: Schematic of Integration of DIRAC with FTS.

To allow DIRAC DMS to use FTS new methods were developed within the Replica Manager and additional functionality added to the Transfer Agent to deal with bulk operations. Previously, data management operations within the DMS were performed in a blocking fashion for individual files. The FTS allows bulk asynchronous transfer of files therefore new functionality was required to store monitoring information on the progress of each FTS job within the Request DB.

OPERATION OF DIRAC BULK TRANSFER MECHANICS

The interaction of the DIRAC DMS with the SC3 machinery can be seen in Fig. 2. This system was deployed centrally on a managed machine at CERN and serviced all LHCb data replication jobs for SC3. The lifecycle of an individual bulk replication request is sketched now.

Bulk Transfer Submission

Bulk transfer requests are submitted to the DIRAC WMS (discussed in [6]) in the form of a JDL file with an input sandbox of an XML file containing important parameters required to perform bulk transfer operation i.e. list of LFNs, source SE and target SE. The DIRAC WMS then populates the Request DB of the central machine with the XML files. The Transfer Agent checks the Request DB periodically for waiting requests and processes bulk replication operations as shown schematically in Fig. 3.

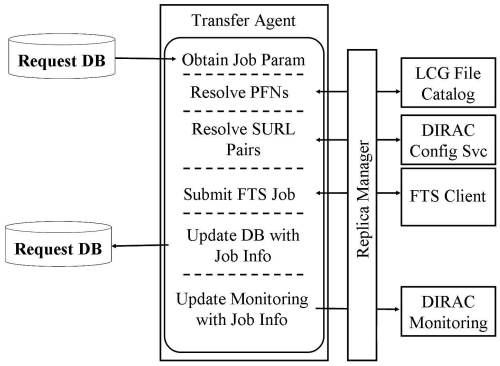

Figure 3: Schematic of Submission of Bulk Transfer Job.

The required data replication parameters are obtained by the Transfer Agent from the XML file populated to the Request DB. The replica locations stored in the catalog for the supplied LFNs are obtained by the Replica Manager by interaction with the File Catalog Client. This list of replicas is then compared with the source SE and target SE to determine whether the file is to be replicated. If the file already exists at the target site it will be omitted from the transfer operation and similarly for LFNs which don't exist at the source SE.

The DIRAC Configuration Service (CS) stores endpoint and relative path information for each of the available protocols on a particular SE. This data is used to resolve SURLs of the form 'srm://host/path/lfn' in accordance with LHCb conventions. The source and destination SURLs required by FTS are resolved using SRM protocol information in the CS. These SURL pairs are then submitted via the FTS Client command line operation 'glite-transfer-submit' which returns an unique FTS Job GUID. This GUID can be used to monitor the transfer job asynchronously and is therefore stored back in the Request DB file, along with other information on the FTS job useful for monitoring.

Bulk Transfer Monitoring

The Transfer Agent is executed periodically using the 'runit'[7] set of daemon scripts. Thus, if active replication requests exist in the Request DB at the point of execution

the Transfer Agent retrieves the XML file and extracts the relevant information. Using the FTS GUID the status of the FTS job is obtained via the FTS Client 'glite-transfer-status -l' command line call. The output of this poll includes a job state (Active, Pending, Submitted, Failed etc.) and status information for each of the files in the job. This output is parsed to obtain the status of interest (Active, Done, Failed) of individual files. Information on the progress of each of the files in the transfer request is updated in the request XML file and monitoring information sent to the DIRAC Monitoring Service to allow web based tracking of jobs (see [8]).

When a FTS job reaches terminal state successfully completed files are registered in the file catalogs while failed files are constructed into a new bulk replication request and placed back in the request XML file. These resubmitted files will be picked up on the next loop of the Transfer Agent. Transfer accounting information is sent to the DIRAC Transfer Accounting Server to allow effective bandwidth measurements to be made. These operations can be seen schematically in Fig. 4.

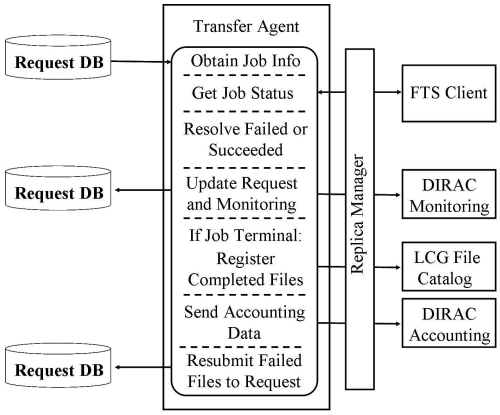

Figure 4: Schematic of FTS Job Monitoring.

PERFORMANCE OBTAINED DURING T0-T1 REPLICATION

The required rate to meet LHCb's first two SC3 goals was a combined rate of 40MB/s from CERN out to all 6 LHCb Tier1s sustained over two week period. From Fig. 5 the aggregated daily over a four week period during October and November 2005 is shown. It can be seen that the required rate was obtained for five days but, because of the overall instability of the SC3 system during this time, not sustained over the required period. The first half of this period was characterised by problems with the deployment of the new Castor2 system at CERN causing extensive transfer failures.

The Figure displays an average daily rate over a 24 hour period and doesn't give a fine grained rate within this period. Peak rates of 100MB/s were observed over several hours during successful running demonstrating the ability of the system to support the eventual rate required by the LHCb Computing Model. A rerun of this exercise is planned to demonstrate the required rates for success in SC3.

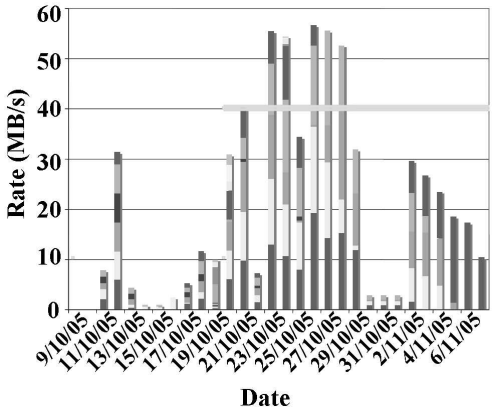

Figure 5: Aggregated Daily T0-T1 Throughput.

BULK FILE REMOVAL OPERATIONS

Once T0-T1 replication phase completed bulk removal of these files was performed. Additional functionality was required to be added to Replica Manager and Storage Element modules to allow utilisation of bulk 'srm-advisory-delete' command line call available at CERN. This tool takes a list of SURLs and performs an advisory delete (as outlined in SRMv1 Specification [9]) on the physical file. Replica Manager machinery developed for SURL resolution was reused and small additions were required for the SRM Storage Element plug-in.

During the file removal exercise problems were encountered with their genesis in varying interpretations of the SRMv1 specification. Different underlying behavior between SRM solutions was observed leading to the possibility of inconsistencies between physical storage and file catalog.

The performance of bulk removal operations executed by a single central agent showed the SC3 goal of 50K replicas in 24 hours to be unattainable. To meet the goal several parallel agents were instantiated each performing physical and catalog removal for a specific SE. With this setup 10K replicas were removed from 5 sites in 28 hours. With multiple agents accessing the LCG File Catalog (LFC)[10] concurrently a performance loss was observed in replica deletion because of unnecessary SSL authentications. In Fig. 6 the time taken for the removal of 100 replica entries from the LFC can be seen for various phases of the file removal. In each phase removal operations were performed at varying numbers of sites (see Table 1) to gauge LFC performance against the number of parallel agents.

This problem has since been remedied by the addition of 'sessions' whilst performing multiple catalog operations and the addition of bulk methods.

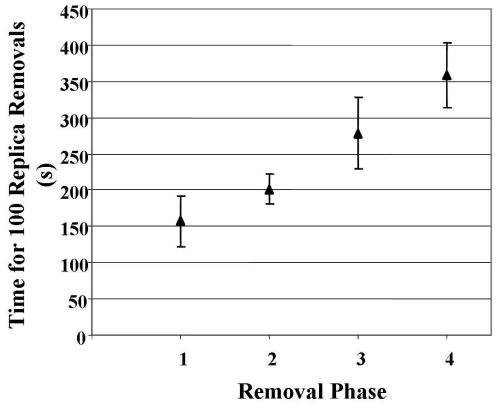

Figure 6: Degradation of Removal Operation Time with Increasing Parallel Agents Accessing File Catalog.

Table 1: Parallel Access to LFC

Phase	Sites	Parallel Agents
1	RAL	1
2	GRIDKA, IN2P3	2
3	GRIDKA, IN2P3, CNAF, PIC	4
4	GRIDKA, IN2P3, CNAF, PIC, RAL	5

TIER1-TIER1 REPLICATION ACTIVITY

During T0-T1 replication it was found that FTS was only efficient when replicating files pre-staged on disk. For this reason files to be used as seed data for LHCb's SC3 T1-T1 exercise were transferred to dedicated disk pools at the LHCb T1 sites. Since FTS provides a point-to-point service it was also required that FTS Servers on the Tier1 sites were installed to allow channels to be setup directly between sites. The current status of this setup is shown in Fig. 7. Replication activity is on going with this exercise to achieve the goals of SC3.

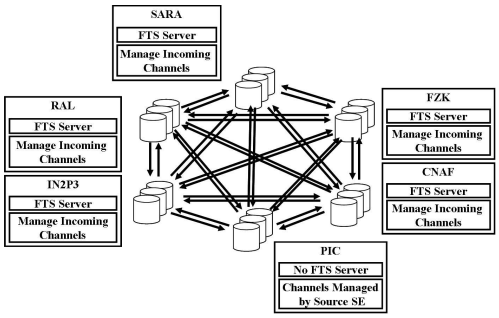

Figure 7: Overview of Tier1-Tier1 FTS Channel Matrix.

FUTURE WORK

The T1-T1 exercise is poised to begin to test the performance of the matrix of channels connecting all LHCb Tier1 sites over public internet connections. On completion of this exercise bulk file removal tests will repeated making use of new LFC capabilities and upgraded SRM deployments at the sites. A re-run of T0-T1 replication exercise will be performed to demonstrate the SC3 goal.

CONCLUSIONS

DIRAC DMS was integrated with gLite's FTS to allow reliable bulk file transfer operations. The system developed has been shown to be performant and capable of meeting the requirements of the LHCb Computing Model for T0-T1 data replication. But, to achieve this the stability of underlying platform must be maintained. Bulk removal operations were also incorporated into the DMS and initially proved difficult due to differing interpretations of SRM specifications. The burden on the LFC due to unnecessary SSL authentication overhead was discovered and new methods deployed by the developers to resolve the problem.

ACKNOWLEDGMENTS

I wish to thank the DIRAC development team and the LHCb Computing group as a whole. I would like to thank the Marie Curie Project for providing the opportunity to perform my current research and to attend CHEP06. Finally, I would like to thank the University of Edinburgh's School of Physics and the members of the Particle Physics Experiments group for their support.

REFERENCES

[1] DIRAC, the LHCb Data Production and Distributed Analysis system, A. Tsaregorodtsev et al. Proceedings of CHEP06, Mumbai, India, Feb 2006.

[2] https://uimon.cern.ch/twiki/bin/view/LCG/LCGServiceChallenges

[3] LHCb Computing, Technical Design Report, June 2005.

[4] http://egee-jra1-dm.web.cern.ch/egee

[5] The gLite File Transfer Service: Middleware Lessons Learned from the Service Challenges, P. Kunszt et al. Proceedings of CHEP06, Mumbai, India, Feb 2006.

[6] DIRAC Infrastructure for Distributed Analysis, S. Paterson et al. Proceedings of CHEP06, Mumbai, India, Feb 2006.

[7] http://smarden.org/runit/

[8] http://lhcb01.pic.es/DIRAC/Monitoring/Test/

[9] Common Storage Resource Manager Operations, Ian Bird et al. Oct 2001.

[10] Evolution of LCG-2 Data Management, J-P Baud, J. Casey, Proceedings of CHEP04, Interlaken, Switzerland, Sep 2004.