Journal of Physics: Conference Series **664** (2015) 042008 doi:10.1088/1742-6596/664/4/042008

An Analysis of Storage Interface Usages at a Large, Multi-Experiment Tier 1

S. de Witt[1] **and M. Reggler**[1,3]

[1] Senior Engineer, Science and Technology Facilities Council, UK

E-mail: shaun.de-witt@stfc.ac.uk

Abstract. Within WLCG much has been discussed concerning the possible demise of the Storage Resource Manager (SRM) and replacing it with different technologies such as XrootD and WebDAV. Each of these storage interfaces presents different functionalities and experiments currently make use of all of these at different sites. At the RAL Tier-1 we have been monitoring the usage of both SRM and XrootD by all the major experiments to assess when and if some of the nodes hosting the SRM should be redeployed to support XRootD. Initial results were previously presented which showed the SRM still handles the majority of requests issued by most experiments but with an increasing usage of XrootD, particularly by ATLAS. This updates these results based on several months of additional data and show that the SRM is still widely used by all the major WLCG VOs. We break down this usage according by read/write and by geographic source to show how a large Tier 1 is used globally. We also analyse usage by 'use case' (archival storage, persistent storage and scratch storage) and show how different experiments make use of the SRM.

1. Introduction

The original purpose of this work was to demonstrate that the use of the Storage Resource Manager (SRM)[1] is decreasing over time within WLCG [2]. This has been expected with the increasing use of XRootD [3], WebDAV and with the new File Transfer Service, FTS3 now not requiring SRM support [4]. Both XRootD and WebDAV are capable of supporting Wide Area transfers, particularly for read operations. In this paper, we look at detail from the production systems at one site. It should be noted that both sets of data were taken during first LHC long shutdown (LS1). During this period, both ATLAS and CMS have been testing federated access to both reduce the burden of data management and improve job efficiency. Currently, the CMS AAA service is in production while the ATLAS FAX service is still undergoing testing and has been throughout the two periods reported on in this paper.

1.1. Production Setup at the RAL Tier 1

The Tier 1 site at Rutherford Appleton Laboratory (RAL) is unusual in supporting all four major LHC experiments, as well as a number of smaller particle physics experiments including T2K, NA62, SNO+ and MICE. Operationally, three of the four LHC experiments have dedicated storage systems

[3] Matthew Reggler was supported by the Nuffield Foundation Research Placement Scheme

(ATLAS, CMS and LHCb), while ALICE shares an instance with the other smaller particle physics experiments. For ATLAS, CMS, LHCb and smaller experiments, the main interface to the storage system has always been the SRM for wide area transfers and, in some cases, even for local area transfers between the batch farm and the storage, with most of the transfers being mediated by the FTS. In addition, for the CASTOR storage system [5] deployed at RAL, the SRM provides the only secure interface. The SRMs run on dedicated hardware with 8 physical cores running hyper threading (16 effective cores) running at 2.2GHz and with 8GB of RAM under load balanced aliases; 2 machines for CMS and ALICE, 3 for LHCb and 4 for ATLAS who make most use of the SRM. Even with this number of processors available, the physical boxes often consume almost all of the available CPU cycles, primarily dealing with handling of the security contexts – the remaining operations being comparatively lightweight. It should be noted that although there are two SRMs dedicated to the 'ALICE' instance, these are for the small experiments since ALICE itself has always used XRootD for both wide and local area transfers and have never used the SRM in production.
In comparison, each storage system instance has a single XRootD manager process. This acts as a redirector, rather than a proxy server, in that connections are forwarded to an appropriate disk server from where the data is served. This single manager deals with all xroot client requests to the storage system.

2. Results

2.1. Request Rates

Figures 1 and 2 show the number of requests per day for the SRM and XrootD during the period 2013/2014. From this, it is immediately clear that during this period, considerably more use was made of the SRM than the XRootD manager. While there is some evidence of increased usage of the latter towards the end of this period, it is clear that the SRM was the dominant interface to the storage system. The difference varies by VO, with LHCb making approximately 16 times more requests to the SRM that the xroot system, while the ratio for CMS is closer to 2:1. This reflects the fact that CMS are moving to xrootd for all LAN traffic between the worker nodes and the storage system, only using the SRM for scheduled WAN transfers.

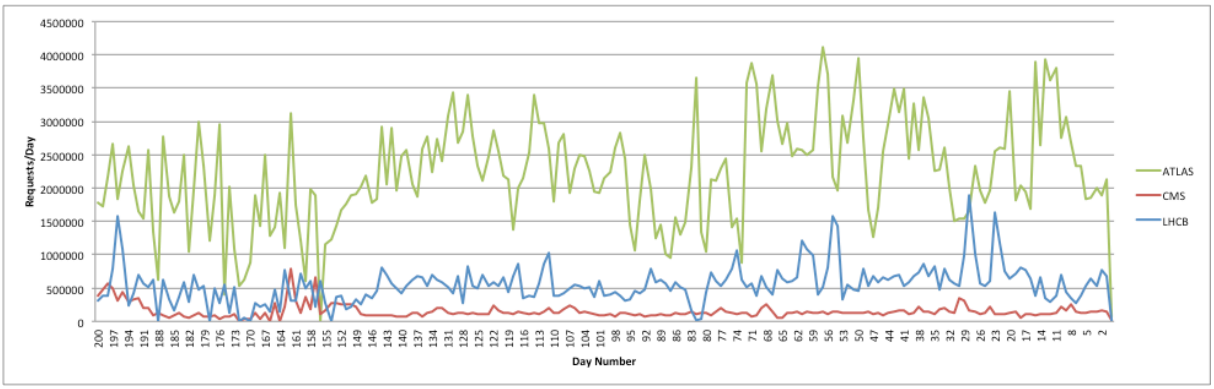

Figure 1:SRM Daily Total Request Rate 2013/2014

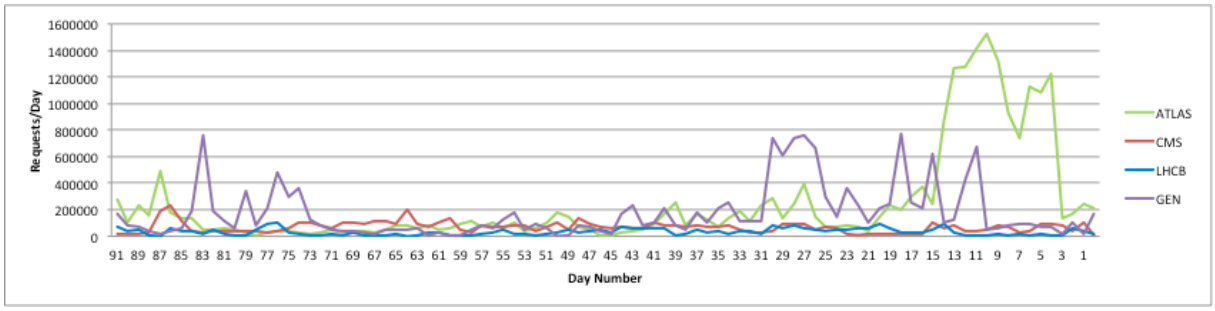

Figure 2: XrootD Daily Request Rate 2013/2014

Figures 3 and 4 show the same information about one year later. It is immediately obvious there was significantly less activity over this period, despite the fact that the coverage represents approximately the same dates over the two periods. SRM usage by CMS has decreased by about 60%, with the other VOs showing a drop in average usage to a smaller extent. Note that an LHCb reprocessing campaign is clearly shown by an increased usage between days 30 to 80. In general, XRootD requests also do not show any significant increase in usage with the exception of ATLAS which has seen steady increase in usage. For CMS and LHCb, the ratio of usage has changed significantly, with much more use of XrootD compared with the SRM, but it is likely that this is caused by less overall usage of the storage system rather than a long term difference in the data management frameworks. This assumption will be validated following LHC start-up when storage system usage will change to it's normal operational mode.

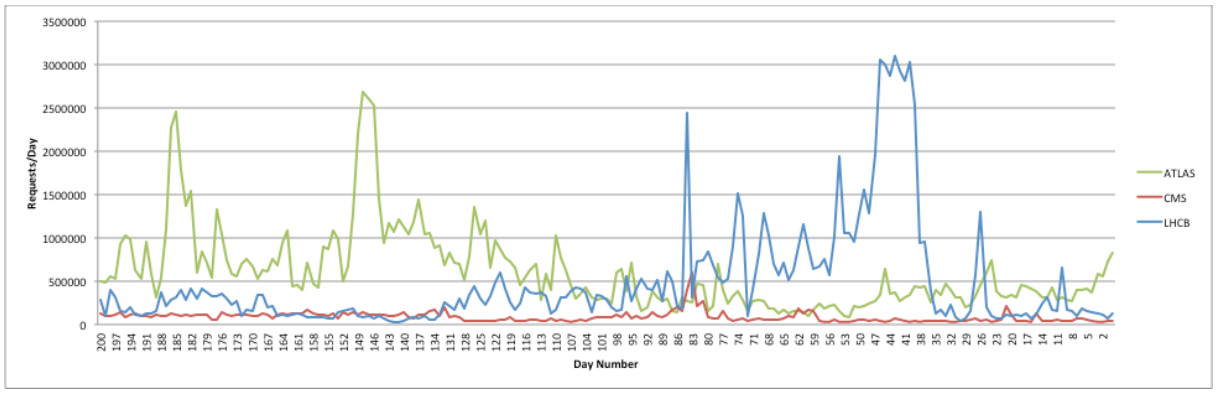

Figure 3: SRM Daily Total Request Rate 2014/2015

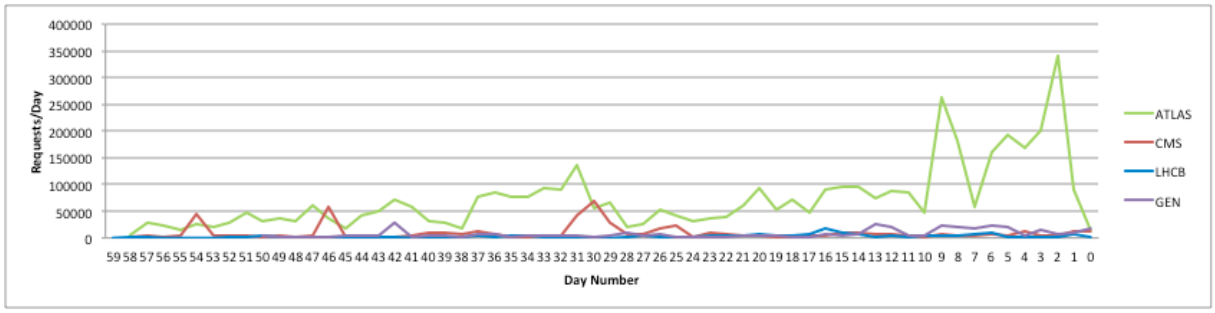

Figure 4: XrootD Daily Request Rate 2014/2015

2.2. Request Type Distribution

Having determined that the SRM is still a widely used interface, it was decided to investigate how the storage system was being used in the period 2014/2015. These results include all requests (including some internal requests). These are shown in figure 5 below. In these diagrams, the inner rings represent a 200 day period between 2013 and 2014, while the outer ring represents approximately the same period in 2014/2015. The requests have been divided into 6 distinct types of query as described below:

- *File Query Requests*: File information requests (ls, stat and checksum operations),
- *File Read Requests*: Attempts to open a file for reading, regardless of success or any actual read,
- *File Write Requests*: Attempts to open a file for writing, regardless of success,
- *File Deletions (system)*: Garbage Collection by the storage system to remove diskcopies which can either be recovered or represent inaccessible data (so called 'dark data'),
- *File Deletions (user)*: User requested file deletions,
- *Space Queries*: User requests to check the amount of space left in a space token.

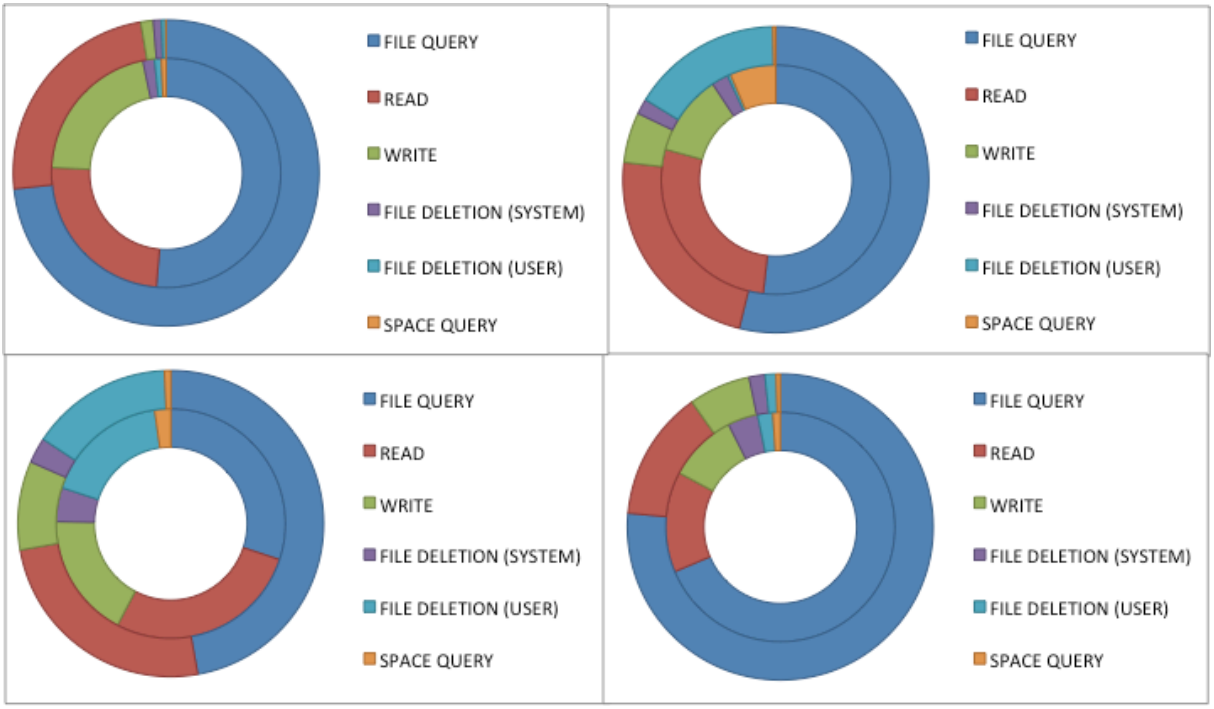

Figure 5: Distribution of Request Types at the RAL Storage System (clockwise from top left: ALICE, ATLAS, CMS, LHCb)

As can be seen, in the case of ATLAS, significantly more deletion requests from external clients was seen in the second period compared to the first which fits with ATLAS using significant amounts of disk space and needing to remove old data to make way for new. CMS seem to be increasingly relying on external catalogues to check the existence of files since the proportion of file query requests has decreased. This may also be down to increasing use or testing of AAA where jobs rely on xroot fallback rather than the existence of an input file locally. In the case of ALICE, the number of writes has reduced significantly as a result of disk space filling up and no campaign to remove older files having taken place, unlike ATLAS. Finally the storage system usage by LHCb had not changed significantly over the periods investigated.

One further point of information regarding usage of the SRM is to compare the 'efficiencies' over the same period. During this time, not only have experiments changed their usage pattern, but the back-

end storage system has been upgraded without changing the SRM. In the case at STFC both read and write requests to the SRM are asynchronous, writing a file requires an *srmPrepareToPut* request followed by a number of *srmStatusOfPut* requests until the transfer URL is available (and similarly for read requests). If the SRM efficiency is considered simply as the ratio of *srmStatusOfXXX* requests compared to *srmPrepareToXXX* requests, the RAL storage system between the two analysis periods is shown in Table 1 below (lower values representing better efficiencies). Again, ALICE are omitted since they do not make use of the SRM.

Table 1 SRM Efficiencies for Supported LHC Experiments

	Read Efficiency		Write Efficiency	
	2013/14	2014/15	2013/14	2014/15
ATLAS	7.01	6.99	4.79	4.82
CMS	1.72	6.21	1.61	4.38
LHCb	1.03	1.88	1.11	2.99

While it appears that for ATLAS, the situation has not changed significantly, this is not true for either LHCb or, particularly, CMS. The underlying reasons for this are not yet understood, but it has prompted further investigation into bottlenecks in the underlying storage system. A number of different possibilities exist, each of which are being investigated separately including the impact of using fewer, large storage servers, slowness introduced during an upgrade, database performance and use of FTS. The latter is particularly relevant since FTS3 has been introduced during the second period and it may have shorter times between queries compared to earlier releases. This was known to be the case in early releases of FTS3 where *StatusOfXXX* requests were received every 100 ms, rather the back-off strategy previously employed (check immediately, after 5 seconds, after 30 seconds and after 60 seconds).

2.3. Geographic Distribution of Requests

The geographic requests arriving at the RAL Tier 1 SRM over the two periods analysed is shown in Figure 6 below, with the inner circle representing 2013/14 and the outer representing 2014/15 periods. In the case of ATLAS and CMS, there has been a significant change with transfers dominated by three countries; the UK, US and Switzerland with over 99% of connections coming from these countries. This is caused by the rationalisation of FTS deployments, where FTS production instances are now only deployed at three sites (RAL, CERN and BNL).

For non-FTS transfers, there are some interesting observations showing the evolution of WLCG as new sites come on-line and smaller or less reliable sites are used less. For CMS, FTS transfers dominate with only two countries making significant non-FTS connections over this period, Brazil and Korea. In both these cases, the number of transfers has increased significantly from the same period the previous year. In the case of ATLAS, many countries still make a significant number of direct connections. Comparing the two years, sites in Rumania (339%), Australia (216%) and Russia (107%) show the largest increase in direct traffic to the RAL SRM, while sites such as Taiwan (-99%) and Austria (-100%) make significantly less direct usage. For LHCb, there is significantly less traffic between CERN and RAL due to less RAW transfers taking place with a drop of some 80%, but increased connections to Germany (2033%), Rumania (203%) and Brazil (641%). Data analysis centres in Cyprus seem to have seen the biggest decrease between the periods, dropping from more than 10,000 requests in 2013/14 to zero in 2014/15.

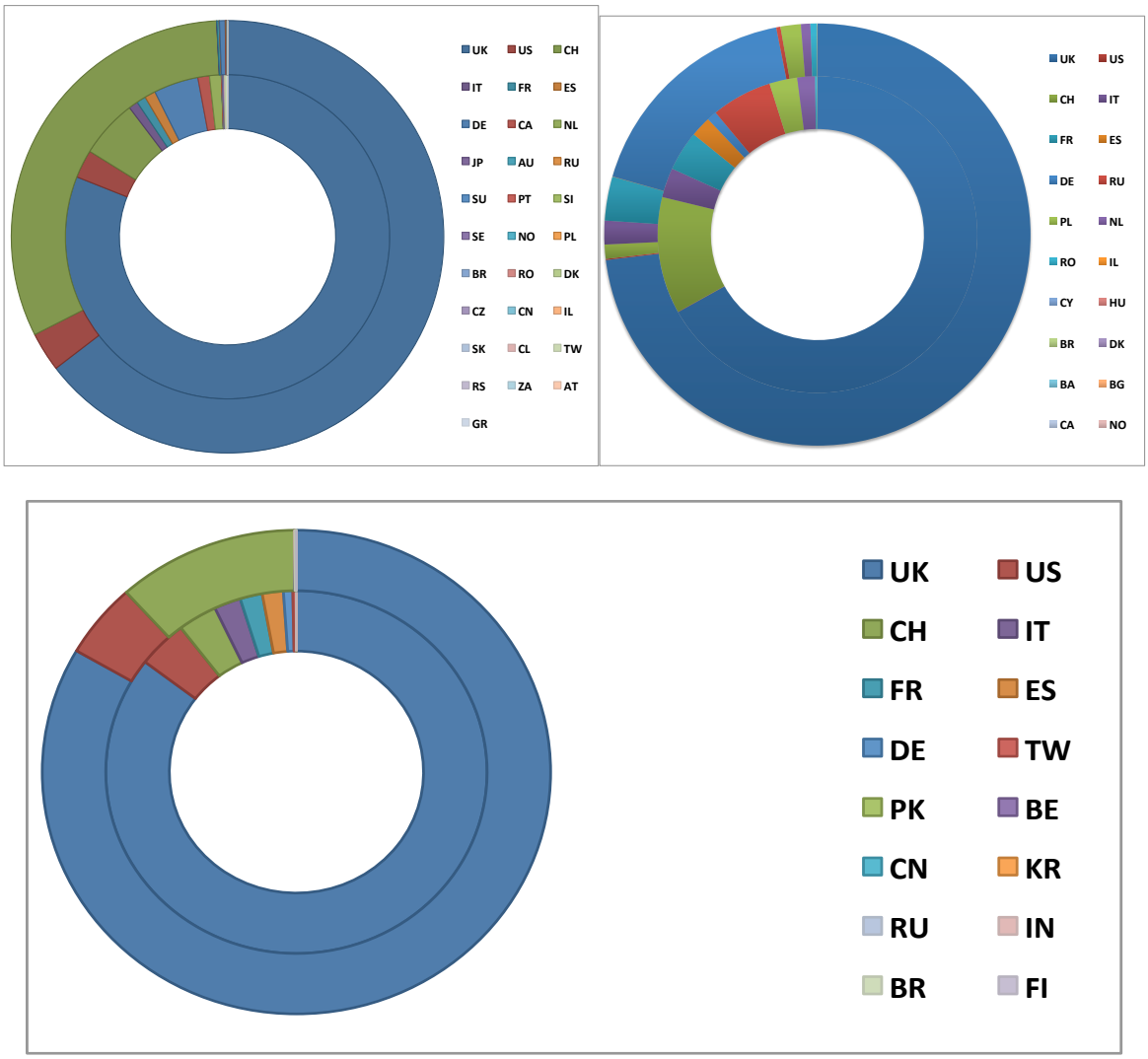

Figure 6: Geographic Distribution of SRM Requests From the Periods 2013/2014 (inner ring) and 2014/15 (outer ring) for ATLAS (top left), LHCb (top right) and CMS (bottom)

3. Conclusions

The original purpose of this work was to investigate whether SRM usage is decreasing in the new WLCG configuration. It would certainly seem that this is not the case. There is little evidence that the user of XRootD in AAA and FAX is having any significant impact on SRM usage and, even though FTS3 supports SRM-less end points, the LHC experiments are not currently making significant use of this feature. While the total number of SRM requests has decreased over the two periods analysed, there was no corresponding increase in the number of requests using xroot directly. In some cases, the pattern of usage of the SRM has changed over the reporting period, with some experiments not using the SRM to check the status of a particular instance of a file, but overall it is likely that the SRM will still be required at Tier 0 and Tier 1 sites for the duration of LHC run 2.

4. References

[1] Sim A, Shoshani A, et al, https://sdm.lbl.gov/srm-wg/doc/SRM.v2.2.html
[2] http://wlcg.web.cern.ch/
[3] http://xrootd.org/
[4] Ayllon AA, Salichos M, Simon M K and Keeble O, 2014 *J. Phys.: Conf.*

Ser. **513** 032081 doi:10.1088/1742-6596/513/3/032081

[5] Baud J-P, et al, 2003 Computing in High Energy and Nuclear Physics (CHEP03), La Jolla, CA, USA, arXiv:cs.oh/0303241