

The ATLAS DQ2 Accounting and Storage Usage Service

Fernando H. Barreiro Megino, Vincent Garonne, Stephane Jezequel,
Miguel de Oliveira Branco, Mario Lassnig, Pedro Salgado

E-mail: fernando.harald.barreiro.megino@cern.ch, vincent.garonne@cern.ch,
jezequel@lapp.in2p3.fr, miguel.branco@cern.ch, mario.lassnig@cern.ch,
pedro.salgado@gmail.com

Abstract. The DQ2 Distributed Data Management system is the system developed and used by ATLAS for handling very large datasets. It encompasses data bookkeeping, managing of largescale production transfers as well as endusers data access requests. The multi-Petabyte ATLAS data volume already under management requires an accounting and monitoring service that collects different data usage informations in order to show and compare it from the experiment and application perspective. In this paper we will describe the design and implementation of the DQ2 Accounting and Storage Usage Service, built to meet the monitoring requirement.

1. Introduction

The ATLAS DDM - Distributed Data Management - (also known as DQ2 [1]), manages over 10 Petabytes of data, corresponding to 120 million files and 1.3 million datasets (see figure 1) that are distributed over more than 500 Grid storage endpoints.

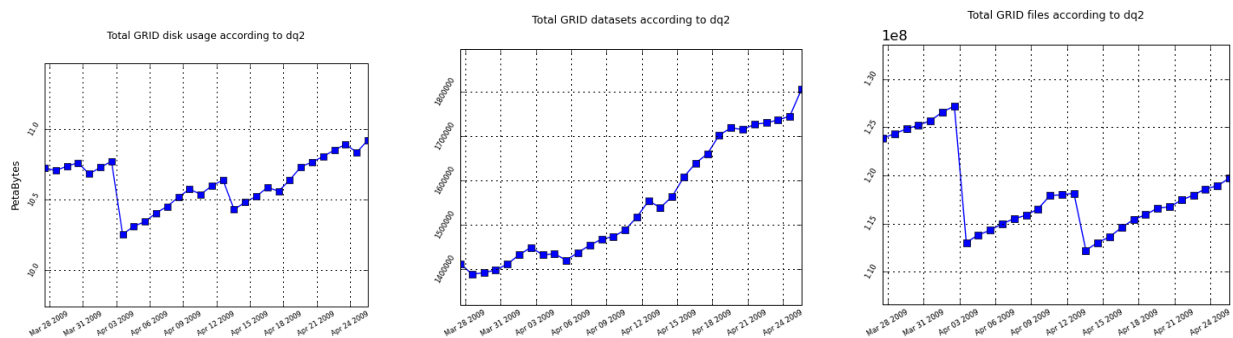


Figure 1. Plots showing the evolution of the amount of data managed by DDM between the 27th March and 25th April 2009.

Given the complexity of the system, it is necessary to develop a monitoring tool, capable of showing the accounting and disk space information from different perspectives and, where

possible associating the available metadata, for example information required by physics groups. As the information available in the DDM catalogs can be incorrect, it is vital to compare the DDM registered data volume with the real data volume on the sites available through SRM¹.

2. System architecture

The system developed consists of two agents that collect the SRM and DQ2 information on a daily basis and store them in the accounting database. The front-end of the system consists of a web interface, capable of generating dynamically the plots required in each view. These elements will be described in the following subsections and are also summarized in the high level overview in figure 2.

2.1. DQ2 agent

The DQ2 agent is in charge of querying daily the location catalog about the files and datasets registered in each of the Grid sites. This information is then stored in the local accounting database via the DQ2 Accounting API. With the information collected by the agent, the disk space occupied at each site can be calculated. However, given that the information in the DQ2 location catalog is not always correct due to the different reasons explained in section 4, it is always important to compare its disk space information with the disk space information available directly from the SRM.

2.2. SRM agent

Querying the used and total disk space at each site through its SRM interface is the work of the SRM agent. The steps followed by the SRM agent are the following:

- (i) First, the agent has to update the list of sites each day, reading from the central information system, also known as Tiers of ATLAS.
- (ii) The more than 500 Grid storage endpoints are queried using the `lcg-stmd` command, using parallel threads to reduce running time.
- (iii) Finally the information is stored in the DQ2 accounting database through the DQ2 Accounting API.

As well as populating the DQ2 accounting database, the SRM agent is also used for other systems:

- the SLS monitoring pages [2]
- the PanDA production and workload management system [3]
- based on the SRM information from each site, the system also sends out emails to the ATLAS operations team, alerting them to those sites that are close to being full. A site will be included in the alert email if it does not pass one of the following conditions:
 - For ATLASMCDISK or ATLASDATADISK in Tier1s: at least 10 Tera Bytes free disk space
 - For ATLASMCDISK or ATLASDATADISK in Tier2s: at least 2 Tera Bytes free disk space
 - For ATLASPRODDISK or ATLASSCRATCHDISK in Tier1s or Tier2s: at least 20% free disk space

The first two conditions are based on the maximum amount of data transfer foreseen to the site. The behaviour of the alert system is currently being improved to send out the emails directly to cloud administrators, avoiding continuous alarms about the same site.

¹ SRM Storage Resource Manager- is a common protocol for managing the different storage element independently of their implementation technology (HPSS, CASTOR, dCache, DPM, StoRM).



Figure 2. Plots showing the evolution of the amount of data managed by DDM between the 27th March and 25th April 2009.

2.3. Accounting Web Interface

The system front-end will be a dynamic web interface, developed in Python and using the Django web framework. The plots are generated through the Matplotlib libraries, checking if there is already an updated plot before generating a new one. Additionally, some plots require the CMS Graphtool extension for Matplotlib (<http://t2.unl.edu/documentation/using-graphtool>).

The general appearance of the user interface is shown in 2. All pages show a common structure, including a header with an extensible menu, where the desired view can be selected. The user can also select the period for the displayed information; the options are for the last week, the last month or for all the information available in the accounting database.

The information displayed in the different views of the system will be described in the next

section.

3. Views of the system

The system offers different plots and structural overviews in order to facilitate the work of the DDM operations team and the different cloud or site administrators. The following subsections will describe the content in each view.

3.1. Site view

The Site view shows the plots related to a particular spacetoken at a site. This view presents the DQ2 versus SRM plot, which compares the used disk space according to the DQ2 accounting and according to the SRM interface to the spacetoken (see figure 3.). This plot is interesting for estimating the coherence of the information in the location catalog.

In order to monitor the disk space that is still available, the plot also includes the total disk space available for that spacetoken. The curve for the total disk space is skipped in cases where the total disk space is over ten times the SRM used disk space, so that the used disk space curves are displayed in more detail.

Other plots in this view show the DDM evolution for the number of files and datasets registered for a particular spacetoken at a site.

At the bottom of this view there are three cumulative plots that show the break down by physics group of the occupied disk space (figure 4), the number of files and the number of datasets registered for the site.

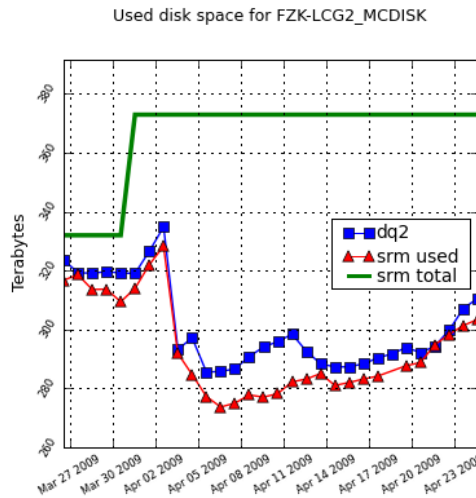


Figure 3. DQ2 vs SRM plot for FZK-LCG2_MCDISK.

3.2. Cloud view

In this view, a summary table of a cloud is presented, including downsized plots for all the sites belonging to the cloud. The rows in the table correspond to the site and the columns correspond to the different spacetokens. The example in figure 5 shows part of the Spanish cloud.

3.3. Tier1 summaries by spacetoken

For operational issues in DDM, it is very practical to have a summary view of the most important spacetokens (ATLASDATADISK, ATLASDATATAPE, ATLASMCDISK and

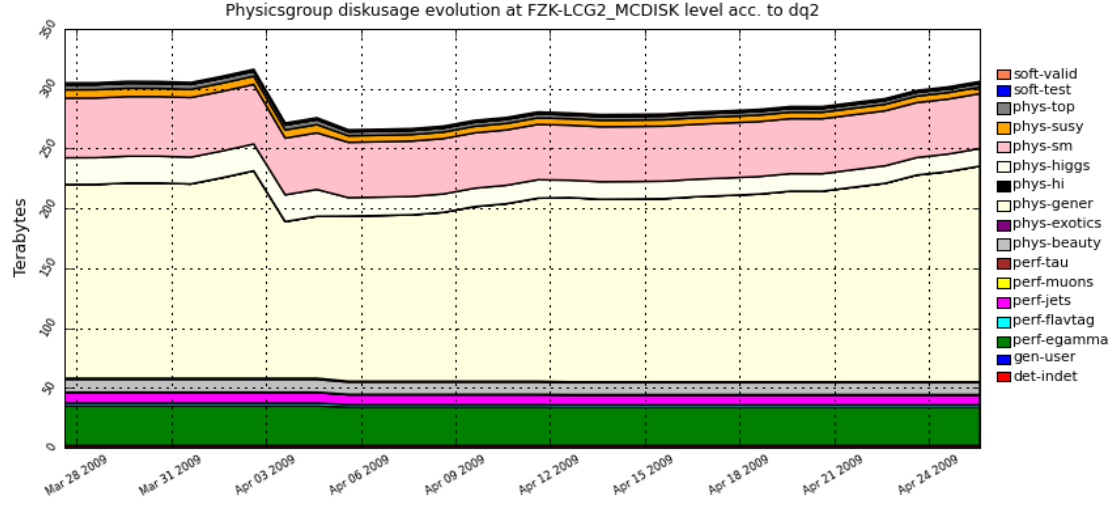


Figure 4. Cumulative plot showing distribution of occupied disk space by physics groups in FZK-LCG2_MCDISK.

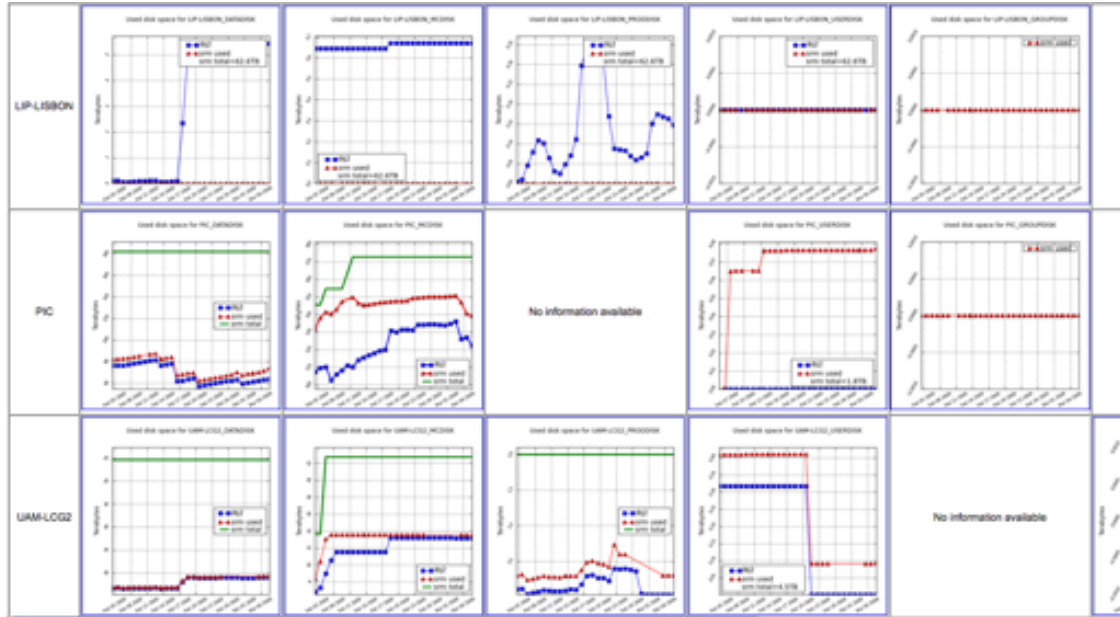


Figure 5. Part of the Cloud View for the Spanish cloud.

ATLASMCTAPE) for all Tier1 sites. In this view, the pertinent DQ2 versus SRM plots are grouped together and the information is also presented in cumulative plots to enable the comparison of the amount of data registered at each site (see figure 6).

3.4. Tier1 and Tier2 reports

Additionally, on a weekly basis, summary tables and comparative plots are generated for Tier1s and Tier2s. These reports allow, at a glance, an overview of the disk space status at the different sites. The reports for Tier1s are more complete and include summary plots, as can be seen in

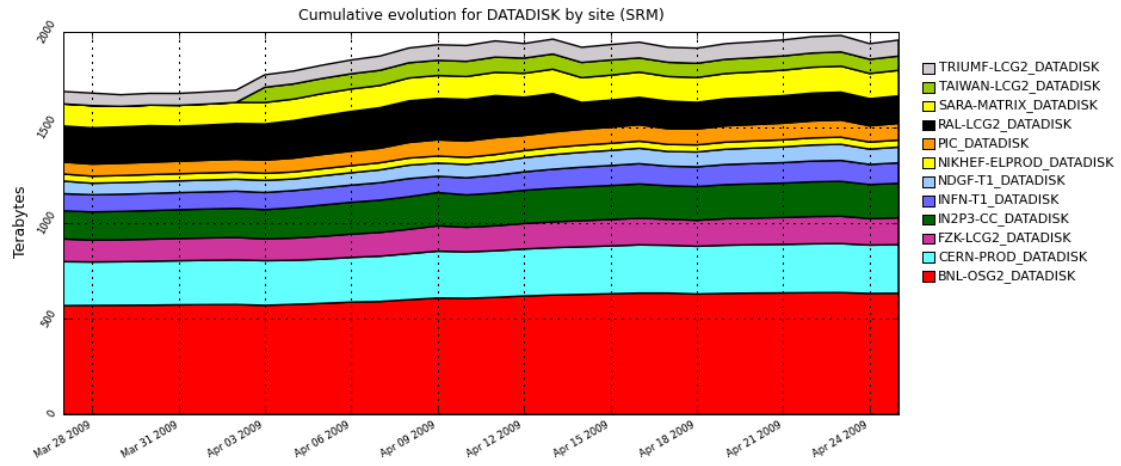


Figure 6. Cumulative plot showing distribution of occupied diskpace by Tier1s.

figures 7 and 8.

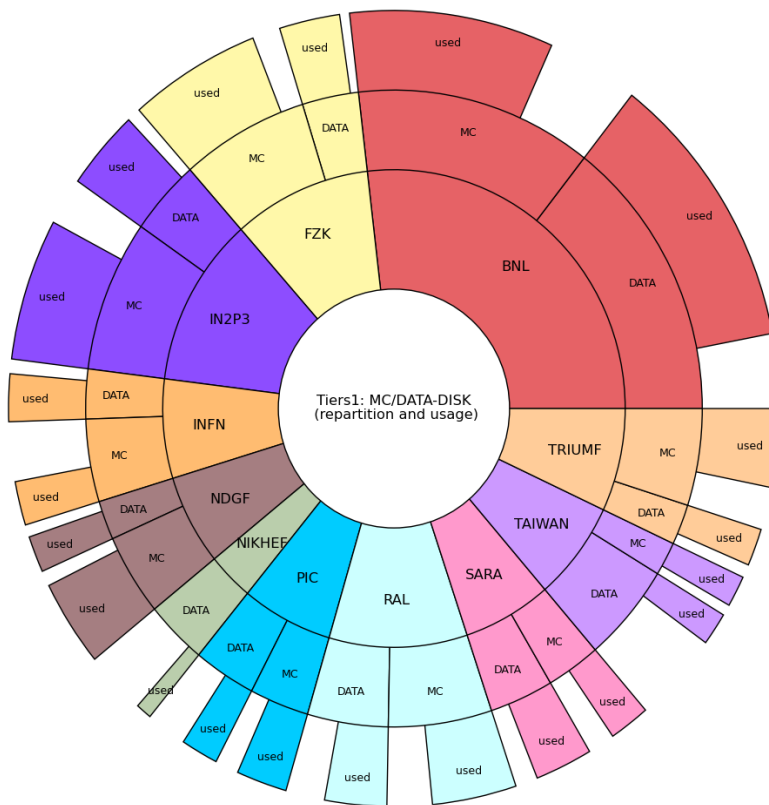


Figure 7. Comparison used/total and MC/DATA-DISK per Tier1.

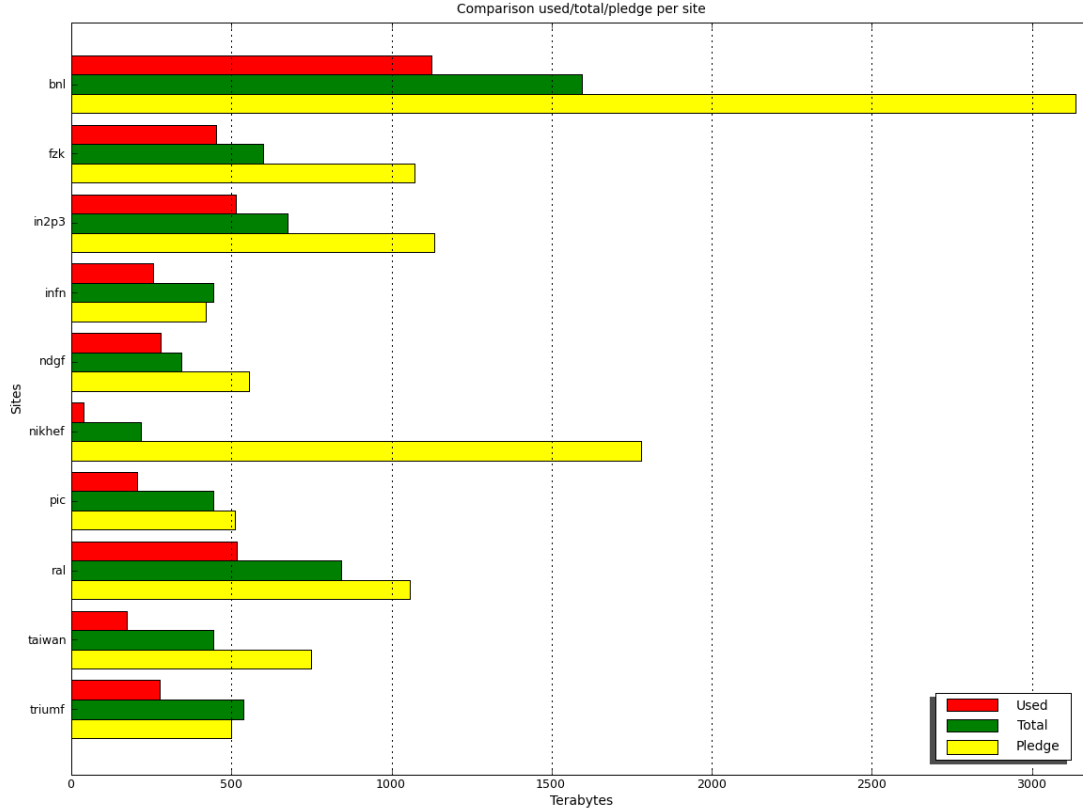


Figure 8. Comparison used/total/pledge per Tier1.

3.5. Predictions

Using the data from the previous seven days, the system tries to predict the behavior of the disk occupancy for the next week and shows the information on a simple Grid plot summary by cloud. The objective of the plot is to try to alert cloud administrators about which sites are becoming full.

3.6. Grid view

Under the Grid view, the user can see the data volume registered in DDM distributed over the whole Grid in the form of disk space, the number of files and number of datasets (see figure 1). The information is also broken down by physics group in a cumulative plot.

4. Discussion

The system has been developed in a close collaboration between the ATLAS DDM development and operations team and has helped to spot different types of problem:

- **Dark data:** Dark data is data not known by the location catalogs. It can be due to bad deletions, where a file has been deleted from the location catalog, but not effectively removed from the site. As the location catalog is not aware of this data, it will not be accessed and is a waste of disk space. Figure 9 shows sites with dark data:
 - The first image presents the typical aspect of dark data, having a constant shift of the red over the blue curve.

- In contrast, the second image shows a site where the site administrator has made the effort to delete dark data and thus made the two curves coincide by the 16th October 2008.

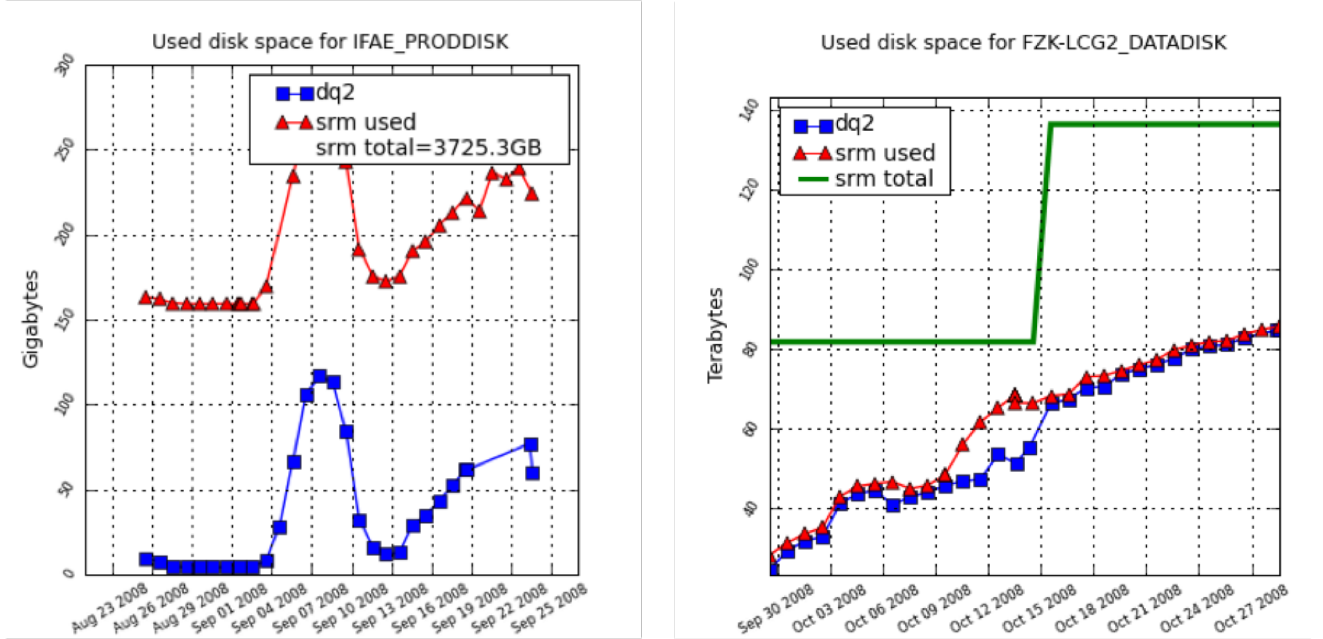


Figure 9. Dark data at sites.

- Multiple registration of files in overlapping datasets: Handling of overlapping datasets clearly complicates the system and in previous versions of DQ2, it could happen that files, belonging to different datasets, were registered more than once. This lead to DDM reporting a data volume higher than that according to SRM. The example in figure 10 shows an example of a site having this multiple registration problem and therefore the blue DQ2 curve has a constant shift with respect to the red curve.
- Publication problems: In some cases, sites are incorrectly configured and do not respond correctly to SRM commands. This problem has been corrected for most of the sites after dedicated effort from the DDM operations team. The example in 11 shows a past example, where the red curve with the SRM occupied disk space is always zero, although the site was publishing correctly the SRM total value (which appears in the legend).

5. Conclusions

The system is regularly used by the DDM operations and many site and cloud administrators. In particular, the German and French cloud use the DQ2 Accounting and Storage Usage Service to maintain the coherence in their sites. Thanks to their constant work and the ongoing development of the DQ2 system, the DDM and SRM values are becoming more and more consistent.

Acknowledgments

Work supported by the Spanish Ministry of Science and Innovation (MICINN).

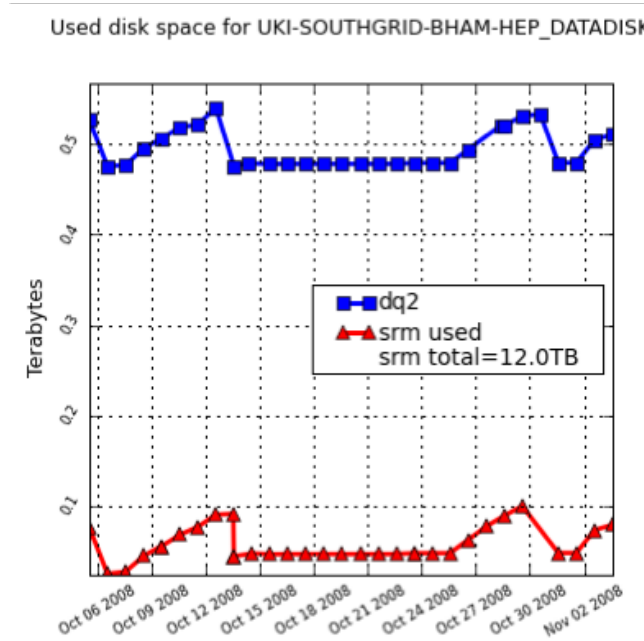


Figure 10. Multiple file registrations.

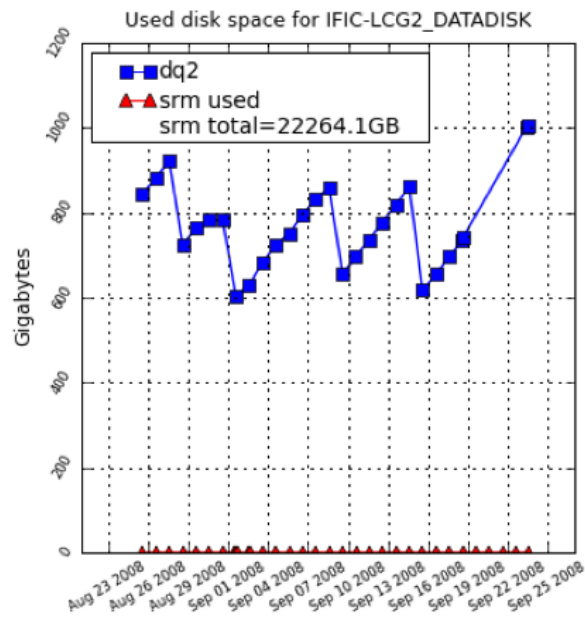


Figure 11. Publication problems.

References

- [1] Branco M, Zaluska E, de Roure D, Salgado P, Garonne V, Lassnig M and Rocha R 2008 Managing Very-Large Distributed Datasets *Proceedings of the OTM 2008 Confederated International Conferences*. 775 - 92
- [2] Di Girolamo A, Campana S, Santinelli R, Sciaba A and Koblitiz B 2009 The Service Level Status monitoring of the LHC Experiments Distributed Computing *Proceedings of the 17th International Conference on Computing in High Energy and Nuclear Physics*

- [3] Nilsson P, Caballero J, De K, Maeno T, Potekhin M, Wenaus T 2008 The PanDA System in the ATLAS Experiment *XII Advanced Computing and Analysis Techniques in Physics Research* .