New data access with HTTP/WebDAV in the ATLAS experiment

J Elmsheuser[1], R Walker[1], C Serfon[2], V Garonne[3], S Blunier[4], V Lavorini[5] and P Nilsson[6] on behalf of the ATLAS collaboration

[1] Ludwig-Maximilians-Universität München, Germany
[2] CERN, Switzerland
[3] University of Oslo, Norway
[4] Pontificia Univ. Catolica de Chile, Chile
[5] Universita della Calabria, Italy
[6] Brookhaven National Laboratory, USA

E-mail: johannes.elmsheuser@physik.uni-muenchen.de

Abstract. With the exponential growth of LHC (Large Hadron Collider) data in the years 2010-2012, distributed computing has become the established way to analyse collider data. The ATLAS experiment Grid infrastructure includes more than 130 sites worldwide, ranging from large national computing centres to smaller university clusters. So far the storage technologies and access protocols to the clusters that host this tremendous amount of data vary from site to site. HTTP/WebDAV offers the possibility to use a unified industry standard to access the storage. We present the deployment and testing of HTTP/WebDAV for local and remote data access in the ATLAS experiment for the new data management system Rucio and the PanDA workload management system. Deployment and large scale tests have been performed using the Grid testing system HammerCloud and the ROOT HTTP plugin Davix.

ATL-SOFT-PROC-2015-002
28 April 2015

1. Introduction

The LHC at CERN is colliding protons or heavy ions at unprecedented centre-of-mass energies and these collisions are recorded by several experiments including the ATLAS experiment [1]. The large number of collision events and resulting data volumes require an analysis on distributed computing resources [2]. The Worldwide LHC Computing Grid (WLCG) is the default and established way to process these large amount of data. The fast and reliable access in reasonably long Grid jobs is essential for any kind of data analysis job not only in particle physics.

The ATLAS experiment uses the PanDA workload management system [3] based on the pilot job paradigm to schedule jobs on WLCG resources. The detector and simulation data are distributed to the different sites and managed by the ATLAS data management system Rucio [4]. The paper describes the deployment and testing of HTTP/WebDAV for local and remote data access using these two systems. Deployment and large scale tests have been performed using the Grid testing system HammerCloud [5] and the ROOT [6] HTTP Davix [7] plugin.

2. Setup data access with WebDAV

The input data access for jobs submitted by the PanDA workload management system is configured on a site by site basis. Sites with dCache storage element usually use direct access to

the input data with the dcap protocol. Jobs at sites with DPM storage elements usually copy the full input data to the worker node scratch disk before starting the actual analysis code. Lately more and more sites with DPM have moved to direct data access using the XRootD protocol [8]. Since both of these very popular storage elements also offer the possibility to access data through the WebDAV/HTTP protocol in their latest versions, this access mechanism to the data has been explored.

Contrary to the more particle physics community specific dcap and XRootD protocols, WebDAV/HTTP is a widely used industry standard protocol. Using this protocol would allow to use the same client and access method in a unified way across all jobs submitted with the PanDA system.

PanDA analysis jobs usually consist of two job types: one initial so called build job that compiles the C++ analysis code and stores the created executable and libraries in a relocatable archive file on the local storage element where the job was executed. When this job is finished several subsequent actual analysis jobs are started that process the input data in parallel jobs. These jobs first have to stage-in and unpack the previously created source code archive file. As a first step to fully use the WebDAV access this PanDA stage-in mechanism has been altered to use aria2c [9] and the WebDAV data access protocol to download the source code archive file to the local worker node in the jobs. Then in the subsequent analysis jobs the input data access mode has been changed as well to use the Davix plugin when using ROOT I/O to enable WebDAV access within ROOT.

The WebDAV access at a site works through a direct connection to the WebDAV door of the local data pool that hosts the data on the storage element. To allow a unified input file naming schema at all sites a Rucio WebDAV re-director service has been setup. A secure RESTful API call with HTTP redirection is used by the Rucio server to query the internal dataset replica table and redirect the query results to a specific replica URL. The site specific WebDAV door names are stored in the ATLAS Grid information system AGIS [10] and are used by the re-director to translate the generic file name to a specific local WebDAV file path URL. The translation mechanism can be configured by adding different parameters to the URL: `?site=SITENAME` to force the usage of a file replica from a specific site, `?site=SPACETOKENNAME` to force the usage of a file replica from a specific site storage area, and `?select=geoip` to force the usage of a file replica that is closest to the job location. Using these options allows not only local data access but also remote reading of the input data in the jobs. A generic re-director URL looks like the example given in Table 1 (a). This URL is translated by calling the re-director in an actual WebDAV URL accessible with aria2c or the ROOT Davix plugin and looks like the example given in Table 1 (b).

Table 1. Examples for a generic URL before the re-director call (a) and a translated and re-directed local WebDAV URL (b).

(a) Generic re-director URL:

```
https://rucio-lb-prod.cern.ch/redirect/mc14_8TeV/
AOD.01507240._010001.pool.root.2?site=LRZ-LMU
```

(b) Translated local WebDAV URL:

```
https://lcg-lrz-dc66.grid.lrz.de:443/pnfs/lrz-muenchen.de/data/atlas/dq2/
atlaslocalgroupdisk/rucio/mc14_8TeV/99/dd/AOD.01507240._010001.pool.root.2
```

Figure 1 shows the schematic work flow of a WebDAV data access for a Grid job submitted

with the PanDA workload management system and using the Rucio data management system.

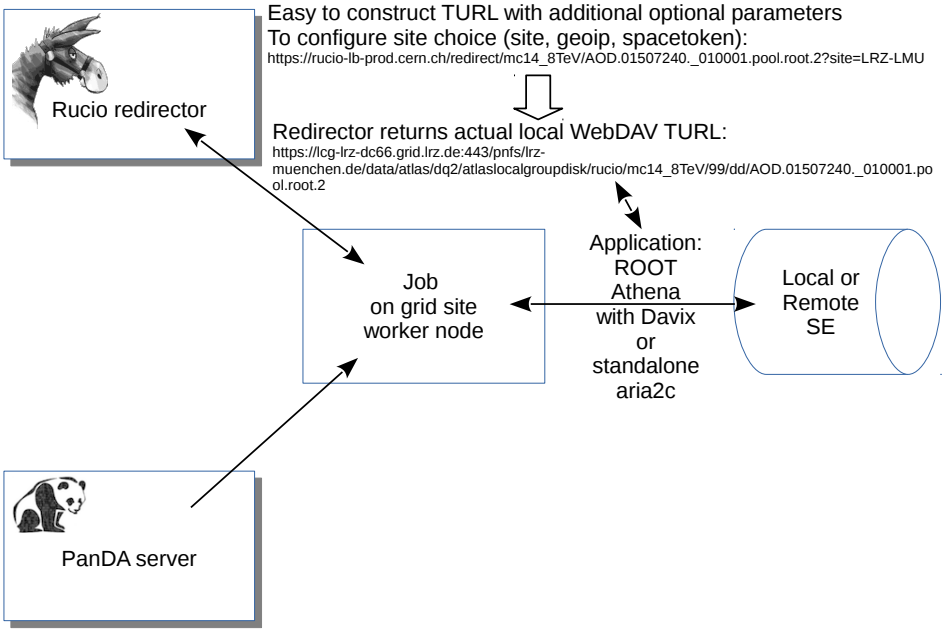

Figure 1. Schematic view of the WebDAV access work flow in a Grid job submitted with the PanDA workload management system and using the Rucio data management system.

The PanDA analysis jobs use as input files the generic WebDAV URLs which are then redirected, opened and accessed through the ROOT Davix plugin. This plugin was intensively tested by the paper authors and feedback provided to the code authors. It is deployed and enabled in the central ATLAS ROOT distribution mechanism through the CVMFS file system [11]. The current version (0.4.0) works reliably for DPM and dCache storage elements. The files downloaded with aria2c are checked for file integrity using the adler32 checksum stored for all files managed by the Rucio system. The support of this checksum format was requested to the aria2c development team and is available since the aria2c version 1.18.0.

3. Functional tests with WebDAV

To enable the workflow as described in the previous section several steps had to be taken. First the ATLAS information system AGIS had to be updated with the correct site name vs. WebDAV door mapping for all sites with WebDAV capabilities. This was an iterative process to verify all configurations at all sites. The second step was to instrument the PanDA pilot [12] code to create a meta-link XML for the aria2c file download as given in the example in Table 2.

In addition the PanDA pilot code was updated to create the generic input file list with WebDAV syntax as described before to allow WebDAV enabled ROOT I/O with the Davix plugin. These setups and code changes have been verified using a continuously running HammerCloud functional test. This test is probing the aria2c and Davix ROOT access at the local storage elements of currently about 70 different sites. Every 2 hours a PanDA Grid job is submitted which executes a lightweight ROOT based analysis reading a couple of variables from an input file via WebDAV. This functional test performs a full analysis workflow and

Table 2. Example of a meta-link XML file for download with aria2c.

```
<metalink>
  <file name="user.gangarbt.0302032633.699303.5577.lib.tgz">
   <identity>user.gangarbt:user.gangarbt.0302032633.699303.5577.lib.tgz</identity>
   <hash type="adler32">22b8cbbd</hash>
   <size>11505</size>
   <url location="LRZ-LMU_SCRATCHDISK" priority="1">
   https://lcg-lrz-dc66.grid.lrz.de:443/pnfs/lrz-muenchen.de/data/atlas/dq2/
atlasscratchdisk/rucio/user/gangarbt/9d/66/
user.gangarbt.0302032633.699303.5577.lib.tgz
   </url>
   </file>
</metalink>
```

probes all failure possibilities of a regular user analysis job. Several issues of storage element configurations or outdated storage element software versions at different sites were detected and consequently fixed. Currently about 30% of the sites require updates to the latest stable storage element software versions. Figure 2 shows a screen shot of a HammerCloud web page for a test configured with WebDAV input data access.

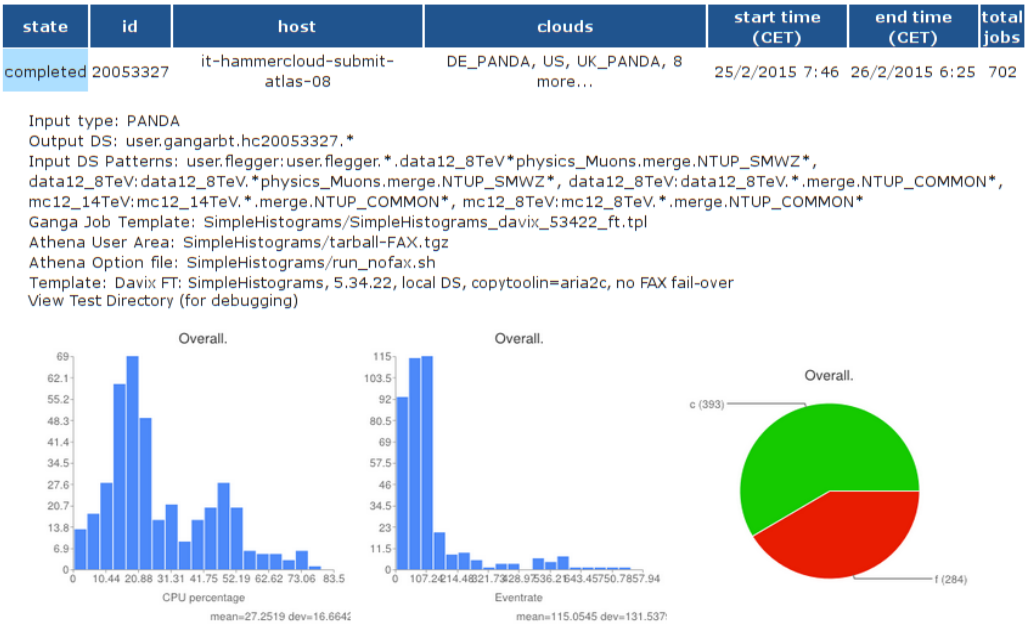

Figure 2. View of the HammerCloud web page for a test configured with WebDAV input data access.

4. Access protocol comparisons

To gain experience with and to measure the performance of the WebDAV protocol, several ATLAS analyses based on the new ATLAS xAOD data format [13] have been setup. The

analysis was performed in three modes accessing and processing different amounts of physics quantities of ATLAS detector objects like muons, electrons and jets and additionally store the results in new output files. These analysis modes have been named slow, medium and fast, and vary between a rather I/O intensive (slow) and a rather I/O moderate (fast) processing. The analysis code allows to choose between two access modes of these physics quantities: branch and class access mode. In the first mode only the actual variables used in the analysis are read from the input file, while in the latter mode the whole physics object with all its quantities is read. All tests have been executed using ROOT 5.34.24 in 64 bit mode on Scientific Linux 6.4 and the C++ analysis code was compiled with gcc 4.8. The read ahead data caching with ROOT TTreeCache is used and configured with a training phase of 10 events and 10 MB cache size. In addition the settings `TTREECACHE_PREFILL=1` and `TTREECACHE_SIZE=1` have been used, to mitigate and bypass the missing pre-fetch buffer mode in Davix 0.4.0.

The different analyses have been repeated several times and are accessing the same input files at the dCache storage element of LRZ-LMU through different access protocols: NFS, dcap, XRootD and WebDAV. Table 3 shows the measured event rates for the different analysis types and the different access protocols. Within the uncertainties of $\sim 10\%$ there is no significant performance difference visible among the different protocols.

Table 3. Measured analysis event rates for different analysis types and local access at LRZ-LMU using different access protocols. The measurements have been repeated several times and uncertainties are $\sim 10\%$.

class access mode	Slow	Medium	Fast	Fast w/o init
local/nfs [Hz]	80	210	195	200
dcap	75	196	147	155
XRootD	69	193	182	194
Davix	72	192	190	205
branch access mode				
local/nfs [Hz]	89	230	550	1050
dcap	94	228	495	850
XRootD	86	211	497	830
Davix	84	205	483	815

Figure 3 shows the access transfer rates for different protocols and access modes for the medium analysis (left) and the fast analysis (right). Reading throughput peaks at 10-12 MB/s are achieved for the medium analysis and no significant differences are visible for the branch and class access mode. For the fast analysis where only very small amount of physics variables are used there is a larger difference visible in the throughput rate of 10-16 MB/s for the branch vs. 15-30 MB/s for the class access mode. No significant differences are visible among the different access protocols in all these tests.

In a second set of tests the analyses have been repeatedly used in several HammerCloud tests and executed in PanDA jobs at LRZ-LMU, FZK, Prague/FZU, BNL, Lyon and Oslo. The default PanDA access modes are dcap for the dCache SEs at LRZ-LMU and FZK, copy-to-scratch at the DPM site Prague/FZU, copy-to-scratch at the dCache site BNL, XRootD at the dCache site Lyon and copy-to-scratch at the ARC site Oslo. Figure 4 shows the overall performance of the fast analysis at all sites for the default PanDA access and the WebDAV access. At the time the test was executed the WebDAV access at BNL was not fully working, which explains the higher failure rate for the WebDAV tests and the lower WebDAV average event rate. Within the errors there is no significant difference visible for the default PanDA

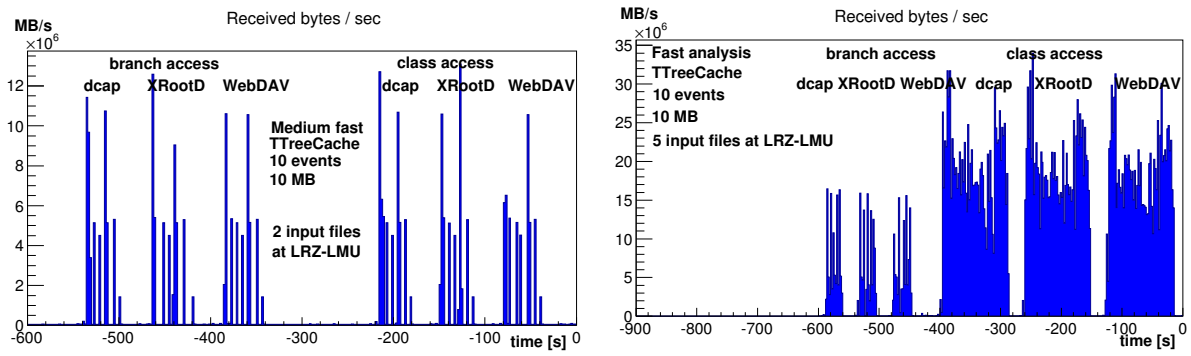

Figure 3. Access transfer rates for different protocols and access modes for the medium analysis (left) and the fast analysis (right).

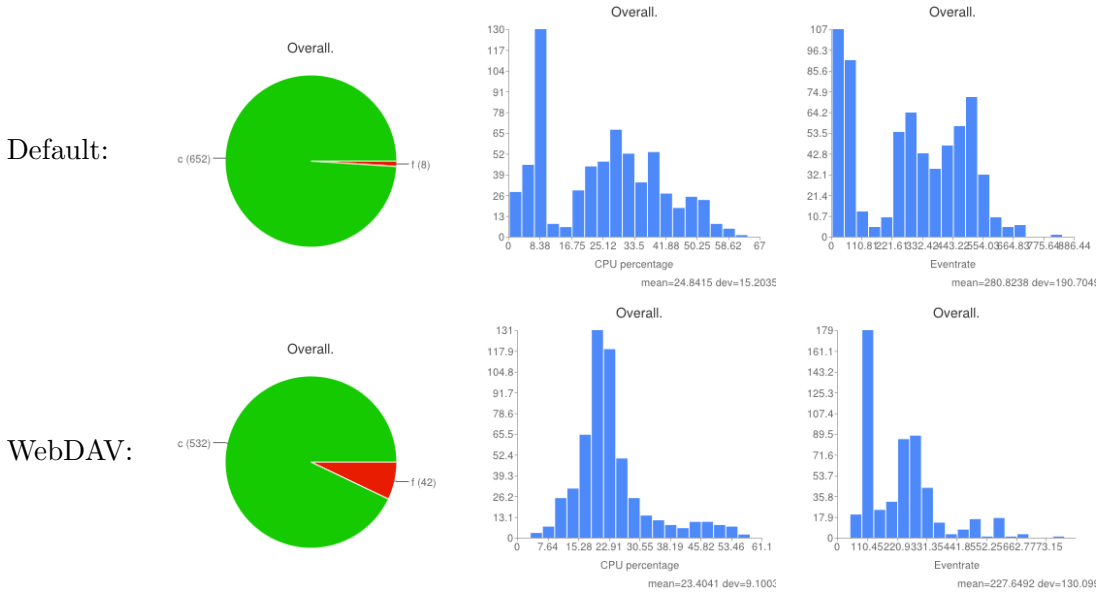

Figure 4. Overall performance of the fast analysis at all sites for the default PanDA access (top) and the WebDAV access (bottom). Shown are the fraction of successful and failed jobs (left), the CPU utilisation fraction (middle) and the event rate (right).

access modes and the WebDAV access mode.

Figure 5 shows the event rate for the fast analysis for LRZ-LMU, FZK and Prague/FZU for different access modes. For LRZ-LMU the event rate is the same for all three protocols within the uncertainties. The CPU utilisation fraction is 24±15%. At FZK there is a slight performance drop for WebDAV visible compared to the dcap and XRootD access modes. At Prague/FZU, WebDAV and XRootD show similar performance. The copy-to-scratch result at this site is too optimistic since the initial time to transfer the full file to the worker node is missing in this statistics of the plot. Figure 6 shows the event rate for remote reading from LRZ-LMU for jobs at BNL, FZK, and Prague/FZU. As a gauge, the local access at LRZ-LMU is also visible. As expected for the remote access to LRZ-LMU the event rate drops significantly with a CPU utilisation fraction of 20±7% but due to the TTreeCache and vector reading access the performance is still reasonable.

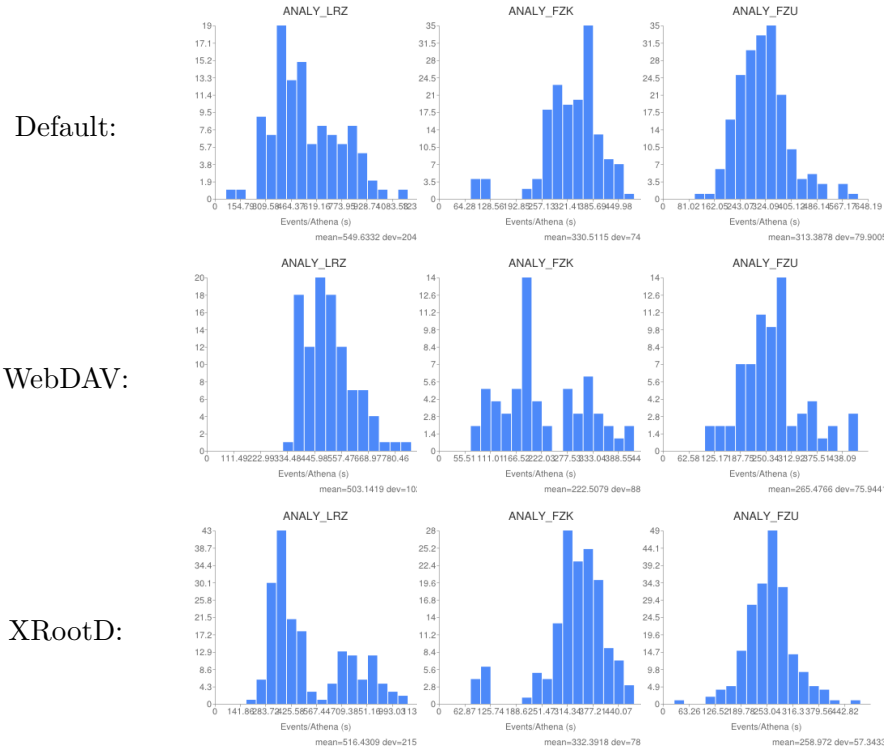

Figure 5. The event rate for the fast analysis for LRZ-LMU (left), FZK (middle) and Prague/FZU (right). Shown are the default PanDA access modes (top), WebDAV (middle) and XRootD (bottom).

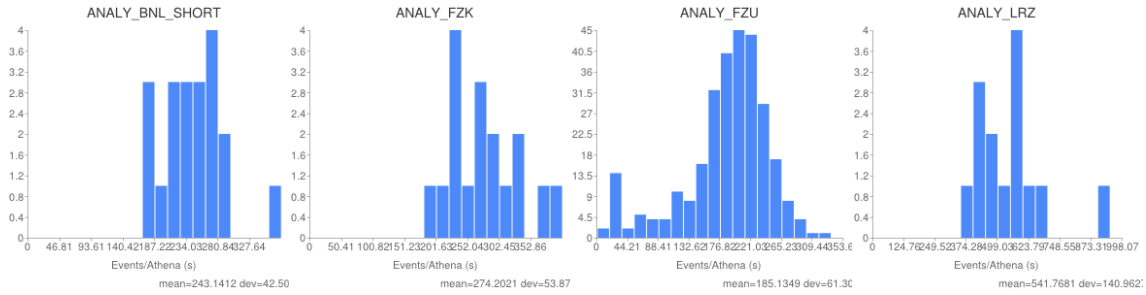

Figure 6. The event rate of the fast analysis for remote reading from LRZ-LMU for jobs at BNL (left), FZK (middle left) and Prague/FZU (middle right). As a reference the local reading at LRZ-LMU is shown as well (right).

Figure 7 shows the event rate for local reading at LRZ-LMU, FZK and Prague/FZU for the default, WebDAV and XRootD access with a CPU utilisation fraction of 63±14%. No significant differences in the event rates are observed. Again the copy-to-scratch result at Prague/FZU is too optimistic since the initial time to transfer the full file to the worker node is missing in this statistics of the plot.

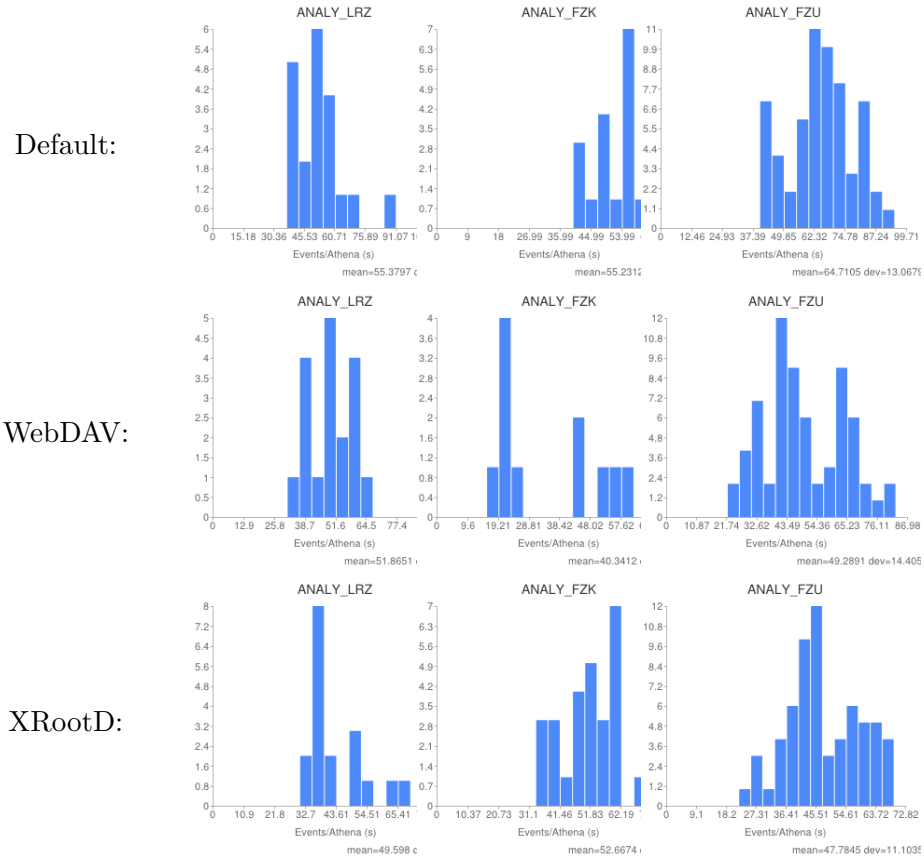

Figure 7. The event rate of the slow analysis at LRZ-LMU (left), FZK (middle) and Prague/FZU (left) for the default access (top), WebDAV (middle) and XRootD (bottom).

5. Conclusions

This paper presented the setup of the WebDAV usage within the ATLAS PanDA workload management system. The key components are the Rucio re-director, aria2c and Davix for data access. An iterative process was necessary to stabilise the different components and to have a fully functional workflow at many sites. As shown there are no differences in the tested different data access protocol visible within the uncertainties and scale of the tests. All in all WebDAV is a very good candidate for a unified access mode at WLCG sites and beyond.

References

[1] ATLAS Collaboration 2008 JINST **3** S08003
[2] ATLAS Collaboration 2005 CERN-LHCC-2005-022, https://cds.cern.ch/record/837738
[3] Maeno T on behalf of the ATLAS Collaboration 2008 J. Phys. Conf. Ser. **119** 062036
[4] Garonne V *et al.* on behalf of the ATLAS Collaboration 2014 J. Phys. Conf. Ser. **513** 042021
[5] van der Ster D C, Elmsheuser J, Ubeda Garcia M and Paladin M 2011 J. Phys. Conf. Ser. **331** 072036
[6] Antcheva I *et al.* 2009 Comput. Phys. Commun. **180** 2499
[7] Davix project webpage: https://dmc.web.cern.ch/projects/davix/home
[8] Gardner R *et al.* on behalf of the ATLAS Collaboration 2014 J. Phys. Conf. Ser. **513** 042049.
[9] Aria2c project webpage: http://aria2.sourceforge.net/
[10] Anisenkov A *et al.* on behalf of the ATLAS Collaboration 2014 J. Phys. Conf. Ser. **513** 032001
[11] Blomer J *et al.* 2012 J. Phys. Conf. Ser. **396** 052013
[12] Nilsson P *et al.* on behalf of the ATLAS Collaboration 2014 J. Phys. Conf. Ser. **513** 032071
[13] Maier T *et al.*, these proceedings