New data access with HTTP/WebDAV in the ATLAS experiment

Johannes Elmsheuser
on behalf of the ATLAS collaboration

Ludwig-Maximilians-Universität München

14 April 2015
21st International Conference on Computing in High Energy and Nuclear Physics (CHEP2015)

Co-authors: Rodney Walker, Cedric Serfon, Sylvain Blunier, Vincenzo Lavorini, Paul Nilsson

Outline

Introduction

Fast and reliable access in reasonably long jobs essential for any kind of (HEP) data analysis job

ATLAS Grid setup:

- Workload management system (PanDA) with pilot jobs
- Input data access is configured on a site by site basis depending on the storage element type:
  - dcap, copy-to-scratch, file, xrootd, FAX (federated xrootd)
- DPM usually copy-to-scratch or lately xrootd
- dCache uses usually dcap

→ Explore webdav/http input access:

- Industry standard
- Same client and access protocol everywhere
- Use aria2c for stage-in
- Use Davix plugin for ROOT I/O

PanDA data access

2 access modes in analysis jobs:

- PanDA analysis usually consists of two job types: one build job for source code compilation and many subsequent analysis jobs processing the input data:
- First local stage-in of analysis source code dataset from previous build jobs → aria2c
- Then ROOT/Athena (ATLAS offline software) event loop with local (or remote) input ROOT file access → Davix

Setup:

- aria2c stage-in has been successfully used at a couple of DE sites for Monte Carlo production for some time
- Davix: remote I/O library for Webdav/http/S3 including ROOT I/O plugin
- ATLAS data management system (Rucio) provides a redirector with very easy syntax
- Redirection rules per site in ATLAS grid information system (AGIS)

Workflow schema

Rucio redirector

Easy to construct TURL with additional optional parameters
To configure site choice (site, geoip, spacetoken):
https://rucio-lb-prod.cern.ch/redirect/mc14_8TeV/AOD.01507240._010001.pool.root.2?site=LRZ-LMU

Redirector returns actual local webdav TURL:
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

Job
on grid site
worker node

Application:
ROOT
Athena
with Davix
or
standalone
aria2c

Local or
Remote
SE

PanDA server

Total Rucio: 705 endpoints, 159.8 PB, 637 mio files

Webdav accessible: 442 endpoints, 82.6 PB, 491 mio files

Outline

Setup Rucio and PanDA I

Step I: Information system:

- Mapping of sitename vs. webdav door,
  e.g. LRZ-LMU: https://lcg-lrz-dc66.grid.lrz.de
- Verify information system webdav configurations
- ATLAS DDM started in the past in SE renaming campaign
- Full table: `http://atlas-agis.cern.ch/agis/service/table_view/?type=SE/HTTP&state=ACTIVE`

Step IIa: PanDA pilot update - aria2c:

- Add aria2c wrapper method to PanDA pilot to create and use metalink xml file for source code download in the job:

```
<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/
   </url>
   </file>
</metalink>
```

- Since release 1.18.10 support adler32 checksums

Setup Rucio and PanDA II

Step IIb: PanDA pilot update - ROOT with Davix:

- Iterated with Davix developers over several versions, current version 0.4.0 working stable and fine
- Deployed and enabled Davix library with ATLAS ROOT installation in CVMFS
- Usage: `TFile::Open("https://myse/scope/myfile.root")`
- Rucio allows together with Davix to redirect to any webdav configured ATLAS SE
- Options:
  - `?site=LRZ-LMU`
  - `?rse=LRZ-LMU_LOCALGROUPDISK`
  - `?select=geoip` to select best replica
- Currently in use in PanDA pilot: `?site=LRZ-LMU`
- Add a configurable wrapper to create TURL input file list for ROOT/Athena job:

```
...
https://rucio-lb-prod.cern.ch/redirect/data12_8TeV/NTUP_COMMON.01306922._000007.root.1?site=LRZ-LMU
https://rucio-lb-prod.cern.ch/redirect/data12_8TeV/NTUP_COMMON.01306922._000013.root.1?site=LRZ-LMU
...
```

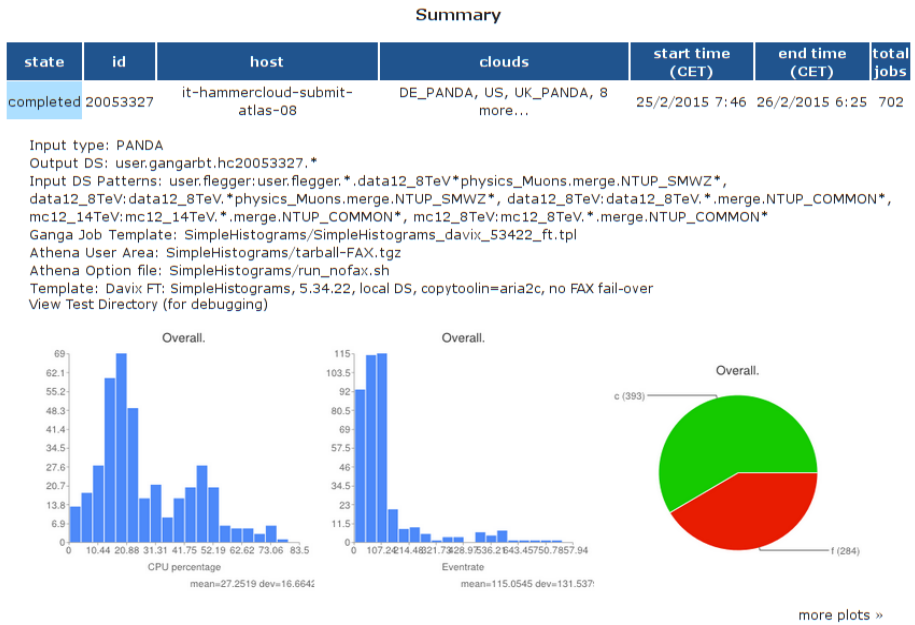
- Future: Option for metalink usage - this is used in the PanDA aria2c sitemover - allows multiple streams from different sites & automatic failover

See also talks by Vincent Garonne on Rucio (#205) and Tadashi Maeno on PanDA (#144)

Webdav functional tests with PanDA

Step III: Continuous HammerCloud tests:

- Probing aria2c/Davix access to local SE via webdav at more than 70 sites using a HammerCloud functional test with ATLAS PanDA jobs
- Short PyROOT 5.34.22 job that reads a couple of variables from ROOT ntuple
- Iterative improvement process - about 1/3 of site systematically fail due to old SE software versions or configurations
  → update needed at site level



See also poster by Michael Böhler about HammerCloud (#159)

Outline

Access protocol comparisons

Comparison of different input file protocols:

- Comparison of nfs, dcap, xrootd/FAX, webdav/Davix
- Use "real-life" analysis based on new ATLAS xAOD data format, EventLoop frameworks, reads/processes muons, jets, stores some output
- Local tests at LRZ-LMU with access to dCache SE and in Grid using PanDA jobs at different sites
- Using different analysis modes: Rather CPU intensive analysis and rather fast I/O analysis with ROOT 5.34.24-x86_64-slc6-gcc48-opt
- Access with xAOD `class` or `branch` access mode
- `TTreeCache` is enabled with 10 events/10 MB learning in code
- Using in addition: `ROOT_TTREECACHE_PREFILL=1` `ROOT_TTREECACHE_SIZE=1` to "patch" missing pre-fetch buffer in Davix

See also talks by Thomas Maier (#171), Scott Snyder (#182), James Catmore (#164)

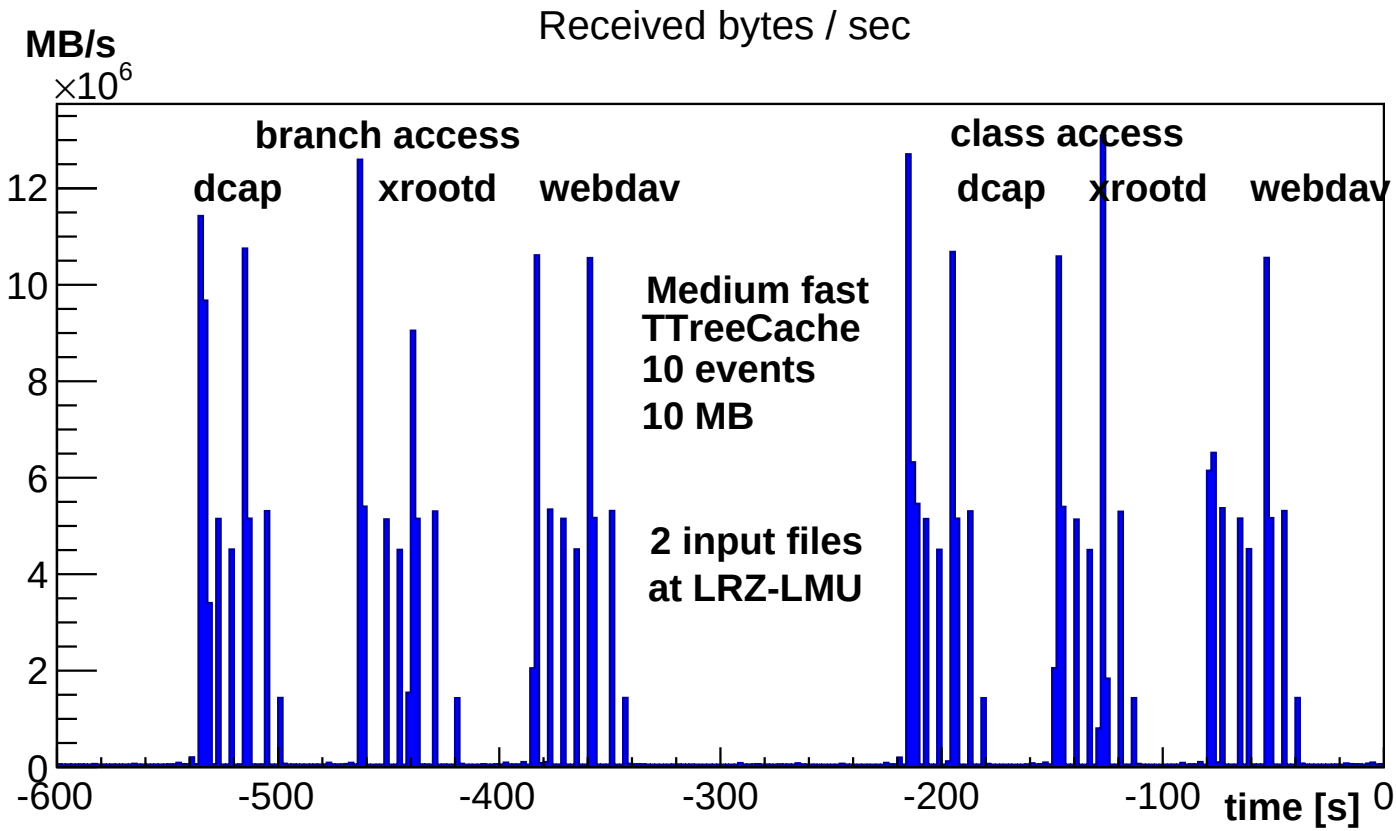
Local access to LRZ-LMU

- Measured event rate for local access at LRZ-LMU
- Repeated several times, uncertainties $\sim 10\%$

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

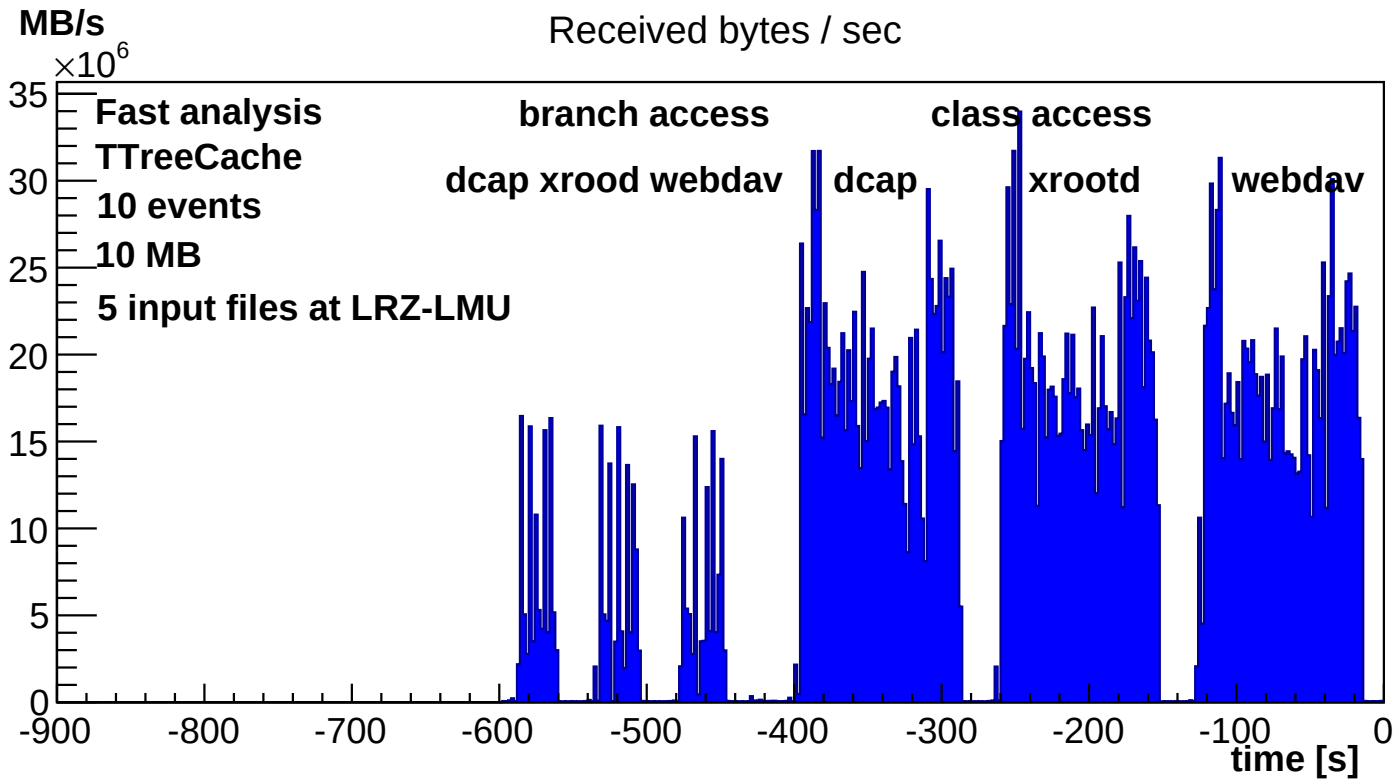
- Davix suffers from missing buffering in start-up - PREFILL=1 cures the problem
- No differences in access modes within uncertainties and scale of tests

Local access to LRZ-LMU II - eth0 rate



10-12 MB/s throughput, no large difference between branch/class access, reading spike for every file open, not much protocol difference

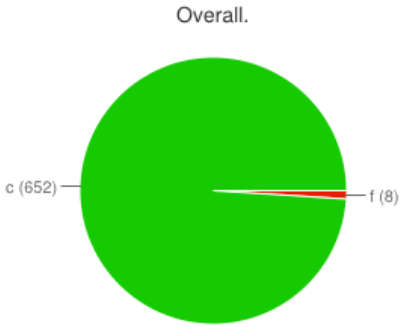
Local access to LRZ-LMU III - eth0 rate

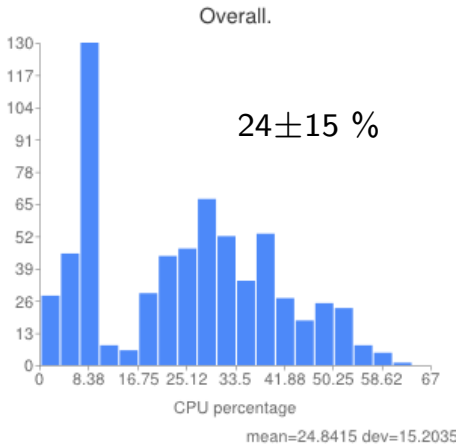


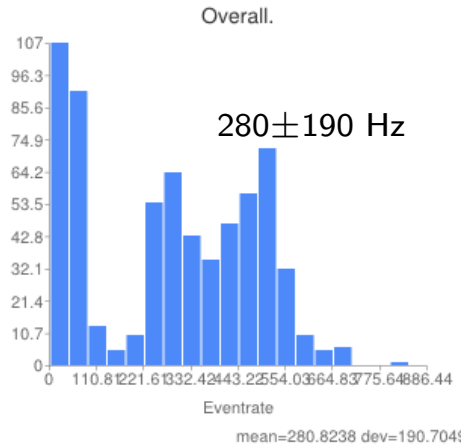
10-16 and 15-30 MB/s throughput, large difference between branch/class access, not much protocol difference

HammerCloud - Fast analysis, local replica I

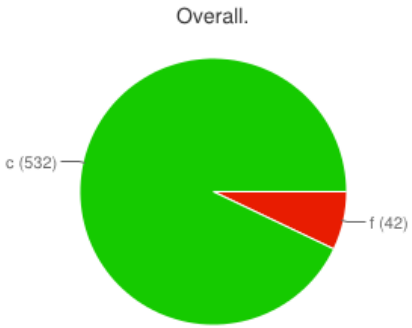
Default access:

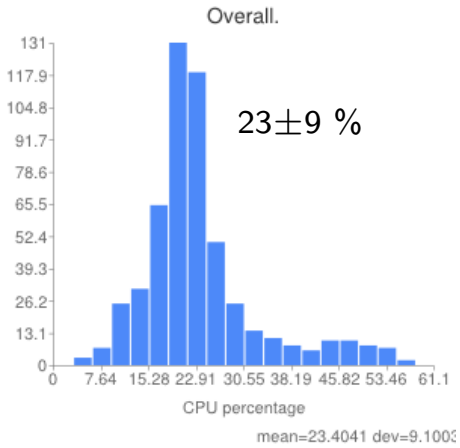


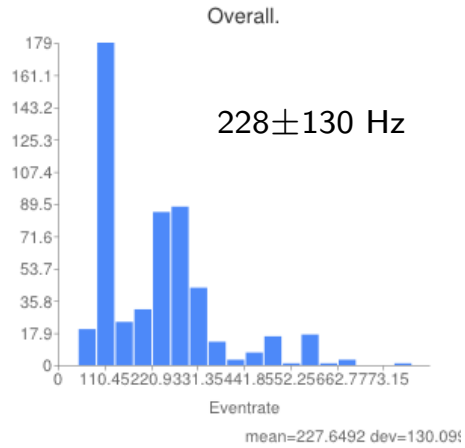




Webdav access:





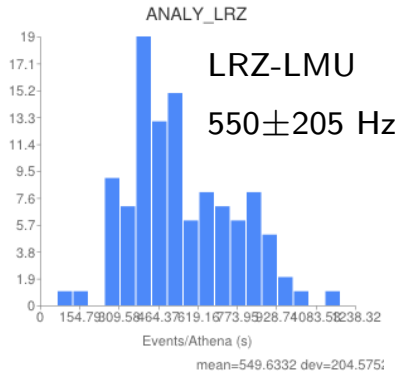


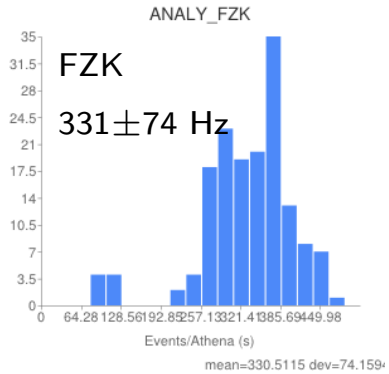
LRZ-LMU (dcap), FZK (dcap), Prague/FZU (DPM/copy-to-scratch)
(BNL (failed in the webdav test), IN2P3-CC, UIO)

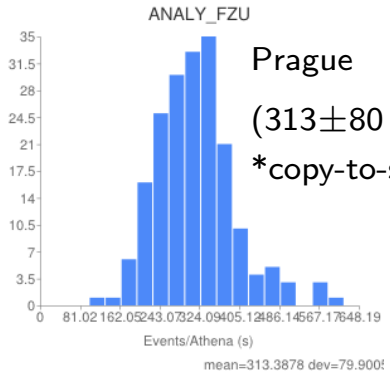
HammerCloud - Fast analysis, local replica II

CPU: $\sim$ 24 $\pm$ 15%

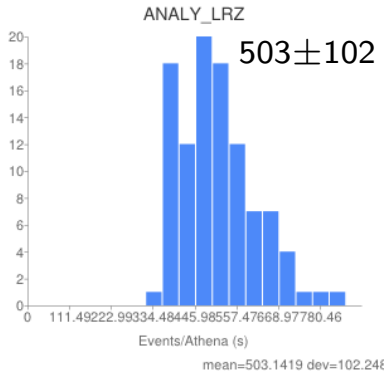
Default:

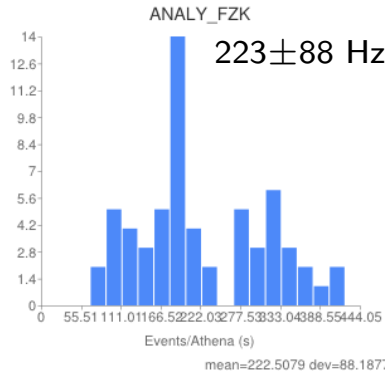


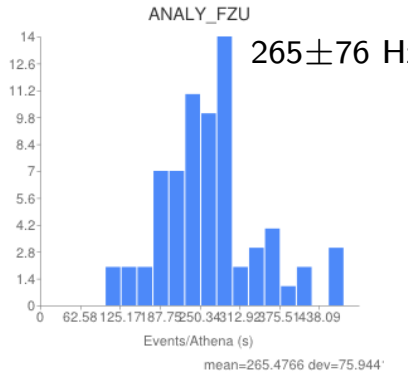




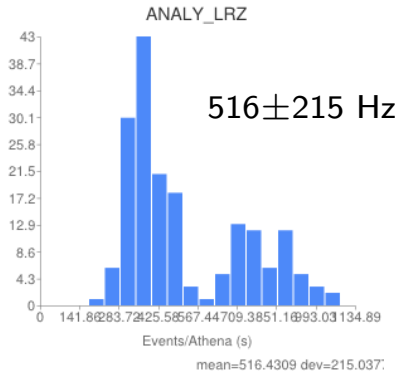
Webdav:

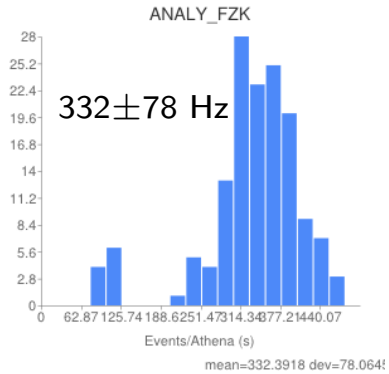


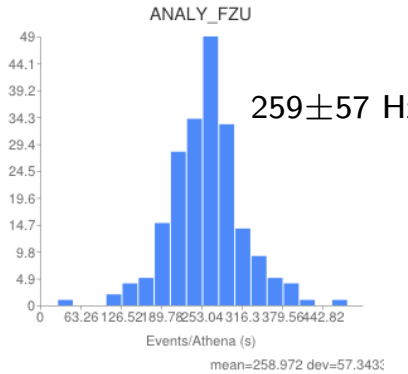




XRootD:



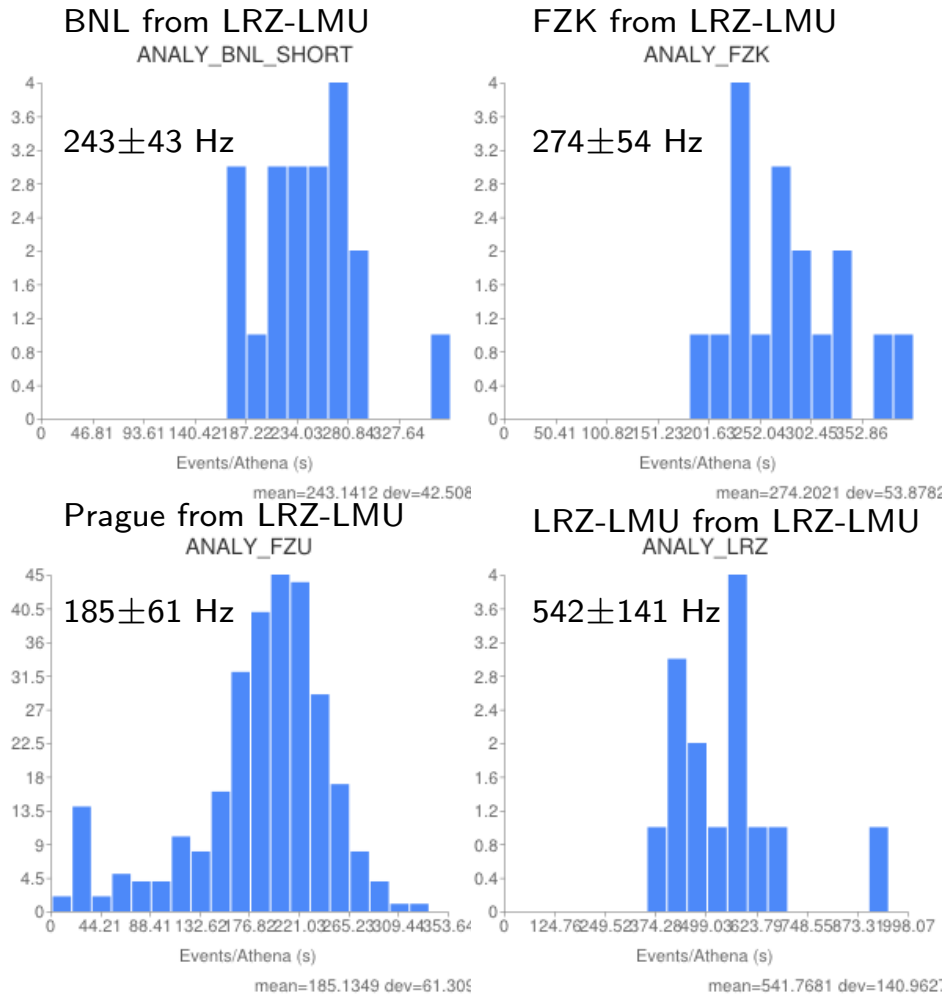




HammerCloud - Fast analysis, Replica at LRZ

Webdav access to remote replica at LRZ-LMU

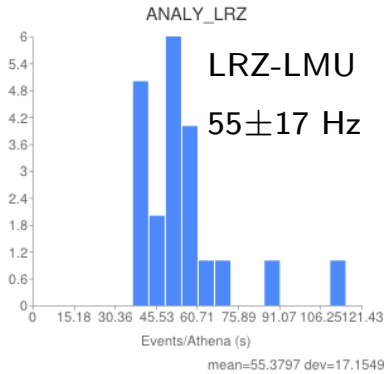
CPU: $\sim 20 \pm 7\%$

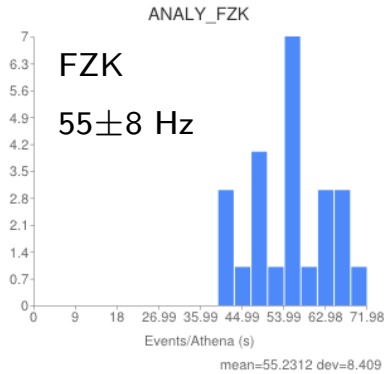


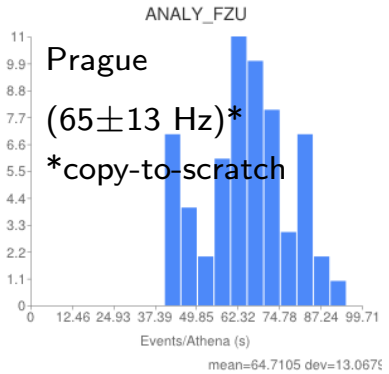
HammerCloud - Slow analysis, local replica

CPU: $\sim 63 \pm 14\%$

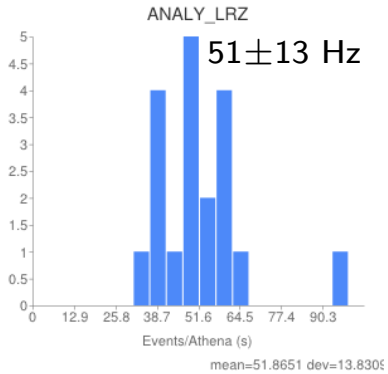
Default:

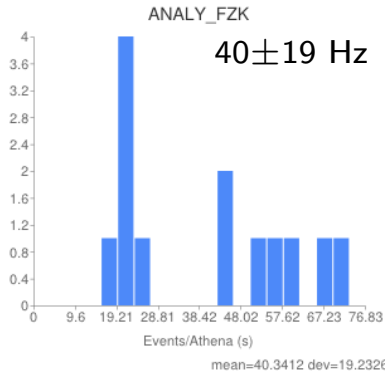


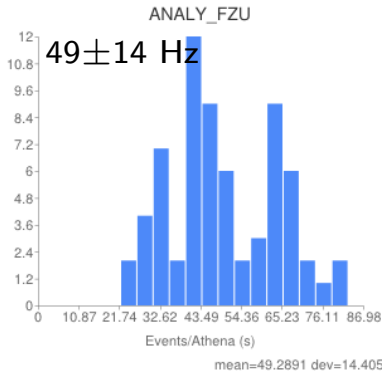




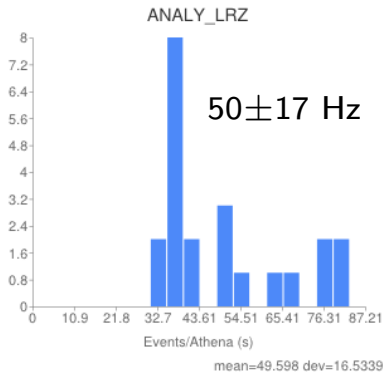
Webdav:

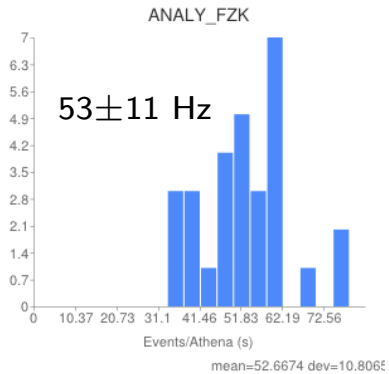


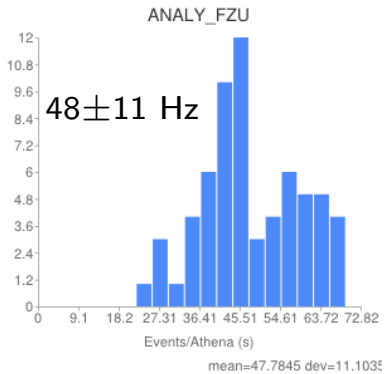




XRootD:







Summary and Conclusions

- Presented the setup and how to use webdav access for PanDA job in the ATLAS experiment
- aria2c and Davix together with Rucio redirector key components
- Several and long process of iterations to stabilise external code components
- Now webdav is working fine at up-to-date sites
- No differences in tested access protocols visible within uncertainties and scale of tests

$\rightarrow$ Webdav is very good candidate for unified access mode at HEP sites