

# NEW DATA ACCESS WITH HTTP/WEBDAV IN THE ATLAS EXPERIMENT

Johannes Elmsheuser  
on behalf of the ATLAS collaboration

Ludwig-Maximilians-Universität München

13 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

- ① INTRODUCTION
- ② SETUP STATUS AND WEBDAV PANDA FUNCTIONAL TESTS WITH HAMMERCLOUD
- ③ ACCESS PROTOCOL COMPARISONS

# 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

## 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



Easy to construct TURL with additional optional parameters

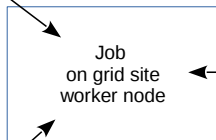
To configure site choice (site, geiop, spacetoken):

[https://rucio-lb-prod.cern.ch/redirect/mc14\\_8TeV/AOD.01507240\\_010001.pool.root.2?site=LRZ-LMU](https://rucio-lb-prod.cern.ch/redirect/mc14_8TeV/AOD.01507240_010001.pool.root.2?site=LRZ-LMU)

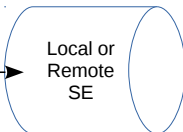


Redirector returns actual local webdav TURL:

[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](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)



Application:  
ROOT  
Athena  
with Davix  
or  
standalone  
aria2c



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

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

- ① INTRODUCTION
- ② SETUP STATUS AND WEBDAV PANDA FUNCTIONAL TESTS WITH HAMMERCLOUD
- ③ ACCESS PROTOCOL COMPARISONS

# 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](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

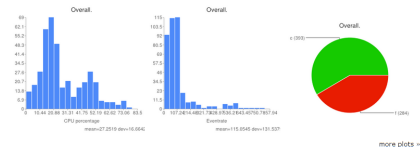
## Step III: Continuous HammerCloud tests:

- 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

Summary

state	id	host	clouds	start time (CET)	end time (CET)	total jobs
completed	20055327	it-hammercloud-submit-atlas-08	DE_PANDA, US, UK_PANDA, 8 more...	25/2/2015 7:46	26/2/2015 6:25	702

Input type: PANDA  
Output DS: user.gangarbit.hc20053327.\*  
Input DS Patterns: user.flegger.user.flegger.\*.data12\_8TeV\*physics\_Muons.merge.NTUP\_SMWZ\*, data12\_8TeV:data12\_8TeV\*physics\_Muons.merge.NTUP\_SMWZ\*, data12\_8TeV:data12\_8TeV\*.merge.NTUP\_COMMON\*, mc12\_14TeV:mc12\_14TeV\*.merge.NTUP\_COMMON\*, mc12\_8TeV:mc12\_8TeV\*.merge.NTUP\_COMMON\*  
Ganga Job Template: SimpleHistograms/SimpleHistograms\_davix\_53422\_ft.tpl  
Athena User Area: SimpleHistograms/tarball-FAX.tgz  
Athena Option file: SimpleHistograms/run\_noFAX.sh  
Template: Davix FT: SimpleHistograms, 5.34.22, local DS, copytool=aria2c, no FAX fail-over  
View Test Directory (for debugging)



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

- ① INTRODUCTION
- ② SETUP STATUS AND WEBDAV PANDA FUNCTIONAL TESTS WITH HAMMERCLOUD
- ③ ACCESS PROTOCOL COMPARISONS

# 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\_TTREETCACHE\_PREFILL=1  
ROOT\_TTREETCACHE\_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  $\sim 10\%$

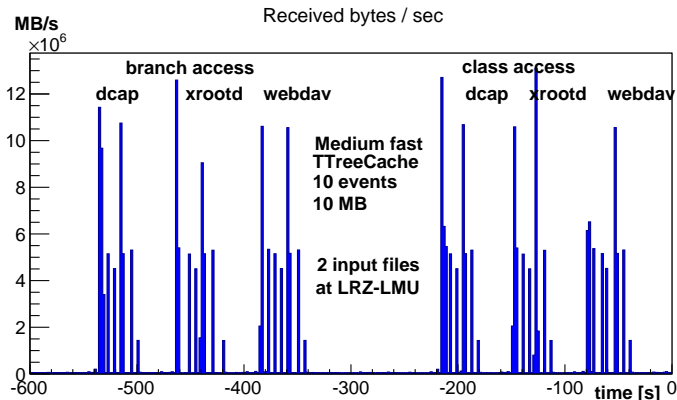
class access mode	Slow	Medium	Fast	Fast w/o init
local/nfs [Hz]	80	210	195	200
dcap	75	196	147	155
FAX	69	193	182	194
Davix	72	192	190	205

---

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

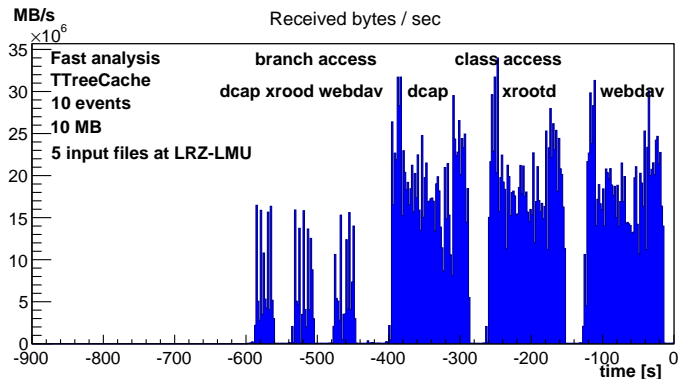
- Davix suffers from missing buffering in start-up - PREFILL=1 cures the problem
- No differences in access modes within uncertainties and scale of tests

# 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

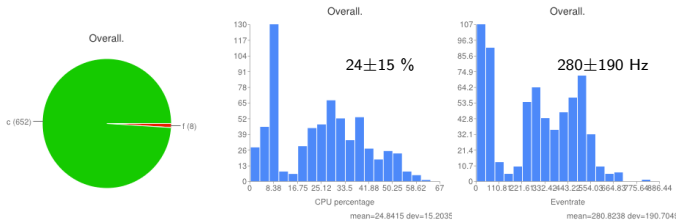
# LOCAL ACCESS TO LRZ-LMU III - ETH0 RATE



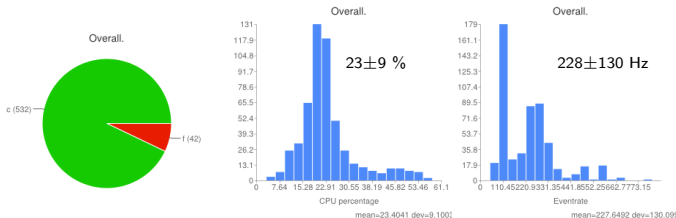
10-16 and 15-30 MB/s throughput, large difference between branch/class access, not much protocol difference

# HAMMERCLOUD - FAST ANALYSIS, LOCAL REPLICAS I

## Default access:



## Webdav access:

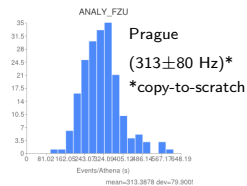
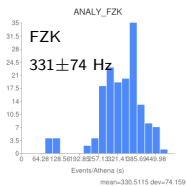
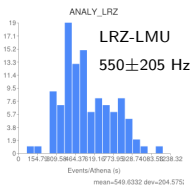


LRZ-LMU (dcap), FZK (dcap), Prague/FZU (DPM/copy-to-scratch)  
(BNL (failed in the webdav test), IN2P3-CC, UIO)

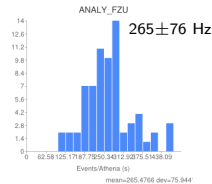
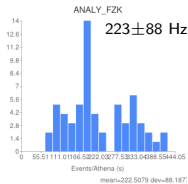
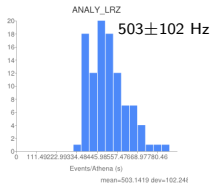
# HAMMERCLOUD - FAST ANALYSIS, LOCAL REPLICA II

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

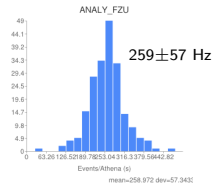
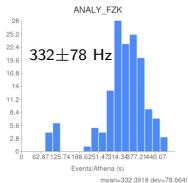
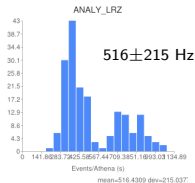
Default:



Webdav:



XRootD:

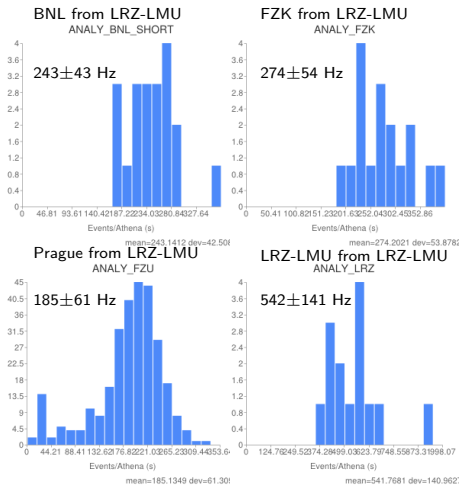




# HAMMERCLOUD - FAST ANALYSIS, REPLICA AT LRZ

## Webdav access to remote replica at LRZ-LMU

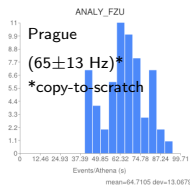
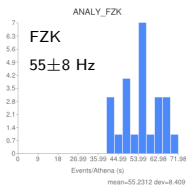
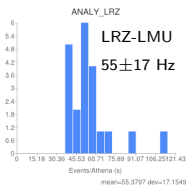
CPU:  $\sim 20 \pm 7\%$



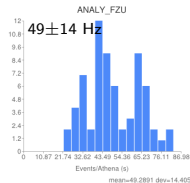
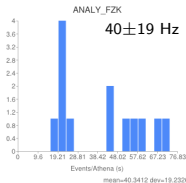
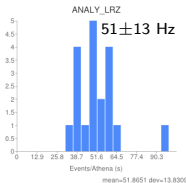
# HAMMERCLOUD - SLOW ANALYSIS, LOCAL REPLICA

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

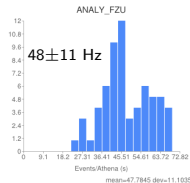
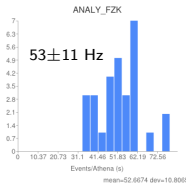
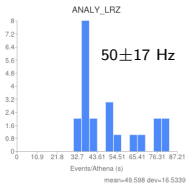
Default:



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

→ Webdav is very good candidate for unified access mode at HEP sites