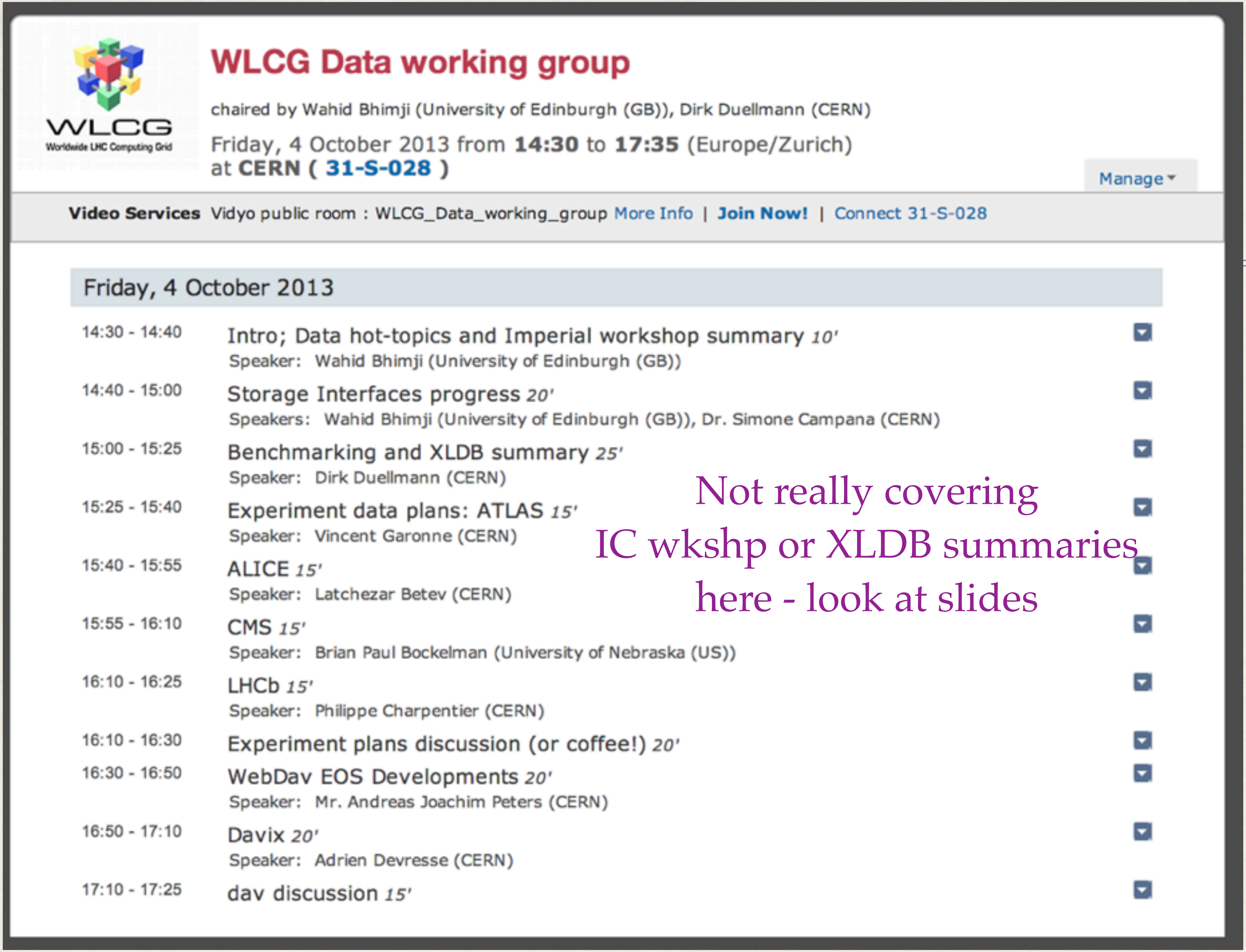
# WLCG Data Working Group (hastily prepared) Summary

Wahid Bhimji

*9th October 2013*

# Introduction / history

- Last WG meeting we agreed to merge the “Storage Interfaces” and “Benchmarking” working groups and broaden remit to the main data issues within WLCG. WLCG MB agreed with this.
  - Backup slides for some extra topics - suggestions welcome
- “ongoing”, “lightweight” WG, specific topics should have finite timescale, report to GDB and be passed for actions to WLCG ops WG.
- Membership: experiments, sites and developers, experts on topics
  - Small and not necessarily complete, important items go to GDB.
- Fri Agenda: https://indico.cern.ch/conferenceDisplay.py?confId=273201



# WLCG Data working group

chaired by Wahid Bhimji (University of Edinburgh (GB)), Dirk Duellmann (CERN)

Friday, 4 October 2013 from **14:30** to **17:35** (Europe/Zurich)
at **CERN ( 31-S-028 )**

**Video Services** Vidyo public room : WLCG_Data_working_group More Info | **Join Now!** | Connect 31-S-028

## Friday, 4 October 2013

| | |
|---|---|
| 14:30 - 14:40 | Intro; Data hot-topics and Imperial workshop summary *10'*<br>Speaker: Wahid Bhimji (University of Edinburgh (GB)) |
| 14:40 - 15:00 | Storage Interfaces progress *20'*<br>Speakers: Wahid Bhimji (University of Edinburgh (GB)), Dr. Simone Campana (CERN) |
| 15:00 - 15:25 | Benchmarking and XLDB summary *25'*<br>Speaker: Dirk Duellmann (CERN) |
| 15:25 - 15:40 | Experiment data plans: ATLAS *15'*<br>Speaker: Vincent Garonne (CERN) |
| 15:40 - 15:55 | ALICE *15'*<br>Speaker: Latchezar Betev (CERN) |
| 15:55 - 16:10 | CMS *15'*<br>Speaker: Brian Paul Bockelman (University of Nebraska (US)) |
| 16:10 - 16:25 | LHCb *15'*<br>Speaker: Philippe Charpentier (CERN) |
| 16:10 - 16:30 | Experiment plans discussion (or coffee!) *20'* |
| 16:30 - 16:50 | WebDav EOS Developments *20'*<br>Speaker: Mr. Andreas Joachim Peters (CERN) |
| 16:50 - 17:10 | Davix *20'*<br>Speaker: Adrien Devresse (CERN) |
| 17:10 - 17:25 | dav discussion *15'* |

Not really covering
IC wkshp or XLDB summaries
here - look at slides

# Storage Interfaces

- “Storage Interfaces” topic concerns move away from SRM for disk-only sites.
- Topic now includes new uses of the replacement interfaces (dav / xrootd etc.) as well as data access interfaces.
- Focus on action items:
    - For SRM these were: gFal2 for deletion; and service discovery Spacetokens, ATLAS SRM-free (demo) site;
    - Very short summary here - see talk for more details.

# SRM: Progress

- ATLAS exploring dav deletion (with/without gFal2) with Rucio
- Spacetokens:
    - Atlas use decreasing (could use only rucio quotas but needs devel.). Will also persue namespace only option with DPM
    - Still need to capture non-ATLAS ST use-cases and make sure they are covered
- ATLAS non-SRM site:
    - FTS pure gridftp transfers needs some work for performance on some servers side. http and xrootd are starting to be tested .
    - Stage-out: testing DAV put with spacetoken
- LHCb developing xrootd/http in Dirac and using FTS3 (as mentioned this morning)

# Other storage interfaces

- xrootd and dav increasing in availability and use
  - Need these as services in GOCDB- something for ops WG to push ?
- Rfio retirement (on DPM) in favour of above interfaces progressing "naturally" but
  - We think its time for a "deadline" - e.g end of year ?

# Benchmarking

- Activity on EOS monitoring already interesting results - stalled due student leaving - new student from Nov.
  - Log collection - also interface for xrootd federation records
  - micro-benchmarks will be developed
- Also some manpower within DPM for this in future

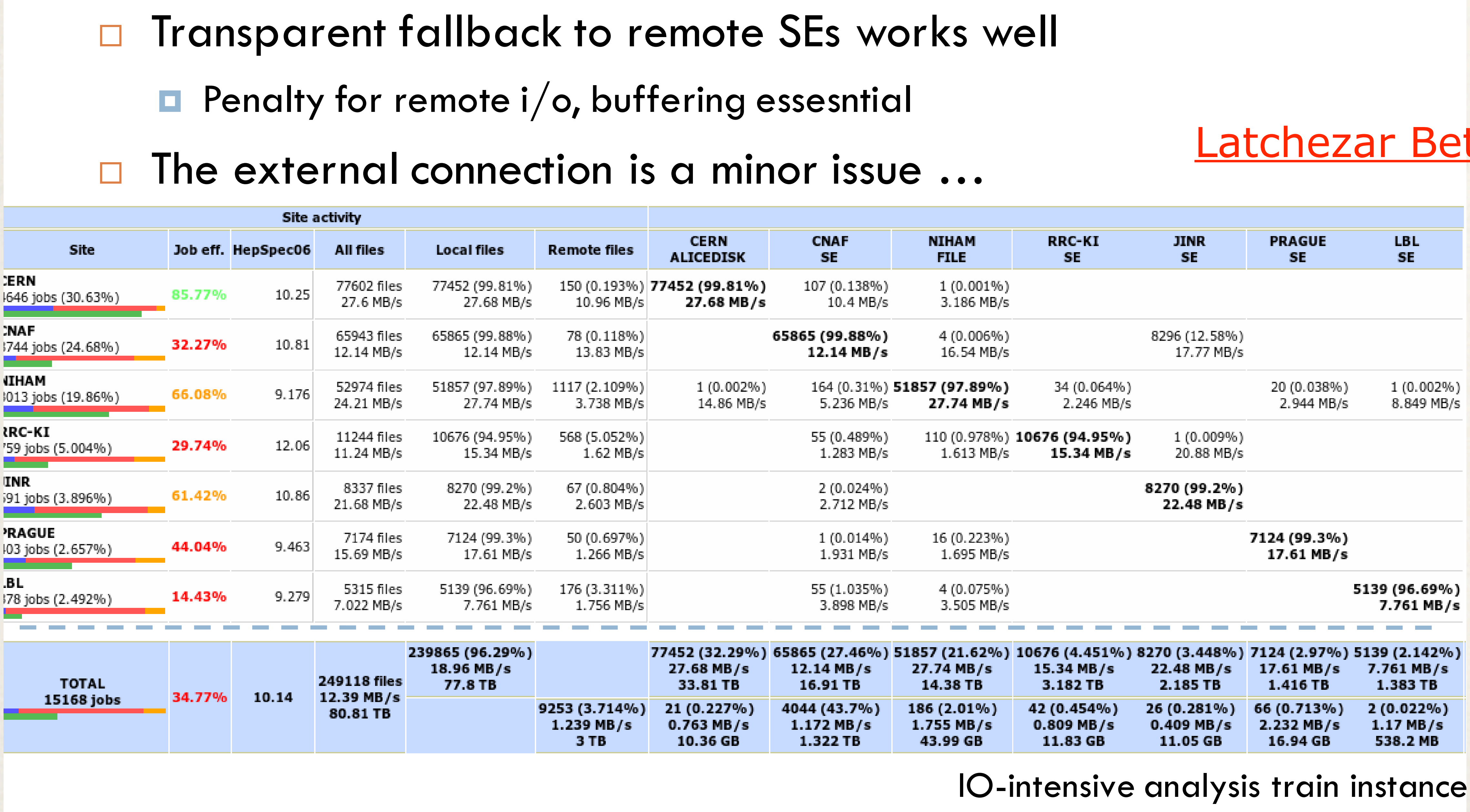
# ATLAS - rucio

Vincent Garonne's slides

- DQ2 will not scale for Run2 -> Rucio. Target in prod by early 2014
- Better quotas; move towards open / widely-adopted protocols.
- Renaming campaign (mostly using Dav) - going well .
- Rucio integration testbed with upload, FTS3, deletion
  - Multiple access protocols - priority to webdav.
  - spacetoken not required if alternative for space usage
    - WG will work on reasonable "common" method for publishing

# ALICE - WAN transfers

- Uniform use of xrootd; Lots of interesting real data from alimonitor.



- □ Transparent fallback to remote SEs works well
  - Penalty for remote i/o, buffering essesntial
- □ The external connection is a minor issue …

Latchezar Betev Talk

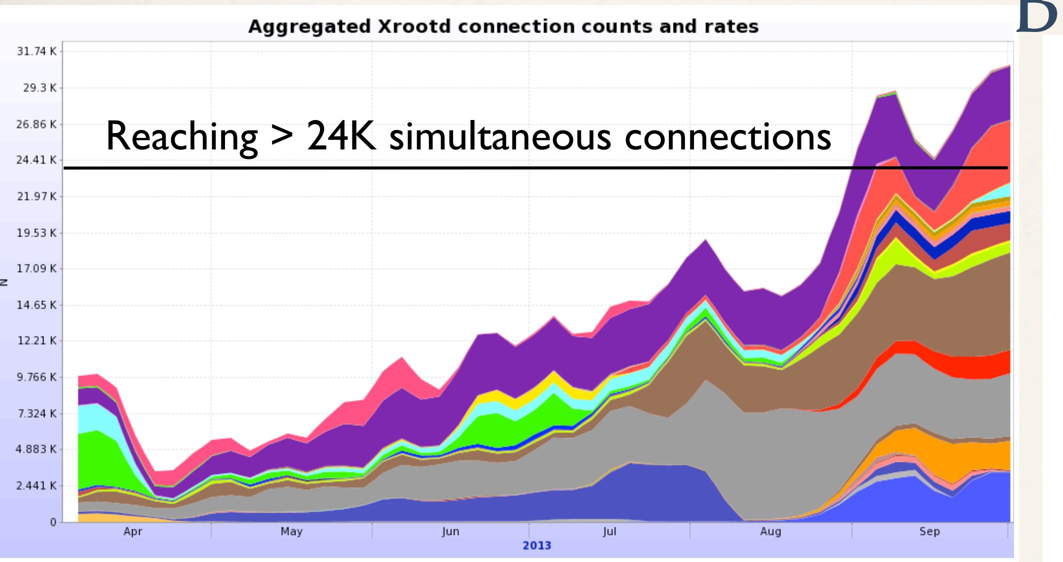
| Site activity | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Site | Job eff. | HepSpec06 | All files | Local files | Remote files | CERN ALICEDISK | CNAF SE | NIHAM FILE | RRC-KI SE | JINR SE | PRAGUE SE | LBL SE |
| ERN<br>646 jobs (30.63%) | 85.77% | 10.25 | 77602 files<br>27.6 MB/s | 77452 (99.81%)<br>27.68 MB/s | 150 (0.193%)<br>10.96 MB/s | **77452 (99.81%)<br>27.68 MB/s** | 107 (0.138%)<br>10.4 MB/s | 1 (0.001%)<br>3.186 MB/s | | | | |
| NAF<br>744 jobs (24.68%) | 32.27% | 10.81 | 65943 files<br>12.14 MB/s | 65865 (99.88%)<br>12.14 MB/s | 78 (0.118%)<br>13.83 MB/s | | **65865 (99.88%)<br>12.14 MB/s** | 4 (0.006%)<br>16.54 MB/s | | 8296 (12.58%)<br>17.77 MB/s | | |
| IHAM<br>013 jobs (19.86%) | 66.08% | 9.176 | 52974 files<br>24.21 MB/s | 51857 (97.89%)<br>27.74 MB/s | 1117 (2.109%)<br>3.738 MB/s | 1 (0.002%)<br>14.86 MB/s | 164 (0.31%)<br>5.236 MB/s | **51857 (97.89%)<br>27.74 MB/s** | 34 (0.064%)<br>2.246 MB/s | | 20 (0.038%)<br>2.944 MB/s | 1 (0.002%)<br>8.849 MB/s |
| RC-KI<br>59 jobs (5.004%) | 29.74% | 12.06 | 11244 files<br>11.24 MB/s | 10676 (94.95%)<br>15.34 MB/s | 568 (5.052%)<br>1.62 MB/s | | 55 (0.489%)<br>1.283 MB/s | 110 (0.978%)<br>1.613 MB/s | **10676 (94.95%)<br>15.34 MB/s** | 1 (0.009%)<br>20.88 MB/s | | |
| INR<br>91 jobs (3.896%) | 61.42% | 10.86 | 8337 files<br>21.68 MB/s | 8270 (99.2%)<br>22.48 MB/s | 67 (0.804%)<br>2.603 MB/s | | 2 (0.024%)<br>2.712 MB/s | | | **8270 (99.2%)<br>22.48 MB/s** | | |
| RAGUE<br>03 jobs (2.657%) | 44.04% | 9.463 | 7174 files<br>15.69 MB/s | 7124 (99.3%)<br>17.61 MB/s | 50 (0.697%)<br>1.266 MB/s | | 1 (0.014%)<br>1.931 MB/s | 16 (0.223%)<br>1.695 MB/s | | | **7124 (99.3%)<br>17.61 MB/s** | |
| BL<br>78 jobs (2.492%) | 14.43% | 9.279 | 5315 files<br>7.022 MB/s | 5139 (96.69%)<br>7.761 MB/s | 176 (3.311%)<br>1.756 MB/s | | 55 (1.035%)<br>3.898 MB/s | 4 (0.075%)<br>3.505 MB/s | | | | **5139 (96.69%)<br>7.761 MB/s** |
| **TOTAL<br>15168 jobs** | **34.77%** | **10.14** | **249118 files<br>12.39 MB/s<br>80.81 TB** | **239865 (96.29%)<br>18.96 MB/s<br>77.8 TB** | | **77452 (32.29%)<br>27.68 MB/s<br>33.81 TB** | **65865 (27.46%)<br>12.14 MB/s<br>16.91 TB** | **51857 (21.62%)<br>27.74 MB/s<br>14.38 TB** | **10676 (4.451%)<br>15.34 MB/s<br>3.182 TB** | **8270 (3.448%)<br>22.48 MB/s<br>2.185 TB** | **7124 (2.97%)<br>17.61 MB/s<br>1.416 TB** | **5139 (2.142%)<br>7.761 MB/s<br>1.383 TB** |
| | | | | | **9253 (3.714%)<br>1.239 MB/s<br>3 TB** | **21 (0.227%)<br>0.763 MB/s<br>10.36 GB** | **4044 (43.7%)<br>1.172 MB/s<br>1.322 TB** | **186 (2.01%)<br>1.755 MB/s<br>43.99 GB** | **42 (0.454%)<br>0.809 MB/s<br>11.83 GB** | **26 (0.281%)<br>0.409 MB/s<br>11.05 GB** | **66 (0.713%)<br>2.232 MB/s<br>16.94 GB** | **2 (0.022%)<br>1.17 MB/s<br>538.2 MB** |

IO-intensive analysis train instance

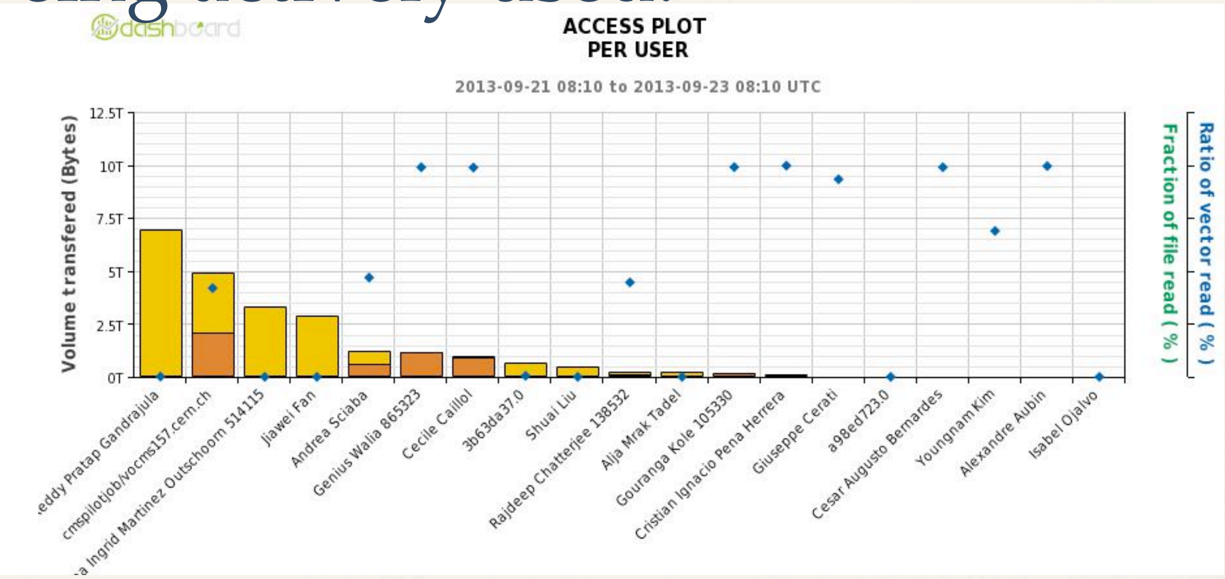
Plan to work with sites to improve infrastructure..

# CMS- xrootd federations- AAA

- Goal to have all CMS T1s and many T2s in AAA by Run 2: Currently 2/7 T1 39/51 T2s (> 95% unique data sets)

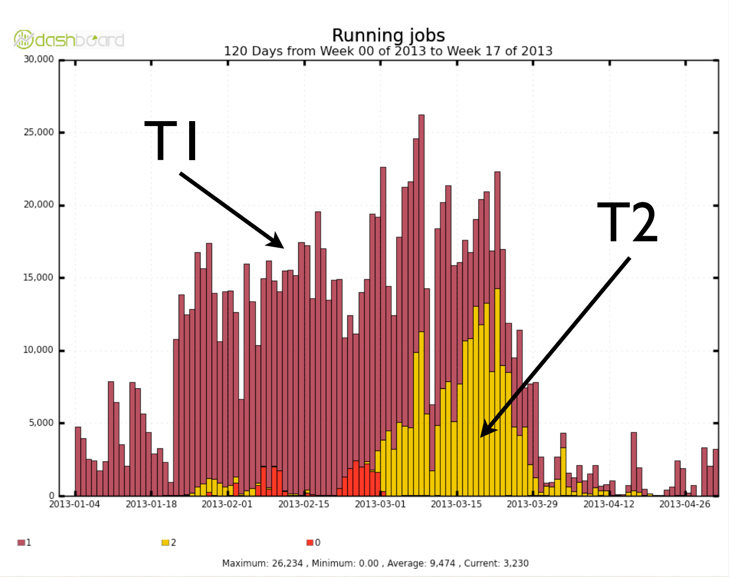


Being actively used:



Ken Bloom's slides

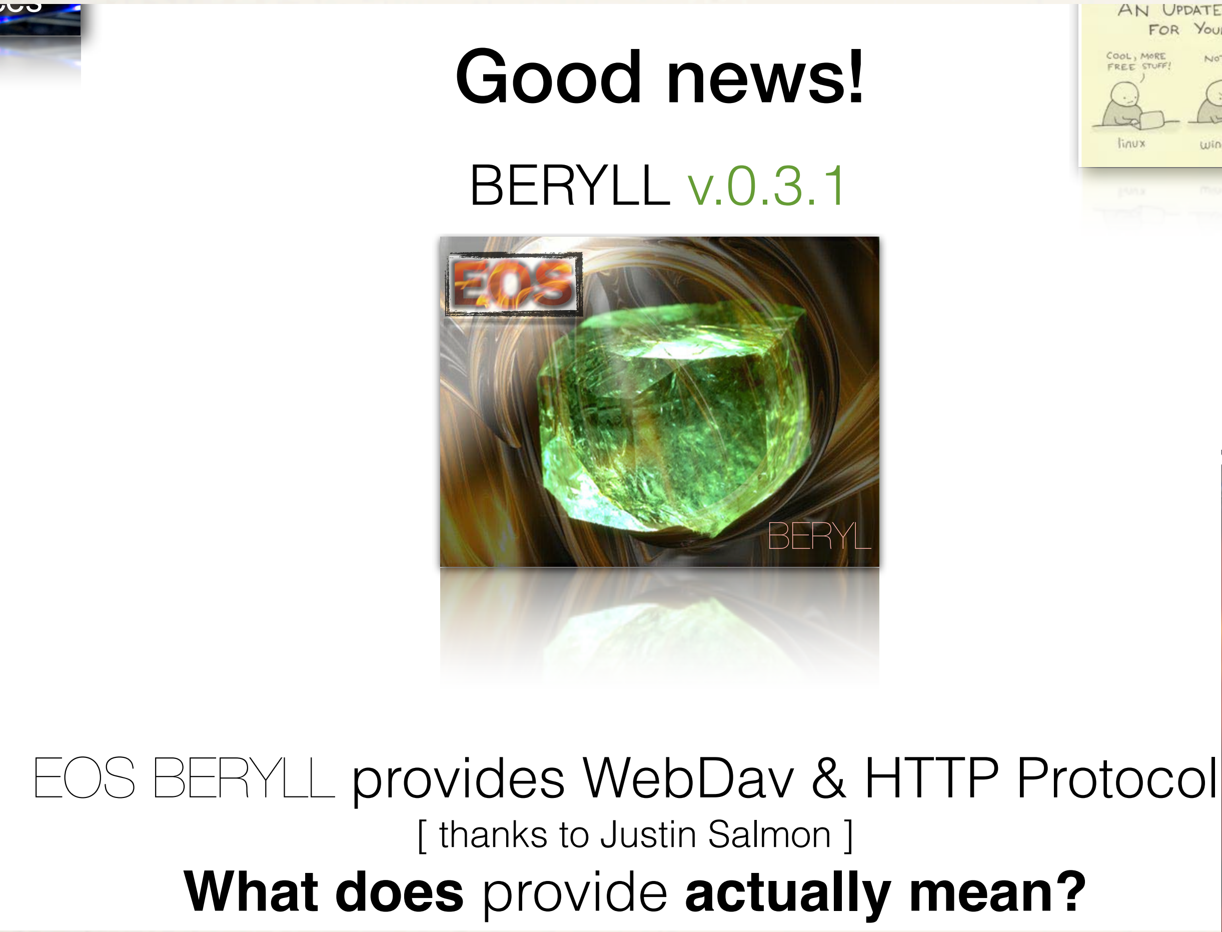
Production with remote data:



- Fallback and diskless sites
- Scale tests underway
- Ideas for healing; caching; load-balancing etc.

I think: useful experiences to share with other VOs - possible discussion at CHEP and fed mtg in Amsterdam

# Dav - eos



## Good news!

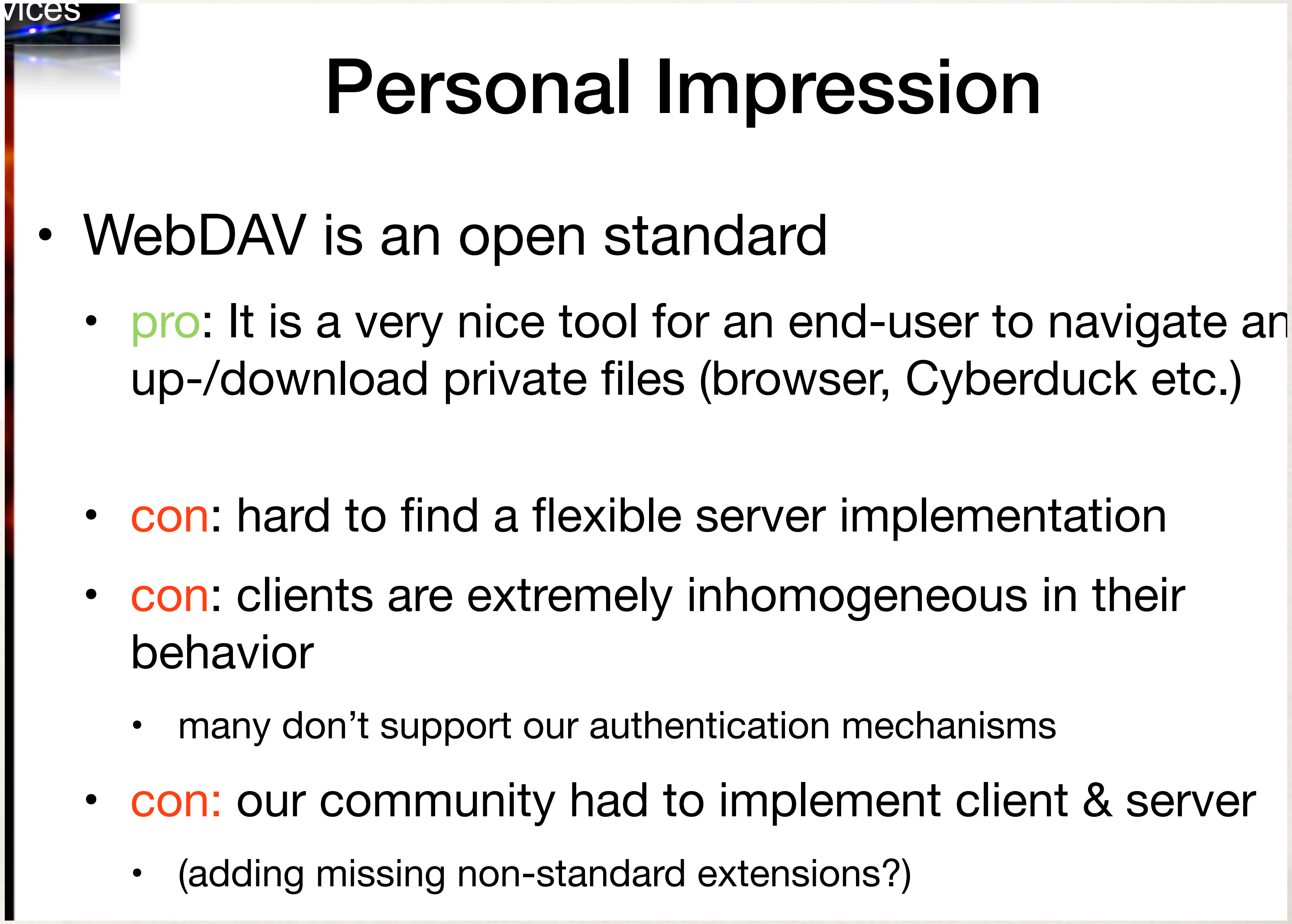
BERYLL v.0.3.1

EOS BERYLL provides WebDav & HTTP Protocol

[ thanks to Justin Salmon ]

**What does** provide **actually mean?**

[Andreas Peters's Slides]



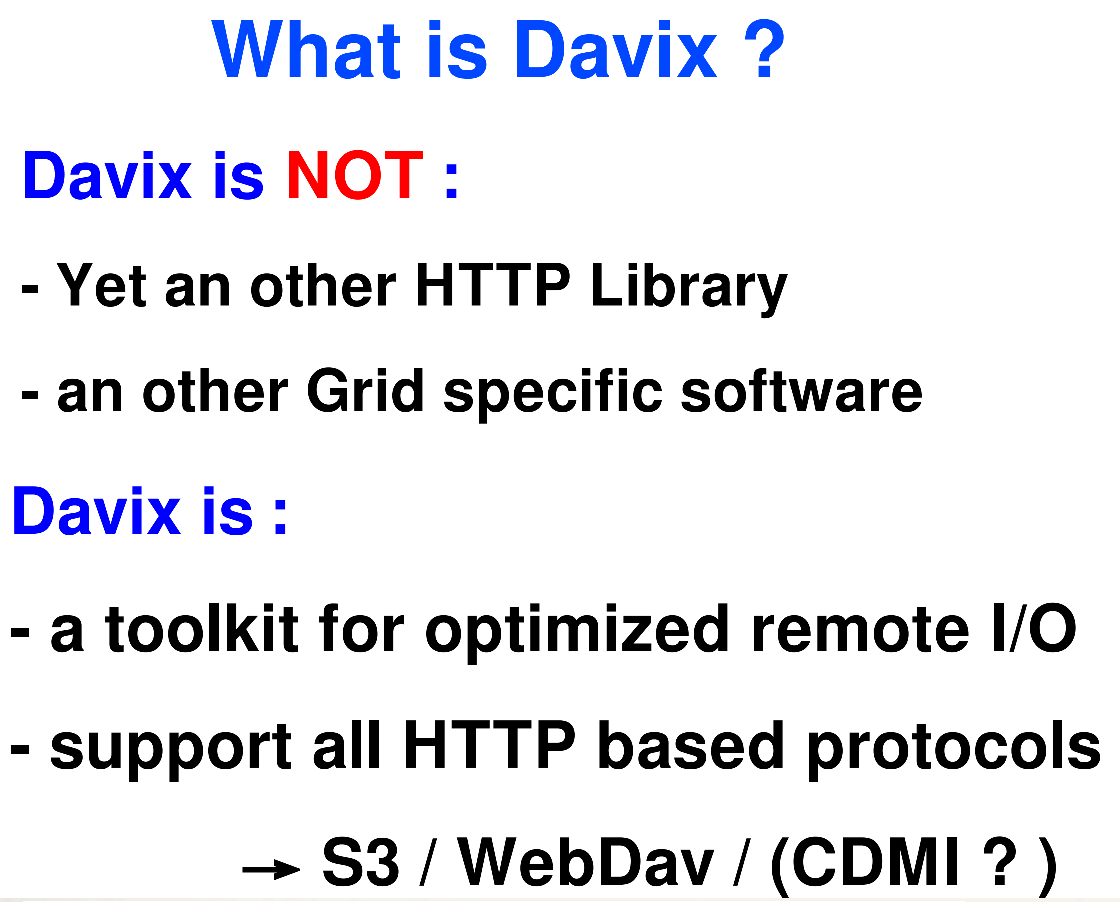
## Personal Impression

- WebDAV is an open standard
  - pro: It is a very nice tool for an end-user to navigate an up-/download private files (browser, Cyberduck etc.)
  - con: hard to find a flexible server implementation
  - con: clients are extremely inhomogeneous in their behavior
    - many don't support our authentication mechanisms
  - con: our community had to implement client & server
    - (adding missing non-standard extensions?)

- Providing basic functionality now
- Adapt to XRootD 4.0 when released
- Currenty XRootD native protocol recommended for best performance
- TDavixFile and other testing...

# Davix

Adrien Devresse's slides



## What is Davix ?

**Davix is NOT :**

- Yet an other HTTP Library

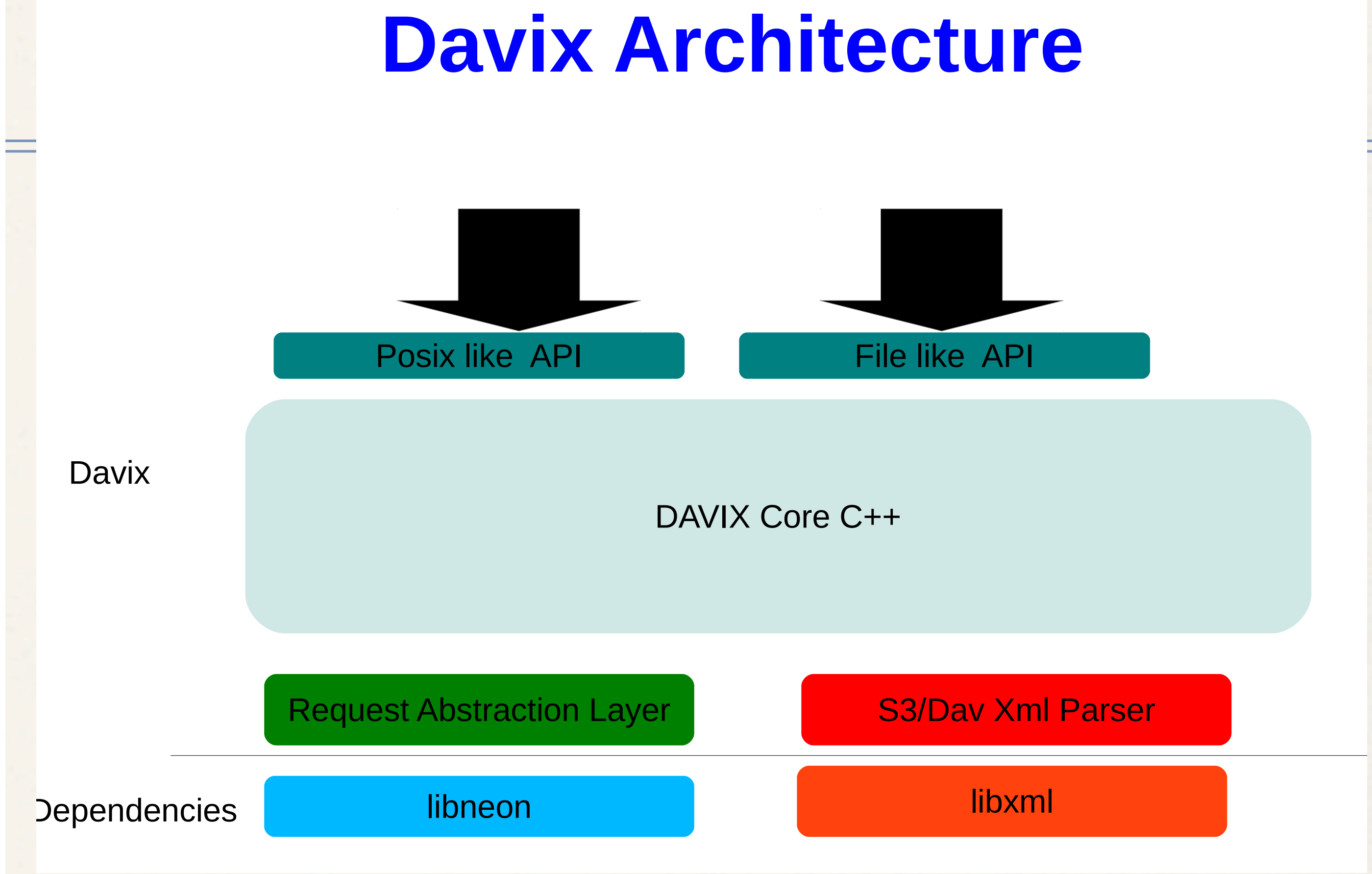
- an other Grid specific software

**Davix is :**

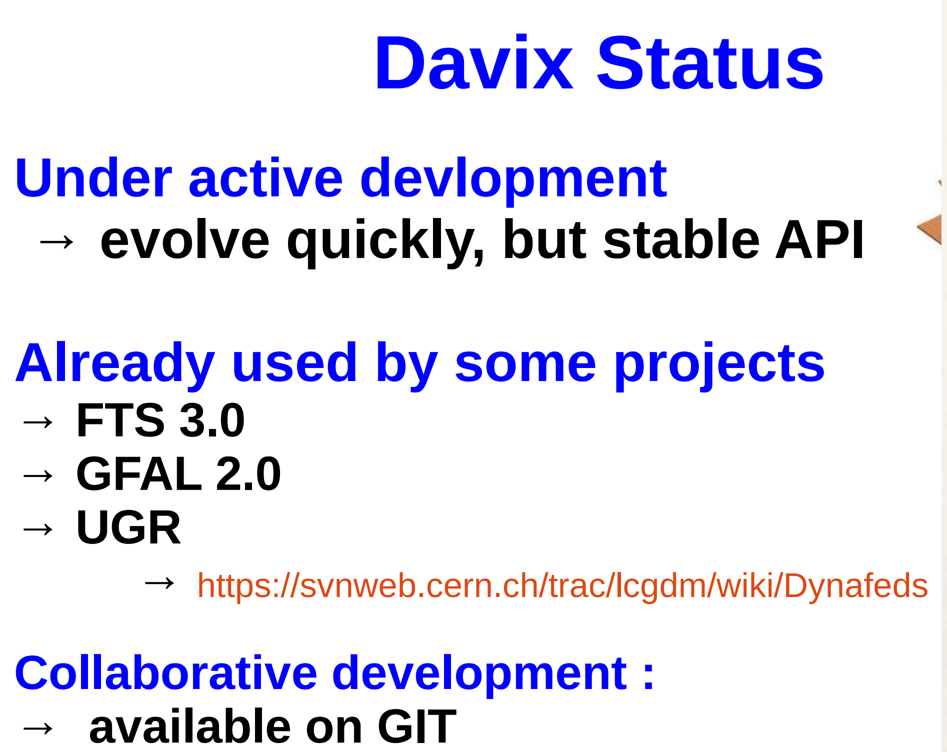
- a toolkit for optimized remote I/O

- support all HTTP based protocols

→ S3 / WebDav / (CDMI ? )



## Davix Architecture

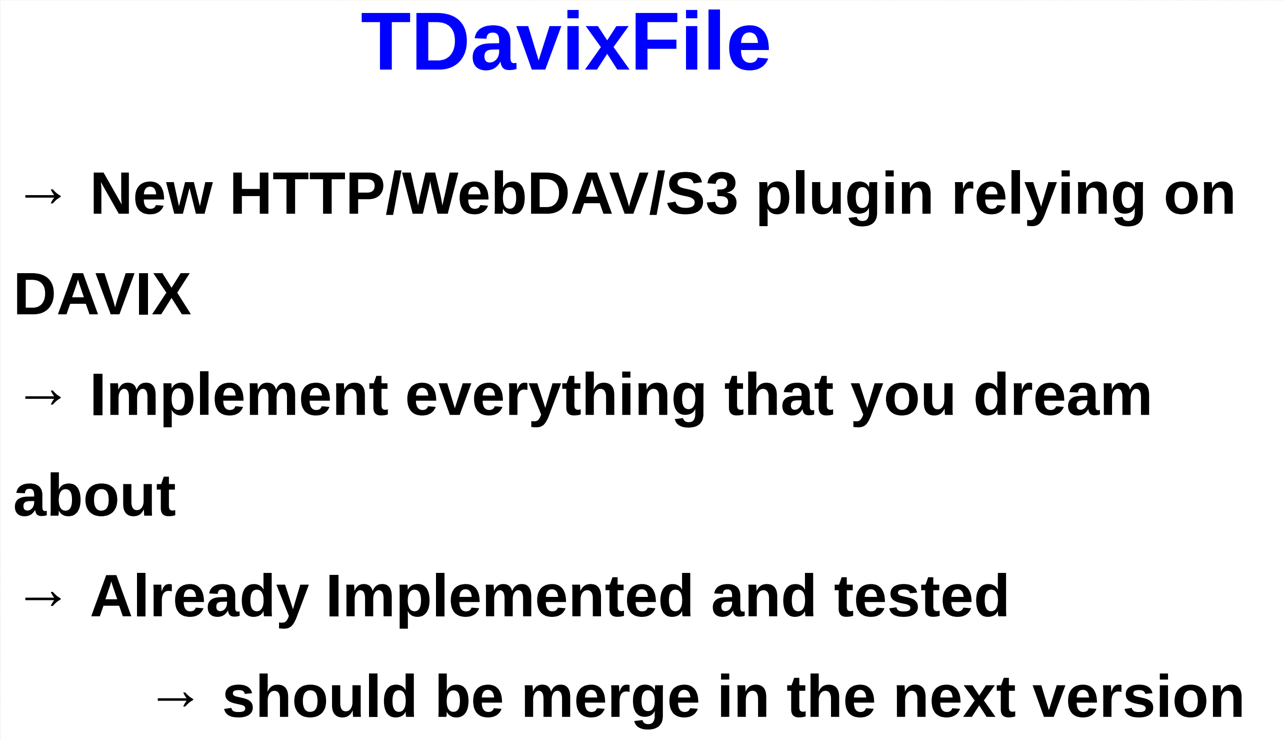


## Davix Status

**Under active devlopment**
→ evolve quickly, but stable API

**Already used by some projects**
→ FTS 3.0
→ GFAL 2.0
→ UGR
→ https://svnweb.cern.ch/trac/lcgdm/wiki/Dynafeds

**Collaborative development :**
→ available on GIT



## TDavixFile

→ New HTTP/WebDAV/S3 plugin relying on DAVIX

→ Implement everything that you dream about

→ Already Implemented and tested

→ should be merge in the next version

Interest in WG
e.g. rucio and other testing

This week ..

# Conclusions

- Valuable discussions in WLCG "Data" Working Group
    - particularly on WebDav and Xrootd
- Many issues still to track in "interfaces" - to make sure srm functionality is covered by xrootd and Dav and in a coherent way.
- ~Few month interval for these meetings is OK :
    - Next time a face-to-face meeting (perhaps with wider attendance) (at a pre-GDB?) may be a good idea.. ..

# Backup slides

# Possible activity areas for WG: refer to TEG recommendations

TEG list - already old -

other suggestions welcome

**1.2 Security [Already in Security activities]**

**1.3.2.2 Storage Management interfaces [Existing (SI) WG]**

**1.3.2.3 "Future Interfaces" [Recommended activity]**

As of today, broader industry storage interfaces (such as cloud storage) have not proven all the functionality required for these to be widely utilized. The development of these needs to be monitored and different approaches to integrate cloud-based storage resources need to be investigated. Experiments, middleware experts and sites to should work together in this exploration phase.

**1.3.3.1 Benchmarking and I/O requirements [Existing (IO) WG]**

**1.3.3.2 [Local Access] Protocol Evolution [Recommended activity]**

... move towards remote IO should be encouraged by both experiments and storage solution providers, and should be accompanied by an increase in resilience of protocols. LHC experiments are able to support all protocols supported by ROOT and expect to be able to continue to do so in the future. This support should be maintained but the current direction of travel towards fewer protocols (in particular the focus on file://, xrootd and http://) is encouraged ...

**1.3.3.3 I/O error management [Possible activity (added to above)]**

Storage errors returned by the system and how they are handled should be more explicitly determined. The client libraries should add high level "intelligence" to recover from transient storage failures and whether this can be achieved in the ROOT layer should be determined by the ROOT I/O working group.

**1.3.3.4 Future Technology review [Possible activity]**

New storage technology or hardware should be investigated and employed (where sensible) by WLCG sites. We recommend a thorough technology review, possibly in collaboration with the HEPiX Storage WG, to consider what low level technologies could be exploited by WLCG sites. The body carrying out this review should provide a mechanism to ensure that evaluations currently carried out by sites or interested vendors – can be communicated and discussed.

**1.3.3.5 High-throughput computing research [Recommended activity]**

Possibilities for much higher throughput computing should be investigated. This research should not be restricted to ROOT data structures and should fully utilise cutting edge industry technologies, such as Hadoop data processing or successors, building on existing exploration activity.

**1.3.4.3 Improved activity monitoring [Possible activity]**

Monitoring of files accesses, access frequency, etc. should be provided at the application and catalogue level [...] We recommend building a working group composed of technology providers and experiment representatives to study the available file access monitoring on both the SE and the application side. Both the ROOT I/O and HEPiX groups could play a role here and involvement wider than WLCG would be welcome. The group shall propose a standardized format to make this information available.

**1.3.4.4 Storage accounting [Already activity reporting to GDB]**

**1.1.1.3 Federation [Completed WG – but possibly continue some oversight]**

Launch and keep alive topical storage working groups to follow up a list of technical topics in the context of the GDB.