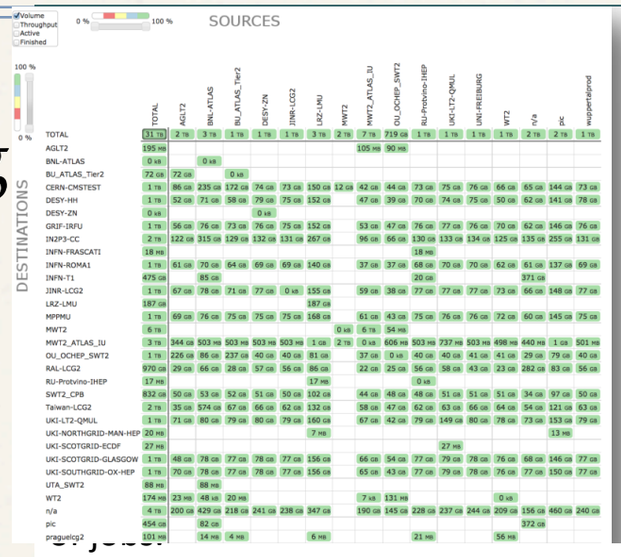
WAN data access and federations

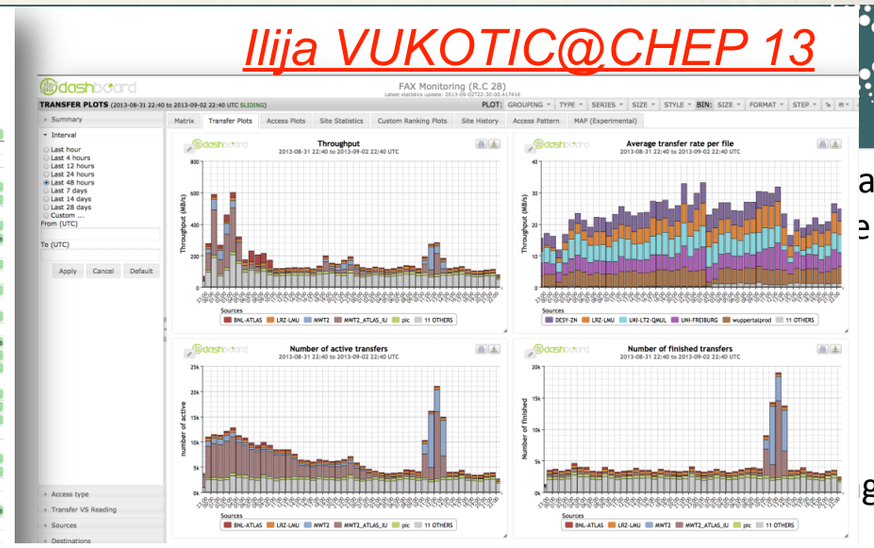
Wahid Bhimji

24 March 2014

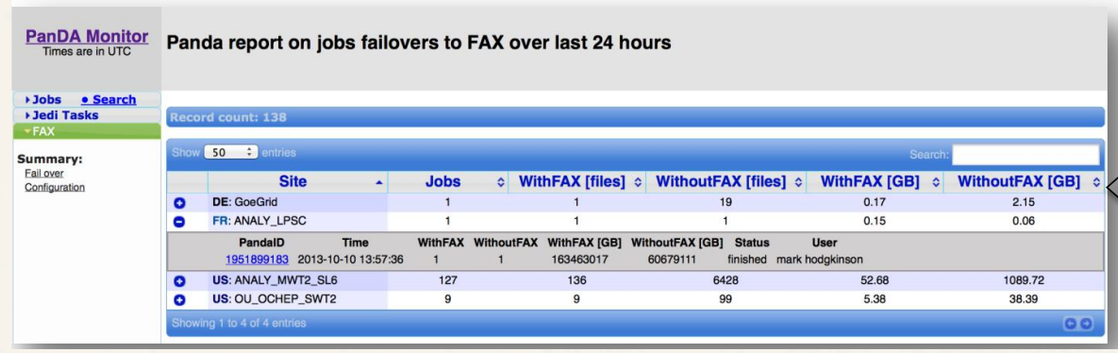
Data Federations

- Now a production service on both ATLAS and CMS offering
  - Fallback: if a job fails then reads / copy over the WAN
  - "Overflow" use if site busy
  - Planned remote production
  - Exploiting opportunistic resources
- ALICE already do WAN fallback for some time

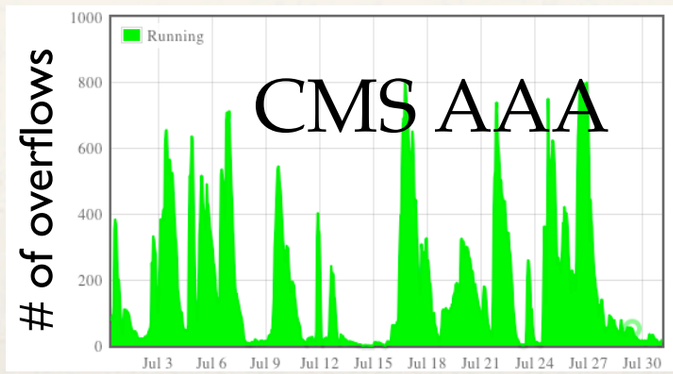




Ilija VUKOTIC@CHEP 13



- Small number of jobs failing
- But these failures cost the most in terms of user's tur around time



Kenneth BLOOM @ CHEP 13

Data federations - open Qs for us

- NOW is a good time to:
  - Compare experiences across experiment
  - Evaluate (projected) use (TEG said <~10% of bandwidth; now >10%)
  - Determine the impact on our infrastructure
    - Stress / infrastructure tests and monitoring
    - Track managed/opportunistic transfers (FTS/xrootd)
- HTTP / DAV - promising ideas that will be realized during Run 2: how will this heterogeneous (xrd/http) landscape evolve ?...

Overview of this talk / session

- **ATLAS FAX testing**
  - Stress / infrastructure testing - current activity for UK.
  - Plan for “diskless” site(s) ...
- **CMS AAA** - see David C’s talk etc. for more info .. but from me:
  - a word on ATLAS / CMS comparisons...
- **WebDav**
  - Davix status
  - Small VOs.

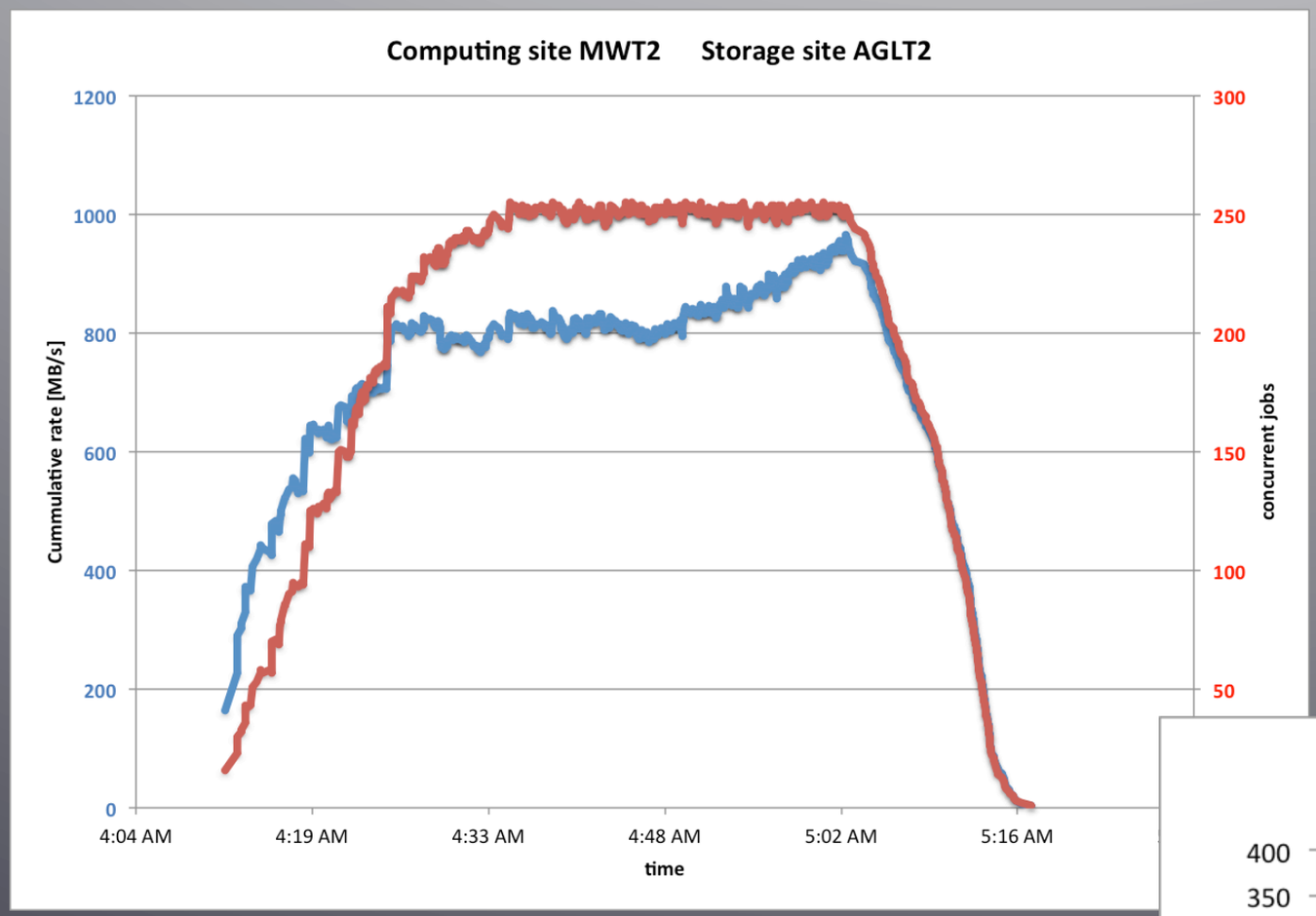
ATLAS Infrastructure testing

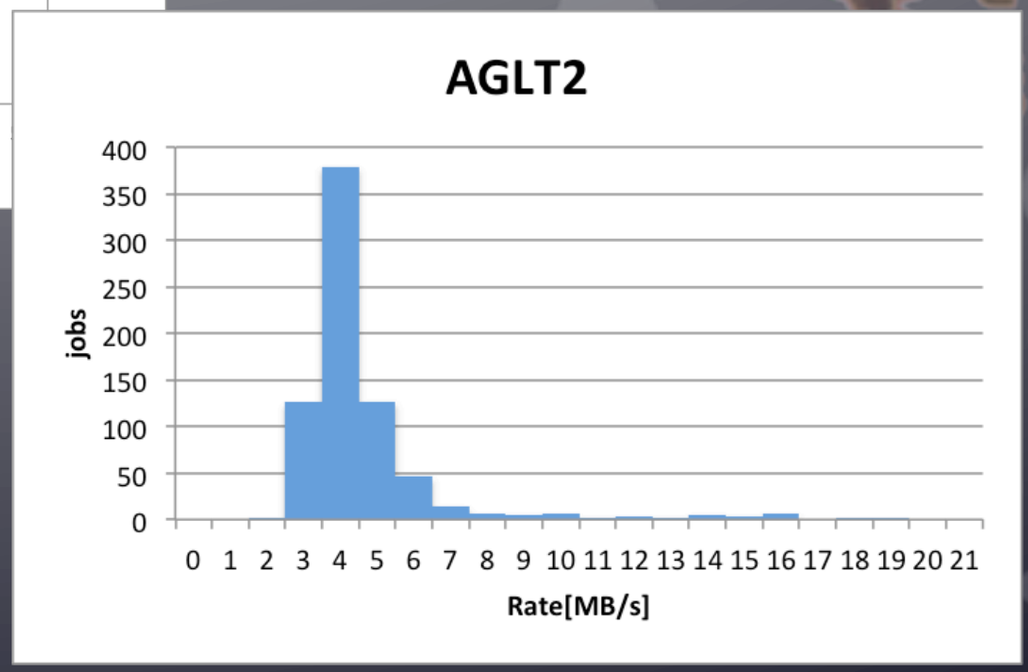
- Extensive testing in US
  - Ilija's description from ATLAS S/W week
- Doing similar tests in UK with ECDF; Lancs; Ox; QMUL (to start)

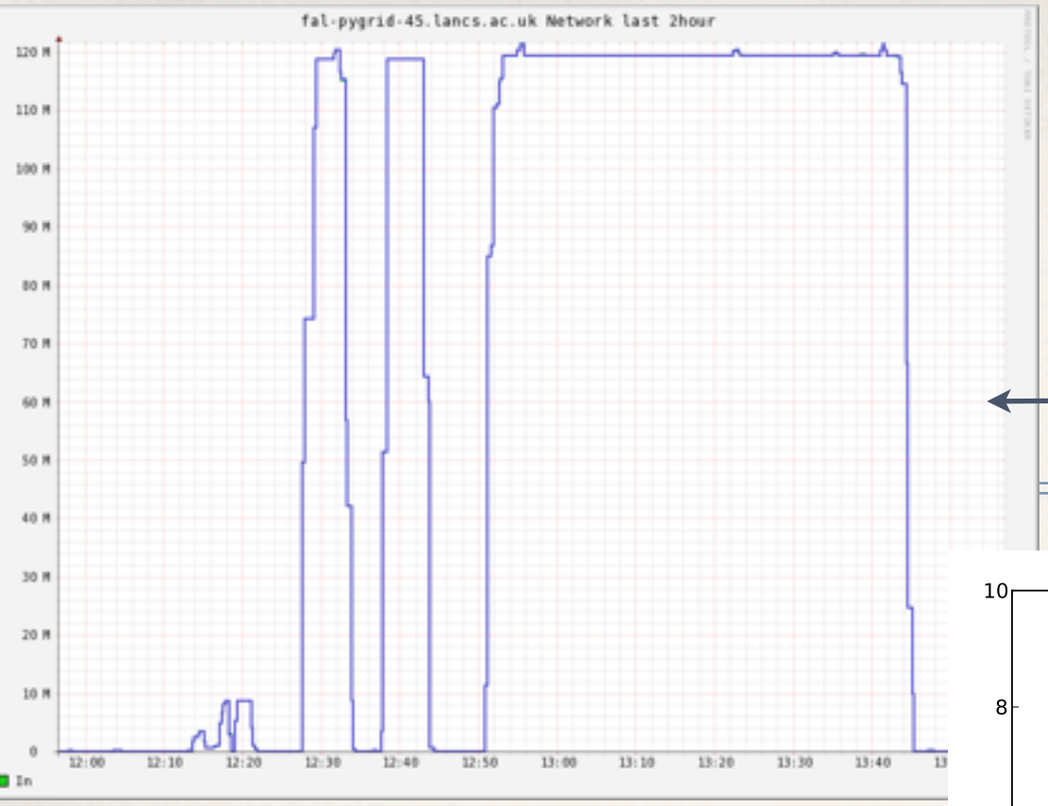
Infrastructure testing

- Test:
  - Using the site specific 744 FDR files (2.7 TB)
  - Reading 10% events, all branches, using 30MB TTC
  - Jobs submitted using ATLAS connect [condor to local batch system, I guess]
  - Running at MWT2 UC only
  - One job one file
  - Average number of transactions ~80
  - Theoretical maximum bandwidth for this kind of job on this kind of CPU is 53MB/s. Limited by uncompression and building of objects in memory.

Infrastructure testing– results – AGLT2 Stress 1







Running on Lancs WNs reading from ECDF

100 simultaneous jobs

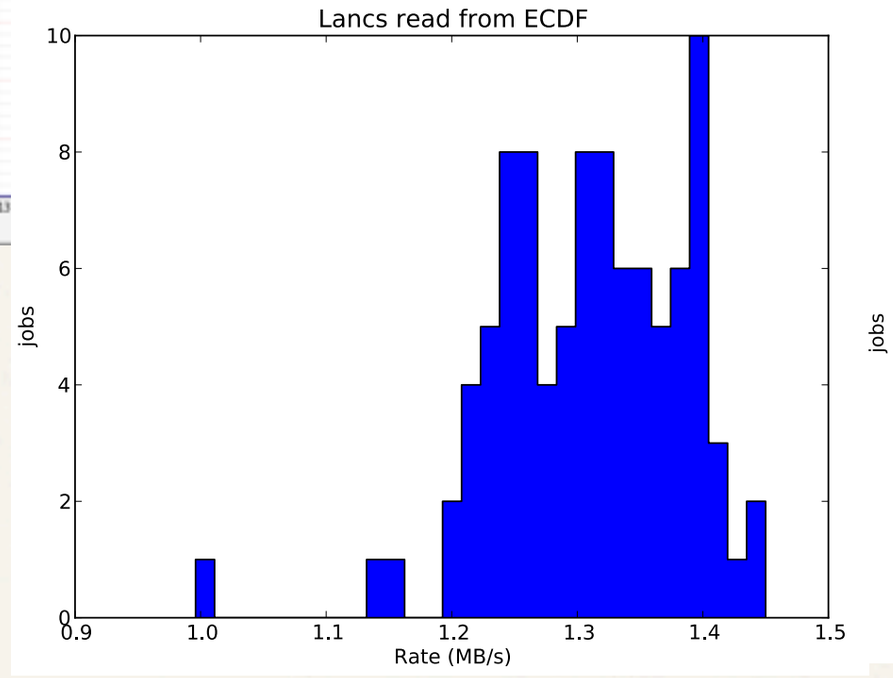
Easily saturate 1 Gig NAT (could also saturate 10Gig)

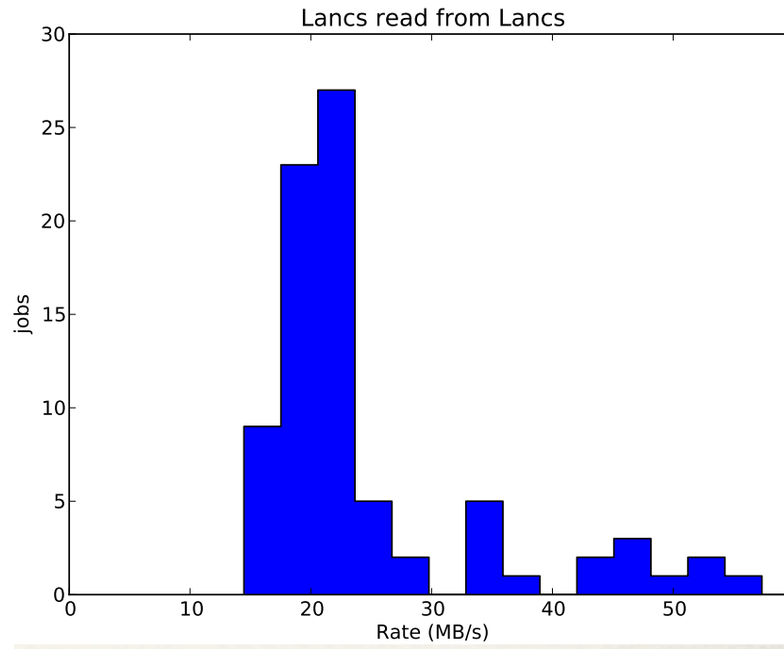
Via local batch system: Thanks Matt

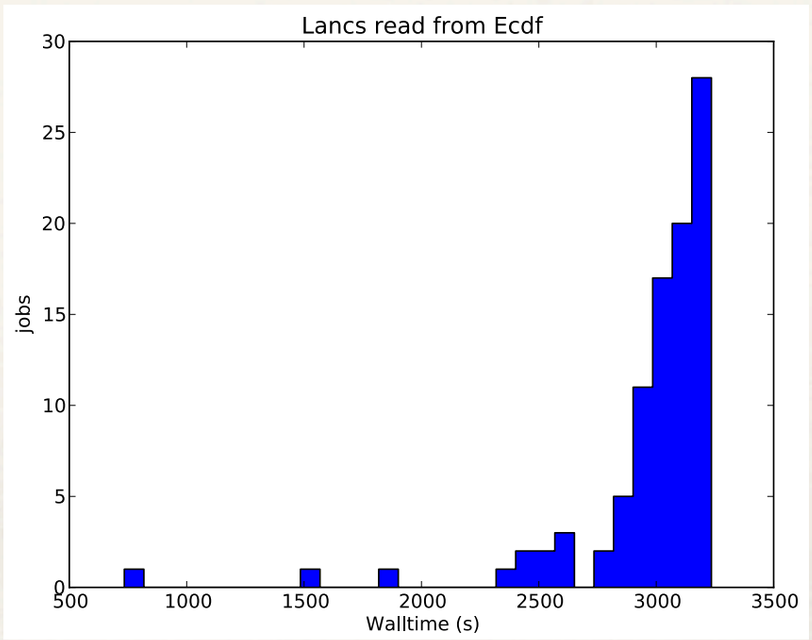
Impact on job data rate: ~1.3 MB/s compared to > 20 MB/s for local access

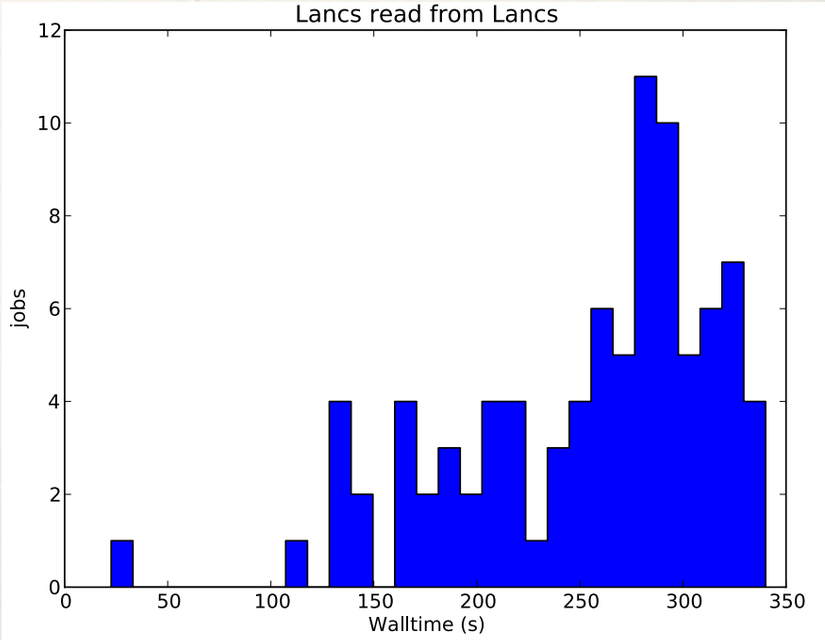
Similar impact as expected on walltime and cpu eff.

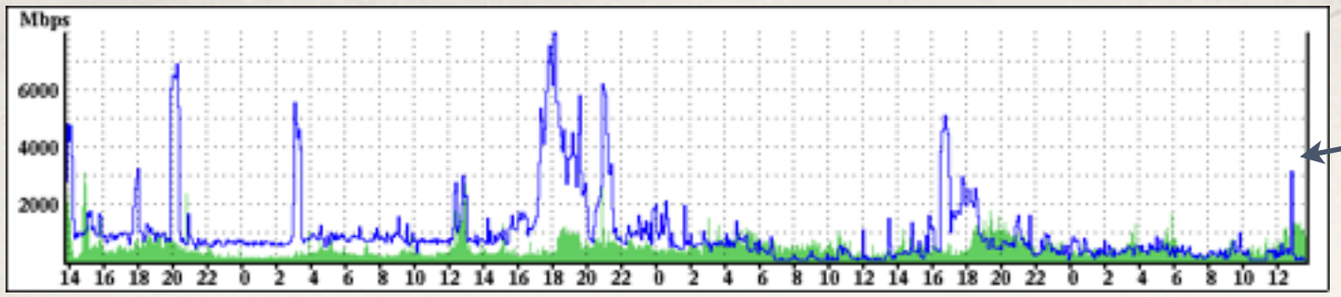
Not “real” ATLAS analysis - but that will have a similar impact (see later)





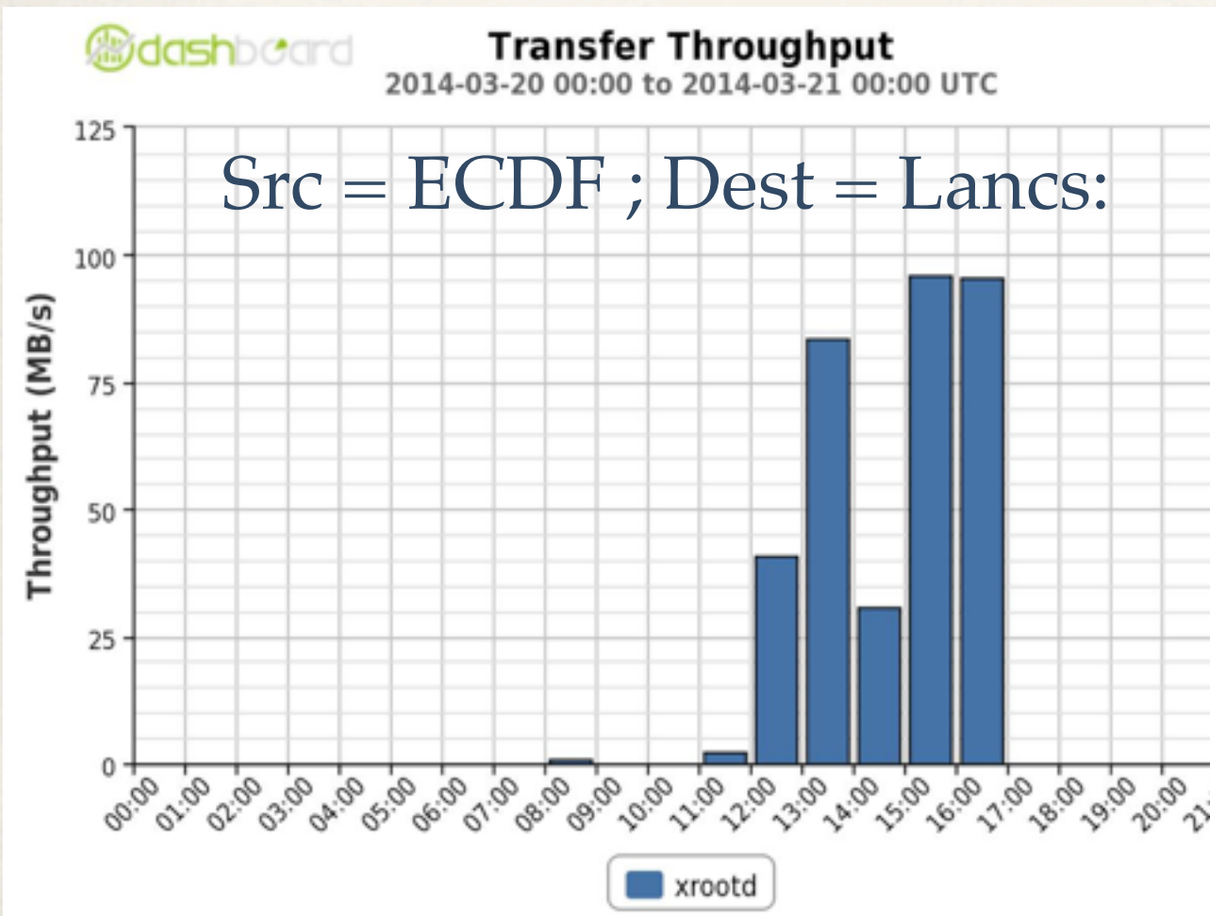


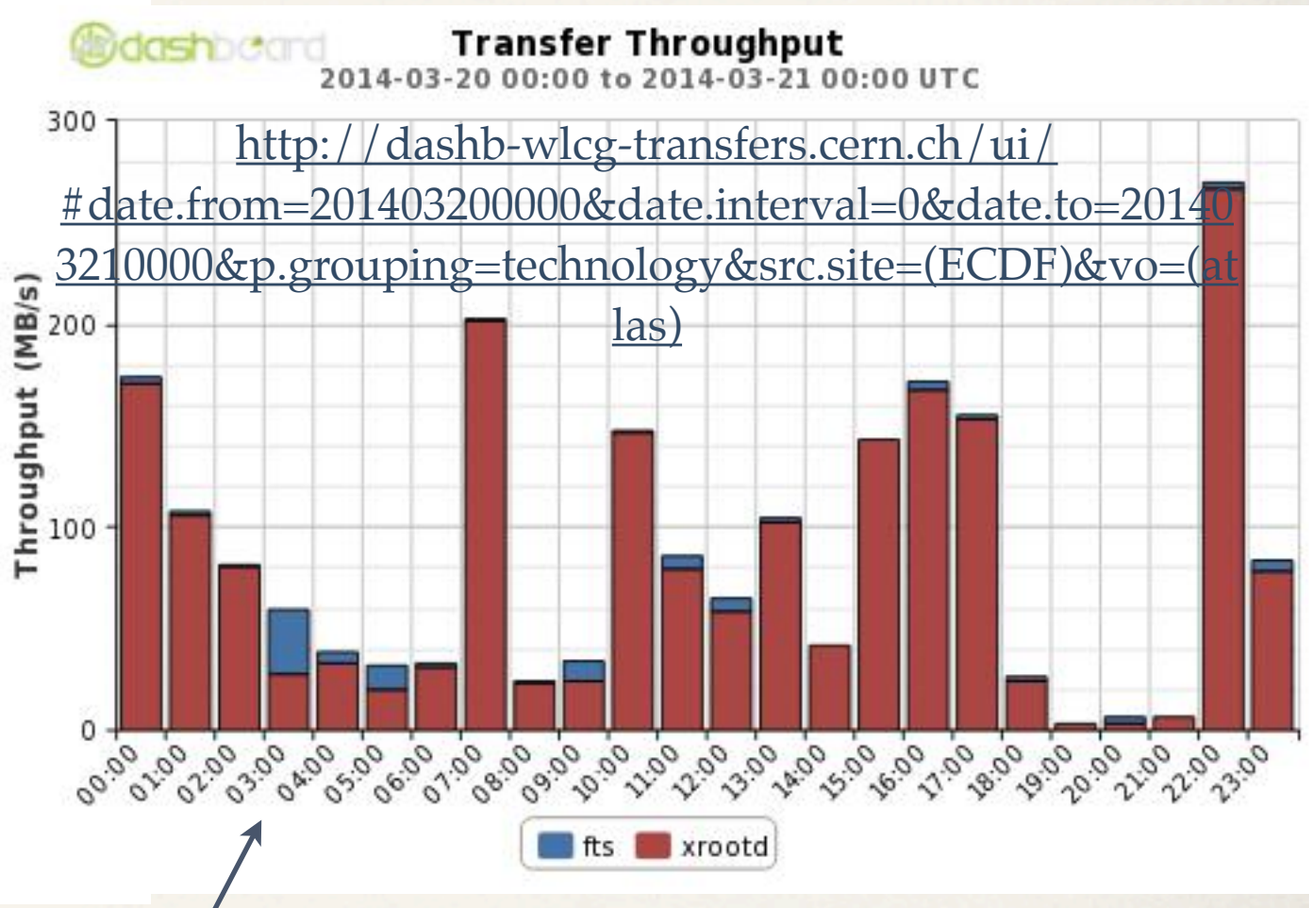




Impact at ECDF switch (Out is blue)
Not an issue compared to FTS rate

Fts v Xrootd from WLCG dashboard:





- This is supposed to be remote transfers- not working for ECDF due to WN and storage domain names being different
- This dashboard needs validation with other monitoring before really trusted for tracking: some more examples in backup slides ...

Plans and discussion

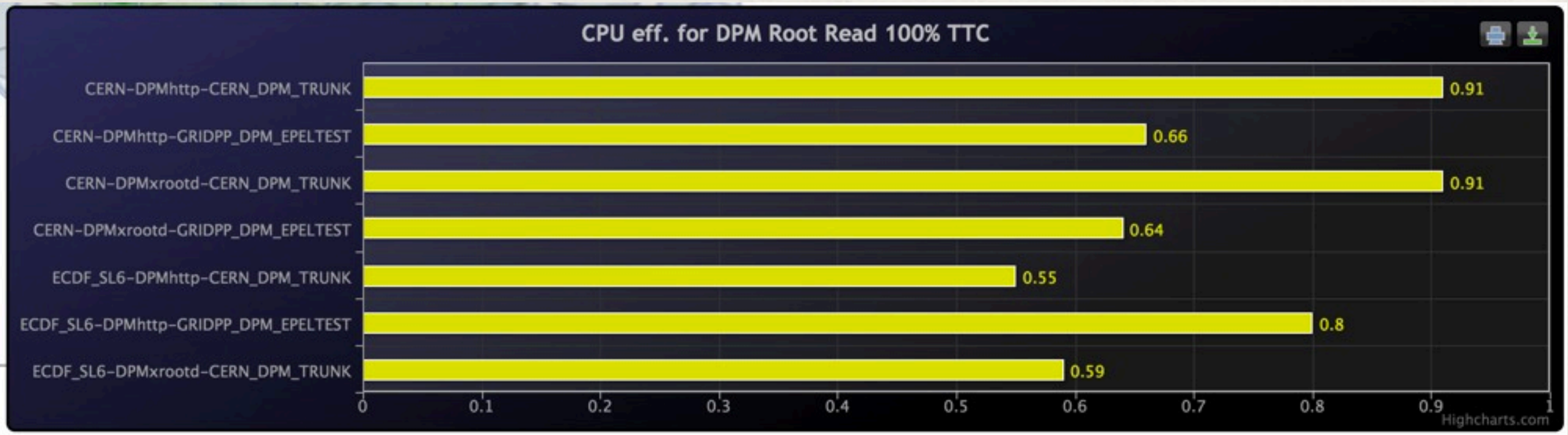
- Will run a "Realistic analysis" (H->WW) code (same as used in HammerCloud)
  - See backup slides for info and US results (data rate lower but still saturate 10 gig)
- Plan a test of ECDF, OX, QMUL, LANCS in different directions (Grid submission tedious)
  - Expose a few different bottlenecks
  - But a Bottleneck is a natural Bandwidth Limit - so if removed we may need another: eg. Proxy server (not in DPM sites by default) or to push promised xrootd Plugin
- Diskless site - save the need for "smaller" sites to run storage
  - Planed a test using UCL reading from QMUL and rest of UK
  - Should be configured only on ATLAS side (but current brokering Qs/issues)

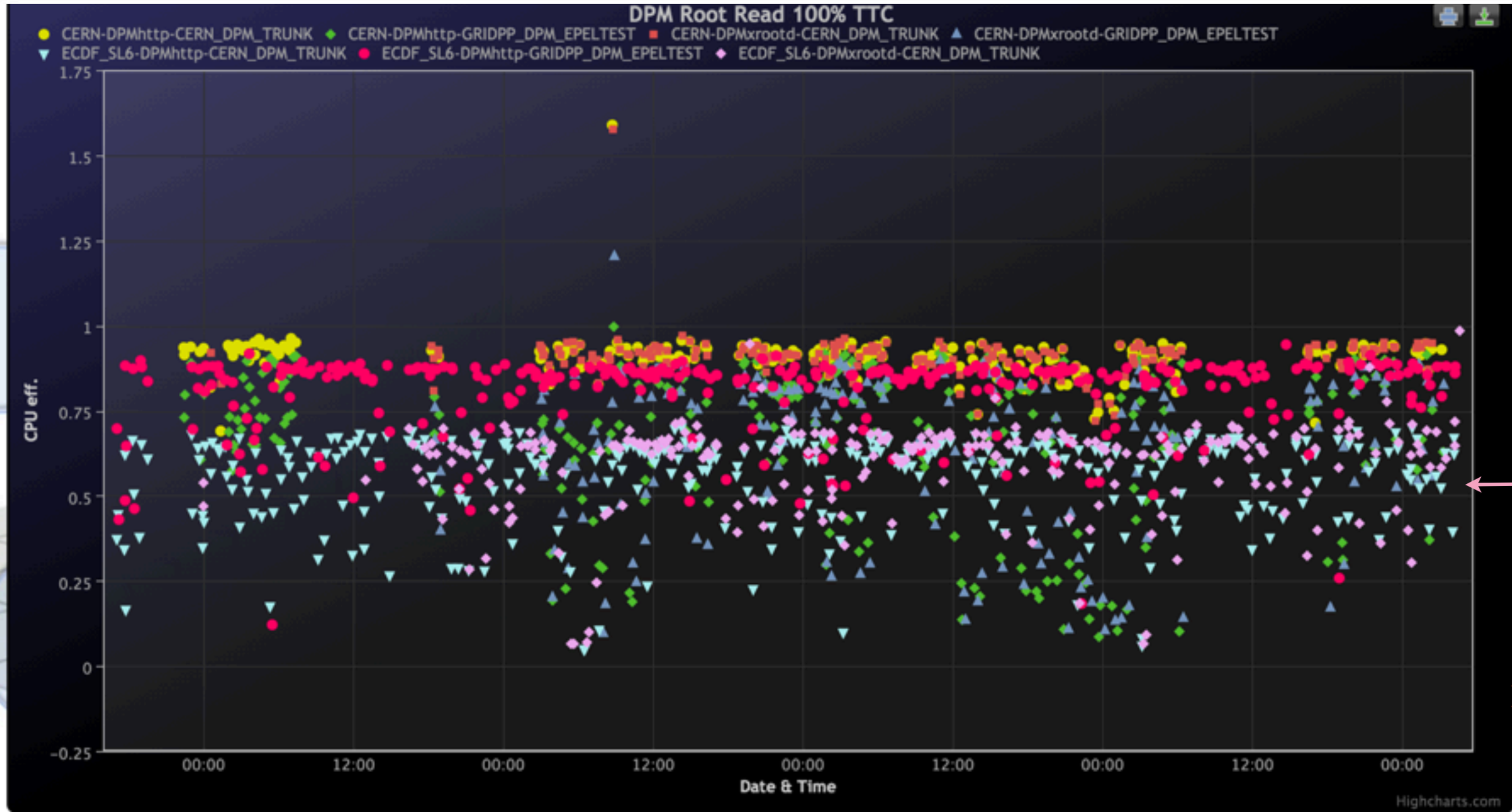
CMS and ATLAS "analysis"

- Typical CMS job (input from Brian Bockelman etc.):
  - We tell sites to plan for 1MB/s per analysis job; usage tends to average 500KB/s.
  - CPU efficiency is around 75-80%.
- Not the same as ATLAS ("H->WW" code - 20 MB/s from memory)
- Discussion in Monday's ROOT IO workshop that helped clarify
- CMS "analysis" can involve Reconstruction - higher CPU usage.
  - Its not a flaw or surprise if ATLAS get 10% eff where CMS have 100%
  - Not the same impact and issues in using federations.

HTTP/ Dav for data access

- Davix: Adrien's GDB talk - Performant http data access
  - Integrated into ROOT since 5.34-15 (but not built by default)
    - Essential for https running; Others hit some issues in test...
- ATLAS Rucio will use Dav for user download (inc. using metalink)
- Next page tests (similar to that used before, 100% of file) from DPM tests (Using TWebFile not TDavixFile ..)
  - Between ECDF (epel-test) and CERN (trunk) TEST boxes
  - Just to illustrate comparable http and xroot data access performance





Large spread in WAN results

Thanks to Ilija / Fabrizio Furano for recent improvements

Dav for storage management

- Davix also used in gFal2
  - Therefore in FTS3 (potentially for third party transfers)
  - And gfalFS - fuse mount your storage:

```
yum install gfal2-all gfalFS gfal2-plugin-http
usermod -a -G fuse wbhimji
/usr/bin/gfalFS /tmp/mnt2 davs://srm.glite.ecdf.ed.ac.uk//dpm//
```

```
[wbhimji@gridpp09 ~]$ find /tmp/mnt2/atlasscratchdisk/ | grep -v rucio | grep -v SAM | wc -l
6437

 time rmdir /tmp/mnt2/atlasscratchdisk/!(rucio|SAM)
```

BE CAREFUL - rmdir will remove non-empty directories ..

Dav for small VOs

Thoughts from Chris Walker...

Make remote access easier for interactive users:

- Looking like a filesystem would be ideal to help T2Ks of this world work out what data is where.
- Chris tried Gfalfs for this:
  - filed some tickets - but these are being fixed
- Pushing Dav at UK sites for non-LHC VOs (Atlas using it helps this happen of course)
  - LFC now dav read-only (OK)
  - SEs most support - though Storm support buggy

Conclusions

- Just starting to test WAN infrastructure in UK
  - for ATLAS... (would like CMS results + a script I can run myself.)
- ATLAS and CMS analysis use cases cannot be directly compared and ATLAS will be limited to some extent by current bandwidth (not necessarily a problem)
- Dav / Http will be used in some ways for ATLAS
  - and offers interesting potential for management and small-VOs as well as data access

backups...

Realistic analysis tests

- Friedrich's Higgs to WW analysis
  - Using RootCore framework
  - All corrections applied
  - Reads 512 from 8543 branches (13% of total size)
  - Writes out <1% events and a number of histograms/ TTrees
  - 10MB TTC
  - Default learning phase (100 events)

ealistic analysis tests – results

Having the data file cached in memory

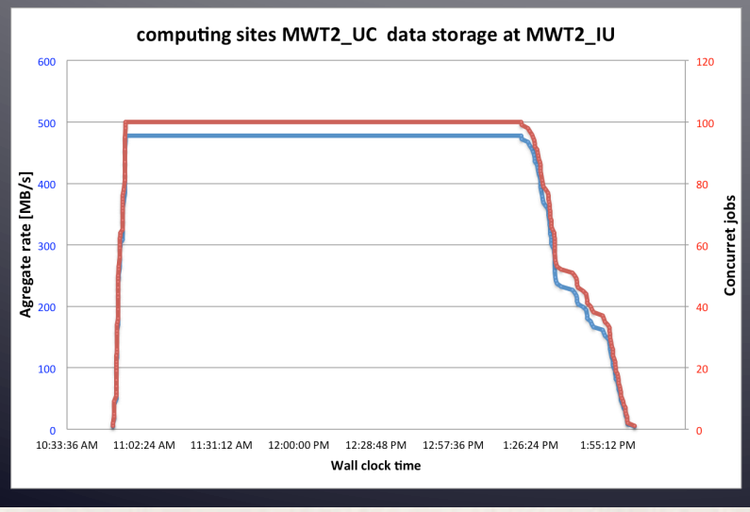
- linux keep parts of the recently accessed files in memory. Simple copy to dev/null will cache the full file
- This informs us on how CPU intensive is the analysis code relative to unzipping the data (code doing only uncompressing can do 40-50MB/s) and sets the maximum this analysis can do on that CPU
- 100% CPU efficient
- **1300** ev/s
- **20.2** MB/s of useful data (data that will be transferred and uncompressed)

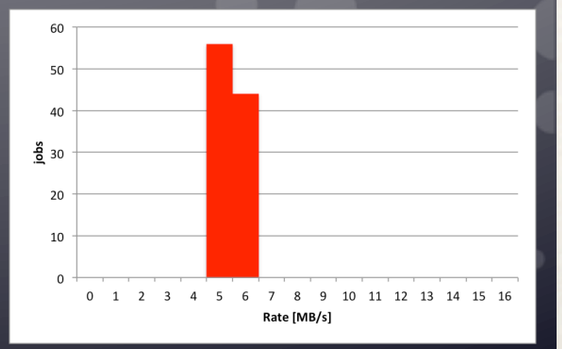
Realistic analysis tests - results

WAN access

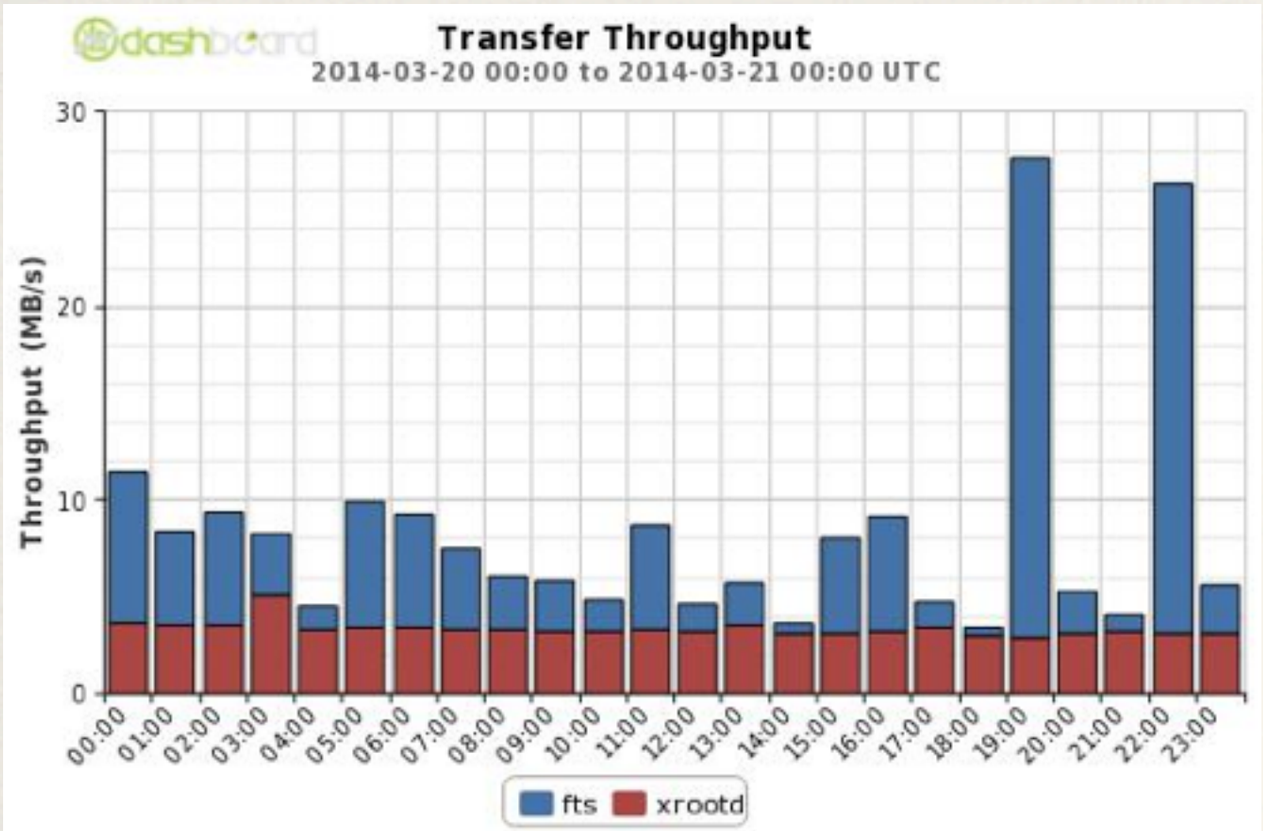
- run at UC reading data from IUC (5ms RTT)
- Each job analyzing:
  - 100 files
  - 356 GB
- 3.54 Mevents

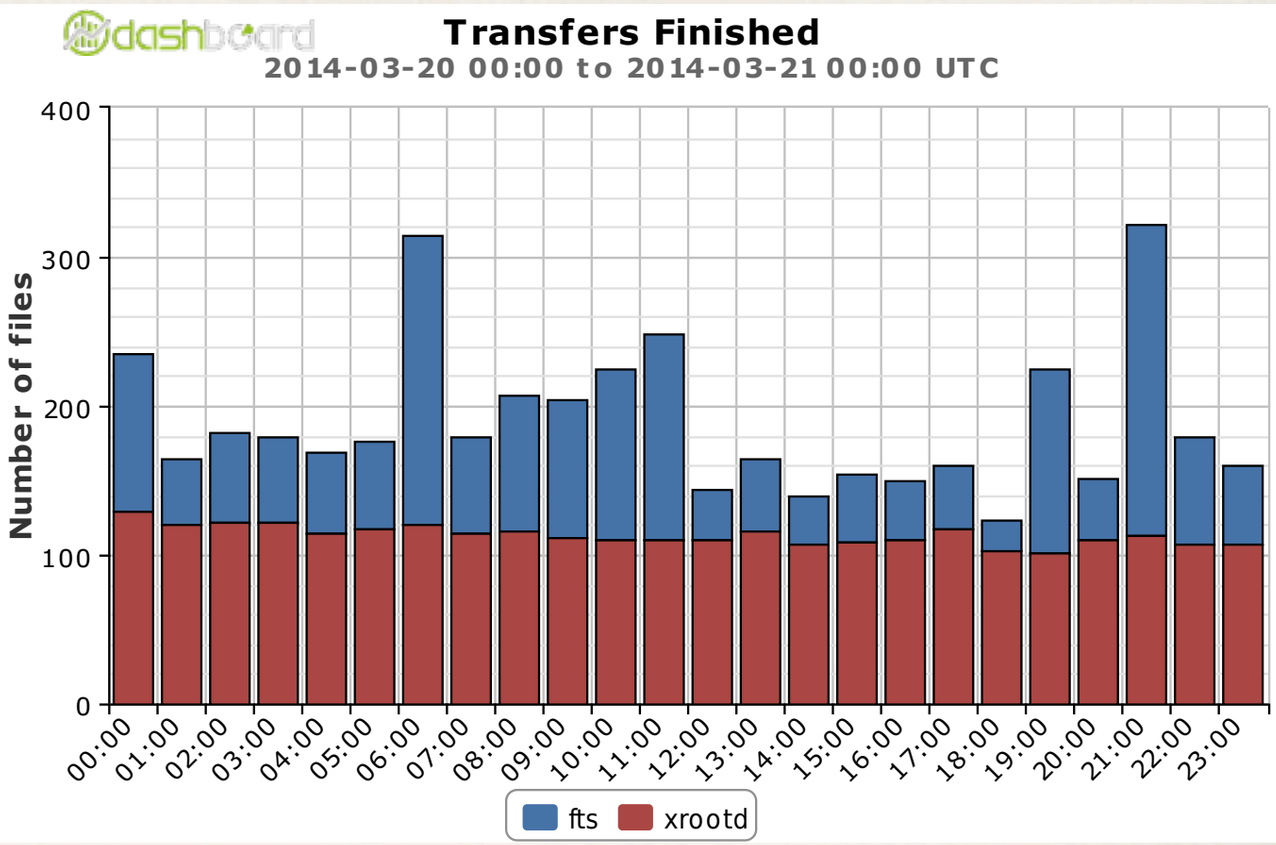
- Much larger scale needed to saturate the link that we have (60+ Gbps)
- Not much slower than local disk
- Next to try:
  - Newer ROOT version
  - Asynchronous Prefetch





Oxford:





Whole UK CMS activity according to dashboard (is it right ??)