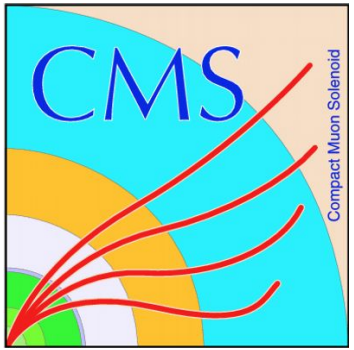
AAA: The Storage Federation in CMS

Ken Bloom, Federica Fanzago, Marian Zvada
for the AAA team



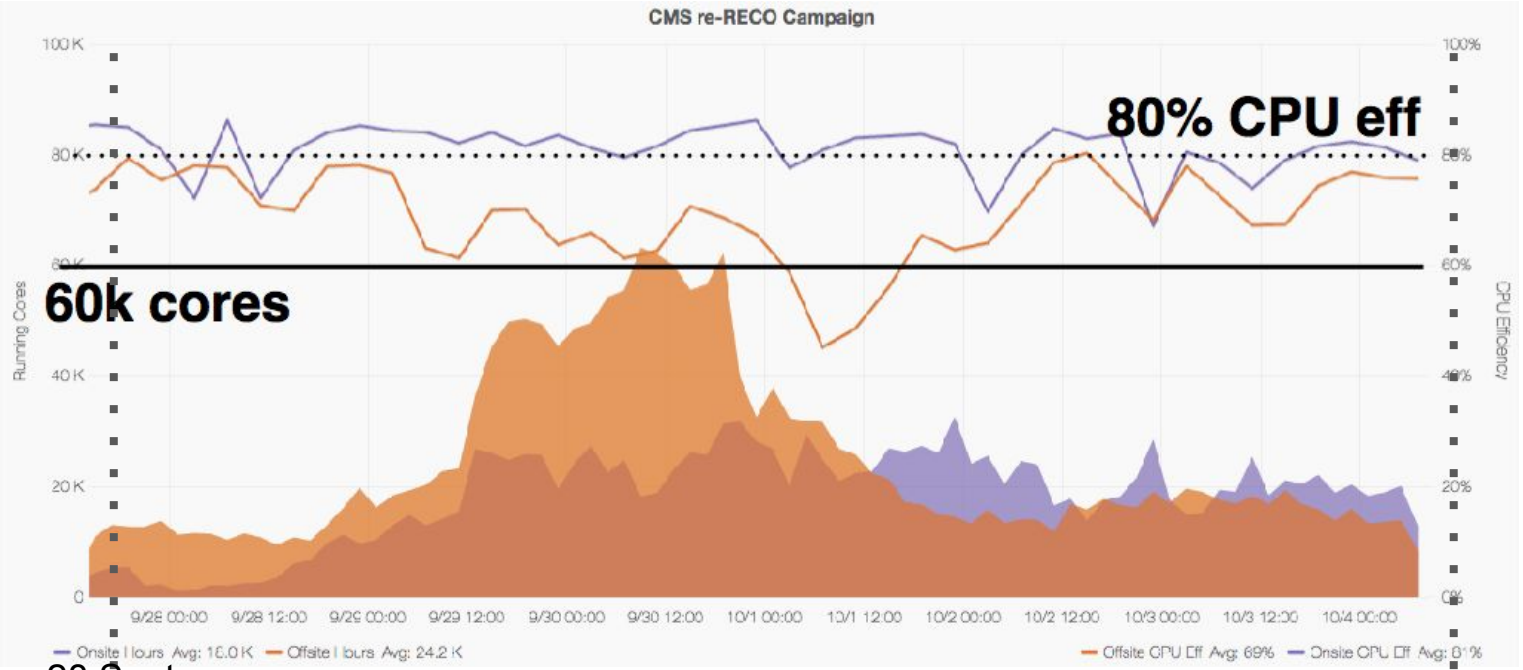
Introduction to AAA

- AAA what is it?
  - AAA = Any data, Anytime, Anywhere.
  - An effort to create a storage federation of the CMS sites.
- Why AAA?
  - To provide data access transparent toward users at any CMS sites.
  - Evolution of data-driven approach of CMS infrastructure.
    - Fallback solution if access of local data fails
    - Possibilities of opportunistic (not only) resources without storage do job processing on data not stored locally

AAA successes at scale

- AAA has become an increasingly integral part of CMS's computing strategy
  - Through the fallback mechanism, any CMS job naturally seeks input files within the federation when they are not found locally; no need to take special steps
  - Already regularly used by analyzers
  - But now successfully integrated into large processing campaigns such as the just-completed re-reconstruction of 2016 data
- Some recent performance numbers (~Jan-Aug 2016)
  - 14% of production and analysis jobs use AAA
  - CPU efficiency ~5% better for local rather than remote access in production jobs

Experience of re-reco campaign

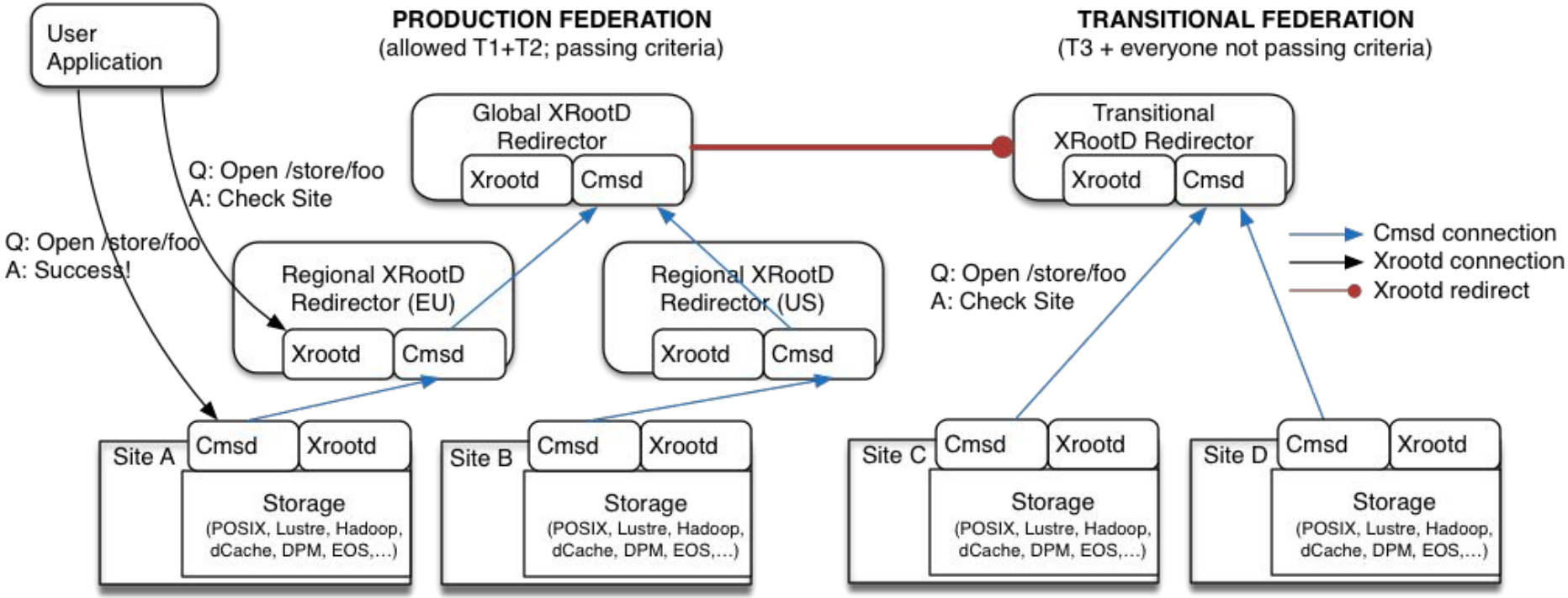


28 Sept

4 Oct

- AAA allows more sites to participate; more processing offsite than onsite
- Manageable hit on CPU efficiency, little dependence on data rates

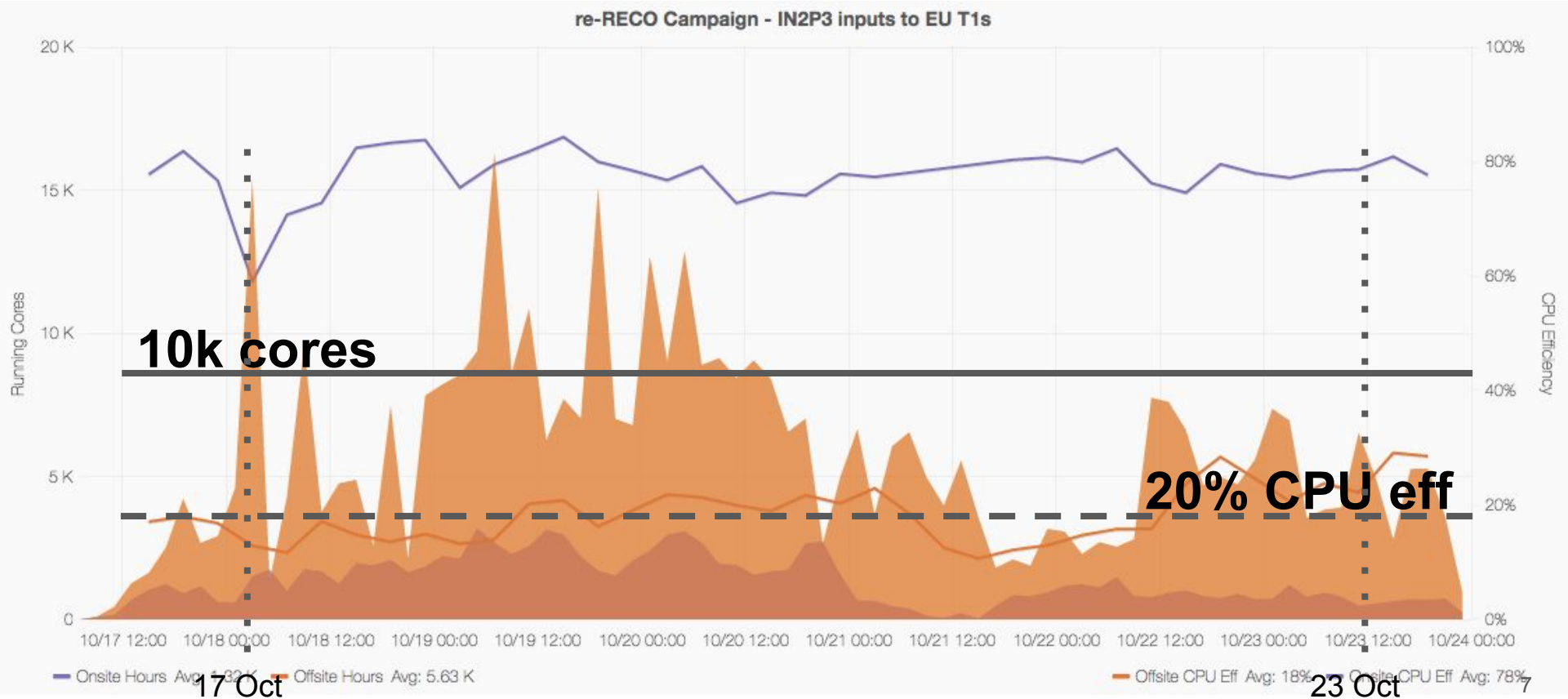
AAA federation - production and transitional



What might limit the scale of AAA?

- Operational issues
  - managing membership of the transitional federation
  - site validation through scale testing, SAM and HC tests, Kibana
  - monitoring of xrootd traffic local versus external
- Infrastructural issues
  - Issues in redirectors themselves
  - Transient cmsd subscription losses
  - WAN connectivity between regions (HW issues like faulty routers)
- Site issues
  - Can sites sink data at sufficient rates?
  - Can sites source data at sufficient rates?
  - Does site have optimal xrootd cluster configuration?

Oh no - what happened?



17 Oct

23 Oct

Operational issues

- Transfer Team (TT) & Site Support Team (SST)
  - shared effort with AAA developers
- TT & SST acting as ‘front line’ for GGUS tickets
  - If the issue is global or concerns regional redirectors, Ops refers the issue to AAA devs
  - If it is site-specific
    - Ops attempt to debug the issue themselves with the site admin
    - Usually involves a change in the site’s configuration - subscription or redirector versus server configuration
    - If the problem is deeper, Ops refers to AAA devs
    - check up on xrootd reports (site name, xrootd version) and open GGUS
    - inform sites of changes necessary to enter AAA (production or transitional federation)
- Central management of membership of transitional federation
  - should be possible once we have ALL sites up to xrootd >=v4
  - will save one operational step - do not ask site admin switch subscription as we do now

Are sites ready for the production federation?

- First step: scale tests that allow to discover unoptimized sites
  - File-opening test: access total rate of 100 Hz at a site via regional redirector
    - test runs up to 100 jobs simultaneously, opening files at rate of 2Hz each one
      - more than one year of tests demonstrate opening rate is strongly dependent from SE system and site setup so the target for tests was changed in “at least 10 Hz”
  - File-reading test: 600 jobs reading average rate of 2.5MB every 10s at a site. Reading total rate of 150MB/s via regional redirector
    - test runs up to 800 jobs simultaneously, each one reading data blocks of 2.5MB from a file
- Continuous check of reliability
  - Cross-check with other system of tests and monitoring
- Criteria to move the site from production to transitional federation
  - AAA-related ticket in GGUS open for longer than two weeks
  - SAM xrootd access and fallback test < 50% for two weeks
  - Hammer Cloud (HC) test success rate < 80% for two weeks

Status of the scale tests

- Currently tests run once a week on a subset of CMS sites with different storage backend.
  - List of input files obtained via PhEDEx.
  - Condor pool for test in Wisconsin.
- 2 regional redirectors, US and EU (dns alias including Pisa, Bari and Paris)
  - Sites provided special path “/store/test/xrootd/<cms_site_name>/LFN” to allow the match.

non-US sites	US sites
19 dCache (4 tier1)	1 dCache (tier 1)
2 Hadoop/BeSTMan	6 Hadoop/BeStMan
11 DPM	1 Lustre/BeStMan
6 StoRM (1 tier 1)	1 LStore/BeStMan
1 Castor (1 tier1	

Last DPM scale test results

- Weekly results are published http://www.pd.infn.it/~fanzago/TEST/summary.html and are included in CMS dashboard views.
  - Usefull to compare results with other sites and evaluate if there is a common problems

    https://dashb-ssb.cern.ch/dashboard/request.py/siteviewhistory?columnid=221

    https://dashb-ssb.cern.ch/dashboard/request.py/siteviewhistory?columnid=222

Summary of AAA opening and reading tests

SITES castor dcache dpm storm hadoop/BeStMan LStore/BeStMan Lustre/BeStMan

TEST STATUS	OPENING TEST	READING TEST
FAILED	completely failed test, no plots produced	completely failed test, no plots produced
PROBLEM	the opening rate is lower than 10 Hz	the reading rate is lower than 150 MB/s and the number of simultaneous clients is lower than 600
WARNING	the number of simultaneous clients is lower than 90	the reading rate is lower than 150 MB/s even if the number of simultaneous clients reaches 600
OK	the opening rate reaches 10 Hz and the number of simultaneous clients is bigger than 90	the reading rate reaches 150 MB/s

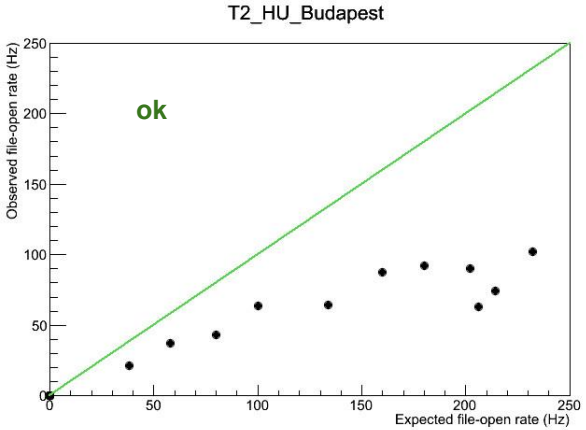
T2_AT_Vienna-xrootd-cms.infn.it_20_11_16	OPENING	READING
T2_FR_GRIF_IRFU-xrootd-cms.infn.it_20_11_16	OPENING	READING
T2_FR_GRIF_LLR-xrootd-cms.infn.it_20_11_16	OPENING	READING
T2_FR_IPHC-xrootd-cms.infn.it_20_11_16	OPENING	READING
T2_HU_Budapest-xrootd-cms.infn.it_20_11_16	OPENING	READING
T2_IN_TIFR-xrootd-cms.infn.it_20_11_16	OPENING	READING
T2_PK_NCP-xrootd-cms.infn.it_20_11_16	OPENING	READING
T2_PL_Swierk-xrootd-cms.infn.it_20_11_16	OPENING	READING
T2_RU_INR-xrootd-cms.infn.it_20_11_16	OPENING	READING
T2_UA_KIPT-xrootd-cms.infn.it_20_11_16	OPENING	READING
T2_UK_London_Brunel-xrootd-cms.infn.it_20_11_16	OPENING	READING

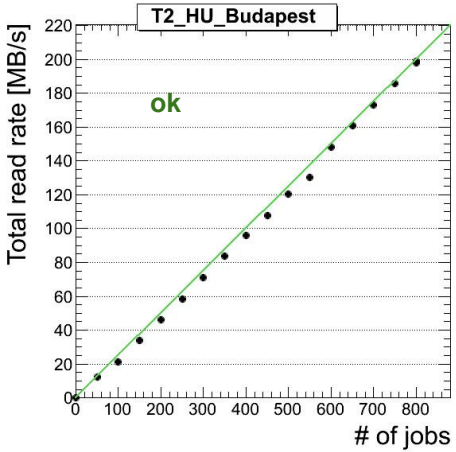
All the sites have installed DPM 1.8.11

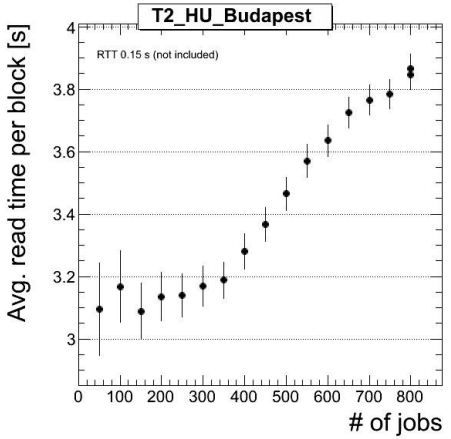
How to debug problems

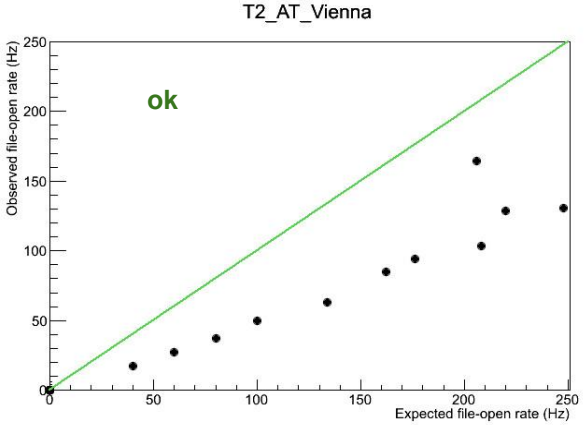
- Complete failure of opening and reading tests.
  - The special path isn't correctly defined at site, check the trivial file catalog setup.
  - Wrong subscription to regional redirector.
  - Temporary problems with regional redirector or condor testbed (generally they affect more sites).
- Poor results with opening tests.
  - Performance depends on the storage setup and hardware. If a site is supporting multi-VO, a slow rate can be due to contention with other VO's.
  - Even if no failure with input files, file opening may take very long time (up to 200s).
- Poor results with reading tests.
  - The reason could be a network or filesystem or disk bottleneck. Only the site monitoring of nodes' load can say something.

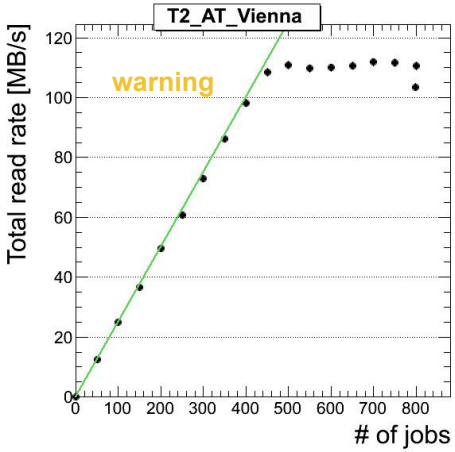
Some DPM results

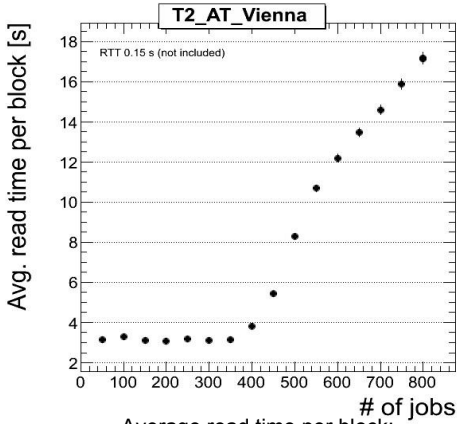






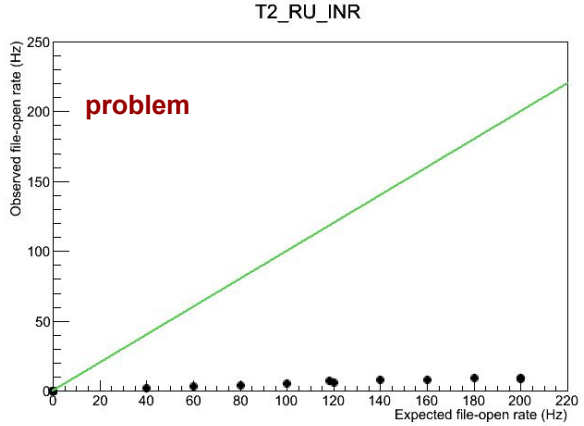


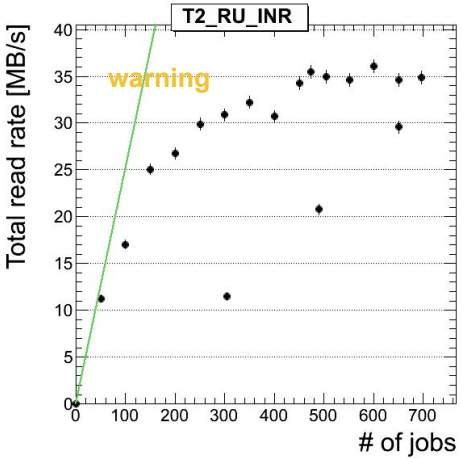


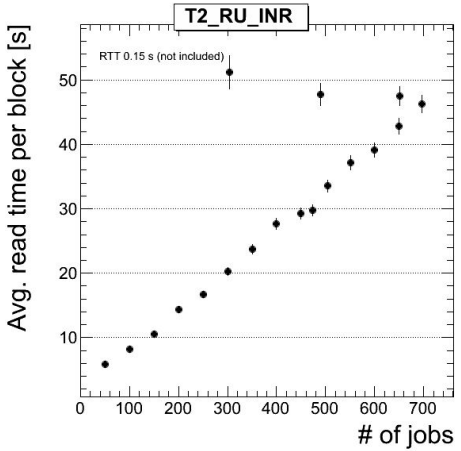


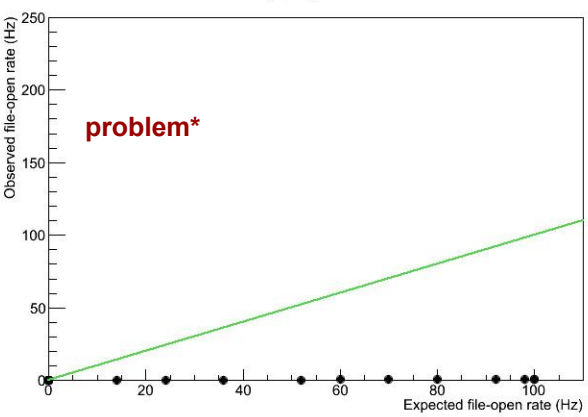
Average read time per block:
lower time (around 2 s) and flat trend is better

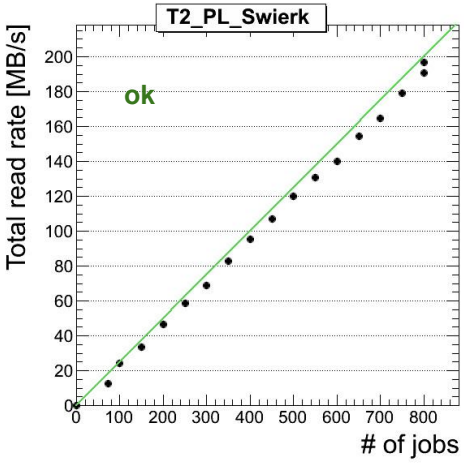
Continue ...

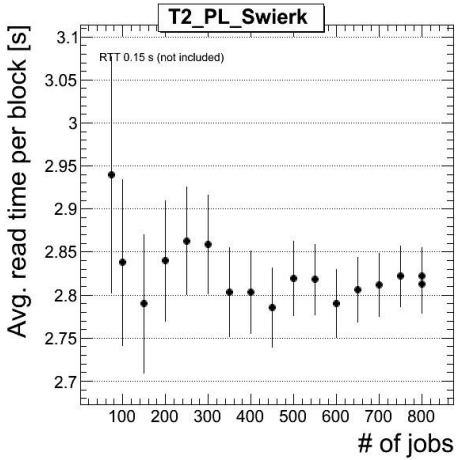












* Site currently registered in transitional federation. Opening test has to run with transitional redirector

What information can help the debug?

- Info about dpm site hardware and configuration to be compared in order to suggest optimal hw vs sw setup.
  - Needed feedback from sites.
    - Is it a good idea to collect hardware and configuration info in a webpage for operators?
- Dpm parameter setup
  - Help from developers about the tuning of DPM cluster.
    - Is the https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm/Admin/TuningHints page updated?
- Is the available documentation about how to join the AAA federation clear enough?

Conclusion

- AAA is adopted by CMS as a system to access remote data if not available locally.
  - Its usage continues to grow past the original use case (fallback) from >5 years ago.
- Good performance of infrastructure is reached by monitoring and testing sites regularly.
- The correct evaluation of test results and the debug of problems require the collaboration of site-manager and backend developers.
  - Would like to increase sites' connectivity to AAA to guarantee success going forward!