Performance Tests of DPM Sites for CMS AAA

Federica Fanzago on behalf of the AAA team

Introduction to AAA

- AAA="Any data, Anytime, Anywhere"
- an effort to create a storage federation of the CMS sites
- AAA makes CMS data access transparent at any CMS sites
- sites' content is federated on the fly using the native clustering of the xrootd framework

Why AAA

- The CMS data discovery system lets us know where data are stored around the world
- CMS sends "jobs to where the data is", for analysis and reprocessing
  - design decision from 10 years ago, still valid
- some pitfalls:
  - jobs at a remote site may fail for data access reasons, e.g. a file "disappeared"
  - if a small number of sites has the right data, queue waiting time can be long
  - a data transfer preparatory step may be necessary before submitting the jobs

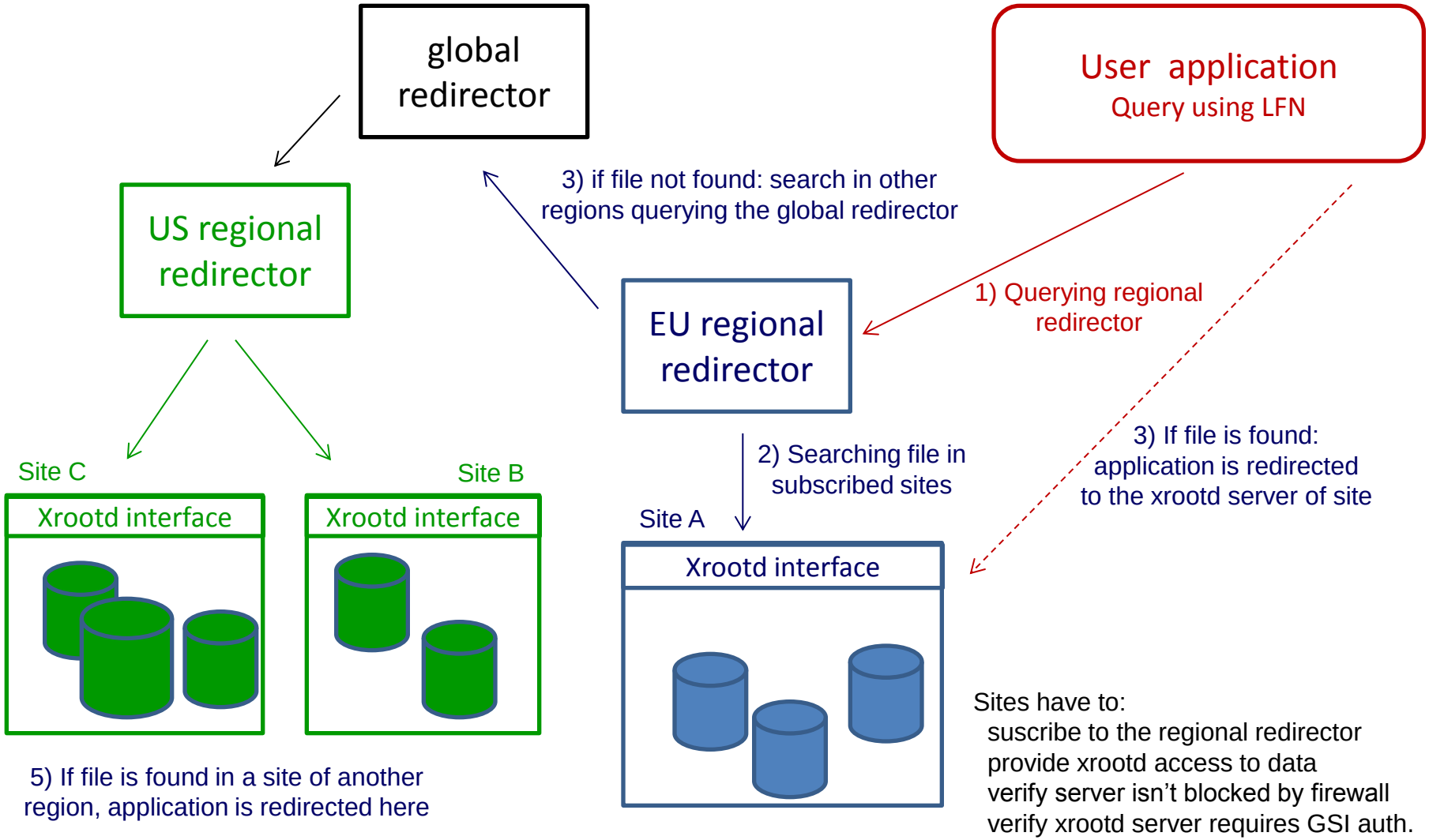
AAA goals

- Make all data available to any CMS physicist, anywhere, interactively
  - reliability: no access failure
    - improvement of process efficiency, CPU utilization
  - transparency: never notice where data actually reside
    - run jobs independently from data location
  - universality: fulfill the promise of opportunistic grid computing
    - much more flexible use of resources globally
    - allow jobs to run on sites not hosting data, only providing CPU

How it works

- In AAA the underlying technology is xrootd
  - interface with different storage backends
- sites in data federation subscribe to a hierarchical system of redirectors (local, regional and global)
- sites provide common namespace for data (LFN)
- applications access data by querying their regional xrootd redirector using LFN
- if the file is not found, the search falls back to the global one

Redirectors schema



Sites have to:
- suscribe to the regional redirector
- provide xrootd access to data
- verify server isn't blocked by firewall
- verify xrootd server requires GSI auth.

AAA deployment

- First deployment at US sites, demonstrating AAA functionality and improving CMS analysis and reprocessing
  - fallback, error rate, queue time, etc.
- additional sites later joined federation
  - currently 60 sites (8 tier1s)
    - 18 DPM
    - 22 dCache
    - 5 Castor
    - 7 Hadoop/BeStMan
    - 2 Lustre / BeStMan
    - 6 StoRM

Xrootd protocol is the common access interface

Performance assessment

- To evaluate the potential of data federation, CMS needs to understand the current performance of each site
  - how are the sites performing? Is their performance and quality of service sufficient?
- through "File opening and reading scale tests" CMS checks if sites are able to sustain the expected load for LHC Run2

File opening and reading tests

- Tests emulate CMS jobs running at CMS sites, choosing the site through regional redirectors
- CMS target for initial tests:
  - File-opening test: access total rate of 100 Hz at a site
  - File-reading test: 600 jobs reading average rate of 2,5 MB every 10 s at a site → reading total rate of 150MB/s
- these numbers come from internal CMS analysis, based on historical figures
- tests reveal sites that need further optimization and possible improvements
- these are not meant to be a stress test for the site

How tests are run

- Sites have to provide a "special" path to allow redirector to match only the site we want to test: /store/xrootd/test/<cms_site_name>/LFN
- tests are submitted from a condor pool in Wisconsin
  - necessary to correctly manage the ramp-up of running jobs
- submission to US sites goes through the Fermilab redirector; others through Bari
- list of input files obtained via Phedex (dataset required to be complete and on disk)

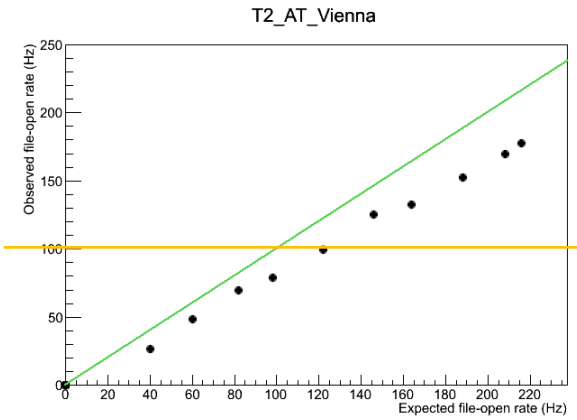
DPM sites in AAA

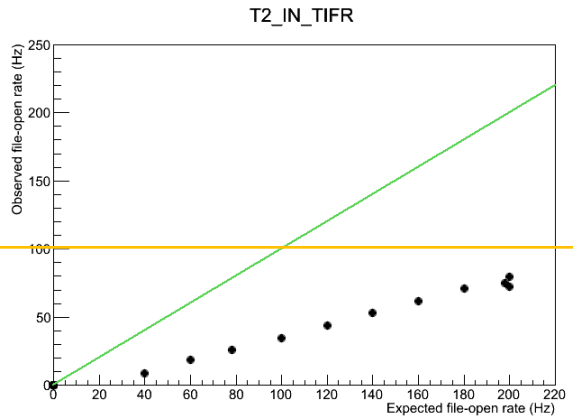
CMS_SITE_NAME		CMS_SITE_NAME
T2_AT_Vienna		T2_RU_PNPI
T2_FR_GRIF_IRFU		T2_RU_RRC_KI
T2_FR_GRIF_LLR		T2_RU_SINP
T2_FR_IPHC		T2_TH_CUNSTDA
T2_GR_Ioannina		T2_TR_METU
T2_HU_Budapest		T2_TW_Taiwan
T2_IN_TIFR		T2_UA_KIPT
T2_PK_NCP		T2_UK_London_Brunel
T2_PL_Warsaw		T1_TW_ASGC

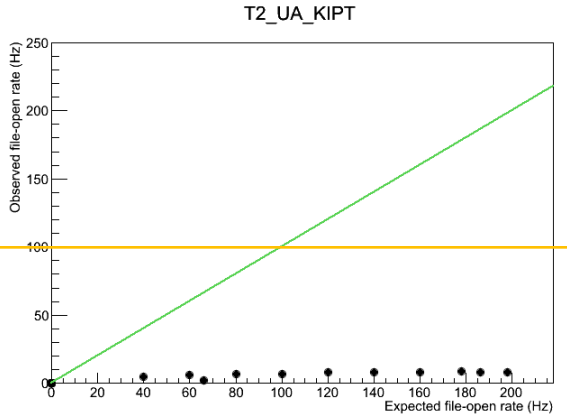
Sites in red are not ready for testing because 'special path' isn't available

Opening plots for some DPM sites

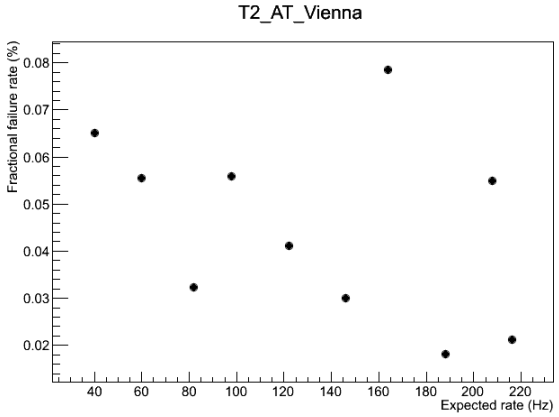
– Tests run up 100 jobs simultaneously, opening files at rate of 2 Hz each.

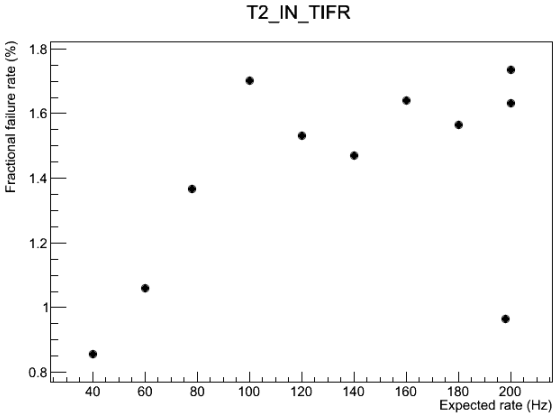


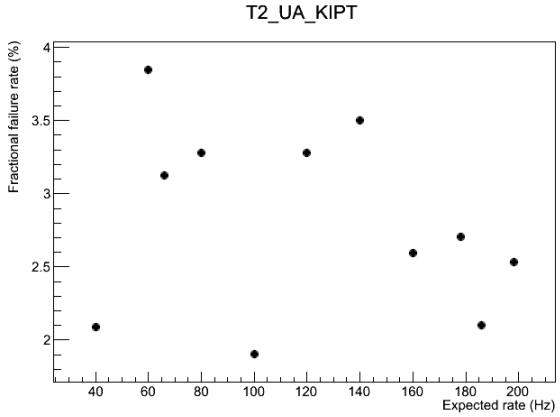




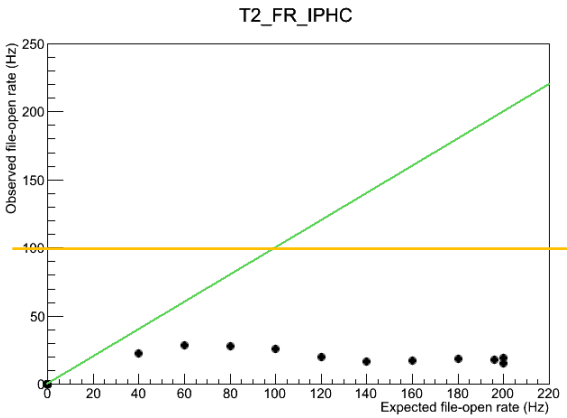
TEST TARGET IS 100 Hz

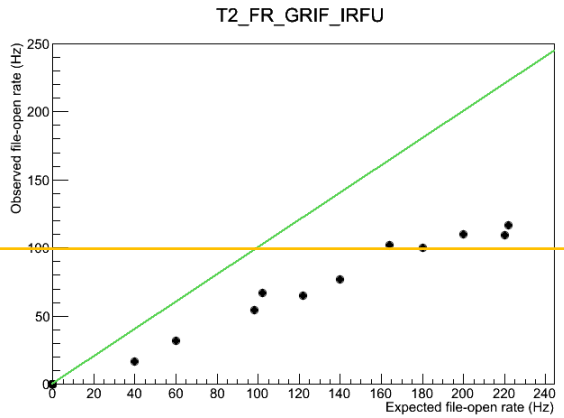


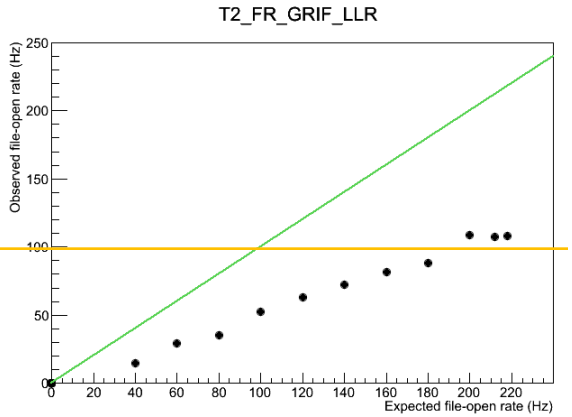


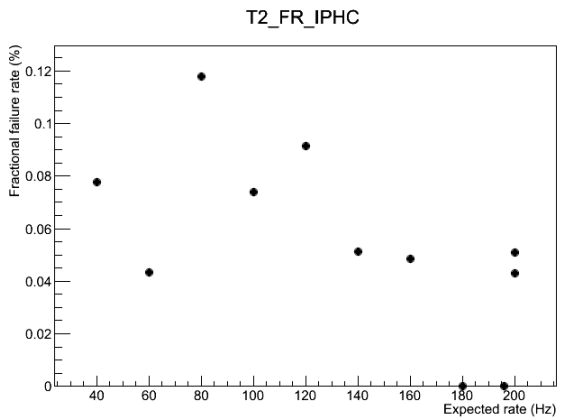


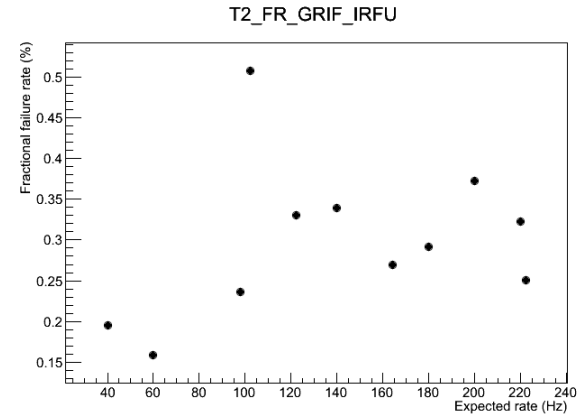
Opening plots

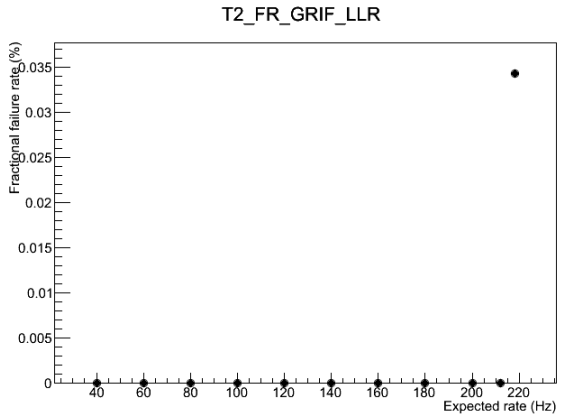




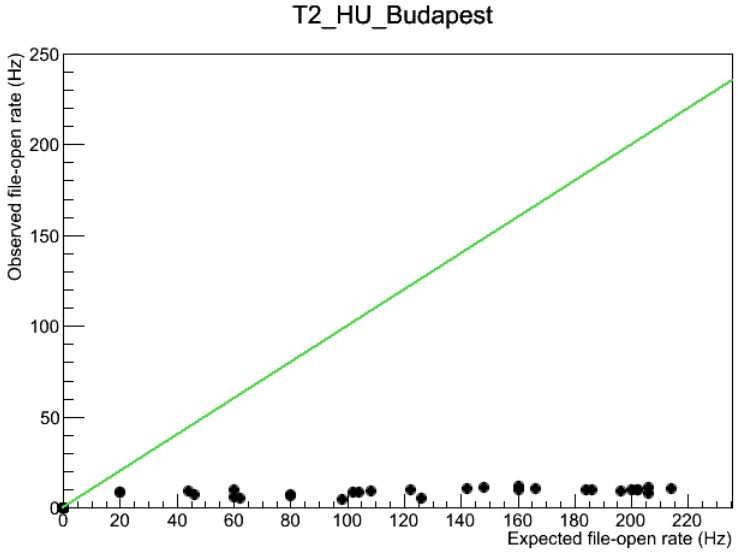


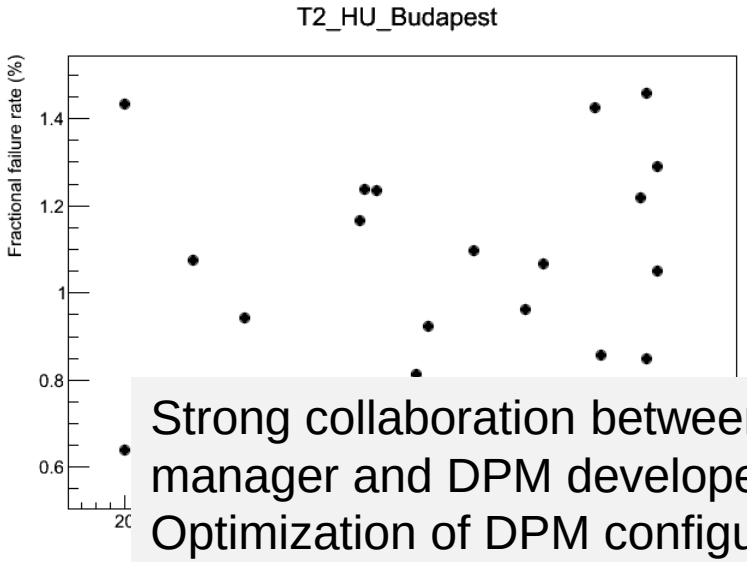






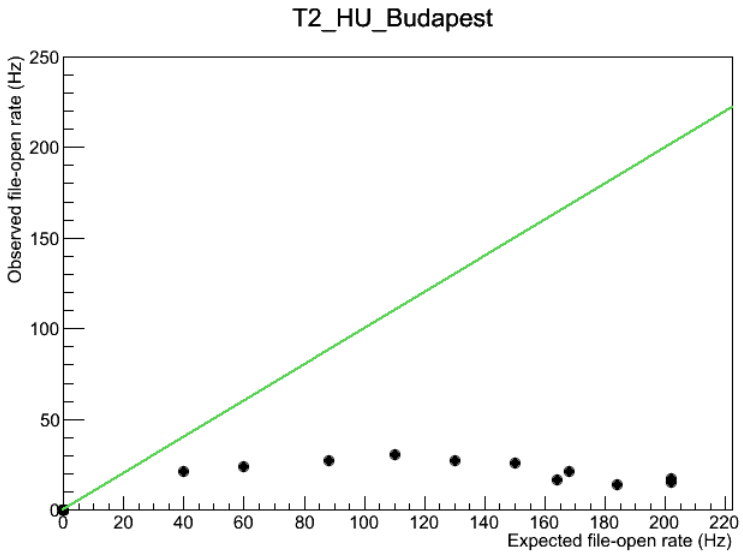
Debugging Budapest

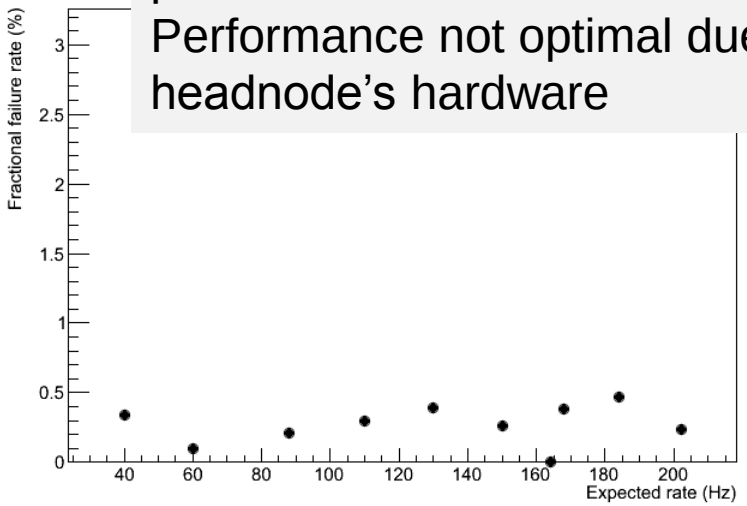




Before tuning

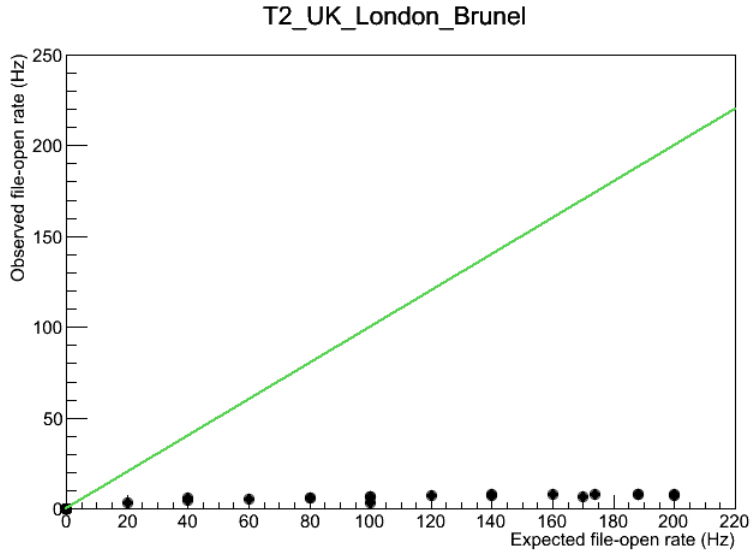
Strong collaboration between site manager and DPM developers
Optimization of DPM configuration parameters
Performance not optimal due to headnode's hardware

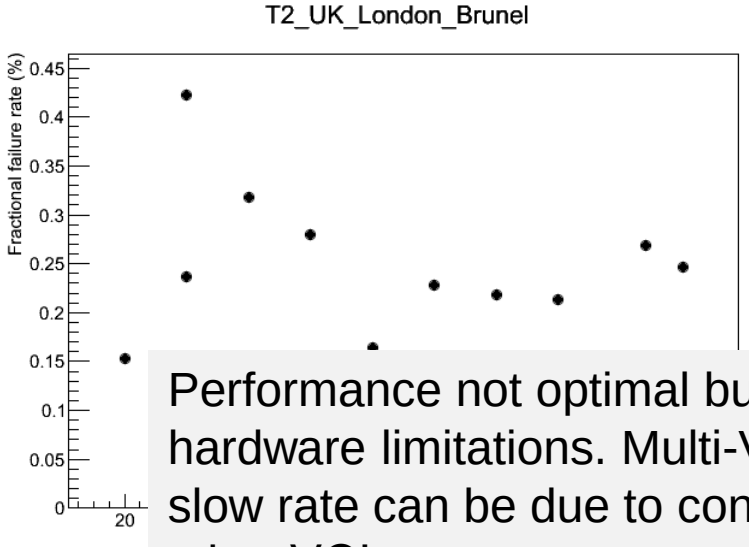




After tuning

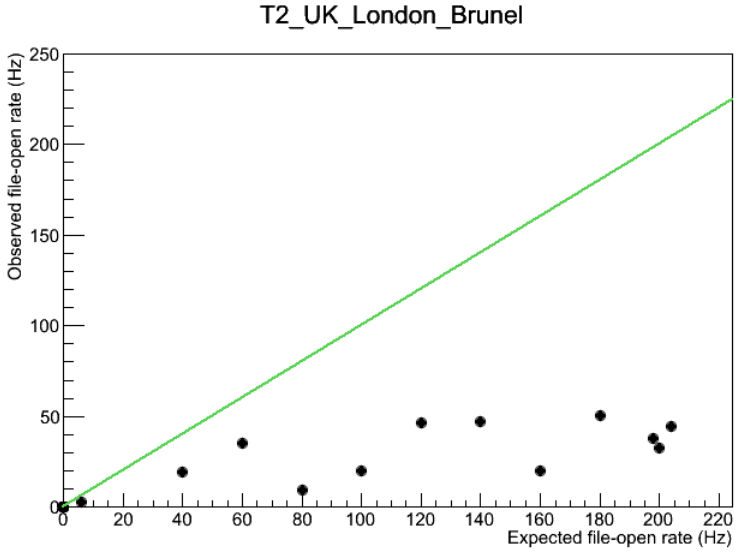
Debugging Brunel

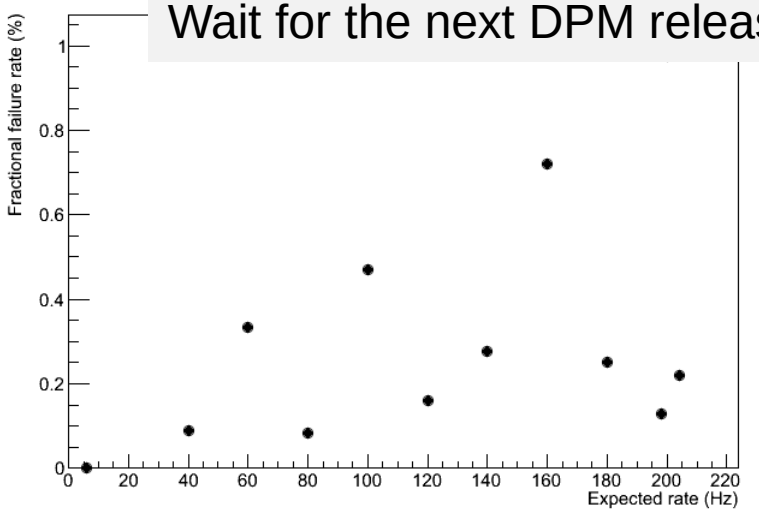




Before tuning

Performance not optimal but not due to hardware limitations. Multi-VO site so a slow rate can be due to contention with other VO's
Wait for the next DPM release?

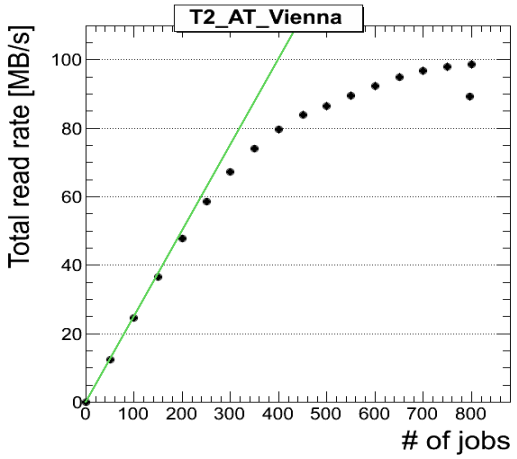


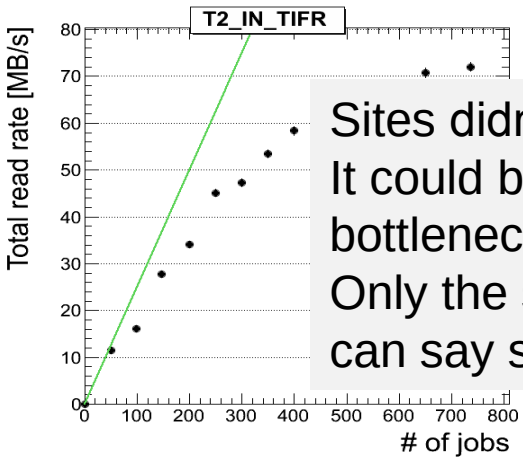


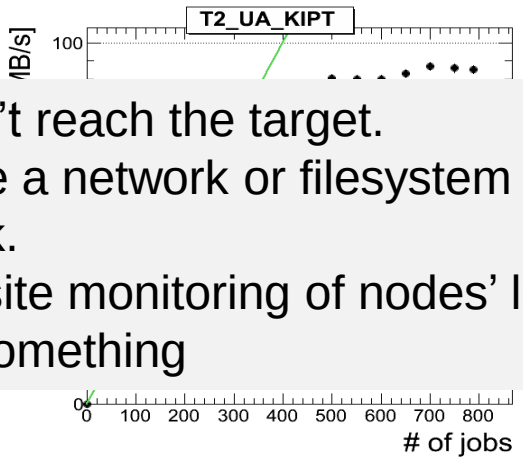
After tuning

Reading plots for some DPM sites

- Tests run up to 800 simultaneously jobs reading block of 2,5 MB every 10 s

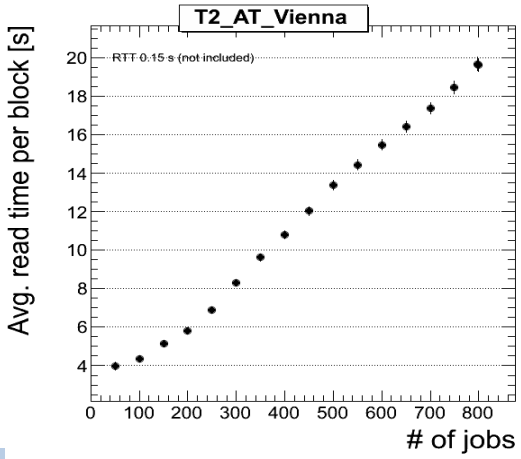


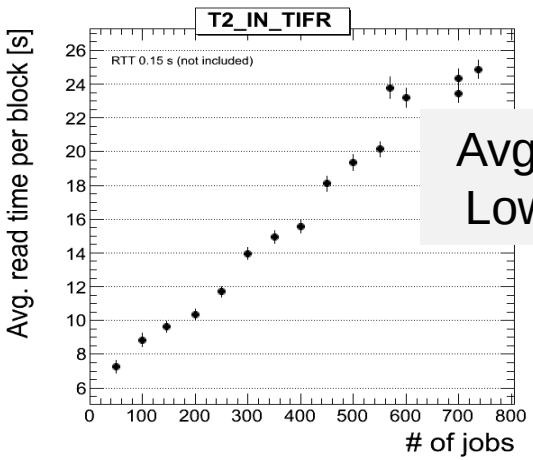


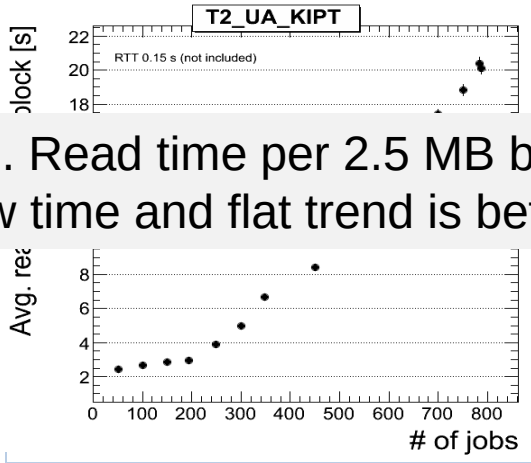


Sites didn't reach the target.
It could be a network or filesystem or disk bottleneck.
Only the site monitoring of nodes' load can say something

TEST TARGET is 600 jobs, reaching 150 MB/s

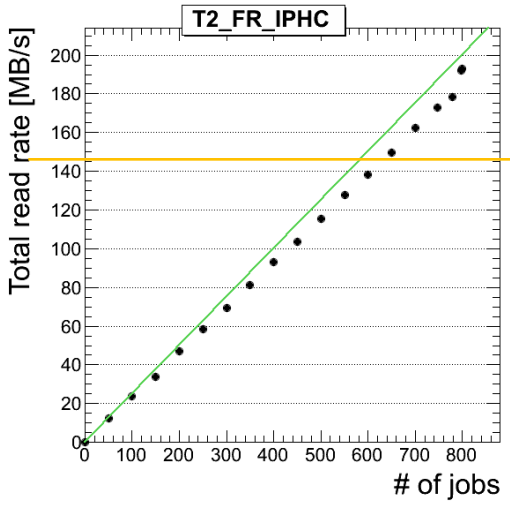


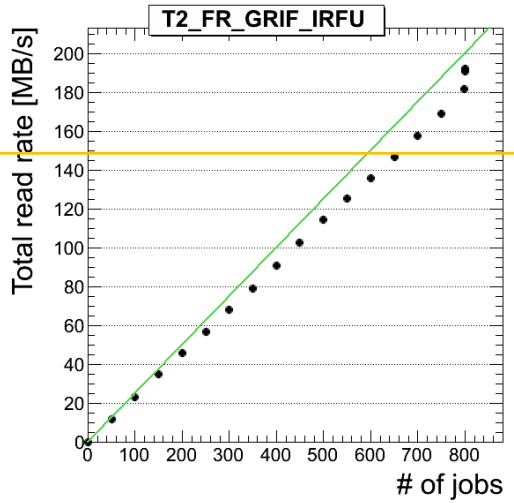


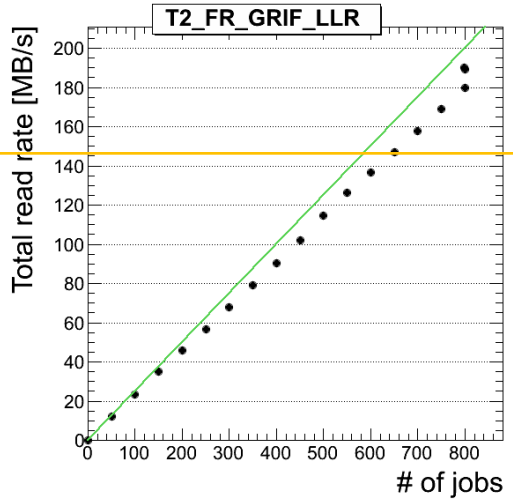


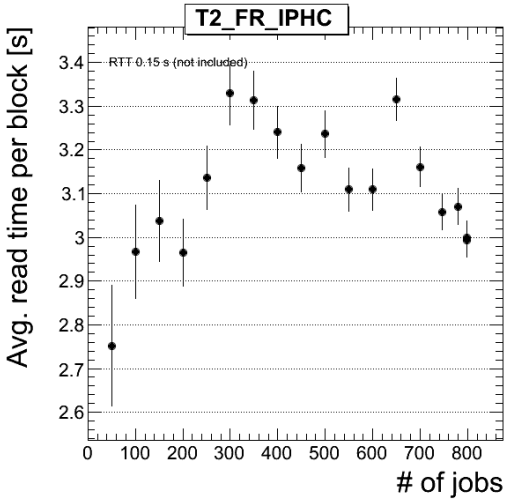
Avg. Read time per 2.5 MB block
Low time and flat trend is better

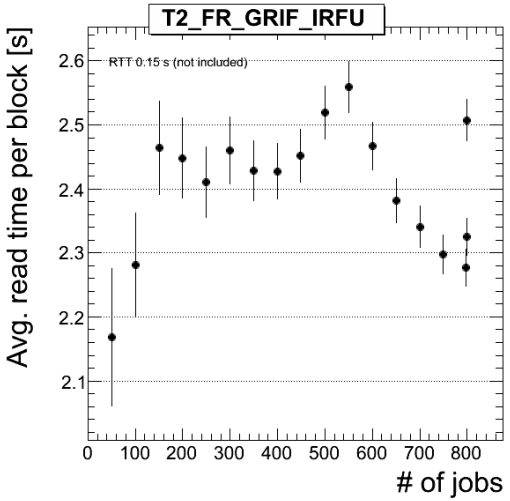
Reading plots 2

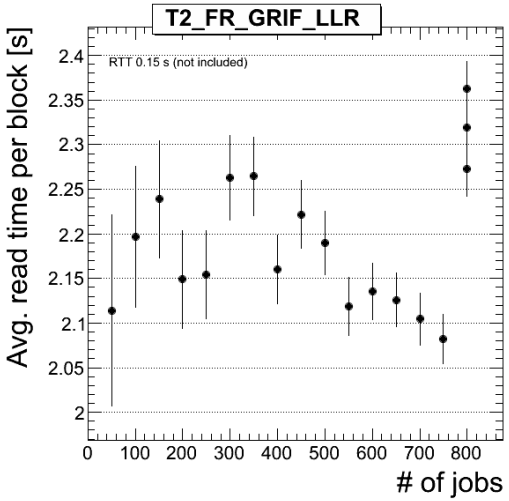




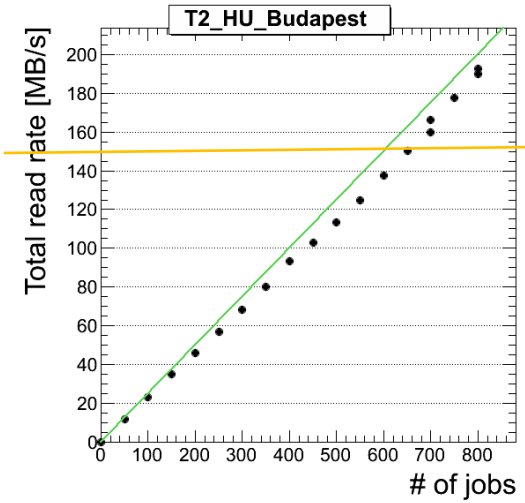


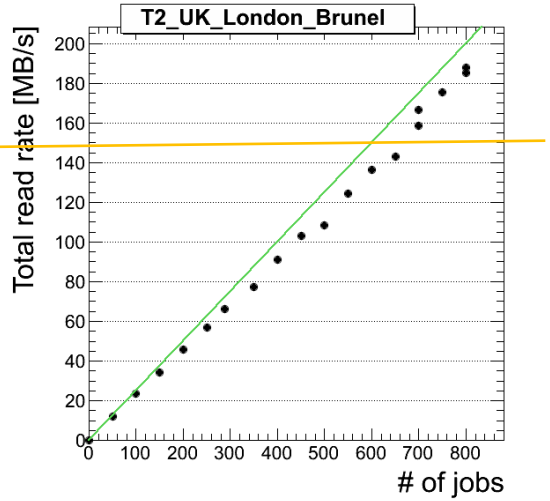


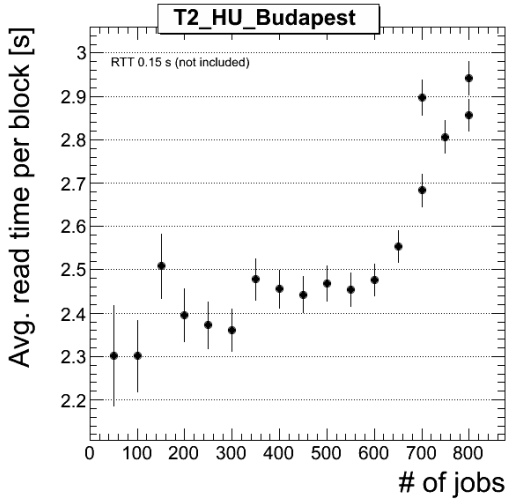


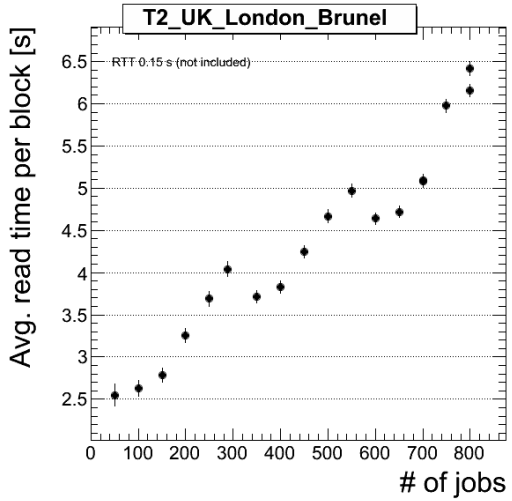


Reading plots 3









Issues found

- Some files are missing ("file not found") giving a performance penalty up to 40 s
  - not synchronized info between files really stored in a site and CMS catalogues
- some false negatives ("files not found" even if they are in the storage)
- sometimes some files need more than 200 s to open
- the first file needs generally more time to be opened (redirector caching system of info)
- "no connection available": comunication error with server
- failure during the file read (xrootd server goes down)

An evaluation attempt

- An attempt to classify sites as "good" or "to debug" is done using as threshold
  - opening test: access rate < 10Hz, flat trend and/or access failure > 6 %
  - reading test: the number of jobs < 600 and/or reading time > 5 s

CMS_SITE_NAME		CMS_SITE_NAME		CMS_SITE_NAME	
T2_AT_Vienna	td	T2_FR_IPHC	td	T2_TH_CUNSTDA	td
T2_FR_GRIF_IRFU	ok	T2_HU_Budapest	ok	T2_IN_TIFR	td
T2_FR_GRIF_LLR	ok	T2_IN_TIFR	td	T2_UK_London_Brunel	ok

The complete table with all the sites is on twiki:
https://twiki.cern.ch/twiki/bin/view/Main/CmsXrootdOpenFileTests#Summary_table_of_EU_tests

Open questions

- Can sites reach the needed figures with the resources they have?
  - access rate 100 Hz, reading rate 150MB/s
- Can we improve the overall performance that we see?
- Suggestions:
  - Make sure that the site is well tuned
  - Make sure that the site has updated sw
- What can DPM sysadmins do to help us?
  - Answer to the survey that has been sent to site managers
  - Suggestions are welcome

Other ideas

- These tests are useful to “debug” remote sites
- May be practical to run tests from a condor pool in EU
- An automatic system to run these tests every night (currently done in US, not in EU)

Conclusion

- CMS is exploring the current performance of remote sites joined AAA federation
- This exercise spots site configuration problems and infrastructure weaknesses
- With the collaboration of site managers, storage backend developers and the AAA team a lot can be done
- This common effort is a important component of readiness for the LHC Run2

Thanks a lot to all the collaborators

A special thanks to Fabrizio Furano for helping us in site debugging and providing useful suggestions for this talk

Documentation and Feedback

- AAA tests twiki page

  https://twiki.cern.ch/twiki/bin/view/Main/CmsXrootdOpenFileTests
- Survey for site managers

  https://twiki.cern.ch/twiki/bin/view/Main/Sites_setup
- DPM tuning hints

  https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm/Admin/TuningHints
- Test code

  https://github.com/cvuosalo/xrootd_scaletest
- AAA support mailing list

  hn-cms-wanaccess@cern.ch

Backup

Test benchmarks

- Minimum benchmarks were determined based upon a historical study of CMS jobs.
- An average CMS job opens a new file once per 1000 s and reads from a file at an average rate of 0.25MB/s.
- Assuming the worst case of 100000 jobs opening files at a site at once gives a benchmark of 100 Hz for file-opening rates
- For file reading, as assumption of 600 simultaneous jobs actively reading files from a site gives a total rate of 150 MB/s

Budapest details

- MySQL daemon: /etc/my.cnf → 1500 (was not set)
- DMLite MySQL plugin: /etc/dmlite.conf.d/mysql.conf NoPoolSize → 256 (was 32)
- Memcached → 2GB (was non installed)
- Dpmmgr account has enough file descriptors
  - ulimit –n → 65000 (was 1k)
- Mysql user
  - Ulimit –n → 65000
- Restarting daemons
- How fast are CPU and mysql disk
  - Pentium D @3.40GHz, 8 year old (headnode)
  - During test: user CPU 30%, system CPU 25% and iowait 10% (what is producing iowait during metdata exercise?)