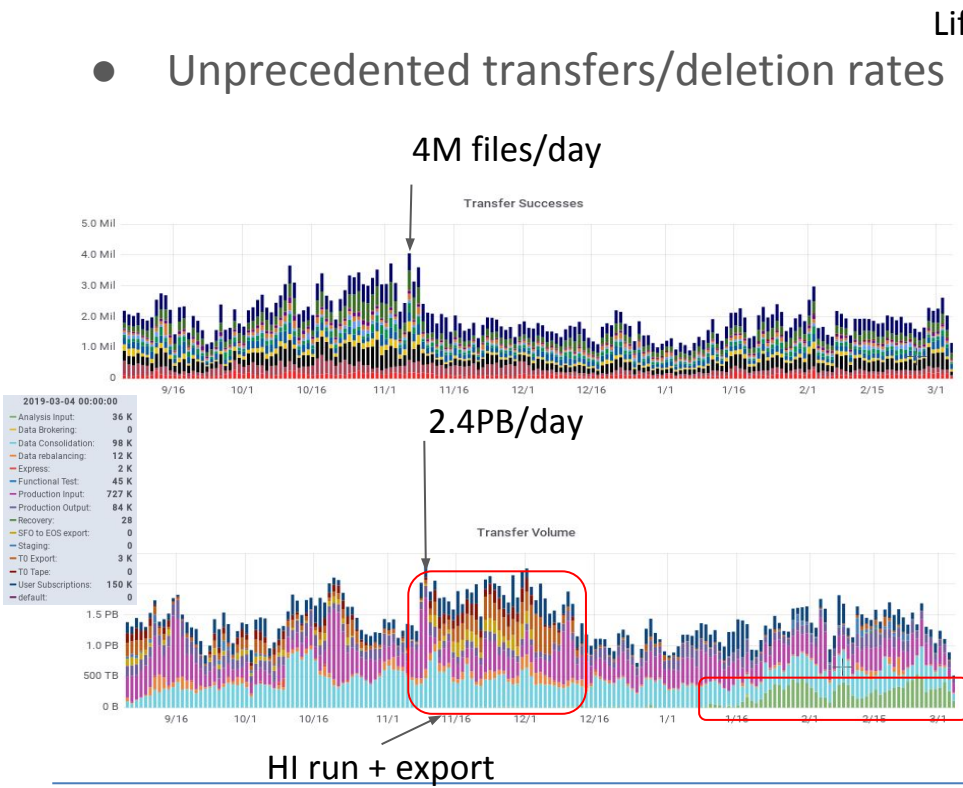
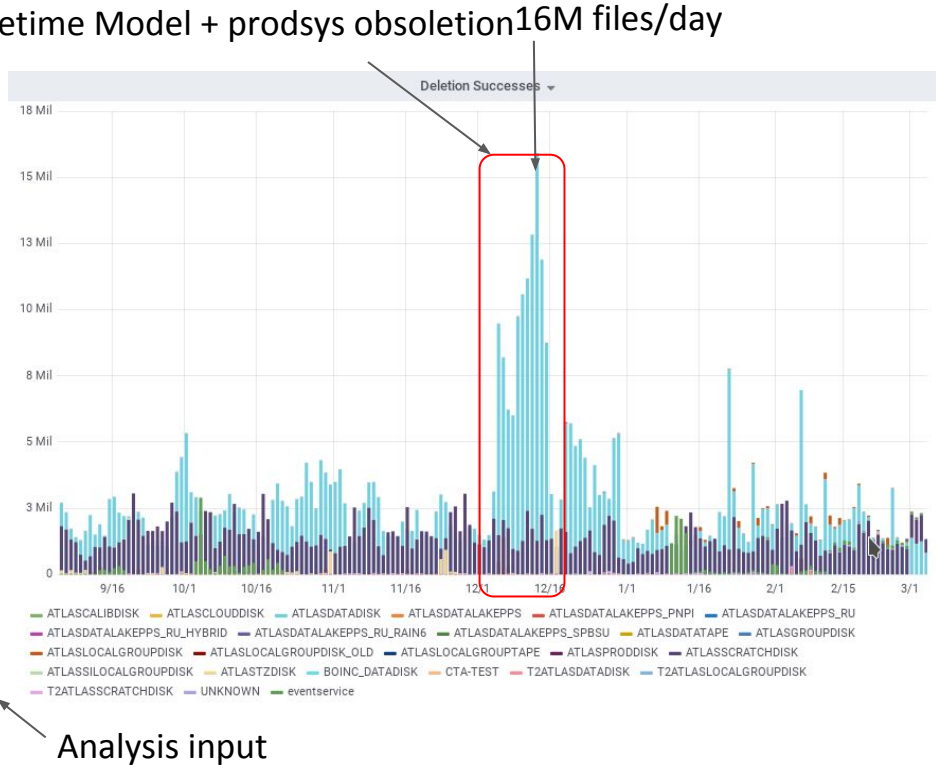
DDM ops report

cedric.serfon@cern.ch
for the DDM ops team

Overall view

- Unprecedented transfers/deletion rates



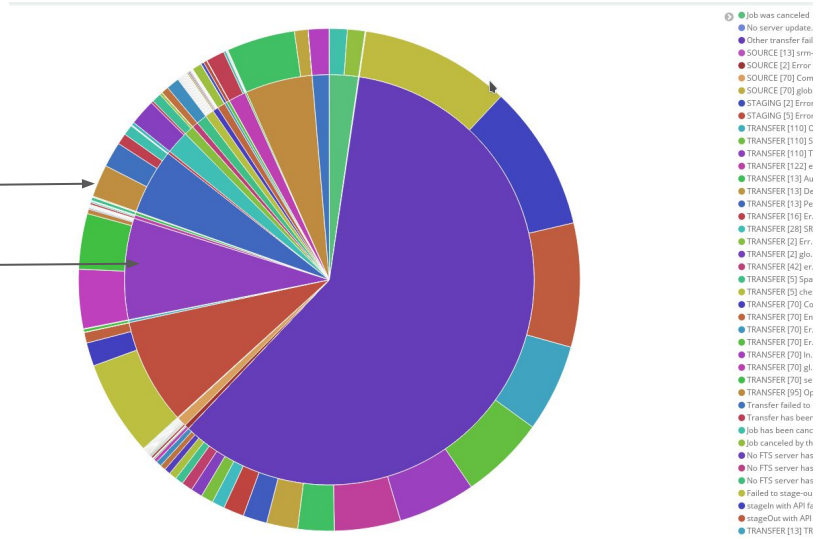


Understanding errors

- Over the last months we have an irreducible flow (<10 %) of errors
- Work ongoing to categorize these errors and try to fix the underlying problems
- Some of the problems are detailed in the next slides

“Raw” Errors

Categorized errors



Lost files/Suspicious

- Lost/corrupted files has been a recurrent problem for many months. Produced from different sources :
  - Rucio (double submission to FTS)
  - FTS (2 FTS servers submitting the same transfers at the same time)
  - Panda/arc (double job submission)
- Suspicious replicas (i.e. files that failed to be transferred many times) :
  - Possibility to enable notifications to the cloud squad (e.g. weekly)
  - New WebUI to find the suspicious files/declare them bad https://rucio-ui.cern.ch/suspicious_replicas
  - Feedback from cloud squads on these new tools is welcome
- Automatic recovery enabled for suspicious files with more than 1 replica
- Log files automatically declared bad if declared suspicious too frequently
- Don’t forget to provide monthly (or at least quarterly) storage dumps

Files temporarily unavailable

- Feature requested by some sites last year. Available in the last Rucio feature release
- Possibility to declare O(1M) files in less than a minute
  - One daemon (horizontally scalable) takes care of changing the state asynchronously
  - The temporary replicas can become automatically online after a defined timeout
  - Same permissions than the one to declare bad replicas
- Still at an early stage :
  - Tested on 2 sites so far
  - API available, no CLI yet
  - Need to check that Panda/Jedi can handle these files properly for the brokering
- https://twiki.cern.ch/twiki/bin/view/AtlasComputing/DDMDarkDataAndLostFiles#Temporary_unavailable_files for more info

Network

- Network seems to be one frequent issue (~10% of the failures are due to timeouts)
- During the last months saturation observed leading to timeouts :
  - On some links to particular sites with low bandwidth
    - Will apply some dynamic scaling of the FTS parameters to decrease the number of timeouts
  - On some links between NRENs and Géant. No easy way to circumvent this on the Rucio/FTS layer
    - Have we been too far in getting rid of the Monarc Model ?
- Ideally, we would like to put some minimum network requirements (as we do for the disk space) for the sites hosting data so that we have no blackholes
  - Sites not fulfilling these requirements could become lightweight sites
- Tests successfully done in Autumn to spill over transfers from LHCOPN to LHCOne between CERN and SARA.
  - Would be good to test on more sites + testing spillover LHCOne → LHCOPN

Space reporting

- WLCG is now pushing for json space reporting.
  - ATLAS DDM infrastructure is able to use this new json as well as the old ATLAS one
  - If you want to deploy this json, please do it
  - Disclaimer : for DPM sites it requires to run in DOME mode, but there are still some issues to address (c.f. Petr's talk)
- Requirements is that the json needs to be updated every hours, or even better every 30 minutes
  - The frequency is important since Rucio relies on the space value reported by the site for the cleanup. If not updated frequently, deletion will lag behind and the site might get full

Space management

- We asked the sites to scale their SCRATCHDISK size accordingly to the number of analysis slots : Recommended value 100TB per 1k analysis job slots
- When we run large deletion campaigns, we can have problems because the deletion is not fast enough
  - Problem due to very small files O(10MB)
  - The deletion agent (reaper) in its current configuration cannot delete more than 16M files/days
  - Work ongoing to allow the reaper to be faster in case of high deletion activity
- Some sites complained that we are not leaving enough free space
  - e.g. BNL tot space : 18.8 PB vs free space 0.3 PB
  - But we have little room to increase the min free space value (we are scrutinized on our resources use and we try to maximize the space used)

Other problems

- Different problems pop up regularly that generate a significant load of DDM ops/shifters :
  - IPv6 (Routing, firewall problems)
  - Network (already mentioned)
  - Deletion failures (in particular DPM)
  - Space reporting
- Most of these problems are detected by our pROblem Detector (aka ROD)
  - As we did recently for the lost/corrupted, we rare putting in place new tools to identify these problems before they are reported
- Some of them can take months to be solved (e.g. network)

DOMA activities

- There are a lot of new activities where DDM ops is involved :
  - SRMless TPC (See Alessandra’s talk)
  - QoS (idem)
  - TAPE carrousel (See Xin’s talk tomorrow)
  - Test of new CTA instance at CERN
  - Diskless and lightweight sites (See Stephane+Ilija’s talk)

Conclusion

- Development of new tools to help reducing the number of transfer errors and detect errors proactively
- Start to see some bottlenecks (e.g. deletion, network)
  - Being addressed
- New activities driven by Doma will keep us busy during LS2
- Big thanks to the shifters (dast, adcos, CRC) who are of great help